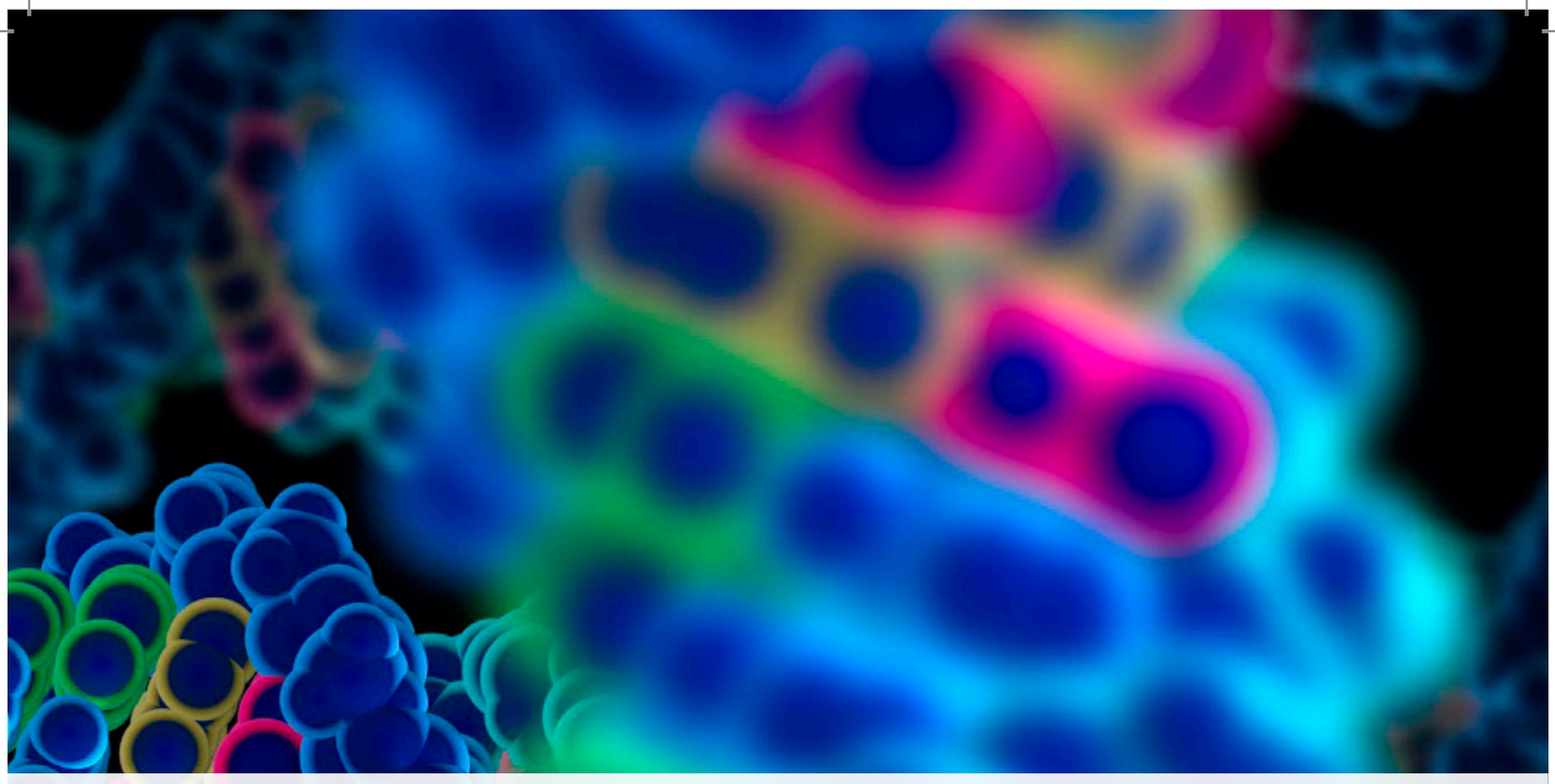


ABACBS

2017 Australian Bioinformatics & Computational Biology Society (ABACBS) Annual Conference and Combine Student Symposium

Monday 13 - Friday 17 November

Platinum sponsors

Silver sponsors

Bronze sponsors

Sponsors

The ABACBS 2017 Organising Committee and ABACBS Executive Committee would like to express our sincere and grateful thanks to our generous sponsors, without whom this meeting would not be possible. One of our platinum sponsors, Illumina, will also be holding a workshop on Tuesday November 14th (1-2pm) on ***NovaSeq® The New Era in Sequencing. From Big Data to Bigger Breakthroughs!*** Please support our sponsor by attending.

Platinum sponsors

illumina®

Centre for Cancer Biology

University of South Australia

THE UNIVERSITY of ADELAIDE

Flinders UNIVERSITY

Silver sponsors

EMBL Australia

SAHMRI South Australian Health & Medical Research Institute

Australian Genomics Health Alliance

Bronze sponsors

PACBIO®

millennium® science
enabling next-generation research™

scientifix

agrf

SHOTTESBROOKE

About ABACBS 2017

The 2017 Australian Bioinformatics & Computational Biology Society (ABACBS) National Conference and the Combine Student Symposium will be held at the South Australian Health and Medical Research Institute (www.sahmri.org) in Adelaide, November 13-15 2017. The meeting will showcase the cutting edge of research in the field through an excellent program of invited international and national speakers, and selected oral and poster presentations. The ABACBS conference will be followed on Thursday November 16 and Friday November 17 by several optional satellite workshops and events including a Bioconductor masterclass, the 3rd Asia-Pacific Bioconductor meeting, a workshop on Using Open Science in Bioinformatics Training and a Train-the-Trainer workshop.

Organising Committee

We would like to thank all of those that have helped organise ABACBS 2017 including the ABACBS national executive committee, the SAHMRI Marketing, Facilities, IT, and Finance teams, Sharon Macgowan (SAHMRI) and especially the 2017 conference organising committee:

- **David Lynn** (EMBL Australia Group Leader, SAHMRI/Flinders) – Chair/Conference Convener
- **David Adelson** (Chair of Bioinformatics, University of Adelaide)
- **Karin Kassahn** (Head Technology Advancement Unit, SA Pathology)
- **Andreas Schreiber** (Head of Bioinformatics, Centre for Cancer Biology)
- **John Williams** (JS Davies Research Professor, University of Adelaide)
- **Mirana Ramialison** (Group Leader, Monash) - ABACBS National Conference Co-ordinator
- **Ute Baumann** (Group Leader, Australia Centre for Plant Functional Genomics)
- **Katherine Pillman** (Research Associate, Centre for Cancer Biology)
- **Atma Ivancevic** (PhD Candidate, University of Adelaide)
- **Steve Pederson** (Bioinformatics Hub, University of Adelaide)
- **Klay Saunders** (PhD Candidate, University of South Australia) – COMBINE Student Symposium Representative
- **Alan Rubin** (Postdoctoral Researcher, WEHI) – Representative of the 2018 organising committee

Contact Details and Social Media

Email: conference@abacbs.org
Twitter: #abacbs17 and the handles @abacbs and @combine_au

Wifi

The Eduroam network is available at SAHMRI https://www.sahmriresearch.org/eduroam-at-sahmri. Participants may also connect to the SAHMRI-Guest network using the password : Sahmr1guest. Please note that SSH access is disabled on this network and it is heavily used.

Laptop Charging

Please note that power sockets to charge laptops in the auditorium are extremely limited. Please come with a fully-charged battery!

Code of Conduct at ABACBS 2017

ABACBS is committed to making its science, training and public outreach activities productive and enjoyable for everyone. We will not tolerate inappropriate behaviour by individuals associated with our activities in any form. This code applies to all participants, instructors and organisers in ABACBS organised or sponsored events, including talks, workshops, conferences, social media and event-related social activities.

Expected Behaviour

All event participants are expected to behave in accordance with both the ABACBS Code of Conduct as well as their respective employer's policies governing appropriate workplace behaviour, and applicable laws. All communication, including online, should be appropriate for a professional audience including people of many different backgrounds. Sexual or sexist language and imagery is not appropriate. Be kind to others. Treat everyone with respect. If requested not to tweet/photograph/video or otherwise disseminate the content of a presentation, do not do so! Requests to remove photos or videos published in social media should be respected.

Unacceptable Behaviour

Harassment and sexist, racist or exclusionary comments or jokes are not appropriate. Harassment includes sustained disruption of talks or other events, inappropriate physical contact, sexual attention or innuendo, deliberate intimidation, stalking, and inappropriate photography or recording of an individual without consent. It also includes, but is not limited to, offensive comments related to gender, sexual orientation, disability, physical appearance, body size, race or religion. Anyone who witnesses or is subjected to unacceptable behaviour should notify an event organizer at once or contact a member of the ABACBS executive committee.

Consequences of Unacceptable Behaviour

Individuals asked to stop any inappropriate behaviour are expected to comply immediately. ABACBS event organisers may take any action they deem appropriate, ranging from issuance of a warning to expulsion from the event with no refund, depending on the circumstances. ABACBS reserves the right to exclude any participant found to be engaging in inappropriate behaviour from participating in any future ABACBS conference, events, workshops or other activities, and may take disciplinary action as described in the model rules of association.

Please note that photographs will be taken during ABACBS 2017 and may be used in future promotional material.

If you do not want a photo of you used in this way please email conference@abacbs.org

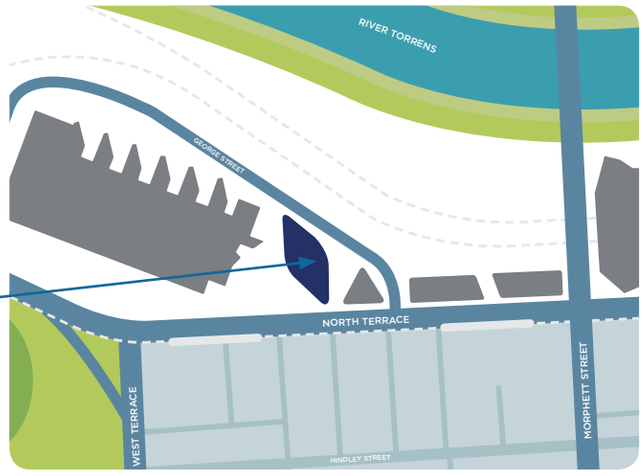
The SAHMRI building and Precinct are smoke free. Strictly no alcohol is to be taken outside during the conference reception/dinner.

ABACBS 2017 Program at a Glance

Monday November 13	
8:00am - 8:50am	COMBINE Symposium Registration Open
9:00am - 5:30pm	COMBINE Symposium
6:00pm - 7:00pm	ABACBS 2017 Opening Keynote Lecture
7:00pm - 7:30pm	ABACBS Honorary Senior Fellow Awards
7:45pm - 10:00pm	COMBINE student, ABACBS Postdoc & ABACBS Professional Bioinformatician Social Events
Tuesday November 14	
8:00am - 8:30am	ABACBS 2017 Registration Open
8:30am - 10:30am	ABACBS 2017 Session 1 - Personalised Medicine & Cancer Genomics
11:00am - 12:00pm	ABACBS 2017 Session 2 - Natural Language Processing
12:00pm - 2:00pm	Lunch & Poster Session 1
1:00pm - 2:00pm	Illumina Workshop
2:00pm - 3:30pm	ABACBS 2017 Session 3 - Single Cell Genomics
4:00pm - 5:30pm	ABACBS 2017 Session 4 - Clinical Genomics
5:30pm - 8:00pm	Wine Reception & Conference Dinner (Ticket Holders Only)
Wednesday November 15	
9:00am - 10:30am	ABACBS 2017 Session 5 - Molecular Evolution
11:00am - 12:00pm	ABACBS 2017 Session 6 - Systems/Network Biology
12:00pm - 2:00pm	Lunch & Poster Session 2
1:00pm - 2:00pm	ABACBS AGM
2:00pm - 3:30pm	ABACBS 2017 Session 7 - Neurogenomics
4:00pm - 5:00pm	ABACBS 2017 Session 8 - Animal & Plant Omics
5:00pm	End of ABACBS 2017 Meeting
Thursday November 16	
9:00am - 5:00pm	Using Open Science in Bioinformatics Training
9:00am - 5:00pm	Bioconductor Masterclass
Friday November 17	
9:00am - 5:00pm	Train-the-Trainer Workshop
9:00am - 5:00pm	Third Asia-Pacific Bioconductor Meeting

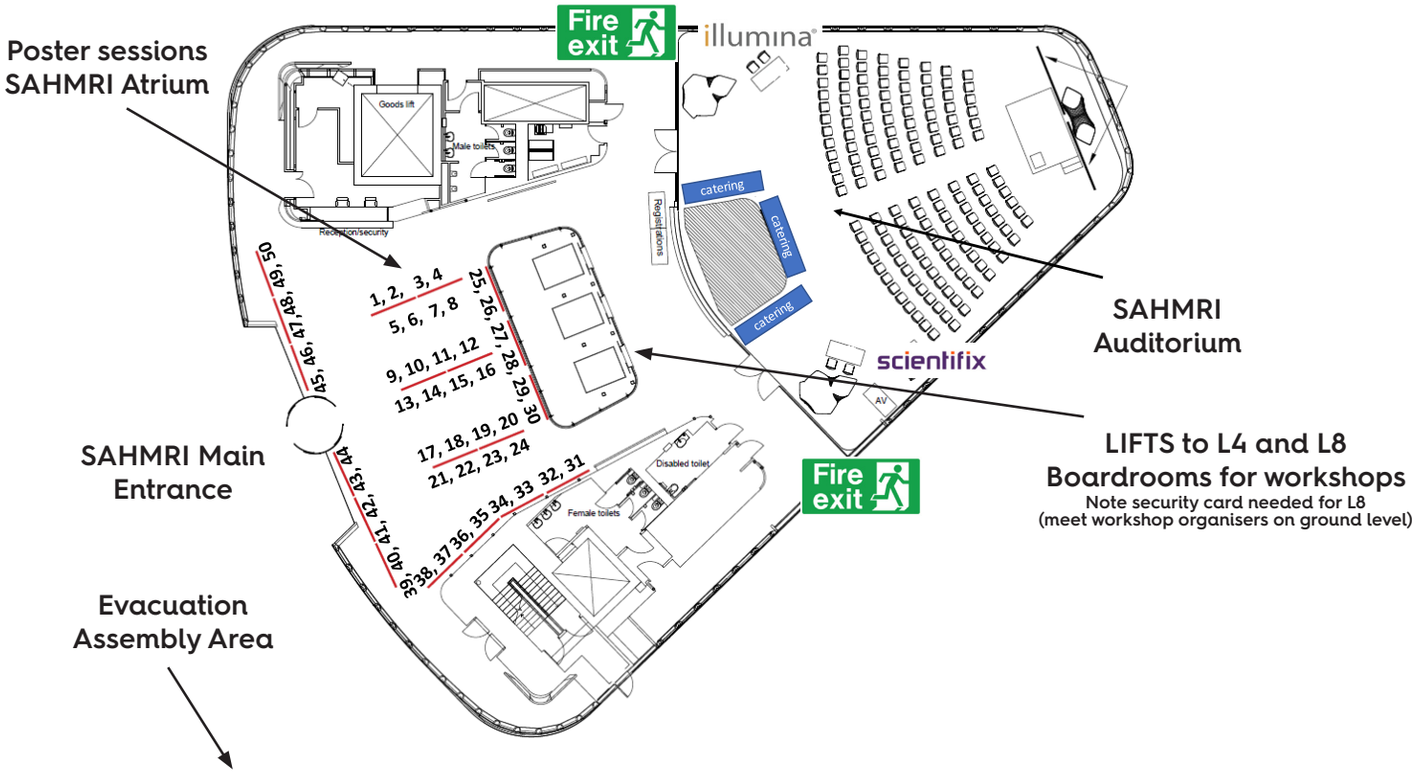
Location

The 2017 ABACBS Conference and COMBINE symposium will be held in Adelaide at the South Australian Health and Medical Research Institute (SAHMRI). Located on North Terrace in the CBD of Adelaide, SAHMRI is right in the center of South Australia's new health and medical precinct.



Venue Layout

Figure 1: Venue and poster board layout



Invited Speakers

The organising committee would like to thank our panel of international and national speakers for attending ABACBS 2017 and presenting their work:

Professor Des Higgins
Professor of Bioinformatics at University College Dublin

Speaking Monday November 13 at 6pm.
"Everything you ever wanted to know about multiple alignments".
@HigginsDes

Des Higgins is Professor of Bioinformatics at University College Dublin, Ireland where his lab works on "omics" data analysis and sequence alignment algorithms. He has been carrying out bioinformatics research since 1985 and is most well-known for the Clustal package for multiple sequence alignment. He wrote the first Clustal program in 1988. ClustalW and ClustalX came from a later collaboration with colleagues in EMBL Heidelberg, where he worked from 1990 to 1994. The papers describing those programs are very highly cited with the ClustalW paper being the tenth most cited paper of all time. Currently, he has run out of letters and his group support and develop the Clustal Omega package.

Professor Su-In Lee
Associate Professor in the Department of Computer Science & Engineering and the Department of Genome Sciences at the University of Washington

Speaking Tuesday November 14 at 8.45am.
"From big data to precision medicine using interpretable machine learning".

Professor Su-In Lee is an Associate Professor in Genome Sciences at the School of Computer Science & Engineering and the School of Medicine, at the University of Washington, in the USA. She completed her PhD in 2009 at Stanford University and was a visiting professor at Carnegie Mellon University before joining the University of Washington. Her lab seeks to develop powerful Artificial Intelligence and Machine Learning techniques to learn from big data novel biological mechanisms, how to improve human health, and how to treat challenging diseases such as cancer and Alzheimer's disease. She has major research grants from the American Cancer Society, the NIH, and the National Science Foundation (NSF).

Invited Speakers

Professor Karin Verspoor
Professor in the School of Computing and Information Systems at the University of Melbourne

Speaking Tuesday November 14 at 11am.
"Examining biological database quality through the lens of the scientific literature".
@karinv

Karin Verspoor is a Professor in the School of Computing and Information Systems and Deputy Director of the Health and Biomedical Centre at the University of Melbourne. Trained as a computational linguist, Karin's research primarily focuses on text mining of clinical texts and the biomedical literature to support biological discovery and clinical decision support. Karin held previous posts as the Scientific Director of Health and Life Sciences at the NICTA Victoria Research Laboratory, at the University of Colorado School of Medicine, and the Los Alamos National Laboratory. She also spent 5 years in start-ups during the US Tech bubble.

Dr Joseph Powell
Group Leader at the Institute for Molecular Bioscience, at the University of Queensland.

Speaking Tuesday November 14 at 2pm.
"Using ultra high-throughput single cell sequencing to understand cellular heterogeneity".
@JosephPowell_UQ

Since 2015 Joseph Powell has been the head of the Single Cell and Computational Genomics Lab at the University of Queensland's Institute for Molecular Bioscience and an NHMRC Career Development Fellow. He obtained his Ph.D. from the University of Edinburgh in 2010 and subsequently completed a postdoctoral position in Peter Visscher's (FAA) group at QIMR. His research involves the use of large-scale transcriptomic and DNA sequence data from both bulk tissues and single cells, focusing on understanding the genetic mechanisms by which heritable variants contribute to disease susceptibility at a cellular level, and ultimately achieve therapeutic and diagnostic outcomes.

Invited Speakers

Professor Gavin Huttley
Group Leader in the School of Biology at the Australian National University

Speaking Wednesday November 15 at 9am.
"Statistical methods for detecting and exploiting mutation motifs".

Professor Gavin Huttley is a group leader in the School of Biology at the Australian National University. He obtained a B.Sc from Macquarie University in Sydney, a PhD in Molecular Population Genetics from the University of California, Riverside and undertook postdoctoral research in the Laboratory of Genomic Diversity, National Cancer Institute (USA). He is a recipient of the Howard Florey Young Investigator award. Professor Huttley's research is focussed on genome decryption — identifying regions of the genome which encode functions that influence susceptibility to disease — through analysis of genetic variation.

Professor Melanie Bahlo
Division Head at Walter Eliza Hall Institute of Medical Research

Speaking Wednesday November 15 at 2pm.
"Detecting known repeat expansions with standard protocol next generation sequencing, towards developing a single screening test for neurological repeat expansion disorders".
@MelanieBahlo

Professor Melanie Bahlo is a Division Head at the Walter Eliza Hall Institute of Medical Research. Her research focuses on developing new methods to analyse complex genetic data. These are being used to discover the genetic causes of human diseases including epilepsy, ataxia and autism. Hers is a highly collaborative laboratory, working closely with clinician researchers to reach important clinical research outcomes for families with genetic disorders. In recent years our research has focused on brain disorders, identifying genetic causes for diseases that have previously proven intractable to analysis. The software developed in her group is freely available to others, aiding many research fields.

Oral and Poster Presenter Instructions

For all oral presenters

- **There are 4 different talk formats:** Invited international speaker talks; invited national speaker talks; selected long oral presentations; selected short oral presentations.
- **Invited international speakers:** 50 minutes talk time; 10 minutes questions.
- **Invited national speakers:** 40 minutes talk time; 5 minutes questions.
- **Selected long oral presentations:** 25 minutes talk time; 5 minutes questions. These longer format selected talks are to encourage senior researchers or postdocs to submit abstracts and to present. In some cases, a substantial body of work by a student may be appropriate.
- Selected short oral presentations: 12 minutes talk time; 3 minutes questions.
- **All speakers are expected to keep to time!** Session chairs will ruthlessly police this.
- Powerpoint is preferred; PDF also acceptable. Other software at presenter's own risk!
- Aspect ratio of slides 4:3
- Presenters may use the provided Windows laptop or their own laptop (via a VGA or HDMI interface). If using your own laptop – any set-up time will eat into your presentation time so pre-loading on the provided laptop is strongly encouraged. All presentations will be subsequently deleted.
- If using the provided Windows PC, please bring your presentation on a USB stick.
- All presenters should **upload and check their presentation** with the AV team **prior to their session.**

Combine oral presentations

- **Selected Oral Presentations:** 12 minutes talk time; 3 minutes questions.
- The best student presentation will be given an oral presentation slot at the ABACBS conference.

For all poster presenters

- **All posters** will need to be printed in **A0 portrait format i.e. (841mm width x 1189mm height).** Posters printed outside of these dimensions may be unable to be presented.
- There are **two different** ABACBS poster sessions: Tuesday Nov. 14th 12 – 2pm and Wednesday Nov. 15th 12 -2pm. Please check which session you are presenting in and ensure you hang your poster at the **assigned position** based on the poster board number (not abstract number!) (see Figure 1 on page 7 – venue layout).
- Poster sessions will be held in the SAHMRI **ground level atrium** (right outside the auditorium).
- Posters should be hung in the morning of the day you are presenting and **taken down the same evening.**
- **Prizes** will be awarded for the best posters.

Oral and Poster Presenter Instructions

For poster presenters selected for the fast forward session

- **10 posters** from the session on Tuesday Nov. 14th **and 10** from the session on Wednesday Nov. 15th have been selected for a fast forward oral presentation prior to the poster session.
- These talks provide an excellent opportunity to explain the **highlights** of your research to the entire conference and to **advertise** your poster presentation.

Fast-forward talk format

- **1 slide only**
- No additional slides – but props encouraged!
- Your slide must be **pre-loaded** on the morning of the fast forward session – if not your presentation will be skipped.
- **2 sessions:** Tuesday Nov. 14th and Wednesday Nov. 15th at 11.45pm. Same day as your poster session. Check which day you are presenting!
- SAHMRI Auditorium.
- Strict **90 secs** talking time – no questions.
- There will be **prizes!**

F1000Research

We invite you to deposit (for free) your posters and slides that you presented at the ABACBS-2017 National Conference to the ABACBS collection through the quick and simple online submission form on the page: https://f1000research.com/collections/ABACBS/for-authors/publish-your-research. ABACBS are committed to open access and would like to bring the attention of a much wider audience to the work you presented. Deposition will extend the visibility of you work and enable all researchers with an interest to contact you and ask questions even if they are unable to attend the meeting in person. We prefer to receive all submissions by the 15th December, although you will still be able to deposit your poster(s)/slides after this date should you be unable to make the deadline. If you have any questions about this invitation, then please do not hesitate to contact the F1000Research team at research@f1000.com.

Combine Student Symposium 2017 Program

Monday November 13		
8:00am - 8:50am	**Registration Open**	SAHMRI Entrance Atrium
8:50am - 9:00am	Symposium Welcoming Address	SAHMRI Auditorium
9:00am - 10:30am	Session 1	SAHMRI Auditorium
10:30am - 11:00am	**Morning Tea**	SAHMRI Entrance Atrium
11:00am - 12:30pm	Session 2	SAHMRI Auditorium
12:30pm - 1:45pm	**Lunch and Poster Session**	SAHMRI Entrance Atrium
1:45pm - 2:00pm	Group Photo	TBA
2:00pm - 3:15pm	Session 3	SAHMRI Auditorium
3:15pm - 5:15pm	Career Panel and Afternoon Tea	SAHMRI Auditorium
5:15pm - 5:20pm	**Awards and Closing Address**	SAHMRI Auditorium
5:00pm - 6:00pm	ABACBS Registration	SAHMRI Entrance Atrium
6:00pm - 7:00pm	ABACBS Keynote Lecture	SAHMRI Auditorium
7:00pm - 10:00pm	**Combine and ABACBS Social Events**	TBA

ABACBS 2017 Program

Monday November 13		
9:00am - 5:30pm	COMBINE Student Symposium	SAHMRI Auditorium
5:00pm - 6:00pm	ABACBS Registration Open	SAHMRI Entrance Atrium
6:00pm - 6:10pm	Opening Remarks - Assoc. Prof. David Lynn - Chair ABACBS 2017 & SAHMRI Exec. Director Prof. Steve Wesselingh	SAHMRI Auditorium
6:10pm - 7:10pm	Opening Keynote Talk - Prof. Des Higgins (University College Dublin) Everything you ever wanted to know about multiple alignments	SAHMRI Auditorium
7:10pm - 7:30pm	ABACBS Honorary Senior Fellow Awards	SAHMRI Auditorium
8:00pm - 10:00pm	COMBINE, ABACBS Postdoc, & ABACBS Professional Bioinformaticians Social Events	COMBINE @ The Edinburgh Castle Hotel 233 Currie St; Professional Bioinformaticians @ Cumberland Arms Hotel 205 Waymouth St; ABACBS Postdocs @ Duke of York Hotel 82 Currie St

Tuesday November 14		
8:00am - 8:30am	ABACBS Registration Open	SAHMRI Entrance Atrium
8:30am - 8:45am	Welcome Address	SAHMRI Auditorium
8:45am - 10:30am	ABACBS Session 1: Personalised Medicine & Cancer Genomics (Chair Andreas Schreiber)	SAHMRI Auditorium
8:45am - 9:45am	Day 1 Keynote Talk - Prof. Su-In Lee (University of Washington) From Big Data to Precision Medicine using Interpretable Machine Learning	SAHMRI Auditorium

9:45am - 10:15am	Selected Long Talk: Abstract #13 Nicola Roberts (WTSI) Patterns of somatic genome rearrangement in 2500 human cancers.	SAHMRI Auditorium
10:15am - 10:30am	Selected Short Talk: Abstract #3 Marek Cmero (U. Melb.) SVclone: inferring structural variant cancer cell fraction.	SAHMRI Auditorium
10:30am - 11:00am	Morning Tea/Coffee	SAHMRI Entrance Atrium
11:00am - 12:00pm	ABACBS Session 2: Natural Language Processing (Chair Steve Pederson)	SAHMRI Auditorium
11:00am - 11:45am	Prof. Karin Verspoor (University of Melbourne). Examining biological database quality through the lens of the scientific literature.	SAHMRI Auditorium
11:45am - 12:00pm	Fast Forward poster presentations (Chair Dr. Jimmy Breen)	SAHMRI Auditorium
12:00pm - 2:00pm	Lunch & Poster Session 1	SAHMRI Entrance Atrium
1:00pm - 2:00pm	Illumina Workshop: NovaSeq® The New Era in Sequencing. From Big Data to Bigger Breakthroughs!	SAHMRI Auditorium
2:00pm - 3:30pm	ABACBS Session 3: Single Cell Genomics (Chair Katherine Pillman)	SAHMRI Auditorium
2:00pm - 2:45pm	Dr. Joseph Powell (University of Queensland) Using ultra high-throughput single cell sequencing to understand cellular heterogeneity	SAHMRI Auditorium
2:45pm - 3:00pm	Selected Short Talk: Abstract #4 Joshua Ho (VCCRI) Fast and scalable analysis of single-cell RNA-seq data.	SAHMRI Auditorium
3:00pm -3:15pm	Selected Short Talk: Abstract #113 David Koppstein (UNSW) VDJPuzzle: A computational method for BCR and TCR reconstruction from single-cell sequencing data.	SAHMRI Auditorium
3:15pm - 3:30pm	COMBINE best speaker talk	SAHMRI Auditorium
3:30pm - 4:00pm	Afternoon Tea/Coffee	SAHMRI Entrance Atrium
4:00pm - 5:30pm	ABACBS Session 4 - Clinical Genomics (Chair Karin Kassahn)	SAHMRI Auditorium
4:00pm - 4:30pm	Selected Long Talk: Abstract #72 Ismael Vergara (Peter Mac) Large-scale chromosomal changes dominate the genomic landscape of end-stage melanoma.	SAHMRI Auditorium

4:30pm - 4:45pm	Selected Short Talk: Abstract #49 Patricia Graham (U. Tas). Whole exome sequencing and linkage analysis of extended pedigrees to identify glaucoma susceptibility genes	SAHMRI Auditorium
4:45pm - 5:15pm	Andrew Lonie (EMBL-ABR) Plans for an Australian Bioscience Data Cloud	SAHMRI Auditorium
5:30pm - 8:00pm	Wine Reception & Conference Dinner: choose from the Moorish Bites (Moroccan) or Sookii La La (Asian Fusion) food trucks (3 courses). GF and Vegan options available. Available only with pre-purchase of a conference dinner ticket - sold out!	SAHMRI Entrance Atrium & Auditorium

Wednesday November 15		
9:00am - 10:30am	ABACBS Session 5 - Molecular Evolution (Chair Ute Baumann)	SAHMRI Auditorium
9:00am - 9:45am	Prof. Gavin Huttley (Australian National University). Statistical methods for detecting and exploiting mutation motifs	SAHMRI Auditorium
9:45am - 10:15am	Selected Long Talk: Abstract #1: Kevin Downard (UNSW). A Mass Perspective of Molecular Evolution	SAHMRI Auditorium
10:15am - 10:30am	Selected Short Talk: Abstract #12 Polina Yu (VIB) Adaptation to the whole genome duplications in plant and animal systems.	SAHMRI Auditorium
10:30am - 11:00am	Morning Tea/Coffee	SAHMRI Entrance Atrium
11:00am - 12:00pm	ABACBS Session 6 - Systems/Network Biology (Chair Alan Rubin)	SAHMRI Auditorium
11:00am - 11:30am	Selected Long Talk: Abstract #15 Sriganesh Srihari (SAHMRI). Adaptive rewiring of protein-protein interactions and signal flow in the EGFR signalling network in RAS mutated colorectal cancer cells.	SAHMRI Auditorium
11:30am - 11:45 am	Selected Short Talk: Abstract #104 Monika Mohenska (Monash) A Systems Biology Approach to Investigate SRSF7 Functions in RNA Regulation of Autism Spectrum Disorder.	SAHMRI Auditorium

11:45am - 12:00pm	Fast Forward poster presentations (Chair Dr. Jimmy Breen)	SAHMRI Auditorium
12:00pm - 2:00pm	Lunch & Poster Session 2	SAHMRI Entrance Atrium
1:00pm - 2:00pm	ABACBS AGM	SAHMRI Auditorium
2:00pm - 3:30pm	ABACBS Session 7 - Neurogenomics (Chair Atma Ivancevic)	SAHMRI Auditorium
2:00pm - 2:45pm	Prof. Melanie Bahlo (Walter & Eliza Hall institute of Medical Research). Detecting known repeat expansions with standard protocol next generation sequencing, towards developing a single screening test for neurological repeat expansion disorders.	SAHMRI Auditorium
2:45pm - 3:15pm	Selected Long Talk: Abstract #61 Ellis Patrick (U. Sydney) Deconstructing a molecular network of the aging frontal cortex.	SAHMRI Auditorium
3:15pm - 3:30pm	Selected Short Talk: Abstract #97 Benjamin Goudey (IBM) A blood-based signature of cerebral spinal fluid AB_1-42 status.	SAHMRI Auditorium
3:30pm - 4:00pm	Afternoon Tea/Coffee	SAHMRI Entrance Atrium
4:00pm - 5:00pm	ABACBS Session 8 - Animal & Plant Omics (Chair John Williams)	SAHMRI Auditorium
4:00pm - 4:30pm	Selected Long Talk: Abstract #45 Terence Speed (WEHI) Direct Determination of Mouse Genome-Wide, Allele-specific DNA Methylation from Nanopore Long-Read Sequencing.	SAHMRI Auditorium
4:30pm - 4:45pm	Selected Short Talk: Abstract #18 Nikeisha Caruana (La Trobe). A slimy situation: Using de novo 'Omics and computational methods to identify the biochemical and biophysical properties of the slime of the striped pyjama squid, Sepioloidea lineolata.	SAHMRI Auditorium
4:45pm - 5:00pm	Selected Short Talk: Abstract #27 Alicia Oshlack (MCRI) Using superTranscripts for RNA-seq analysis in cancer and non-model organisms.	SAHMRI Auditorium
End of ABACBS Meeting		

Thursday November 16		
9:00am - 5:00pm	Using Open Science in Bioinformatics Training	SAHMRI Level 8 Boardroom
9:00am - 5:00pm	Bioconductor Masterclass	SAHMRI Level 4 Boardroom

Friday November 17		
9:00am - 5:00pm	Train-the-Trainer workshop	SAHMRI Level 8 Boardroom
9:00am - 5:00pm	Third Asia-Pacific Bioconductor Meeting	SAHMRI Level 4 Boardroom

ABACBS 2017 Poster Session 1: Tuesday 14 November 12:00pm - 2:00pm

Abstract #	Title	Poster Board #
2	Genome-wide study of 10,539 cancer samples reveals 27 novel associations between mutational processes and somatic driver mutations	1
10	Comparative genomics suggest Mycobacterium ulcerans migration and expansion is aligned with rise of Buruli ulcer in south-east Australia.	2
17	Visualisation and analysis of spatially-resolved transcript data using InsituNet	3
26	Understanding the Mechanism of Action of In-feed Antibiotics for Chicken	4
31	STRetch: detecting and discovering pathogenic short tandem repeat expansions	5
32	Cloud-based single-cell transcript reconstruction using Falco	6
43	Physical coherence and network analysis to identify novel regulators of exosome biogenesis.	7
46	A pan cancer hypoxic gene signature – highlighting temporal changes that lead to poor patient survival.	8
51	Optimising intrinsic protein disorder prediction for short linear motif discovery	9
54	A dynamical systems simulator to evaluate methods for inferring co-expression networks	10
56	Multi-omic Characterisation of a Novel Xylose Metabolising Strain of Saccharomyces cerevisiae	11
58	VPOT: a customisable tool for the prioritisation of annotated variants.	12
62	Spatial statistics analysis of super-resolution protein co-localization data	13
64	Predicting the outcome of breast cancer using novel RNA-Seq analysis	14
66	The causative role of Serine and Glycine on Macular Telangiectasia - a Mendelian randomization approach	15
82	SeqScrub: A web tool for automatic cleaning of FASTA file headers.	16

Abstract #	Title	Poster Board #
105	Reference-free methods for genomic prediction and selection	17
36	Comparative analysis of phosphoethanolamine transferases involved in polymyxin resistance across ten clinically relevant Gram-negative bacteria	18
19	Epigenetic Differential DNA Methylation Analysis in Monozygotic Twins Discordant for Depression	19
23	Analysis of melanoma data with a mixture of survival models, utilising multiclass DQDA to inform mixture class	20
24	bcGST - an interactive bias-correction method to identify over-represented gene-sets in boutique arrays	21
42	Investigating computational analysis pipelines and genomic proximity interactions in T lymphocytes	22
69	Utilising mixture models for unveiling patterns in scRNA-Seq data	23
83	Small data bioinformatics: identifying leaderless secretory proteins in plant cell walls with limited sample data	24
89	MicroRNA Regulatory Networks in Cancer Progression	25
96	SWATH-MS Spectral Reference Library Species Conversion with the R Package "dialects"	26
98	Precision Medicine: A clinical perspective on genome data	27
99	Integrative Analysis of Lipid Metabolic Pathways in Prostate Cancer Reveals DECR1 as a Key Cancer-Related Gene that Promotes Tumour Cell Survival	28
5	Examining differences in genome wide chromatin architecture with Hi-C	29
6	Simulation and analysis tools for single-cell RNA sequencing data	30
8	Confident effect sizes controlling FDR provide an ideal ranking of differentially expressed genes	31
9	Detecting cell types reliably with dtangle	32
11	Search for Glioma Direct Binding Site of Alkaloids using PLANTS	33

Abstract #	Title	Poster Board #
20	t-SNE Generated Transcriptome Landscapes Reveal Native Gene Expression Wiring.	34
21	Identifying Positive Selection Associated with Antimalarial Drug Resistance in Plasmodium falciparum using Identity-By-Descent Analysis	35
25	Characterising blood gene expression as a function of gut microbiota composition in preterm infants	36
28	Ximmer: Getting the best out of CNV detection on Exomes	37
29	Creating and exploiting Gene networks	38
30	Targeted Search for Genomic Variants for Clinical Applications	39
33	The histone variant H2A.Z is a master regulator of the epithelial-mesenchymal transition	40
34	Glimma: getting greater graphics for your genes	41
35	Bioinformatic challenges in the analysis of CLIP Experiments	42
37	Reliably Detecting Clinically Actionable Variants Requires Combined Variant Call	43
38	Predictors of damage transition in systemic lupus erythematosus	44
39	IVAT and VariantGrid: integrative annotation and analysis of genetic variants from next-generation sequencing data	45
40	RNA editing in an editing deficient Adar1 mouse model	46
41	Finding optimal regulatory element classifiers using automatic machine learning	47
44	Comprehensive benchmarking of short read structural variant callers	48
47	Bypassing the pseudogenes – How to diagnose with un-mappable genes	49
48	miR-200 regulates widespread changes in alternative splicing by controlling Quaking	50

ABACBS 2017 Poster Session 2: Wednesday 15 November 12:00pm - 2:00pm

Abstract #	Title	Poster Board #
52	Clinker: visualising fusion genes detected in RNA-seq data	1
53	Investigating the evolution of complex novel traits using whole genome sequencing and molecular palaeontology	2
55	Statistical inference in single-cell lineages	3
57	Annotating single-cell RNAseq clusters by similarity to reference single-cell datasets	4
59	PacBio sequencing, de novo assembly and haplotype phasing of diploid yeast strains	5
60	CNVmineR: A novel approach to identification of common and rare CNVs in WGS population studies	6
63	Galaxy training for microbial genomics	7
65	Family–based whole exome sequencing study of childhood apraxia of speech provides insight into the genetic basis of speech disorders.	8
67	People Powered Protein Predictions!	9
68	Gut microbiome changes in T1D during pregnancy and early life	10
70	DECENT: Differential Expression with Capture Efficiency AdjustmeNT for Single-Cell RNA-seq Data	11
71	Molecular Dynamics Modelling of a Variant of Unknown Effect in RAD51D	12
73	CAVALIER: an R package to produce reports for variant interpretation in clinical meetings	13
74	Transcriptome assembly and population differentiation analysis in Echinometra sea urchins, subjected to elevated pCO2 at volcanic vents	14
75	Using single cell RNA-Seq profiles to study heterogeneities in mouse mammary gland	15

Abstract #	Title	Poster Board #
76	Copy number variations from RNA-seq gene expression data	16
77	Performance of analysis software on TCGA exomes	17
78	The Monash Bioinformatics Platform (MBP) – Making bioinformatics accessible.	18
79	Signature-based binning for metagenomic analysis	19
80	Building user friendly applications for biologists	20
81	The Effect of Binding on the Enantioselectivity of an Epoxide Hydrolase	21
84	Using Singular Value Decomposition to alleviate batch effects in RNA sequencing data	22
85	Learning Epistatic Interactions from Sequence-Activity Data to Predict Enantioselectivity	23
86	Development of computational methods to analyse single-cell high-dimensional mass spectrometry data	24
87	Calling variants from RNA-Seq data identifies significant eQTLs in a small cohort of asthma patients	25
88	Identification of epigenetic complexes driving haematopoiesis	26
90	Reducing the impact of batch effects in single-cell RNA-Seq by imputing using Expectation Maximization algorithm	27
91	Integrating RNA, miRNA, DNA methylation and histone modification data uncovers biologically significant TGF-B-induced gene regulation	28
92	Galaxy Training Network - Training Material Repository	29
93	Resolving cardiac stromal-cell diversification through single-cell transcription profiling	30
95	Data Visualisation for Clinical diagnosis	31
100	HIV-1 RNA structure heterogeneity in cells	32

Abstract #	Title	Poster Board #
101	Usage of VLSCI Compute Resources by the Life Sciences Research Community	33
102	scPipe: a flexible data preprocessing pipeline for single-cell RNA-sequencing data	34
103	Identification and characterization of new cell populations using droplet-based single-cell RNA-seq: an example on breast cancer T cell infiltrate	35
106	Systems biology framework as an integrative approach in psychiatric research	36
108	UMID-dedup: A flexible tool for marking duplicate reads in NGS data using unique molecular identifiers	38
110	Learning representations for sequence comparison	39
111	An improved geometric measure for comparison of biological sequences	40
112	TRNDiff: Large-scale Visualisation for Transcriptional Regulation	41
114	Discovering a mechanism of drug-resistance in P. falciparum	42
115	Direct transcriptional regulation by microRNAs	43
116	Integrative genomic assessment of advanced CML reveals widespread genomic instability mediated by the recombination activation gene (RAG) pathway	44

Fast Forward Session 1:
Tuesday 14 November 11:45pm – 12:00pm

Abstract #	Presenting Author
28	Simon Sadedin
8	Paul Harrison
44	Daniel Cameron
6	Luke Zappia
37	Matt Field
43	David Chisanga
66	Roberto Bonelli
69	Shila Ghazanfar
46	Kristy Horan
40	Alistair Chalk

Fast Forward Session 2:
Wednesday 15 November 11:45pm – 12:00pm

Abstract #	Presenting Author
93	Ralph Patrick
108	Paul Wang
88	Yih-Chih Chan
73	Mark Bennett
60	Jacob Munro
80	Jarny Choi
90	Alexander Hayes
106	Liliana Ciobanu
57	Sarah Williams
115	Klay Saunders

ABACBS 2017 Satellite Events - Thursday 16 November and Thursday 17 November

Using Open Science in Bioinformatics Training

Venue: SAHMRI Level 8 Boardroom, Adelaide, Australia
Date: November 16 2017 (9:00am – 5:00pm)
The Using Open Science in Bioinformatics Training day will cover aspects of open science in bioinformatics training and will consist of invited and contributed talks and panel discussions for more informal interactive sessions.

Timetable Outline

9.00am - 10.30am	Introduction and finding online training material (Annette McGrath, CSIRO Data61) Using Standards for greater reuse of material (Sonika Tyagi, Monash University)
10:30am - 11:00am	Morning Tea
11:00am - 12:30pm	Open access and collaboration resources for training (Paul Harrison & Adele Barugahere, Monash University)
12:30pm - 1:30pm	Lunch
1:30pm - 3:00pm	Version Control and Github (Nathan Watson-Haigh, Uni Adelaide)
3:00pm - 3:30pm	Afternoon Tea
3:30pm - 5:00pm	Licensing your material and what you need to know (Pip Griffin, University of Melbourne)

Train-the-Trainer workshop

Venue: SAHMRI Level 8 Boardroom, Adelaide, Australia
Date: November 17 2017 (9:00am – 5:00pm)
The Train-the-Trainer day will consist of a small group workshop for those who are interested in learning more about how people learn, what makes a good workshop and how to design a short course to teach to others.

Timetable Outline
Trainers: Annette McGrath (CSIRO Data61), Ann-Marie Patch (QIMR)

9.00am - 10.30am	Good vs bad training and trainers
10:30am - 11:00am	Morning Tea
11:00am - 12:30pm	How do people learn?
12:30pm - 1:30pm	Lunch
1:30pm - 3:00pm	How to design a good session and a good course
3:00pm - 3:30pm	Afternoon Tea
3:30pm - 5:00pm	Assessment and gathering feedback

This workshop will be delivered by instructors trained to deliver a Train-the-Trainer workshop by the European Bioinformatics Institute.

Bioconductor Masterclass

Venue: SAHMRI Level 4 Boardroom, Adelaide, Australia
Date: 16th November 2017 (9:00am – 5:00pm)
Overview: High-throughput sequencing technologies generate large volumes of data that present challenging bioinformatic and statistical problems. This series of hands-on tutorials introduces established and new R/Bioconductor packages and workflows for analysing high-throughput sequence data.

Course Outline

9.00am - 10.30am	Bioconductor Basics (Martin Morgan, Roswell Park Cancer Institute)
10:30am - 11:00am	Morning Tea
11:00am - 12:30pm	RNA-seq analysis with limma, Glimma and edgeR (Charity Law, The Walter and Eliza Hall Institute of Medical Research)
12:30pm - 1:30pm	Lunch
1:30pm - 2:30pm	Gene-set testing for RNA-seq data with EGSEA (Monther Alhamdoosh, CSL Limited)
2:30pm - 3:30pm	Single cell RNA-seq analysis in Bioconductor (Shian Su, The Walter and Eliza Hall Institute of Medical Research)
3:00pm - 3:30pm	Afternoon Tea
3:30pm - 5:00pm	Package Development in R/Bioconductor / Data Visualisation (Martin Morgan, Roswell Park Cancer Institute)

Prerequisites: The workshop assumes an intermediate level of familiarity with R, and basic understanding of biological and technological aspects of high-throughput sequence analysis. Participants should come prepared with a modern wireless-enabled laptop and web browser installed.

Intended Audience: This workshop is for bioinformaticians and computational biologists interested in using R/Bioconductor for the analysis and comprehension of high-throughput sequence data.

Third Asia-Pacific Bioconductor Meeting

Venue: SAHMRI Level 4 Boardroom, Adelaide, Australia
Date: 17th November 2017 (9:00am – 5:00pm)

Overview: The Bioconductor project has undergone significant growth over the past 16 years, with over 1,300 packages for high-throughput genomic analysis included in the latest release. Many popular software tools from the project are developed by researchers from the Asia-Pacific region. To enhance collaboration and provide an avenue for networking, the Third Bioconductor Asia-Pacific meeting will be held following the ABACBS 2017 conference. This event aims to bring together researchers with an interest in the Bioconductor project to provide a forum for exchanging ideas and future plans for software development. We welcome attendance from both users and package developers (current and prospective). The meeting will consist of a number of longer talks selected from abstracts, short 'lightning' presentations and software demonstrations to maximize the opportunities for participants to highlight their work and share expertise.

Organisers

- Dr Matt Ritchie, Molecular Medicine Division, The Walter and Eliza Hall Institute of Medical Research, Melbourne, Australia
- Dr Charity Law, Molecular Medicine Division, The Walter and Eliza Hall Institute of Medical Research, Melbourne, Australia
- Mr Stephen Pederson, Bioinformatics Centre, The University of Adelaide, Adelaide, Australia

Timetable

9.00am - 10.00am	Project Updates **(Professor Martin Morgan)**
10:00am - 10:30am	Lightning talks **Stuart Lee -** plyranges: a fluent interface to Bioconductor's Ranges infrastructure **Charity Law -** Differential splicing analysis in limma with diffSplice **Anna Quaglieri -** One year of R-Ladies Melbourne **Shian Su -** scPipe: a preprocessing pipeline for single-cell RNA-seq data **Yunshun Chen -** Analysing Fluidigm C1 and 10X Chromium single cell RNA-seq data using edgeR
10:30am - 11:00am	Morning Tea
11:00am - 12:30pm	Research talks / Software demos Session I **Thuc Le -** A dry lab for exploring miRNA functions and applications in cancer subtype discovery **Hannah Coghlan -** The search for differential interactions: differential analysis of Hi-C data with diffHic **Jacob Munro -** CNVmineR: A novel approach to identification of common and rare CNVs in WGS population studies **Maria Doyle -** Turning Bioconductor workflows into Galaxy Tools
12:30pm - 1:30pm	Lunch
1:30pm - 2:20pm	Keynote: **Professor Jean Yang**

2:20pm - 3:40pm	Research talks / Software demos Session II **Luke Zappia -** Splatter, a package for simulating single-cell RNA sequencing data **Alex Garnham -** Pipeline to detect viral RNA in human RNA sequencing data **Stefan Mutter -** Numero: A statistical framework to define multivariable subgroups in complex biomedical datasets **Stephen Pederson -** ngsReports: An R Package for managing FastQC reports and other NGS related log files
3:40pm - 4:10pm	Afternoon Tea
4:10pm - 5:00pm	Discussion and closing remarks

Abstracts

1 A Mass Perspective of Molecular Evolution

Kevin Downard and Elma Akand

The University of New South Wales

Molecular based approaches to phylogenetic analysis, driven by technological advances in gene or whole genome sequencing and computational analysis, have revolutionized our view of evolution. More recently there has been a focus on studying more functional traits of organisms with an emphasis on predicting protein function. We have developed a novel mass (or numbers) based phylogenetic approach that can applied to study the evolution of any organism from the protein perspective using datasets commonly employed in proteomics. When a mutation in a protein gives rise to another, its molecular mass changes. This mass difference can be localised within peptide segments following the digestion of the protein. For most part these differences correspond to unique values, when high mass resolution is employed, so as to enable the nature of the amino acid substitution to be identified. We have advanced a Mass Tree phylogenetic approach, which avoids the need for gene or protein sequences, to identify and chart protein mutations associated with the evolution of the organisms in which they are expressed. The modified MassTree algorithm identifies and displays all such mutations and calculates the frequency of a particular mutation across a tree. Its significance in terms of its position(s) on the tree is scored, where mutations that occur toward the root of the tree are weighted more favourably. A comparison with data generated from a conventional sequence based tree demonstrates the reliability of mutational analysis employing this approach. This presentation will demonstrate application of the approach to study mutation trends and patterns involved in the evolution of the influenza virus, particularly in relation to epistatis. The method, however, has far broader applicability and can be applied to investigate the evolution of any organism at the molecular protein level.

2 Genome-wide study of 10,539 cancer samples reveals 27 novel associations between mutational processes and somatic driver mutations

Yuen Ting Wong, Rebecca C. Poulos, Regina Ryan and Jason W. H. Wong

Prince of Wales Clinical School, Faculty of Medicine, UNSW Sydney, NSW, 2052

Driver mutations are the genetic variants responsible for oncogenesis, but how these somatic events occur remains poorly understood. Mutational signatures represent trinucleotide frequencies of somatic mutations in a sample, and these can provide an avenue for investigating the mutational processes operative in a given cancer. Here, we analysed somatic mutation data from 10,539 cancer exomes from The Cancer Genome Atlas (TCGA). Using 252 known cancer driver mutations, we performed regression analyses to establish the statistical relationship between driver mutations and mutational signatures across 22 cancer types. Our analyses led to 37 significant associations between driver mutations and mutational signatures ($P < 0.001$), of which 27 are novel associations. As proof of concept, our findings implicate the POLE P286R mutation in driving the mutational landscape of uterine cancer associated with signature 10. We found 30% ($n = 11$) of our significant associations to occur in uterine cancer, and another 30% in colorectal cancer. In addition, we found BRAF V600E mutations to be associated with mutational signatures in 3 different cancer types. Most interestingly, none of these associations are with signature 7 which is a hallmark ultraviolet (UV) light mutagenesis, suggesting that BRAF V600E mutations may develop independently of UV light exposure in skin cancers. We found a total of 16 mutational signatures to be statistically associated with at least one driver mutation, including the APOBEC enzyme-associated signature 2, which was significantly associated with six driver mutations within oncogene PIK3CA. Finally, we observed a negative association between IDH1 R132H and the age-associated signature 1, suggesting that age does not contribute to the formation of this driver event. Our study has uncovered previously unknown relationships between driver mutations and mutagenic processes during cancer formation which can improve our understanding of how cancer develops, and provide new avenues for investigating cancer preventative strategies.

3 SVclone: inferring structural variant cancer cell fraction

Marek Cmero, Cheng Soon Ong, Ke Yuan, Jan Schroeder, Kangbo Mo, Niall Corcoran, Tony Papenfuss, Chris Hovens, Florian Markowetz and Geoff Macintyre

The University of Melbourne

Cancers arise from single progenitor cells that acquire mutations, eventually differentiating into mixed populations with distinct genotypes. These populations, corresponding to mutational profiles at common prevalences, can be estimated using computational techniques applied to bulk samples. Existing methods have largely focused on single nucleotide variants (SNVs), despite growing evidence of the importance of structural variation (SV) in shaping the development of molecular subtypes of cancer. While some approaches use copy-number aberrant SVs, no method has incorporated balanced rearrangements. To address this gap, we present SVclone, a computational pipeline for inferring the cancer cell fraction (CCF) of SV breakpoints from whole-genome sequencing data. SVclone uses heuristic read counting, comprehensive filtering and a Bayesian inference clustering approach to infer SV CCF. We validate our approach using simulated and real tumour samples, and demonstrate its utility on 2,778 whole-genome sequenced tumours from the pan-cancer analysis of whole genomes project. We find a subset of cases with decreased overall survival that have subclonally enriched copy-number neutral rearrangements, an observation that could not have been otherwise discovered with currently available methods.

4 Fast and scalable analysis of single-cell RNA-seq data

Joshua Ho

Victor Chang Cardiac Research Institute

Single-cell RNA sequencing (scRNA-seq) enables researchers to interrogate the genome-wide expression profile of tens of thousands of individual cells – and very soon millions of cells. This rapidly growing data size poses a fundamental challenge in terms of scalability of the scRNA-seq bioinformatics analysis. In this talk, I will discuss two recently published bioinformatics tools from our laboratory that deal with the scalability issue - Falco and CIDR. Falco is a cloud-based framework to enable parallelisation of existing RNA-seq processing pipelines using big data technologies of Apache Hadoop and Apache Spark for performing massively parallel analysis of large scale transcriptomic data. CIDR is an ultrafast dimensionality reduction and clustering tool that alleviates the impact of dropouts using a novel 'implicit imputation' approach.

5 Examining differences in genome wide chromatin architecture with Hi-C

Hannah Coughlan, Timothy Johanson, Aaron Lun, Stephen Nutt, Rhys Allan and Gordon Smyth

The Walter and Eliza Hall Institute of Medical Research

Although chromatin is traditionally viewed in a linear sense, recently it has been recognised that the higher order organisation of the chromatin is extremely relevant to biological function by regulating gene expression, DNA replication and repair, and recombination. We used chromatin conformation capture with high-throughput sequencing (Hi-C) combined with differential analysis to examine 3D chromatin structure at the genome wide level in differentiated immune cells. Of particular interest in the project was to understand how multiple cell types (T cells, B cells and granulocytes) originate from a common lymphoid progenitor. We used the statistically robust R package diffHic to examine whether each cell lineage possess a distinct genome organization. diffHic can identify significant changes in chromatin interaction frequency between biological conditions. Additionally, we explored the role that 3D chromatin architecture plays in B cell differentiation and the effect of the transcription factor PAX5 on establishing and maintaining the genome architecture of B cells.

6 Simulation and analysis tools for single-cell RNA sequencing data

Luke Zappia, Belinda Phipson and Alicia Oshlack

(1) Bioinformatics, Murdoch Children's Research Institute (2) School of Biosciences, University of Melbourne

Single-cell RNA sequencing (scRNA-seq) has opened up a range of opportunities but with the dramatic increase in resolution 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 if methods designed for bulk RNA-seq are appropriate for scRNA-seq data. Researchers have risen to address these challenges and there are now more than 140 scRNA-seq analysis tools available, with the majority released under open-source software licenses. We have catalogued these software tools in the scRNA-tools database (www.scRNA-tools.org). Analysis of this database shows that there are methods available for completing a wide range of analysis tasks, with the biggest areas of development being in clustering cells to identify cell types and ordering of cells to understand dynamic processes. With an ever increasing number of analysis tools available researchers are faced with the challenge of choosing which to use, making it important to be able to assess and compare the performance, quality and limitations of each tool. One common approach is to test methods on simulated datasets where the true answers are known. To aid this process we have developed Splatter, a Bioconductor R package for reproducible simulation of scRNA-seq datasets (bioconductor.org/packages/splatter). Splatter is a simulation framework that provides access to a variety of simulation models, allowing users to estimate parameters from real data in order to easily generate realistic synthetic scRNA-seq datasets. As part of Splatter we also introduce our own simulation model, Splat, capable of reproducing scRNA-seq datasets with multiple groups of cells, differentiation paths or batch effects. Here we discuss some of the trends in scRNA-seq analysis and how Splatter can be used to develop and compare analysis tools.

7 Transcriptomic and proteomic characterisation of a zebrafish model of familial Alzheimer's disease

Nhi Hin, Morgan Newman, Michael Lardelli and Stephen Pederson

University of Adelaide

Identifying the earliest molecular events in Alzheimer's disease pathogenesis is critical for understanding how and why Alzheimer's disease develops. Although the molecular changes in post-mortem brains afflicted with Alzheimer's disease have been characterised with technologies like whole-transcriptome sequencing (RNA-seq), the earliest changes in gene expression patterns that occur decades before Alzheimer's disease onset are still unknown. The brains of animal models of Alzheimer's disease can be theoretically studied at any age, but many animal models of Alzheimer's disease may not accurately model the physiological state of Alzheimer's disease due to overexpressing multiple mutant human familial Alzheimer's disease genes. Because of this, we created the first zebrafish model of a dominant familial Alzheimer's disease mutation in the orthologous zebrafish gene. In this study, we analysed RNA-seq and proteomic (LC-MS/MS) data from this zebrafish model to characterise the changes that distinguish their brains from those of normal aging in zebrafish. In addition, we applied weighted gene co-expression network analysis of the RNA-seq data to compare changes in gene expression in the zebrafish model to those from a human Alzheimer's disease dataset. By evaluating the similarities and differences in gene expression between the zebrafish model and human brains with Alzheimer's disease, there are opportunities to identify early molecular changes in Alzheimer's disease pathogenesis that might contribute to preventing or delaying its onset.

8 Confident effect sizes controlling FDR provide an ideal ranking of differentially expressed genes

Paul Harrison, Andrew Pattison, David Powell and Traude Beilharz

Monash Bioinformatics Platform, Monash University

A method is described for giving a confidence bound on the magnitude of Log Fold Change (LFC) in gene expression while controlling the False Discovery Rate (FDR). We propose this confidence bound as an ideal quantity by which to rank differentially expressed genes when presenting the results of an RNA-seq experiment. The method builds on the TREAT method of McCarthy and Smyth (2009). Unlike TREAT, a minimum LFC of interest does not need to be specified. The only parameter is the desired FDR, for which a reasonable default value can be given. Sorting by p-value is a common default in the output of differential expression software. We compare these two methods of ranking genes using a breast cancer RNA-seq data-set consisting of matched tumor-normal pairs. The top genes as ranked by p-value have small but consistent differential expression, whereas the top genes as ranked by confidence bound have a much larger magnitute of differential expression (but also higher variability). This leads to a difference in biological interpretation, with greater emphasis placed on genes related to the extra-cellular matrix by the confidence bound method. The confidence bound method degrades gracefully on subsets of samples in this data-set. For experiments with low statistical power, the ranking is similar to the p-value ranking, but as the power of an experiment increases the ranking is increasingly determined by the true effect size. Comparing the confidence bound and estimated LFC of top genes provides immediate feedback on whether or not an experiment was under-powered. An R package implementing the method is available at https://github.com/pfh/topconfects

9 Detecting cell types reliably with dtangle

Gregory Hunt, Saskia Freytag, Melanie Bahlo and Johann Gagnon-Bartsch
Walter and Eliza Hall Institute

Shifts in cell type composition are hallmarks of developmental processes of organisms including embryogenesis, morphogenesis, cell differentiation and growth. They can also be a sign of disease and dysfunction. Flow cytometry, which researchers typically employ to establish composition, is time consuming and expensive. It has thus rarely been used in combination with techniques measuring gene expression, despite cell type composition having the potential to confound results from differential gene expression analyses. Given the importance of understanding cell type composition, several methods estimating cell type proportions from gene expression data, also referred to as deconvolution, have been developed. We propose dtangle that can be used to deconvolve heterogeneous tissue using measurements from DNA microarray and bulk RNA-seq platforms. To evaluate dtangle, we assembled the largest collection of benchmark datasets with known cell-type proportions. Bench testing shows that dtangle is competitive with published deconvolution methods. In particular, dtangle has the lowest mean and median error of all the eight tested deconvolution algorithm across the benchmarking datasets highlighting that dtangle is working well across many technologies and tissue types. Furthermore, dtangle's accuracy and efficiency are robust to selection of its tuning parameters. Investigating a real life gene expression dataset on Lyme disease dtangle's estimated proportions reveal a temporal trend consistent with previous findings. Correspondingly, these proportions are important covariates in the associated differential expression analysis. dtangle is available as an R package at https://cran.r-project.org/web/packages/dtangle/index.html

10 Comparative genomics suggest Mycobacterium ulcerans migration and expansion is aligned with rise of Buruli ulcer in south-east Australia.

Andrew H. Buultjens, Koen Vandelannoote, Janet A. M. Fyfe, Maria Globan, Nicholas J. Tobias, Jessica L. Porter, Takehiro Tomita, Benjamin P. Howden, Paul D. R. Johnson and Timothy P. Stinear

The University of Melbourne

Over the past five years, cases of the neglected tropical disease Buruli ulcer have increased dramatically in specific areas around Melbourne (population 4.4 million) in the state of Victoria, a temperate region in south-east Australia. The reasons for this increase are unclear. Here we have used whole genome sequence comparisons on 184 M. ulcerans isolates obtained primarily from human clinical specimens, spanning 70 years, to model the population dynamics of this pathogen from this region. Using phylogeographic and Bayesian approaches, we found that there has been a westward migration of the pathogen from the east of the state, beginning in the 1980s, 300km west to the major human population centre around Melbourne. This move has then been followed by a significant increase in M. ulcerans population size. These analyses inform our thinking around Buruli ulcer transmission and control, indicating that M. ulcerans is introduced to a new environment and then expands, rather than the awakening of a quiescent pathogen reservoir.

11 Search for Glioma Direct Binding Site of Alkaloids using PLANTS

Yusnita Rifai

Hasanuddin University

This research aims to know the best affinity and the best chemical conformation of anticancer compounds from alkaloid groups that have closed-direction to Glioma-associated oncogene using PLANTS® (Protein-Ligand Ant System). The interaction energy and hydrogen bond are included as evaluated targets. In this research, twenty-seven ligands with RMSD (Root Mean Square Deviation) score at 1,614 Å and cyclopamine as a native ligand are used. Meanwhile, Staurosporinone acts as Glioma's directed-binding-site-internal-control. Each ligand is docked in GLI with PDB code 2GLI using 2 methods, GLI contains water and without water. PLANTS® score for the native ligand in the first and the second method is -73,9002 and -73,2700 respectively. Pancracristine, homoharringtonine, and sanguinarine showed PLANTS® score close to the cyclopamine score result but their hydrogen bond interaction differed from native ligand interaction. Evodiamine ligand has a good score and hydrogen bond to the same amino acid of protein GLI, which is GLU 175 and THR 173. This results indicated that evodiamine has the same identical mechanism as staurosporinone.

12 Adaptation to the whole genome duplications in plant and animal systems

Polina Yu. Novikova, Steve Donnellan, Yves Van de Peer and Levi Yant

VIB / Ghent University

Polyploidy plays important role in evolution, providing a 'backup' genetic material and increasing genetic novelty. However, polyploids have to adapt their cellular machinery to the whole genome duplication (WGD) itself. In the autotetraploids, crossovers may occur randomly between each copy during meiosis, compromising regular chromosomal segregation. Adapted to WGDs autopolyploid plants allow only one crossover per chromosome, which leads to successful meiosis. Although recent WGDs have been described in animals, they occur rarely and usually such animals reproduce asexually. Amphibia is the only exception among bisexually reproducing vertebrates with multiple independent occurrences of WGDs, for example, frog genus Neobatrachus consists of 6 diploid and 4 tetraploid species. This project aims to provide the first description of adaptation mechanism in animal kingdom using Neobatrachus frogs. We conducted a pilot experiment comparing nucleotide diversity and differentiation between two diploid N. pictus and one tetraploid N. sudelli mapping Illumina paired end data to the closest available reference genome Nanorana parkeri. Preliminary results showed that potentially selected genes in the tetraploid are enriched for molecular function in microtubule motor activity. This suggests modifications at the homologous pairing process during meiosis. To achieve further clarity on the mechanism of meiotic adaptation to WGD in Neobatrachus, we will assemble and annotate a genome of the diploid N. pictus. Next, we will compare genetic variation along the genome for diploids and tetraploids using pooled sequencing, identifying selected regions in tetraploids and potentially adaptive changes. Integration of the findings in Neobatrachus and plants will shed light on whether the mechanism of adaptation to the WGDs is universal or plants and animals use alternative ways.

13 Patterns of somatic genome rearrangement in 2500 human cancers

Nicola Roberts, Yilong Li, Rameen Beroukhim and Peter Campbell

Wellcome Trust Sanger Institute; University of Cambridge

Cancer development is driven by somatic genome alterations, ranging from single point mutations to larger structural variants (SV) affecting kilobases to megabases of one or more chromosomes. To date, studies of somatic rearrangement have been limited by a paucity of whole genome sequencing data, and a lack of methods for comprehensive structural classification and downstream analysis. The ICGC project on the Pan-Cancer Analysis of Whole Genomes provides an unprecedented opportunity to analyse somatic SVs at base-pair resolution in more than 2500 samples from 30 common cancer types. Here, I present a census of SV classifications across the pan-cancer cohort, including chromoplexy, replicative two-jump structures, and a novel description of templated insertion events connecting as many as eight distant loci. By identifying the precise structure of individual rearrangements and separating out complex clusters, our classification scheme allows for detailed exploration of all simple rearrangement signatures, incorporating size, genome property associations, and patterns of co-occurrence. We find that event size distributions are multi-modal, with striking differences between samples and cancer types. Looking across the genome, most SV classes are enriched in early-replicating, gene- and GC-rich DNA, with the exception of deletions that skew towards late-replicating, AT-rich regions. Bucking these general trends, SVs in some hypermutators exhibit markedly different biases. These findings, along with signatures of their co-occurrence, suggest even the simplest SV structures have distinct mechanistic subclasses. Rearrangement hotspots mark out fragile sites and certain cancer genes under natural selection. Fragile site deletion patterns are remarkably tissue-specific, with particularly high prevalence in gastrointestinal cancers. In exploring the patterns of SV classes around cancer genes, we find templated insertion cycles are a recurrent mechanism of TERT oncogene activation in liver cancer and RB1 tumour-suppressor disruption in breast and ovarian cancers.

14 Genome-wide SNPs modelling improved genetic risk prediction for psoriasis

Wei Lu and Gad Abraham

The University of Melbourne

Psoriasis, a chronic skin disease affecting ~2% of Western populations, is caused by multiple genetic factors. Genome-wide association studies (GWAS) have identified multiple psoriasis susceptibility loci (single nucleotide polymorphisms, SNPs). A simple model that counts the number of significantly associated risk alleles in an individual has been used to predict disease genetic risk. However, this model suffers from poor prediction accuracy since it doesn't include all potential SNPs that are associated with the disease. In this study, we applied penalised linear models to genome-wide SNPs for psoriasis genetic risk prediction. These models employs genome-wide SNPs and estimates their genetic effects simultaneously. To select the optimal model we compared different regression methods (ordinary least squares, least absolute shrinkage and selection operator (LASSO), Ridge) to estimate SNP effects. We also compared the prediction accuracy between models using genotyped SNPs and imputed SNPs. The prediction accuracy is measured by Pearson's correlation coefficient between observed disease risk and predicted genetic risk for case-control study. We tested the models using three large psoriasis datasets (~2500 cases and controls each) for model selection and validation. We found that models including genome-wide SNPs led to increases of ~10% in prediction accuracy compared with models using SNPs from chromosome 6 only. We also found including imputed SNPs into model increased prediction accuracy by a small amount. It suggests that genetic risk prediction should include modelling genome-wide SNPs for complex diseases. Prediction accuracy could also be improved by using alternative SNP effect estimation methods. We also expect that the statistical models can help researchers and clinicians get a better understanding of genetic causes of Psoriasis.

15 Adaptive rewiring of protein-protein interactions and signal flow in the EGFR signalling network in RAS mutated colorectal cancer cells

Sriganesh Srihari, Susan Kennedy, Mohammed-Ali Jarboui, Cinzia Raso, Kenneth Bryan, Theodosia Charitou, Priyanka Pillai, Carlos Montarez, Aleksandar Kristic, David Matallanas, Layal Dernayka, Igor Jurisica, Jasna Curak, Igor Stagljar, Thierry Lebihan, Lisa Imrie, Victoria Wong, Miriam A. Lynn, Erik Fasterius, Cristina Al-Khalili Szigyarto, Christina Kiel, Luis Serrano, Nora Rauch, Ruth Pilkington, Patrizia Cammareri, Owen Sansom, Steven Shave, Manfred Auer, Manuel Bernal-Llinares, Marius Ueffing, Karsten Boldt, David J Lynn and Walter Kolch
EMBL Australia Group, South Australian Health and Medical Research Institute

EGFR signalling plays a major role in colorectal cancer (CRC) and mutations in the GTPase KRAS abrogate the efficacy of anti-EGFR treatments. Here, we present high-quality protein-protein interaction network (PPIN) maps of the EGFR pathway in two isogenic CRC cell-lines, HCT116 and HKE3, which express different levels of oncogenic KRAS. To assemble the PPINs, 95 bait proteins, covering the main functional aspects of the EGFR pathway, were expressed in the two cell-lines and their interactors were detected using an AP-MS based approach. EGFRNet-HCT116 and EGFRNet-HKE3 consisted of 3,162 and 2,788 bait-prey interactions, respectively. ~83% of the interactions were not reported previously. More than 30% of interactions were significantly rewired between the two networks. The rewired interactions were non-uniformly distributed among the bait-prey complexes with the BAD, MAP2K1, RAC1, and SH2D3C complexes being the most rewired. The rewired preys were enriched for functions in RNA splicing, mitochondrial translation, protein folding, and cell migration, and several of the top rewired genes have been identified as synthetic lethal with mutant KRAS. ~10% of the rewired preys were differentially expressed at a protein level and 7.6% were differentially phosphorylated between the two cell lines, which may explain their rewiring. Information flow analysis predicted higher flow to downstream transcription factors (TFs) FOXO1 and MYC in EGFRNet-HCT116, and to STAT1 and FOS in EGFRNet-HKE3, which was confirmed by RNAseq and TF binding site analysis. The 20 most rewired baits were altered in 54% of TCGA CRC patients and these patients displayed significantly poorer survival at 10 years (35% compared to 61%). Our PPIN maps present a unique resource to quantify the ripple effect of the oncogenic KRAS mutation through the EGFR signalling network providing new insight into the cellular mechanisms underpinning the pathogenesis of CRC and other KRAS-mutated cancers. The data can be explored at http://primesdb.eu.

16 Using genome-wide variants to determine the historical migration of chickens through South East Asia to the Pacific Islands

Pei Qin Ng

The University of Adelaide

Chickens were domesticated from wild jungle fowls and were among the commensals transported during human migration eastwards across the Pacific Ocean. However, there is no detailed documentation of the exact origin of domesticated chickens, although several domestication centres have been suggested, based on archaeological and biomolecular evidence. A previous study by Thomson et al. (2014) on chicken mitochondrial DNA suggested the possible origin and ancestry of modern Pacific chickens to be from the South East Asia jungle fowl. As mtDNA is maternally inherited as a single locus, this study could not consider complex evolutionary events such as introgression. To confirm previous findings and identify possible gene flow, three wild and five domestic samples of four different Gallus spp. were sequenced using whole genome sequencing and analysed using a bioinformatics workflow. We confirm the hypothesis that the modern Pacific chickens originated from the Philippines Gallus gallus (red jungle fowl). Our results confirm the phylogeny of the wild species, with Gallus varius (green jungle fowl) basal to both G. lafayettii (Ceylon jungle fowl) and G. sonneratii (grey jungle fowl), consistent with findings of previous studies. Gene flow observed within the domestic chicken samples suggests a pattern of dispersal eastwards across the Pacific. Our analysis also suggests putative introgression of grey jungle fowl into domestic chickens. This could be explained through a recent common ancestor before the Pacific Island radiation that is not identified in this study. Our findings elucidate a possible migration pathway of the chickens, which can be used to infer the potential route of human dispersal to the Pacific Islands and its impact on the genetic diversity of chickens as a commensal. Understanding of the phylogenetic relation and further investigation on the underlying genetic variations between the wild and domesticated samples can provide information on improving the commercial chicken breeds.

17 Visualisation and analysis of spatially-resolved transcript data using InsituNet

John Salamon, Xiaoyan Qian, Mats Nilsson and David Lynn

EMBL Australia Group, South Australian Health and Medical Research Institute

Gene expression studies typically homogenise samples before sequencing, discarding spatial information on where transcripts are expressed. In contrast to this, in situ sequencing is a novel technique for generating spatially-resolved, in situ RNA localization and expression data that preserves the spatial context of transcripts. Gene-specific barcodes allow data for up to 40 different transcripts/genes at an almost single-cell resolution to be generated in situ, resulting in images that display the location and intensity of a million or more individual transcripts in a tissue section. Despite the obvious potential of in situ sequencing, few methods currently exist to analyse 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, I 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-expression relationships between transcripts. InsituNet identifies co-expressions that occur between transcripts both significantly more, and less, than statistically expected given their relative frequencies within the tissue to allow intelligent filtering of the resulting visualisations. 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. When multiple networks are constructed, their layouts may be spatially synchronised to facilitate comparison. Synchronisation allows one to easily observe how transcriptional relationships change across different tissue sections and conditions. InsituNet has been developed for the popular Cytoscape visualisation platform, and is available for download from within the Cytoscape app store.

18 A slimy situation: Using de novo 'Omics and computational methods to identify the biochemical and biophysical properties of the slime of the striped pyjama squid, Sepioloidea lineolata.

Nikeisha Caruana, Jan Strugnell, Pierre Faou, Julian Finn and Ira Cooke

La Trobe University

Cephalopods (squids, octopuses and cuttlefishes) comprise over 800 species, possess advanced nervous, cardiovascular and visual systems, and are masters of camouflage. Sepiadariidae, one family of benthic squids, possess specialised systems of secretion, secreting a viscous slime from their underside. It is believed that these secretions are involved in defending the species, and exhibit unique biochemical and biophysical properties including dramatic volume expansion, adhesion, and antimicrobial defence. In order to further understand the mechanisms behind the secretion of the slime and the proteins strictly involved in it, six tissues from four individuals of the striped pyjama squid (Sepioloidea lineolata), underwent label-free quantification. The proteomic results were combined with a de novo transcriptome and analysed using computational bioinformatics methods, including a proteogenomic pipeline developed in Galaxy. Seventy-eight proteins were highly differentially expressed within the slime and glands of S. lineolata, which included ten completely novel proteins. Gene Ontology term enrichment analysis also indicated a number of extracellular proteins involved in immune defence, catalytic peptidase activity, and protein binding. The identified proteins may have further implications for biomimetics such as the design of new pharmaceuticals and creation of non-toxic glues. A comparison of the proteins in the slime and glands found on the ventral epithelium of S. lineolata reveal that the slime does originate from these glands. While traditionally label-free quantification is not used to analyse complete proteomes on non-model organisms, our work illustrates that the combination of de novo assembled transcriptomes and comprehensive bioinformatic analysis can provide rapid and informative investigation into species, which have high potential for biological applications without the need for a full genome. Our work incorporates the first label-free quantification analysis of a cephalopod and its secretion, and is also the first recording of the origin of the slime of Sepiadariidae from the secretion of specialised glands.

19 Epigenetic Differential DNA Methylation Analysis in Monozygotic Twins Discordant for Depression

Yan Ren, Jimmy Breen, Stephen Pederson and Sarah Cohen-Woods

The University of Adelaide Bioinformatics Hub

Major depressive disorder (MDD) is a pervasive psychiatric disorder characterized by its symptoms that consist of persistent low mood, insomnia, anhedonia, a feeling of guilt and intention of suicide. Studying depression from the epigenetic aspect sheds light on its etiology by revealing how environmental factors regulate gene expression. The genetics of the individual however, often complicates the ability to uncover these epigenetic mechanisms. To account for genetic background and to identify the potential epigenetic mechanisms of MDD, we analysed the DNA methylation profile of whole peripheral blood cells collected from 12 monozygotic twins (MZTs) discordant for MDD using Illumina Infinium 450K methylation array. The analysis was mainly included the identification of differential methylated positions (DMPs), differential methylated regions (DMRs) and gene ontology (GO) terms enrichment analysis. More significantly, we developed a novel DMPs identification method, which is believed more statistically reliable than previous approaches. By introducing a list of empirical blood invariant sites summarised by Edgar et al. (2017), we identified 351 sites that have significant differences in methylation between MZTs. The p-values and q-values of the 10 top-ranked significantly differential methylated sites ranged from 5.551e-15 to 3.06e-09 and 1.76e-09 to 9.702e-05 respectively. Informative GO terms such as 'cranial nerve formation' and 'cranial nerve morphogenesis' were ranked in top 10, and most of them are involved in the mechanism of channel activity. Further, the method was extended to DMRs identification with 178 and 366 DMRs have been produced for contiguous and sliding windows respectively. To validate our discoveries, we sorted 254 differential expressed genes between MZTs using a public dataset (GSE76826), and WNT7B gene was found close to our identified DMPs and DMRs. These outcomes not only support the suspected role of channel activity in causing

MDD but also provide a potential analytical method for epigenetic differential DNA methylation studies.

20 t-SNE Generated Transcriptome Landscapes Reveal Native Gene Expression Wiring.

Michael See, Paul Harrison, David Albrecht, David Powell and Traude Beilharz

Monash University

Background: Every high-content gene expression analysis holds two areas of information: the response to altered experimental conditions intended by the researcher, and the general output of a native repertoire of transcriptional and post-transcriptional expression wiring. The information content now held in public transcriptomic databases, suggests it might be possible to dissect the detail of such expression connectivity in an unbiased manner. That is, the expression of some genes are likely always co-regulated with a set of functionally related genes, because they share either transcription factors, post-transcriptional regulatory elements, or both. Results: We used t-SNE projections to understand native transcriptome wiring from across 1000's of public expression datasets. The highly curated SPELL microarray datasets for baker's yeast, revealed discretely ordered and biologically meaningful islands of co-regulated genes. Analogous landscapes built from Fantom5 promoterome data, also showed that co-regulated genes in mammalian transcriptomes form meaningful islands of expression providing a new format for data visualisation. To make these landscapes useful to the community, we built Shiny apps to search, explore and annotate yeast, mouse and human landscapes and to use these for overlay of new data. Conclusions: Dimensionality reduction can reveal biologically meaningful network wiring by harnessing the information content in 1000s of experiments. These visual landscapes together with interactive search tools can hand back analytical control of complex data-sets to biologists with the domain knowledge to interpret the results. These landscapes are particularly useful when used as a framework for overlay and interpretation of new data.

21 Identifying Positive Selection Associated with Antimalarial Drug Resistance in Plasmodium falciparum using Identity-By-Descent Analysis

Lyndal Henden, Stuart Lee, Ivo Mueller, Alyssa Barry and Melanie Bahlo

Walter and Eliza Hall Institute

Drug resistance in microorganisms is a global health crisis and identifying the mechanisms underlying such resistance is crucial in advancing disease control and elimination efforts. Genes associated with resistance experience selective pressures, creating strong genetic signals in the microorganism's genome. Here we present a novel method, isoRelate (github.com/bahlolab/isoRelate), for identifying loci under recent positive selection in microorganisms using identity by descent analysis. We apply our method to whole genome sequencing data of more than 2,000 Plasmodium falciparum isolates from Africa, Southeast Asia and Papua New Guinea. In doing so we are able to identify many well-known signals associated with antimalarial drug resistance as well as several new loci suspected of being associated with resistance. Identity by descent analysis also allows us to explore population structure through relatedness networks, providing clues as to the number of haplotypes contributing to a selection signal and the distribution of these signals within and between countries. Furthermore, we are able to determine whether a haplotype conferring drug resistance has arisen independently between geographic locations or whether it has spread from other locations.

22 WITHDRAWN

23 Analysis of melanoma data with a mixture of survival models, utilising multiclass DQDA to inform mixture class

Sarah Romanes and John Ormerod

The University of Sydney

Melanoma is a prevalent skin cancer in Australia, with close to 14000 new cases estimated to be diagnosed in 2017. Survival times are markedly dierent from one individual to the next. In particular, there appears to be three classes of survival outcome. This talk considers integrating survival time data with micro-array gene expression data. We construct a hybrid model that seamlessly integrates a three-class quadratic discriminant analysis model, mixture of parametric survival models, and model selection components. We fit this model using a variational expectation maximization (VEM) approach. Our model selection component naturally simplifies as a function of likelihood ratio statistics allowing natural comparisons with traditional hypothesis testing methods. We compare our method with several naïve approaches which only addresses the classification aspect or survival model aspect in isolation.

24 bcGST - an interactive bias-correction method to identify over-represented gene-sets in boutique arrays

Kevin Wang, Jean Yang, Samuel Mueller and Garth Tarr

The University of Sydney Gene annotation and pathway databases such as Gene Ontology Kyoto Encyclopedia of Genes and Genomes are important tools in Gene Set Test (GST) that describe gene biological functions and associated pathways. GST aims to establish an association relationship between a gene set of interest and an annotation. Importantly, GST tests for over-representation of genes in an annotation term. One implicit assumption of GST is that the gene expression platform captures the complete or a very large proportion of the genome. However, this assumption is neither satisfied for the increasingly popular boutique array nor the custom designed gene expression profiling platform. Specifically, conventional GST is no longer appropriate in this new context due to the gene set selection bias induced during the construction of these platforms. We propose bcGST, a bias-corrected Gene Set Test method, by introducing bias correction terms in the contingency table needed for calculating the Fisher's Exact Test (FET). The adjustment method works by estimating the proportion of genes captured on the array with respect to the genome in order to assist filtration of annotation terms that would otherwise be falsely included or excluded. We illustrate the practicality of bcGST and its stability through multiple differential gene expression analyses in melanoma and TCGA cancer studies. The bcGST method is made available as a Shiny web application.

25 Characterising blood gene expression as a function of gut microbiota composition in preterm infants

Max Moldovan, Damian Drew, Lex Leong, Miriam Lynn, Anastasia Sribnaia, Naomi Fink, Irmeli Penttila, Carmel Collins, Maria Makrides, Robert Gibson, Geraint Rogers and David Lynn

South Australian Health and Medical Research Institute

Immediately after birth, the human gastrointestinal tract is colonised with various microorganisms collectively known as the gut microbiota. The gut microbiota performs a range of beneficial functions, outcompeting pathogenic bacteria, programming our immune system and determining our long-term health. The pathways through which the microbiota influences processes outside the gut, however, are poorly understood. We hypothesised that the gut microbiota could modulate host gene expression patterns in blood. To investigate this, we undertook a pilot study and applied linear modelling

(using limma) to correlate variation in the composition of the gut microbiota with variation in blood gene expression in 14 preterm infants. Faecal and blood samples were obtained from the infants at the Adelaide Women's and Children's Hospital. 16S rRNA gene sequencing was used to assess the composition of the faecal microbiota as a proxy for the gut microbiome in each infant. 15 taxa with non-zero abundances in at least 60% of samples were retained for further analysis. For the same 14 infants, RNA sequencing was performed on whole blood samples. Approximately 1.2 billion 100bp reads were sequenced, with an average of 56 million reads per sample. After QC, read counts for 18,149 genes were voom transformed and linearly modelled as a function of normalised ascsin square root transformed relative abundances. Moderated t-statistics were obtained using the eBayes function and gene set enrichment analysis was performed using limma's camera method. We identified a number of gene sets associated with changes in the relative abundance of certain bacteria, such as an inverse association between the relative abundance of Staphylococcus and the expression of interferon response genes in blood (Q-value = 8.4e-12). The current pilot study is limited due to small sample size but, based on the promising findings to date, samples are currently being collected from a larger number of infants.

26 Understanding the Mechanism of Action of In-feed Antibiotics for Chicken

Candida Vaz, Silvia Fibi-Smetana, Gerd Schatzmayr, Vivek Tanavde and Bertrand Grenier

Bioinformatics Institute A*STAR

Background: Feed additives are products used in animal nutrition to improve the quality of feed and to improve the animal's performance and health. Antibiotics have been used since the mid 1940s, but the spread of antibiotic resistance in zoonotic bacteria poses a threat to health and hence have been banned in several countries. This has led to the need of developing viable alternatives to improve performance and protect animal health. To develop effective alternatives it is crucial to understand the mechanism of action of in-feed antibiotics. Methods: In this project we used next generation sequencing and omics technologies to decipher the mechanism of action and to understand what signaling pathways are enriched on the usage of in-feed antibiotics. For this purpose RNA from the mid-ileum tissue of chickens fed with no in-feed additives (control set) and with avilamycin (antibiotics treatment set) for 35 days, was extracted and used for RNA sequencing. Five biological replicates and two technical replicates from each set were used for RNAseq (Illumina HiSeq 4000, PE data, 28-50M reads, 150 bp). Comparison was carried out between the treatment set and control set to obtain the genes changing due to the antibiotic treatment. The pathways enriched with these differentially expressed genes were determined to understand the mechanism of action of antibiotics. Results and Implications: The number of differentially expressed genes was around 237. The most enriched pathways were related to inflammatory and immune responses, such as: IL-6, IL-10, IL-22 signaling, LXR/RXR, FXR/RXR activation, Communication between Innate and Adaptive Immune Cells, Production of Nitric Oxide and Reactive Oxygen Species in Macrophages, Th1 Pathway, Graft-versus-Host Disease Signaling. Such kind of studies, will promote the development of safer alternatives to antibiotics. It would be interesting to study the differences in the mechanism of actions of alternative treatments to antibiotic treatment.

27 Using superTranscripts for RNA-seq analysis in cancer and non-model organisms

Nadia Davidson and Alicia Oshlack

Murdoch Childrens Research Institute

RNA-seq data is usually analysed and visualised either against the genome reference, which contains unnecessary intronic sequence, or the transcriptome reference which includes many copies of the same exons. To provide an alternative, we recently developed superTranscripts, a compact and data driven reference for the transcriptome. SuperTranscripts contain the sequence of all exons of a gene without redundancy. They can be constructed from any set of transcripts, including annotated transcripts and

de novo assemblies, using our Lace software. We demonstrate how superTranscripts enable enhanced visualisation of RNA-seq data and especially of alternative splicing events. SuperTranscripts also allow the detection of variants and differential isoform usage in non-model organisms. Using examples in sheep and chicken, we show that we can create superTranscripts by combining existing annotation with data derived transcriptomes from genome guided and de novo assembly. We have developed a pipeline called Necklace to perform this process and show how these superTranscriptomes are more complete and allow us to perform a more comprehensive differential expression analysis. In addition, superTranscripts can be used for analysis and visualization of RNA-seq data in a variety of other contexts. Specifically, we have developed Clinker, which uses superTranscripts to visualize the RNA-seq data underlying fusion gene calls in cancer data. This allows us to examine fusion genes and their transcripts, along with the RNA-seq reads used to call the fusion. These insightful visualizations of the data allow a more complete understanding of the structure of transcripts driving cancer.

28 Ximmer: Getting the best out of CNV detection on Exomes

Simon Sadedin and Alicia Oshlack
Murdoch Childrens Research Institute

Exome and targeted sequencing have become widely used tools for both clinical and research investigations into human disease. While these technologies are now highly mature for assaying single nucleotide variants (SNVs) and short indels, they struggle to reliably detect other forms of genetic variation such as copy number changes (CNVs). Despite this, the widespread use of exome sequencing has spurred the development of many methods that attempt to detect CNVs from exome data. However numerous evaluations and benchmarks have highlighted highly variable and inconsistent performance of these methods. We show that the performance of CNV detection methods depends not only on the CNV detection method used but also on how the parameters of the methods are configured. In addition, we show that different tools perform differently across dataset generated by different capture technologies. Together these observations explain much of the variability in benchmarking results. We propose that extracting good results and understanding the limitations of CNV detection on HTS data requires a systematic approach involving rigorous quality control, adjustment of method parameters, and calibration of confidence measures for filtering. Here we present Ximmer, a suite of integrated tools which supports the end to end process for applying these procedures to get improved results from CNV detection methods. Ximmer includes a simulation method, an automated CNV detection analysis pipeline, and a visualisation tool which enables inspection of quality control measures and interactive exploration of CNV results. Ximmer is open source software, freely available at http://ximmer.org (example results are viewable at http://example.ximmer.org).

29 Creating and exploiting Gene networks

Philippe Moncuquet

CSIRO

Secondary Cell Wall (SCW) is found in a variety of cell type in plant. It is central to the production of fibers in cotton. Understanding the network behind the formation is key to decipher how the main actors are involved in SCW formation in fiber. Traditional Differential Expression (DE) approach can be complemented by exploring expression profile correlation network. Such gene network can be generated using expression profile from RNASeq experiment, inferring correlation between those profiles and feeding this information along with annotation into Cytoscape. Networks can be designed in many ways to answer specific biological questions. Filters can be applied prior to the network creation to look at specific gene population (eg. based on expression or based on known function) and designed lay out can help explore the network created by visually highlighting experimental conditions. Networks can also be explored using different approaches (eg. based on statistical features of nodes or based on the overall structure of the network) to define potential candidates to

be explored in terms of function and structure of relationship in the model studied (here applied to SCW formation in cotton). Multiple approaches to explore expression landscape is an efficient way to provide more biological insights and discover new potential gene candidates.

30 Targeted Search for Genomic Variants for Clinical Applications

Thomas Conway, Hannah Huckstep, Andrew Fellowes and Ken Doig

Peter MacCallum Cancer Centre

Calling variants in clinical cancer samples has generally been tackled as a discovery problem – where all the putative variants in the sample are called, and those found are analysed for biological relevance and clinical importance. There is inevitably a careful trade-off between false positive and false negative variant calls. In cancer patient testing, this can be particularly fraught with the need for quick turnaround-time conflicting with the imperative of not missing clinically actionable variants. This problem is made worse by the fact that some clinically significant variants confound standard pipelines. However, the routine testing context has some particular features suggesting that an alternative approach may be more useful. Our laboratory performs in the order of 200 clinical gene panels per week, comprised of hybrid capture across about 150 genes, followed by paired-end Illumina sequencing. The samples are referred from a mix of familial cancer, solid tumour, and haematological clinics. In this context, our specific aim is to rapidly answer the primary question – are known pathogenic variants present? To this end we are developing a k-mer based method that quickly answers the question of whether variants from a given list are present. In essence, it works by computing sets of k-mers characteristic of both the wild-type and mutant alleles of each variant, then counts the frequency of these k-mers in the read data. We present some results showing that the method not only has very good sensitivity and specificity for most variants of interest, but also is extremely fast. We also discuss some of the practical considerations for clinical deployment of such a method.

31 STRetch: detecting and discovering pathogenic short tandem repeat expansions

Harriet Dashnow, Monkol Lek, Belinda Phipson, Andreas Halman, Simon Sadedin, Andrew Lonsdale, Mark Davis, Phillipa Lamont, Nigel Laing, Daniel MacArthur and Alicia Oshlack
Murdoch Childrens Research Institute

Short tandem repeat (STR) expansions have been identified as the causal DNA mutation in dozens of Mendelian human diseases. Traditionally, pathogenic STR expansions could only be detected by single locus techniques, such as PCR and electrophoresis. These methods are expensive and most diagnostic tests only genotype the most common known events. In addition these methods do not scale to the whole genome and so cannot be used to identify new pathogenic STR loci. The ability to genotype STRs directly from next-generation sequencing data has the potential to discover new causal STR loci and to reduce both the time and cost to reaching a diagnosis. Most existing tools for detecting STR variation are limited to repeat lengths that fit within the read length, and so are unable to detect the majority of pathogenic expansions. We present STRetch, a new genome-wide method for detecting pathogenic STR expansions and estimate their approximate size directly from short read sequencing. STRetch takes the approach of adding STR decoy sequences to the reference genome prior to mapping reads. Reads mapping to the decoys are assigned back to their genomic position using read-pair information. Each locus is assessed for expansion using a statistical test based on coverage of the decoy chromosome. We apply STRetch to the analysis of 97 whole genomes to reveal variation at known STR loci. We further demonstrate the application of STRetch to solve cases of patients with undiagnosed disease. A key advantage of STRetch over other STR detection tools is that it assesses expansions at all STR loci in the genome and so can be used to detect novel disease-causing STR loci. STRetch is open source software, available from github.com/Oshlack/STRetch. The preprint can be found at http://biorxiv.org/content/early/2017/07/04/159228.abstract.

32 Cloud-based single-cell transcript reconstruction using Falco

Andrian Yang, Abhinav Kishore and Joshua Ho

Victor Chang Cardiac Research Institute

Current bioinformatics tools for analysis of single-cell RNA-seq (scRNA-seq) data mainly focus on quantification of gene expression and clustering of samples into sub-populations, and there are limited tools available for further downstream analysis of sub-populations, such as reconstruction of full-length transcripts and analysis of alternative splicing. Existing tools for transcript reconstruction are designed to work on bulk RNA-seq data and perform poorly on scRNA-seq data due to the low sequencing depth and high technical noise inherent to scRNA-seq. Furthermore, they can be very slow when directly used on scRNA-seq data, which can contain transcriptome information for hundreds of thousands of cells. We need a highly scalable solution for scRNA-seq transcript reconstruction. To leverage existing tools for transcript reconstruction for scRNA-seq analysis, we need to enable sharing of information between samples in order to circumvent the limitation introduced by scRNA-seq. Moreover, we need to utilise a scalable platform in order to enable efficient processing of scRNA-seq data through parallel processing of samples. Here we present a single-cell transcript reconstruction extension for the cloud-based Falco framework. The Falco framework enables highly parallellised processing of scRNA-seq data using big data technologies of Apache Hadoop and Apache Spark. The new transcript reconstruction pipeline allows for sharing of information across samples and are highly scalable to allow for efficient and timely analysis of large number of cells in scRNA-seq data.

33 The histone variant H2A.Z is a master regulator of the epithelial-mesenchymal transition

Sebastian Kurscheid, Renae Domaschenz and David Tremethick

John Curtin School of Medical Research, Australian National University

Epithelial-mesenchymal transition (EMT) is a profound example of cell plasticity that is crucial for embryonic development and cancer. Although it has long been suspected that epigenetic-based mechanisms play a role in this process, no master epigenetic regulator that can specifically regulate EMT has been identified to date. Here, we show that H2A.Z can coordinate EMT by serving as either an activator or repressor of epithelial or mesenchymal gene expression, respectively. Following induction of EMT by TGF-beta we observed an unexpected loss of H2A.Z across both down-regulated epithelial and up-regulated mesenchymal promoters. Strikingly, the repression of epithelial gene expression was associated with reduction of H2A.Z upstream of the transcription start site (TSS), while the activation of mesenchymal gene expression was dependent on removal of H2A.Z downstream of the TSS. Therefore, the ability of H2A.Z to regulate EMT is dependent on its position, either upstream or downstream of the TSS.

34 Glimma: getting greater graphics for your genes

Shian Su, Charity Law and Matthew Ritchie

Walter and Eliza Hall Institute

RNA-sequencing is a popular technology for studying changes in gene expression across tens of thousands of transcripts simultaneously. To make exploration of gene expression data easier, we developed Glimma, an R package which generates interactive plots for gene expression analyses. Glimma plots connect the many layers of information in a single html page using d3.js. For example, a Glimma-style mean-difference plot, allows one to select a point from a display of summary statistics to reveal the sample-wise expression levels alongside the original plot. This feature enables researchers to interrogate the data more easily by allowing searches for genes or samples of interest and zooming 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. Results from Glimma can

be easily shared between bioinformaticians and biologists, enhancing reporting capabilities while maintaining reproducibility. Besides bulk RNA-sequencing data, Glimma can also handle data from microarray, single-cell RNA-sequencing and methylation experiments.

35 Bioinformatic challenges in the analysis of CLIP Experiments

Emily Hackett-Jones, John Toubia, Katherine Pillman, Kate Dredge, Andrew Bert, Cameron Bracken, Andreas Schreiber and Gregory Goodall

Centre for Cancer Biology

MicroRNAs (miRs) are small non-coding RNAs known to bind - often via complementary seed sequences - to messenger RNAs and act repressively on the translation to cellular proteins. Much of the interest in miRs concerns their role in cancer, as dysregulation of miRs is common in cancer. Although miR:mRNA binding sites can be predicted in silico, many of the millions of predicted sites are spurious. CLIP-Seq experiments involve high-throughput next generation sequencing of RNA cross-linked to miRs, and are a relatively new method to determine functional bindings of miRs to mRNAs. We shall outline our development of a bioinformatic analysis pipeline for CLIP-Seq data, including peak calling, identification of PCR duplicates, and downstream motif analysis. We will discuss ways to improve the identification of predicted functional miR:mRNA binding sites.

36 Comparative analysis of phosphoethanolamine transferases involved in polymyxin resistance across ten clinically relevant Gram-negative bacteria

Jiayuan Huang, Yan Zhu, Meiling Han, Mengyao Li, Jiangning Song, Tony Velkov, Chen Li and Jian Li

Monash University

The rapid emergence of Gram-negative 'superbugs' has become a significant threat to human health globally and polymyxins become a last-line therapy for these very problematic pathogens. Polymyxins exhibit their antibacterial killing by the initial interaction with lipid A in Gram-negative bacteria. Polymyxin resistance can be mediated by phosphoethanolamine (PEA) modification of lipid A that abolishes the initial electrostatic interaction with polymyxins. Both chromosome-encoded (e.g. EptA, EptB and EptC) and plasmid-encoded PEA transferases (e.g. MCR-1 and MCR-2) were reported in Gram-negative bacteria; however, their sequence and functional heterogeneity remain unclear. Here, we report a comparative analysis of PEA transferases across ten clinically relevant Gram-negative bacteria species using multiple sequence alignment and evolutionary analysis. Our results show that the pairwise identities among chromosome-mediated EptA, EptB and EptC from E. coli are very low, and EptA shows the highest similarity with MCR-1/2. Among PEA transferases from representative strains of ten clinically relevant species, the catalytic domain is more conserved compared to the transmembrane domain. Particularly, PEA acceptor sites and zinc binding pockets show high conservation among different species, indicating their potential importance for PEA transferase function. The evolutionary relationship of MCR-1/2 and EptA from ten selected bacterial species was evaluated by phylogenetic analysis. Cluster analysis illustrates that 325 EptA from 275 strains of ten species within each individual species are highly conserved, whereas the interspecies conservation is low. Our comparative analysis provides key bioinformatic information to better understand the mechanism of polymyxin resistance via PEA modification of lipid A.

37 Reliably Detecting Clinically Actionable Variants Requires Combined Variant Call

Matt Field, Chris Goodnow and Dan Andrews

Australian Institute of Tropical Health and Medicine. James Cook University. Cairns, QLD; Garvan Institute of Medical Research. Darlinghurst, NSW; John Curtin School of Medical Research. Australian National University. Canberra, ACT

A diversity of tools is available for identification of variants from genome sequence data. Given the current complexity of incorporating external software into a genome analysis infrastructure, a tendency exists to rely on the results from a single tool alone. The quality of the output variant calls is highly variable however, depending on factors like the choice of short-read aligner, variant caller, and variant caller filtering amongst others. Here we present a two-part study first analysing a melanoma cell line derived from a control lymphocyte sample finding that only one of the three clinically important melanoma risk-factor variants is unanimously detected by all software permutations. Second, we use the high quality 'genome in a bottle' reference set to more broadly demonstrate the significant impact the choice of aligner, variant caller, and variant caller filtering strategy has on variant call quality and further how certain software is superior at dealing with increased sample contamination, an important consideration when working with heterogeneous tumour samples. This analysis confirms previous work showing that combining variant calls of multiple tools results in the best quality resultant variant set, for either specificity or sensitivity, depending on whether the intersection or union, of all variant calls is used respectively. Finally, we describe a cogent strategy for implementing a clinical variant detection pipeline; a strategy that requires careful software selection, variant caller filtering optimizing, and combined variant calls in order to effectively minimize false negative variants. While implementing such features represents an increase in complexity and computation the results offer indisputable improvements in data quality.

38 Predictors of damage transition in systemic lupus erythematosus

Kevin Zhang, Sarah Boyd, Rachel Koelmeyer, Alberta Hoi, Francois Petitjean, Eric Morand and Hieu Nim

Monash University

Systemic Lupus Erythematosus (SLE) is a heterogeneous multisystem autoimmune disease with a high burden of morbidity. Organ damage is an important outcome of SLE, that affects morbidity, mortality and quality of life; however, clinical instruments conventionally used to estimate damage risk entail high subjectivity and no objective predictors of short term risk of damage accrual have been validated. To assess the potential for objective laboratory measurements to predict organ damage accrual in SLE. Retrospective analysis of prospectively collected data from the Australian Lupus Registry and Biobank, a multi-centre database of SLE patients, was conducted. Time Adjusted Mean (TAM) values were calculated for 16 routine clinical laboratory parameters and analysed with multivariable logistic regression, adjusting for age, gender, race, prednisolone dose and existing organ damage as confounders. Those values that were significant after Holm-Bonferroni correction were then selected to create odds ratio plots with logistic regression and bootstrapping. Organ damage transition was defined as a period of time preceding accrual of new organ damage, measured by standard scoring system. 323 patients' data were analysed. TAM haemoglobin, estimated glomerular filtration rate (eGFR), erythrocyte sediment rate (ESR), urine protein:creatinine ratio and creatinine were significantly associated with organ damage transition. Moreover, haemoglobin, ESR, eGFR and creatinine's trend with damage transition risk were found to be monotonic, whereas urine protein:creatinine ratio was not. Other laboratory measures, including albumin, anti-DsDNA, C3, C4, white cell count, urine white and red cell counts, lymphocytes, neutrophils, platelets, lymphocyte count and C-reactive protein were statistically insignificant. Objective laboratory parameters predict organ damage transition in SLE. Validation in independent cohorts is justified.

39 IVAT and VariantGrid: integrative annotation and analysis of genetic variants from next-generation sequencing data

Jinghua Feng, David Lawrence and Andreas Schreiber

ACRF Cancer Genomics Facility, Centre for Cancer Biology, SA Pathology and University of South Australia

Next-generation sequencing (NGS) provides unprecedented power to rapidly identify genetic variation, and is increasingly used in research and diagnostics for human diseases. However, interpreting detected variants and isolating the minority of variants underlying disease remains a challenge. To address those issues, we have developed IVAT (Integrated Variant Annotation Tools) and VariantGrid. IVAT integrates published methods and databases and provides comprehensive annotations, including the variant's effect on genes (e.g. synonymous and nonsynonymous mutations) predicted by SnpEff, allele frequencies in public genomic databases (e.g. the 1000 Genomes Project), conservation levels (e.g. PhyloP and GERP++ scores), predicted functional importance scores (e.g. CADD and SIFT), and flagging variants located in the regions less accessible to NGS short reads. Gene level annotation (such as phenotype information from Ensembl and protein information from UniProtKB) is also provided. Annotations are constantly improving with new literature, improved databases and tools. IVAT can be used as command line tools, or implemented within VariantGrid, a variant database with a graphical user interface that allows non-bioinformaticians to filter, analyse and classify variants. Within VariantGrid, variants get automatically annotated by IVAT. Multiple annotation versions are all stored, permitting re-analysis and re-run of previously unsolved cases as new information becomes available. IVAT and VariantGrid further manage the different effects a variant can have on genes with multiple transcripts. Effects are calculated for every known transcript and, while the most damaging is picked as a representative for the purpose of filtering, the effects are actually stored for all transcripts. This allows the user to examine and choose which one to use for classification and reporting. Together, IVAT and VariantGrid provide greatly enhanced flexibility for the analysis of genetic variants from NGS data in modern research and diagnostic laboratories.

40 RNA editing in an editing deficient Adar1 mouse model

Alistair Chalk, Jacki Heraud-Farlow, Joshua White, Brian Liddicoat, Ankita Goradia, Sandra Linder, Qin Li, Scott Taylor, Jin Li and Carl Walkley

St Vincent's Institute of Medical Research

Several recent studies have identified that a feature of absence or reductions of Adar1 activity, conserved across human and mouse models, is a profound activation of interferon-stimulated gene signatures and innate immune responses. Further analysis of this observation has led to the conclusion that editing by Adar1 is required to prevent activation of the cytosolic innate immune system, primarily focused on the dsRNA sensor MDA5 (Ifih1). The delineation of this mechanism places Adar1 at the interface between the cells ability to differentiate self- from non-self dsRNA. We have further characterised our viable adult Adar1 editing deficient (E861A) mice lacking Ifih1. We document gene expression changes, editing and hyper-editing in foetal liver, foetal brain, adult brain and GMCSF immortalised myeloid cell lines using RNAseq. Based on MDA5 dsRNA recognition requisites, the mechanism indicates that the type of dsRNA should fulfil a particular structural characteristic, rather than a sequence-specific requirement. We propose that beyond the structural component, the magnitude of the dsRNA response is based on the amount of (unedited) dsRNA load in the cell. Each cell type will have a different dsRNA load dependent on the cells individual transcriptome, and this dsRNA load may or may not be sufficient to generate an IFN response, when MDA5 is present. We have defined the editing by Adar1 of hyper-edited regions across multiple tissues in the mouse. In mice lacking Adar1 activity, editing within these regions was variable between different tissues and in the same tissue at different ages. We hypothesise that Adar2 can edit these regions when sufficiently highly expressed and that this can compensate for a lack of Adar1. While additional studies are required to molecularly verify the genetic model, the observations to date collectively identify A-to-I editing by ADAR1 as a key modifier of the cellular response to endogenous dsRNA.

41 Finding optimal regulatory element classifiers using automatic machine learning

Liam Fearnley and Melanie Bahlo

Walter and Eliza Hall Institute of Medical Research

Model selection is a difficult, critical step in the analysis of complex data where parameter estimates or state predictions for data are sought. A common approach to modelling biological phenomena is to treat the phenomenon as a black box, choose a model, fit that model's parameters to observations of the phenomenon, then evaluate the model's performance. The choice of modelling strategy, model specification, and fitting methods is usually manual, painstaking and laborious. For applications in computational biology, where data sets such as FANTOM5 and ENCODE are constantly increasing in size, it is important to be able to efficiently select, fit and refit these complex models. Thus, an automated framework is essential. Automatic model selection (AMS) methods treat model selection as a search problem, and allow the automation of this process. In this work, we discuss the application of recent advances in AMS to computational biology, with emphasis on finding deep neural networks for the classification of regulatory elements. We show that the models generated by AMS methods are able to outperform manually specified approaches on enhancer and promoter data sourced from the ENCODE and FANTOM5 databases. We also perform a comparative evaluation of techniques and methods for AMS, from relatively commonly-used grid and random search techniques to recent advances in automatic modelling using Bayesian optimisation and other approaches such as Google's recent AutoML model. We also discuss software engineering challenges (and hardware requirements) that must be met to enable the wider use of such methods.

42 Investigating computational analysis pipelines and genomic proximity interactions in T lymphocytes

Ning Liu, Timothy Sadlon, Stephen Pederson, Simon Barry and Jimmy Breen

University of Adelaide

Chromosome Conformation Capture (3C) technology is a method used for investigating three-dimensional (3D) genome structure, whereby segments of a genome that are in close-proximity can be identified and used to infer their spatial relationship. A 3C-derived method, High-resolution Chromosome Conformation Capture sequencing (HiC-seq) have been used to identify genes that can be affected by distal interactions such as long-range promoter-enhancer contacts that interact with immune system regulators. Although HiC-seq has been widely used to identify 3D interactions genome-wide in many species, many of the analysis tools have yet to be critically assessed. Here, we used publically available HiC-seq data to investigate and compare three major steps of HiC-seq data analysis workflow, including raw HiC-seq data processing, topologically-associated domains (TADs) identification algorithms and visualisation tools. We then applied our validated toolset to a DNaseI-treated, HiC-seq dataset sampled from human conventional T cells (Tconv cells) to investigate the ability of the tools at analysing relative low-coverage datasets. Whilst HiC-seq data analysis requires a significant sequencing coverage, applying HiC-Pro, an insulation score algorithm for TAD identification and HiCPlotter for visualisation, we identified a total of 4,818,855 long-range interactions, leading to the prediction of 3275 TADs genome-wide. Using this HiC-seq data along with other conformation assays (i.e. 4C-seq), we show that an upstream super-enhancer and promoter of the master T cell regulator SATB1 are located within the same TAD region, supporting the hypothesis that long-range interactions regulate the function of SATB1, and that sequence variants in enhancer elements may effect the pathogenicity of autoimmune diseases.

43 Physical coherence and network analysis to identify novel regulators of exosome biogenesis.

David Chisanga, Sushma Anand, Shivakumar Keerthikumar, Suresh Mathivanan and Naveen Chilamkurti

La Trobe University

Exosomes are small (30-150nm in diameter) membranous vesicles of endocytic origin. They have been implicated in a range of biological functions such as intercellular communication through the transmission of macromolecules such as proteins, nucleic acids and lipids, as well as in the pathogenesis and progression of diseases such as cancer. As such, there has been growing interest in understanding the biogenesis, functions, and applications of exosomes in both physiological and pathological conditions. The biogenesis of exosomes has long been associated with the endosomal sorting complex required for transport (ESCRT) machinery together with other accessory proteins. However, the mechanisms behind exosome biogenesis are still poorly understood and the proteins involved in the process of exosome biogenesis have not all been characterised. Here we therefore, attempt to identify novel proteins that regulate the process of exosome biogenesis through the ESCRT pathway and improve our understanding of exosome biogenesis and exosomes in general. To achieve this, network analysis methods are applied to a Protein-Protein Interaction (PPI) network of the ESCRT machinery. To counter the bias that exists in PPIs due to false positives stemming from experimental errors in techniques used to identify them, we extended the network analysis method by using physical coherence, a technique that quantifies the connectedness of a PPI network due to topological changes. Using this technique, STAMBP and NEDD4 were identified as potential novel regulators of exosome biogenesis. It was found that STAMBP increased the physical coherence of the ESCRT machinery network while NEDD4 reduced the physical coherence of the ESCRT machinery network. To validate our findings, SDCBP, a protein that has been previously shown to regulate exosome biogenesis was also found to change the physical coherence of the ESCRT machinery. Further analysis using CRISPR-Cas9 based knockout cells of NEDD4 and STAMBP confirmed their active role in exosome biogenesis.

44 Comprehensive benchmarking of short read structural variant callers

Daniel Cameron and Anthony Papenfuss

Walter and Eliza Hall Institute of Medical Research

In recent years, there has been a proliferation of software packages for identifying structural variants (SVs) using whole-genome sequencing data. While comparisons are made between new and existing methods when published, these tend to be selective, incomplete, possibly over-fitted and invariably show a modest improvement in the new tool, while also overlooking undesirable features of the method. The lack of comprehensive benchmarking across software tools presents challenges for users in selecting methods and for developers in understanding algorithm behaviours and limitations. To address these challenges, we undertook a rigorous selection process to identify representative methods spanning the breadth of SV detection approaches, and performed a comprehensive evaluation and characterisation of 10 SV callers on typical human resquencing data, as well as simulated data. Whilst some of our results capitulate the theoretically expected results based on the SV detection approaches taken, others do not. In some cases, better sequencing can result in significantly worse variant calling performance. To aid in the appropriate selection of tools, and to guide the development of improved callers, we have developed sets of recommendations for both users and methods developers.

45 Direct Determination of Mouse Genome-Wide, Allele-specific DNA Methylation from Nanopore Long-Read Sequencing.

Terence Speed, Scott Gigante, Andrew Keniry, Alexis Lucattini, Christopher Woodruff, Quentin Gouil, Marnie Blewitt, Lavinia Gordon and Matthew Ritchie

Walter & Eliza Hall Institute

Asymmetric expression patterns between the two parental alleles are critical for development of the mammalian embryo. This process known as imprinting involves differential DNA methylation of the parental genomes. We sequence mouse embryonic placenta tissue on the Oxford Nanopore MinION and exploit the long reads to determine, using novel methods, both haplotype and CpG methylation levels. Comparison with matched Reduced-Representation Bisulfite Sequencing data confirms the accuracy of the methylation calls, and highlights the improvement in haplotyping conferred by the longer reads. We successfully identify known imprinting control regions, as well as novel differentially methylated regions. Based on their proximity to hitherto unknown monoallelically expressed genes, we propose that some of these regions could constitute new imprinting control regions.

46 A pan cancer hypoxic gene signature – highlighting temporal changes that lead to poor patient survival.

Kristy Horan, Joseph Cursons, Momeneh Foroutan and Melissa Davis

The University of Melbourne

An insufficient oxygen supply is a common feature of many solid tumours, and the consequential hypoxic microenvironment has been linked to poor patient outcomes as well as and chemo- and radio-therapy resistance. Tumour hypoxia can induce an epithelial-mesenchymal transition and angiogenesis - these changes are linked to a more aggressive cancer progression and perhaps also facilitate metastatic dissemination. A number of high-throughput studies have attempted to develop hypoxic transcriptomic signatures in specific cell types, however, many of these have been restricted in their selection of cell or tissue types and the duration of hypoxia investigated. We have used novel gene-set scoring technique to analyse public data and derive a pan-cancer hypoxia signature which captures both moderate and chronic hypoxia, and appears to be a more accurate classifier of hypoxia than current signatures. Our pan-cancer signature has prognostic abilities when predicting survival across multiple cancer types, and it reveals a temporally-regulated network of genes that may impact on disease progression, and has the potential to identify novel targets for combination therapies.

47 Bypassing the pseudogenes – How to diagnose with un-mappable genes

Adilla Razali, Julien Soubrier, Maely Gauthier, Jacqueline Rossini, Rachel Hall, Wendy Parker, Joel Geoghegan, Scott Grist, Lesley Rawlings and Karin Kassahn

SA Pathology

Despite considerable advancements in genetic testing with the development of high throughput sequencing (HTS) technologies, some regions of the genomes are still difficult to access for routine diagnostics. For example, highly homologous regions are preventing reliable alignment of short sequencing reads (e.g., Illumina data) onto a reference genome, as they represent multi-mapping possibilities. This is the case for some key diagnostic genes which have multiple pseudogenes with highly conserved homologous regions. Although deep whole-genome shotgun sequencing, or the use of different sequencing technology (e.g., PacBio) can help recover information from such regions, it is still not possible to implement these solutions for routine testing due to cost and turnaround time. Instead, we here test the efficiency of amplifying the targeted regions with long-range PCR (LR-PCR), in combination with DNA shearing and HTS, to reliably recover variants from two clinically relevant genes with pseudogene issues: PKD1 (polycystic kidney disease) and PMS2 (colorectal cancer). By

comparing patient results obtained from Sanger sequencing, capture+HTS, and LR-PCR+HTS, we demonstrate that targeted LR-PCR provide a cost-effective way to recover reliable variant calls in highly homologous regions, using pre-existing HTS setups and pipelines. This method outperforms the classic capture+HTS approach, and eliminates the need for Sanger back-filling, by preventing the pseudogenes from being sequenced (and therefore considered for mapping). When re-mapping the LR-PCR reads to references masked for the known pseudogene regions, our results provide insights on the level and consequences of multi-mapping when targeting homologous regions.

48 miR-200 regulates widespread changes in alternative splicing by controlling Quaking

Katherine Pillman, John Toubia, Caroline Phillips, Suraya Roslan, Kate Dredge, Andrew Bert, Yeesim Khew-Goodall, Luke Selth, Gregory Goodall and Philip Gregory

Centre for Cancer Biology

Members of the miR-200 family of microRNAs are critical gatekeepers of the epithelial cell state, restraining expression of pro-mesenchymal genes, which contribute to metastatic progression in cancer by driving epithelial-mesenchymal transition (EMT). We have discovered that miR-200c also exerts widespread control of alternative splicing patterns in cancer cells. This is achieved through strong suppression of the RNA binding protein Quaking (QKI). Deep RNA sequencing and HITS-CLIP revealed that QKI directly regulates hundreds of EMT alternative splicing events, without appreciably affecting gene expression levels. These miR-200/QKI-driven splicing changes converge on targets within the actin cytoskeleton regulatory network and were sufficient to increase key cancer-related cell properties such as cell migration and invasion. Notably, QKI-driven alternative splicing signatures are also found in patient breast tumours and are broadly conserved across many cancer types. These findings demonstrate the existence of a miR-200/QKI axis that controls alternative splicing and has a critical impact on cancer-associated epithelial cell plasticity. Here, we will focus on detection of alternative splicing from RNA-seq data, elucidation of direct targets from QKI HITS-CLIP data and analysis of alternative splicing in patient tumour data from The Cancer Genome Atlas.

49 Whole exome sequencing and linkage analysis of extended pedigrees to identify glaucoma susceptibility genes

Patricia Graham, Juan Peralta, Nicholas Blackburn, John Blangero, Mary Wirtz, Alex Hewitt, David Mackey, Kathryn Burdon and Jac Charlesworth

Menzies Institute for Medical Research, University of Tasmania, Australia

The use of massively parallel sequencing in extended pedigrees has significant potential for identifying functional variants linked with complex disease. We are using whole exome sequencing (WES) of five large, complex families from Tasmania and Oregon (USA) to identify susceptibility genes for primary open-angle glaucoma (POAG), the leading cause of irreversible blindness worldwide. Extended pedigrees, enriched for POAG, provide a powerful tool to search for rare and private genetic variants influencing the disease, where enrichment of rare variants occurs as a function of segregation from the founders. The families in this study range in size from 48 to 201 individuals (28 to 91 sequenced) and span 5 to 7 generations. These families have been used to locate quantitative trait loci (QTLs) for intraocular pressure (IOP), an important glaucoma endophenotype. IOP is a highly heritable (50%) intermediate trait which is correlated with POAG susceptibility in these families (IOP RhoG = 0.80, p = 9.6 x 10-6). WES data were generated for 249 individuals from the five pedigrees using the Illumina Nextera Expanded Exome Capture Kit. After alignment to hg19, over 235,000 variants were identified. Multipoint identity by descent was estimated from a subset of variants using the IBDLD program, which has been specifically developed for dense genotype data. Variance components linkage analysis of IOP was conducted using SOLAR. Suggestive QTLs have been identified on chromosomes 2, 4, 5 and 15. Preliminary analysis of the chromosome 2 locus in two families has identified several rare, potentially deleterious variants. Genes identified within the QTLs will be validated in large POAG case/

control cohorts. Finding genes involved with POAG susceptibility will increase our understanding of the biological pathways involved with the disease process and from that, diagnostic tools and more effective treatments can be developed.

50 Causal Inference Methods and Applications in Bioinformatics

Thuc Duy Le, Lin Liu, Weijia Zhang and Jiuyong Li

University of South Australia

Discovering causal relationships is the ultimate goal of many disciplines, including Bioinformatics. The standard method for causal inference is randomised controlled trials (RCTs). However, it is often infeasible to conduct RCTs. Causal inference with observational data is therefore very important, but it is also a challenging problem. In this talk, I will firstly present an overview of the causal inference approaches and then introduce two applications of causal inference methods in Bioinformatics. The first application uses a causal effect estimation method to infer the regulatory relationships between microRNAs and messenger RNAs from expression data(1). The method simulates a gene knockdown or gene transfection experiment and estimates the micorRNA regulatory effects on genes. The evaluation shows that the method is able to estimate the regulatory effect of each microRNA on each messenger RNA and effectively infer microRNA targets. Applying the same technique to datasets in different biological conditions, we can also detect the microRNAs that are active in a specific condition(2). The second application is about personalised medicine. We developed the Survival Causal Tree (SCT) (3) method to identify the features for stratifying patients into groups that respond differently to a particular treatment, e.g. radio therapy. The trained SCT can be used to predict the treatment response for individual patients.

51 Optimising intrinsic protein disorder prediction for short linear motif discovery

Kirsti Paulsen, Sobia Idrees, Åsa Pérez-Bercoff and Richard Edwards

The University of New South Wales

Short linear motifs (SLiMs) are short stretches of proteins that are directly involved in protein-protein interactions. Identifying SLiMs is important for understanding of the fundamental processes involved in normal cellular function. The functional importance of SLiMs also makes them potential drug targets and possible hotspots for disease causing mutations. SLiMs are commonly only 3 - 10 amino acids in length and form low affinity interactions. This makes them ideal for fast cellular processes, such as cell signalling or response to stimuli, but also difficult to predict experimentally. As a result, many computational SLiM prediction methods have been developed. One major challenge is to extract a significant signal of real SLiMs from the noise of false positive predictions due to randomly recurring sequence patterns. In order to increase the signal to noise ratio of SLiM predictions, different sequence masking techniques have been developed. These attempt to screen out areas that are unlikely to contain SLiMs and thereby preferentially eliminate the random nonfunctional sequence. One widely implemented masking strategy is to remove protein regions that form stable three-dimensional structures; SLiMs are typically found in regions of intrinsic disorder that are natively unstructured in their unbound form. To date, there has been no systematic study of how best to predict these regions for SLiM discovery. Poor quality predictions will not have the desired noise-removal, while over-stringent masking will remove too many true positives. The aim of this study is to compare how a number of different disorder masking approaches affect predictions from the de novo motif discovery tool, SLiMFinder. Prediction performance will be assessed using SLiMBench, which benchmarks the sensitivity and specificity of different methods using datasets of proteins containing known motifs from the Eukaryotic Linear Motif (ELM) database.

52 Clinker: visualising fusion genes detected in RNA-seq data

Breon Schmidt, Nadia Davidson, Anthony Hawkins, Ray Bartolo, Ian Majewski, Paul Ekert and Alicia Oshlack

Murdoch Children's Research Institute

Genomic profiling of cancer has revealed a rich diversity of rearrangements and translocations. Many of these have resulted in oncogenic fusion genes which are emerging as important therapeutic targets. The most efficient way of identifying important fusion genes is through transcriptome sequencing. While there are dozens of different methods available for identifying fusion genes from RNA-seq data, visualising fusion transcripts and their supporting reads remains challenging. Here we present Clinker, a bioinformatics tool written in Python, R and Bpipe, that leverages the superTranscript method to visualise fusion genes. It takes the RNA-seq data from which fusions were called and creates a sample specific fusion superTranscriptome. The reads are then mapped backed to this reference for visualisation of the data and transcripts. In addition, Clinker provides context to the fusion gene by combining: coverage, gene names, protein domains, transcripts, splice junctions, and fusion breakpoints, into a single customisable figure. Clinker's output can also be input into popular genome browsers, such as the Integrated Genome Viewer (IGV). This allows users to have an interactive experience with the fusion gene, the sequencing data and the Clinker generated annotation. As an example, we use Clinker to explore multiple fusion transcripts with novel breakpoints within the P2RY8-CRLF2 fusion gene in six samples of B-cell Acute Lymphoblastic Leukaemia (B-ALL)

53 Investigating the evolution of complex novel traits using whole genome sequencing and molecular palaeontology

Asa Perez-Bercoff, Psyche Arcenal, Anna Sophia Grobler, Philip J. L. Bell, Paul V. Attfield and Richard J. Edwards

The University of New South Wales

Understanding how new biochemical pathways evolve in a sexually reproducing population is a complex and largely unanswered question. We are using PacBio whole genome sequencing and deep population resequencing to explore the evolution of a novel biochemical pathway in yeast over several thousand generations. Growth of wild Saccharomyces cerevisiae (Baker's yeast) strains on the pentose sugar xylose is barely perceptible. A mass-mated starting population was evolved under selection on Xylose Minimal Media (XMM) with forced sexual mating every two months for four years. This produced a population that could grow on and utilise xylose as its sole carbon source. We are now using a novel molecular palaeontology approach to trace the evolutionary process and identify functionally significant loci under selection. Populations at seven key time points during the course of evolution have been sequenced using Illumina paired-end sequencing. In addition, all the parental strains from the founding population have been subject to PacBio de novo whole genome sequencing and assembly. By constructing reliable full genomes of the ancestors of our populations, we can the trace evolution of these populations over time. We can therefore track the trajectory of allele frequencies through time, identifying the contributions of different founding strains and novel mutations. We are using these data to estimate the proportions and regions of the genome that have evolved neutrally (due to genetic drift), under purifying selection, or adaptively in response to xylose selection. To date, much of our understanding of evolutionary processes is derived from theoretical models, and/or by reconstructing theoretical ancestors of extant individuals. Our unique array of both extant and past, but not extinct, populations allows us to put these theories to the test.

54 A dynamical systems simulator to evaluate methods for inferring co-expression networks

Dharmesh Bhuva and Melissa Davis

The University of Melbourne

Inferred gene regulatory networks can provide useful insight around genetic co-regulation during disease progression and such methods have identified novel pathological genes through 'guilt by association'. Numerous methods are available to infer such networks, with most recent approaches attempting to infer context specific or differential co-expression networks. Due to the sparsity of known regulatory interactions, however, there is a need for simulated data to properly assess different methods, especially for the latest inference methods. We have repurposed a simulator that uses models based on Boolean logic and systems of differential equations to simulate expression data. Activation signals are modelled by a single regulator using normalised Hill functions where the dissociation coefficients have been replaced with a more intuitive parameter that reflects the concentration required to achieve half maximal activation (EC50). The simulator provides five alternative classes of activation functions to choose from: linear, linear-like, sigmoidal, exponential and mixed types. Regulation of a target by multiple inputs is modelled using logic equations which specify the regulation mechanism using AND, OR and NOT functions. One major improvement over previous simulators is the simplicity with which a user can specify a model. Older simulators required users to specify a number of parameters which cannot be easily determined, such as dissociation rates for each interaction. We believe that this simulator will enable more thorough evaluation of inference methods under varying conditions, and we are in the process of developing a BioConductor package using S4 classes for easy implementation by end users. The improved inference of regulatory networks in disease may ultimately have implications in drug regimen stratifications and improve our understanding of complex diseases.

55 Statistical inference in single-cell lineages

Damien Hicks, Terence Speed, Mohammed Yassin, Raz Shimoni and Sarah Russell

Swinburne University of Technology

Differentiation patterns in single-cell lineages are a defining feature of embryonic development. In hematopoietic systems any such patterns are difficult to detect above the noise of phenotypic heterogeneity. To improve the signal-to-noise for detecting such differentiation patterns, non-trivial aggregates of lineage measurements have been identified from the symmetry invariance of a binary tree. Examination of these group symmetries has revealed a natural set of variables for describing patterns in a tree and showed that variation across the tree is composed of independent contributions from each division. The technique has been used to look for preferred fate determination stages in T cell lineages mapped using time-lapse microscopy over several generations. For comparison, the method has been applied to previously-published data from C. Elegans, a lineage with clear determination stages, and to a simulated branching process, which has none.

56 Multi-omic Characterisation of a Novel Xylose Metabolising Strain of Saccharomyces cerevisiae

Gustave Severin, Åsa Pérez-Bercoff, Psyche Arcenal, Anna Sophia Grobler, Philip J. L. Bell, Paul V. Attfield and Richard J. Edwards

UNSW

With growing demand for improved biofuel production the need for efficient conversion of xylose to ethanol is vital. Wild Saccharomyces cerevisiae (Baker's yeast) are commonly used in the production of biofuels, however they are unable to efficiently grow on xylose as a sole carbon source. Microbiogen Pty Ltd has evolved a novel xylose metabolising S. cerevisiae using a 15 year process of breeding and selection on xylose. To identify the genes that allow this strain to grow efficiently on xylose we have used a combination of PacBio whole genome sequencing, Illumina population resequencing, and RNA-Seq transcriptomics. Two loci were determined to be under significant positive selection when grown on xylose minimal media in competition with the s288c (reference yeast) variants of these genes. Both loci contain unique variants at the genomic and protein coding level, compared to known yeast genomes. One gene was identified as a master regulator of transcription. The second gene was identified as a dehydrogenase. RNA-Seq analysis of our xylose-metabolising strain was used to identify genes with significantly increased expression on xylose versus glucose minimal media. This highlighted two further candidate genes, previously shown to substitute for the key xylose metabolic proteins xylose reductase (XYL1) and D-xylulokinase (XYL3). Combined with the previously mentioned dehydrogenase, these may explain a complete and novel xylose metabolic pathway. Future work will focus on the confirmation on the role of these genes and their requirement for efficient growth on xylose.

57 Annotating single-cell RNAseq clusters by similarity to reference single-cell datasets

Sarah Williams, Sonika Tyagi and David Powell

Monash Bioinformatics Platform, Monash University

Single cell RNAseq is often used to examine cell types within tissue samples. There are a multitude of methods available for clustering sequenced cells into transcriptionally-similar groups, putatively corresponding to cell type or state. However, once the clusters are defined it can be difficult to determine (a) what cell type each cluster might represent and (b) how well the clustering method has reconstructed the cell-groups at a level relevant to the biological question of interest. This work aims to be able to take pre-computed cell-clusters and annotate possible cell type information to each cluster in a quick, accessible manner on the basis of similarity to publically available single-cell datasets. This will be done by comparing genes differentially expressed in a specific cluster of an experiment (compared to the rest of the experiment), with equivalent pre-computed signatures from publically available reference datasets. Initial experiments show a promising recapitulation of biologist-annotated cell types between public human and mouse brain tissue datasets. Further work will determine how well such an approach might identify shared cell types (e.g. endothelial cells) across different tissue types, and if this technique could be used to evaluate the selection of a particular set of cell cluster definitions in an experiment, on the basis of their biological relevance.

58 VPOT: a customisable tool for the prioritisation of annotated variants.

Eddie Ip, Sally Dunwoodie and Eleni Giannoulatou

Victor Chang Cardiac Research Institute

With the increasing use of Next Generation Sequencing (NGS) methods, whether Whole Exome Sequencing (WES) or Whole Genome Sequencing (WGS), researchers are now faced with an increasing number of variants, from hundreds of thousands to millions, to evaluate. To identify possible disease-causal candidates, annotation of these variants using a deleterious/pathogenicity prediction algorithm is required. There are a plethora of prediction algorithm scores available to be used for annotations. In most cases more than one is used in the annotation process to increase the likelihood of identifying deleterious variant candidates. However, by increasing the number of prediction scores to review, the prioritisation of variants becomes a more labour-intensive and cumbersome task. To simplify this, we have developed VPOT (Variant Prioritisation Ordering Tool) a python-based command line program that allows researchers to create a single deleterious /pathogenicity ranking score from any number of post- annotation values. Using this single score VPOT ranks the variants, thus allowing the researcher to see highly deleterious variants based on data from all the annotation prediction algorithms. By using post-annotation VCFs we still allow researchers to have full control of what annotation predictors they feel is most important to their data. The use of VPOT can be especially informative when dealing with multiple samples, as the prioritisation of variants can allow researchers to select top candidate variants from a multi-samples cohort. VPOT also has a gene list filtering option to allow refinement of any variant priority list.

59 PacBio sequencing, de novo assembly and haplotype phasing of diploid yeast strains

Richard Edwards, Asa Perez-Bercoff, Tonia Russell, Paul V. Attfield and Philip J.L. Bell

The University of New South Wales

PacBio Single Molecule Real Time (SMRT™) sequencing is rapidly becoming the technology of choice for de novo whole genome sequencing. The long read lengths and random error of PacBio data make genome assembly considerably easier and more accurate than short read data. In diploid genomes, heterozygosity – particularly in structural variants – generates some additional challenges for de novo genome assembly. For some regions, homologous chromosomes are assembled together into a chimeric sequence. However, long reads from heterozygous regions will often assemble into two distinct contigs, fragmenting the assembly. Despite these challenges, long reads present new opportunities for assembly of diploid genomes. Where the heterozygosity (i.e. density and number of polymorphisms) is high enough, polymorphisms can be phased into distinct haplotigs derived from a single parent. We have performed genome sequencing of novel diploid Saccharomyces cerevisiae strains using the PacBio RSII at the UNSW Ramaciotti Centre for Genomics. In addition, we have generated in silico diploid strains by combining sequence data from two haploid strains that were previously sequenced and assembled. Here, we report on progress regarding de novo diploid assembly and efforts towards full haplotype phasing of these strains. Our in silico diploids enable us to assess how successfully we can reconstruct known parental chromosomes and haplotypes. Ultimately, the goal is to develop a pipeline for consistent sequencing, assembly and annotation of full-length diploid chromosomes for each strain.

60 CNVmineR: A novel approach to identification of common and rare CNVs in WGS population studies

Jacob Munro and Eleni Giannoulatou

Victor Chang Cardiac Research Institute

There exist numerous tools for CNV identification, however most operate primarily at the individual sample level, and either ignore or make limited use of the shared information available when analysing a large population. We have developed a novel method, CNVmineR, that exploits the read depth distribution across samples to identify genomic regions harbouring common CNVs. Our approach utilises a correlation-based score between adjacent genomic bins as an efficient and effective way to identify the majority of genomic loci overlapping with known common CNVs. The correlation score is subsequently used for segmentation of these CNVs across adjacent bins. After locating common CNVs, non-parametric clustering is used to identify clusters within the samples corresponding to the copy number states. This information is then used to construct a model of the read depth at each genomic bin based on the truncated normal distribution. Samples are then assigned p-values based on their predicted cluster in each bin, and rare CNVs can be identified as stretches of bins with low p-values in a given sample. This method is a work in progress and is planned for release as an R package on Bioconductor in the near future. The package is designed to work with large numbers of samples (hundreds to thousands), and implements parallel computation and shared memory to efficiently process the vast amount of input data.

61 Deconstructing a molecular network of the aging frontal cortex.

Ellis Patrick, Ayla Ergun, Mariko Taga, Marta Olah, Hans-Ulrich Klein, Charles White, Daniel Felsky, Lori Chibnik, Julie Schneider, David Bennett, Elizabeth Bradshaw, Philip De Jager and Sara Mostafavi

The University of Sydney

The Accelerated Medicine Program in Alzheimer's Disease (AMP-AD) is performing multi-level 'omics analysis on post-mortem tissues taken from large numbers of AD patients from several large cohorts. RNA-Seq analysis of the AD brain transcriptome has the promise of providing important mechanistic clues to the molecular etiology of AD, however, analysis of transcriptional profiling data from post mortem brain tissue is complicated because of the significant interindividual variability in cellular composition. We examined transcriptomic profiles of bulk tissue-level profiles from the cortex of 542 subjects from the Rush Memory Aging Project (MAP) and Religious Orders Study (ROS) and developed several methods for predicting cell type proportions from this bulk tissue expression data. We compared these methods and using immunohistochemistry-based image analysis on 60 individuals demonstrated that molecular pathology can be used to effectively describe the proportions of neurons, astrocytes, microglia, oligodendrocytes and endothelial cells in brain tissue. Further to this, we demonstrated that these approaches can also be used to isolate cell sub-types that are associated with disease progression. Specifically, we identified a module of genes that capture the behavior of microglia associated with both the accumulation of tau pathology and cognitive decline. Through two independent image-based approaches we show that these microglia have a distinctly activated morphology. These results demonstrate that by accepting that tissue-level gene expression data captures a complex cocktail of cell-type and disease relevant signal, bioinformatics approaches can be used to deconstruct and model this complexity. Our analysis provides a valuable context for the broader gene expression changes observed in mRNA profiles of the human cortex and specifically highlights the molecular underpinnings of microglia dysfunction in Alzheimer's disease.

62 Spatial statistics analysis of super-resolution protein co-localization data

Greg Bass, Hanneke Okkenhaug, Llew Roderick, Vijay Rajagopal and Edmund Crampin

Systems Biology Laboratory, Department of Biomedical Engineering, University of Melbourne

Protein-protein interaction networks often omit the precise spatial relationships between proteins which may be critical in selectively controlling the behavior of the network. In cardiomyocytes, ryanodine receptors (RyRs) and inositol trisphosphate receptors (IP3Rs) both transmit calcium (Ca^{2+}) signals from intracellular stores into the cytosol. The principal roles of these channels are distinct however with RyR Ca^{2+} release directing the cell to generate mechanical force and IP3R Ca^{2+} release stimulating gene transcription. The activity of both channels is promoted by Ca^{2+}. Prior experiments have shown that IP3R activity is not readily detectable in the absence of active RyRs, meaning that IP3Rs and RyRs may interact in the same cellular compartment at the same time generating the same signaling messenger and yet coordinate distinct phenotypic responses. To explore this relationship, we measured the distributions of RyRs and IP3Rs at nanometer resolution. We then applied a non-hierarchical clustering method adapted from network graph theory to reconstruct spatial signalling islands for each population within the cellular space. The co-associations among these two populations were analyzed using spatial statistics techniques. We found that neither population was randomly distributed. RyRs were highly localized in a stripe-like pattern aligning with the contractile machinery, while IP3Rs displayed a more complex arrangement. In particular, a sub-population of IP3Rs clustered within or around RyR islands, while another sub-population of IP3Rs localized far from RyRs. Analysis of the signalling range between clusters suggested that the specific arrangement of IP3R islands could reduce the spatiotemporal distance between RyR signalling events. Our data and analysis suggest that IP3Rs may play a role in both coordinating cell-wide RyR Ca^{2+} release patterns and directly strengthening Ca^{2+} release at RyR sites, both of which may act to sustain the Ca^{2+} signals required to activate Ca^{2+}-mediated gene transcription.

63 Galaxy training for microbial genomics

Anna Syme, Torsten Seemann, Dieter Bulach, Simon Gladman and Andrew Lonie

Melbourne Bioinformatics

Microbial genomics relies on accurate genome assembly, annotation and variant calling. The Galaxy Platform is a relatively easy yet detailed way to access and integrate relevant bioinformatics tools for many of these tasks. Using these tools, we developed training material for microbial genomics to complement a national project about antibiotic resistance, and we present an overview here.

64 Predicting the outcome of breast cancer using novel RNA-Seq analysis

Andrew Pattison, Paul Harrison and Traude Beilharz

Monash University

With the exception of skin cancers, breast cancer is the most common cancer affecting women. While progress has been made in the detection of primary breast tumours, there are few genomic tests that are able to accurately predict outcome. Current genomic tests such as as Mammaprint and Oncotype DX are not widely available and are only suitable for early stage tumours, with additional restrictions applying depending on the test used. We sought to derive a new predictor of breast cancer outcome from TCGA RNA-Seq data that can provide an accurate indication of prognosis, even in later stage tumours. Alternative polyadenylation (APA) is the process whereby the poly(A) tail is added to the 3′ untranslated region (3' UTR) of a messenger RNA (mRNA) at one of multiple possible sites, changing 3' UTR length and potentially the regulatory elements that bind to it. APA has been suggested to be predictive of tumour outcome and can be inferred from RNA-Seq data. We used elastic net linear modelling to select coefficients that best predict relapse free survival from clinical, APA and gene expression data. The best model was generated using a combination of all 3 data types, with common clinical indicators playing only a small role. Using 10 fold cross validation, patients with a score higher than the median generated by our model were at least 16 times less likely to die of cancer than those with a score below the median ($p << 0.01$). Our ultimate aim is to derive an accurate genomic test for breast cancer outcome that can be applied to all breast tumours and is less reliant on clinical data. This test could potentially be implemented using the in house M-PAT approach, for substantially less than the cost of a full RNA-Seq experiment.

65 Family–based whole exome sequencing study of childhood apraxia of speech provides insight into the genetic basis of speech disorders.

Victoria E Jackson, Thomas S Scerri, Michael S Hildebrand, Ingrid E Scheffer, Olivia van Reyk, Samantha Turner, Angela T Morgan and Melanie Bahlo

Population Health and Immunity Division, The Walter and Eliza Hall Institute of Medical Research, Melbourne

Childhood apraxia of speech (CAS) is a rare developmental disorder of speech motor programming that results in impaired precision and consistency of the movements underlying speech. We conducted whole exome sequencing (WES) on 23 individuals with CAS, and their families. After removal of common variants (population minor allele frequency >5%), and variants not located within exonic or splice regions, a total of 33,346 rare variants were observed across the 23 probands. We searched for mutations, firstly in genes previously implicated in CAS, and secondly through a genome-wide approach, and highlighted variants of interest based on predicted consequences, and known biology. For individuals with WES data available from their parents we prioritised de novo mutations, as well as inherited mutations in probands with an affected parent. Multiple candidate genes were identified. Short tandem repeat expansion analysis was also performed with exSTRa for 21 loci known to cause neurogenetic disorders, some of which overlap CAS in phenotype. No expansions were identified. There are currently no large published genome-wide association studies (GWAS) for CAS. However, the recent release of the UK Biobank data provided an opportunity to examine CAS related phenotypes captured with International Classification of Diseases (ICD) codes. We downloaded summary results of a GWAS of speech disturbances, as defined by ICD-10 code R47, via https://sites.google.com/broadinstitute.org/ukbbgwasresults/. Multiple common SNPs in genes highlighted by the CAS WES analysis showed modest association (p-value<5x10-3), suggesting these genes may have a potential role in other disorders affecting speech. These results suggest that speech disorders are underpinned by both rare and common variants and that more genes are awaiting discovery.

66 The causative role of Serine and Glycine on Macular Telangiectasia - a Mendelian randomization approach

Roberto Bonelli, Luca Lotta, Ferenc Sallo, Traci Clemons, Mactel Consortium, Catherine A Egan, Marcus Fruttiger, Claudia Langenberg and Melanie Bahlo

The Walter + Eliza Hall Institute

Macular telangiectasia type 2 (MacTel), is a rare and often underdiagnosed degenerative eye disease that may result in blindness. In 2017, we published the first five genetic loci involved in this disease discovered from genome-wide association study. Four out of the five loci were previously being connected with the glycine/serine metabolism pathway. In the same study, we identified glycine, serine and threonine to be the mostly differently abundant metabolites in MacTel. Here, we present the causative analyses on the role of these metabolites in MacTel disease. Genetically Predicted Metabolites (GPMs) constructed from SNP data can be used to test causality between metabolites and diseases. By analysing SNP data on 476 MacTel cases and 1733 controls we constructed GPMs for 140 metabolites. From our analysis, serine and glycine were the only two metabolites to have a causative role on MacTel. Genetically predicted serine showed a higher association with the disease (p=1.5E-31) when compared with glycine (p=3.6E-20). Although highly significant in the differential abundant analysis, threonine did not appear to be causally associated with the disease (p=0.902). Further, to assess their effect on disease progression, we performed a case-only analysis in 455 MacTel patients testing the association between GPMs and specific disease phenotypes. Genetically predicted serine was associated with a higher risk for the progression of various macular abnormalities, while glycine only had a small effect on one disease characteristic. Our results confirm that serine and glycine have a causative role in the development and progression of MacTel disease. These results will help elucidate the disease mechanism in future work leading to better prognosis and future treatments for MacTel patients.

67 People Powered Protein Predictions!

Stuart Lee, Roberto Bonelli, Brendan Ansell and Saskia Freytag

Monash University

Computational methods for predicting the three-dimensional (3D) structure of a protein provide biologists with valuable information about its function. Such algorithms generate 3D atomic models from the amino acid sequence, and then align them to solved crystal structures. After applying a protein prediction structure algorithm such as I-TASSER to your favourite organism you are bombarded with many numbers that assess the quality and accuracy of the algorithm's prediction. How can you follow up on these predictions while alleviating your uncertainty about their truth? The I-TASSER suite provides a score that allows you to discriminate high quality from low quality predictions but these are based on unpublished observations. If you kept only putative high-quality structures for human proteins, you would discard 60 per cent of your predicted structures. If you have a more neglected organism this proportion is likely to be higher. We propose to improve these quality assessments by using people power. We will present people (who are not necessarily protein experts) with a simple pattern matching task: they will view a predicted protein structure aligned to its solved crystal structure. We will then ask them whether they think the match is high or low-quality as well as the confidence in their answer. The resulting data in combination with the I-TASSER quality scores will be used to train a machine-learning classifier, with the aim of generating more certain predictions.

68 Gut microbiome changes in T1D during pregnancy and early life

Alexandra J. Roth Schulze, Katrina M. Ngui, Megan A. S. Penno, Lynne Giles, Rebecca L. Thomson, Jenny J. Couper, John M Wentworth, Kelly McGorm, Maria E. Craig, Peter G. Colman, Elizabeth A. Davis, Aveni Haynes, Mark Harris, Andrew M. Cotterill, Peter J. Vuillermin, Claire Morbey, Georgia Soldatos, William D. Rawlinson, Grant Morahan, Simon C. Barry, Richard O. Sinnott, Anthony T. Papenfuss and Leonard C. Harrison

The Walter and Eliza Hall Institute of Medical Research, Parkville, Victoria, Australia

Environmental Determinants of Islet Autoimmunity (ENDIA) (endia.org.au) is a longitudinal study from early pregnancy through early life of environment-gene interactions in 1,400 infants at risk for type 1 diabetes (T1D). Current knowledge supports the idea that the microbiome shapes the development of a normal immune system from birth to adulthood. Accordingly, it has been hypothesized that abnormal development of or a shift away from a healthy microbiome in individuals genetically at risk of T1D predispose to islet autoimmunity and T1D. It is therefore imperative to understand how the gut microbiome matures from birth, how the mother's microbiome and other environmental factors (e.g. mother's diet and antibiotic use during pregnancy) influence the infant's microbiome and ultimately how these changes in the microbiome modify the risk for T1D. We have been investigating longitudinally from birth the bacterial microbiome taxonomic composition of infants at risk for T1D along with the microbiome of their mothers during pregnancy. Briefly, our study has provided evidence of significant differences in the gut microbial community composition of: 1) women with and without T1D during pregnancy and 2) infants of mothers with and without T1D, with particular bacteria being decreased in abundance in mothers with T1D and in their offspring. Based on their taxonomy, these bacteria are potentially involved in the production of anti-inflammatory compounds (e.g. butyrate). How a deficiency of anti-inflammatory gut bacteria is involved in T1D remains to be determined.

69 Utilising mixture models for unveiling patterns in scRNA-Seq data

Yingxin Lin, Shila Ghazanfar, Pengyi Yang and Jean Yang

The University of Sydney

Single cell RNA-Sequencing (scRNA-Seq) has enabled unprecedented insight into the behaviour of individual cells on the scale of the entire transcriptome. Such precision offers an opportunity to explore cell-specific heterogeneity, however two distinct features arise from such data: (1) hyperinflation of identically zero counts for the majority of genes for any given cell, and (2) an apparent bimodal distribution of non-zero counts. Both features are unique to scRNA-Seq, and warrant further development of statistical tools in order to answer biological questions of interest. We propose a mixture modelling framework to classify cells into three transcriptional states for each gene: (1) no, (2) low, and (3) high gene expression. This approach has the potential to reveal the cell-specific dynamics of RNA transcription (bursting) and degradation, as well as acting as a cross-dataset standardisation. We conducted a comparison of four particular models using either gamma-gamma or gamma-normal mixture models, and either performed independently across genes or constrained to ensure the first gamma component (lowly expressed) parameters are common across all genes. Comparison was conducted using metrics such as the Bayesian Information Criterion (BIC) to identify the most parsimonious mixture model type across all profiled genes. As a result, in addition to a standardised dataset, specific gene features can be obtained via the estimated parameters of each mixture model fit and used for further characterisation of genes, e.g. to identify especially highly or lowly variable genes. We utilised a number of publicly available scRNA-Seq datasets, stemming from mouse neuronal cell populations, to perform the mixture model comparison, assess highly and lowly variable genes, and to estimate cell networks via a uniqueness thresholding.

70 DECENT: Differential Expression with Capture Efficiency AdjustmeNT for Single-Cell RNA-seq Data

Agus Salim, Terence P. Speed and Chengzhong Ye

La Trobe University and Walter & Eliza Hall Institute for Medical Research

Recent development in sequencing technology has enabled high-throughput sequencing of transcriptome at single-cell level. The single-cell technology has already led to profound new discoveries that could not be made using data from bulk sequencing. Despite the technological advance current single-cell data are noisy with excess of zero counts a common phenomenon, creating considerable challenges for researchers when performing statistical modelling and inference from these data. In this talk, I will present a statistical framework called DECENT that can be used to impute technical zeroes and differentiate biological from technical zeroes when comparing gene expression across two biological conditions. DECENT uses zero-inflated negative binomial (ZINB) to model the pre-dropout count and uses EM algorithm to estimate parameters and likelihood-ratio test (LRT) is used to perform differential expression analysis. DECENT requires UMI count data as input and it works with and without spike-ins. We demonstrate the performance of our approaches using simulated datasets and three real datasets of breast cancer cells, adipose-derived stem cells and Induced Pluripotent Stem Cells (iPSC). The results demonstrate the advantage of our methods when compared to existing approaches for DE analysis with single-cell RNA-seq data includin SCDE, MAST, zingeR and Monocle.

71 Molecular Dynamics Modelling of a Variant of Unknown Effect in RAD51D

Matthew Wakefield, Micheal Kuiper, Olga Kondrashova, Kristy Shield-Artin and Clare Scott

Melbourne Bioinformatics, The University of Melbourne

High-grade epithelial ovarian carcinomas (OC) containing mutated BRCA1/2 have homologous recombination defects and are sensitive to poly(ADP-ribose) polymerase inhibitors (PARPi). In a clinical trial of the PARPi rucaparib (ARIEL2 Part 1, Clovis Oncology) a patient was observed with a germline truncating mutation in RAD51D (c.770_776del, p.G258Sfs*50) and a secondary mutation (c.770_776delinsA, p.S257_R259delinsK) in a biopsy of a splenic lesion that was progressing on PARPi therapy. Evolutionary analysis and molecular dynamics modelling were used to assess the function of this variant of unknown effect alongside the functional wild-type variant(s). Results indicated that the observed differences in amino acid sequence between the secondary mutation and wild-type RAD51D were unlikely to disrupt normal function and are evolutionarily well tolerated. The secondary mutation (c.770_776delinsA, p.S257_R259delinsK) would likely mirror the function of wild-type RAD51D, thus would restore function and lead to PARPi resistance. This prediction was confirmed by CRISPR directed homology repair introduction of the secondary mutation into a human ovarian cancer cell line, PEO4, which demonstrated a decreased cisplatin and rucaparib sensitivity relative to a PEO4 RAD51D knockout. In conclusion, the secondary RAD51D mutation (c.770_776delinsA, p.S257_R259delinsK) identified in this lesion most likely contributed to or caused the PARPi resistance and lesion progression.

72 Large-scale chromosomal changes dominate the genomic landscape of end-stage melanoma

Ismael A Vergara, Shahneen Sandhu, Lachlan McIntosh, Christopher P Mintoff, Richard J Young, Andrew Colebatch, Xuelin Dou, Stephen Q Wong, Jennifer Mooi, Clare Fedele, Samantha Boyle, Gisela Mir Arnau, Daniel S Widmer, Philip Cheng, Valerie Amann, Mitchell P Levesque, Reinhard Dummer, Nicholas Hayward, Richard A Scolyer, Raymond J Cho, David Bowtell, Heather Thorne, Kathryn Alsop, Sarah-Jane Dawson, Grant McArthur, Graham Mann, Mark Shackleton and Anthony T Papenfuss

Peter MacCallum Cancer Centre

Australia has one of the highest incidences of melanoma in the world and it has been referred to as our national cancer. Survival rates for melanoma are poor if not caught early. Recently, understanding of the molecular events that dominate the landscape of primary disease has benefitted from genomic sequencing, but how melanoma evolves into its metastatic and lethal form is poorly understood. To help rectify this, a rapid autopsy program, CASCADE (CAncer tiSsue Collection After DEath) was established at the Peter Mac that provides multi-region sampling of metastases from patients at time of death. We obtained sequencing data from more than 70 samples from 13 patients, including WES, WGS and RNA-seq. The matricial nature of this dataset prompted us to apply an analysis approach that builds on existing methods and makes use of the multiple samples from each patient. Our analysis reveals striking patterns in the evolution of lethal melanoma. While early melanomas have large numbers of single nucleotide variants, we generally observed limited subsequent SNV and indel gain. Rather, evolution was dominated by large-scale copy number change including a remarkable level of loss of heterozygosity in some patients. In one case, multicore sampling revealed spatial heterogeneity in copy number of the primary tumour. Patterns of copy number change hinted that two mutational processes, aneuploidy and genome doubling, were operating universally. To test this we developed a novel method that models these mechanisms using branching processes. Our findings in lethal melanoma suggest possible biomarkers that might be useful clinically in challenging settings where patients present significant clinical heterogeneity such as stage III disease. We are further developing these using a training set of 55 sequencing datasets from primary disease.

73 CAVALIER: an R package to produce reports for variant interpretation in clinical meetings

Mark Bennett, Yu-Chi Liu, Karen Oliver, Ingrid Scheffer, Samuel Berkovic and Melanie Bahlo

The Walter and Eliza Hall Institute of Medical Research

Next-generation sequencing is commonly undertaken to diagnose rare variants that cause Mendelian diseases. However, an individual typically carries thousands of genetic variants and dozens of rare variants that may plausibly cause the disease. With the large amounts of sequencing data being produced, reducing the burden of interpreting variants of unknown significance is an increasingly important problem, especially in the context of clinical interpretation of variants at multi-disciplinary meetings. Improving the evaluation and prioritization of candidate variants provides a significant benefit as functional studies to determine the biological consequences of a variant require significant resources. We have developed cavalier, a publically available R package designed to assist with variant interpretation. Multiple variant filtering modes have been implemented, relevant to different genetic hypotheses for simple family models including de novo, dominant, parental mosaic, recessive, and compound heterozygous, as well as a targeted search of a list of known disease genes. Numerous resources provide valuable information for interpreting genetic variants, however querying each manually for a shortlist of variants can be time consuming. Cavalier automatically collates and displays information from a number of useful external resources, including functional change predictions, variant frequency in different populations, tissue specific gene expression profile and known gene-phenotype, to enable rapid variant evaluation. Presentation ready results are generated in both PDF and HTML formats, suitable for clinical reporting and discussion. We present several interesting case studies of epilepsy patients showcasing the capabilities of cavalier for evaluating candidate disease

causing variants. We discuss how integrating information from external resources improves the ability to efficiently interpret and prioritise the most promising candidate variants. Cavalier can be downloaded from https://github.com/bahlolab/cavalier. We welcome suggestions for further improvements.

74 Transcriptome assembly and population differentiation analysis in Echinometra sea urchins, subjected to elevated pCO2 at volcanic vents

Nandan Deshpande, Sven Uthicke and Marc Wilkins

Systems Biology Initiative

Sea urchins (Echinoidea) are Echinoderms and are considered model organisms in marine ecosystems to study development and evolution. Burning of fossil fuels and deforestation has resulted in a sharp rise in atmospheric CO2, leading to ocean acidification (OA). In this work we investigate, using next generation sequencing techniques, if OA affects the gene expression profiles of sea urchins and whether genetic variation provides potential for adaptation. This was achieved by i) population genomic analysis from 4 urchin populations including one at a carbon vent with high pCO2 emulating futuristic low pH conditions and ii) exposing larvae from different populations to control and vent conditions for 48h and subsequent genomic analysis. A total of 36 Gb of reads were sequenced using the NextSeq Illumina sequencing technology for 4 adults and 16 libraries of larval Echinometra sea urchins. The larval libraries were assembled individually using Trinity assembler and only those transcripts, which mapped to the sea urchin Strongylocentrotus purpuratus transcriptome, were retained. The algorithm CD-HIT was used for clustering the representative transcripts for the Echinometra sea urchin. The transcriptome was further validated using the BUSCO tool and functionally annotated using Blast2GO. We then used this transcriptome to perform a comprehensive population differentiation analysis. We first mapped the individual samples, both adult and larval against the assembly, followed by computing the allele frequencies at all positions. We used the tool popoolation2 to compare across multiple conditions and identify SNPs and their fixation indices (fst). Outlier SNPs were detected using empirical methods and BAYESCAN across three primary comparisons: potential for adaptation, potential for reverse adaptation and potential intrinsic adaptation from OA. We have also performed a differential expression analysis across the primary conditions to account for and minimize any possible changes in the allele frequency with changes in allele-specific expression.

75 Using single cell RNA-Seq profiles to study heterogeneities in mouse mammary gland

Yunshun Chen

WEHI

Single-cell RNA sequencing (scRNA-seq) has become a widely used technique that allows researchers to profile the gene expression and study molecular biology at the cellular level. It provides biological resolution that can not be achieved with conventional bulk RNA-seq experiments on cell populations. Here we demonstrate a complete bioinformatics analysis of single-cell RNA profiling of mouse mammary epithelial cells. scRNA-seq data from both Fluidigm C1 and 10X Chromium platforms are used to study the degree of heterogeneity within the two primary cellular lineages of the mammary epithelium. We report the epithelium undergoes a large-scale shift in gene expression from a relatively homogeneous basal-like program in pre-puberty to distinct lineage-restricted programs in puberty. In addition, we uncover a luminal transit population and a rare mixed-lineage cluster amongst basal cells in the adult mammary gland.

76 Copy number variations from RNA-seq gene expression data

Stuart Archer

Monash University

Analysis of gene expression in RNA-seq experiments can be complicated by unexpected copy-number variations (CNVs) in the samples under study. Large CNVs, either arising randomly or under selection via experimental treatments, can result in altered gene expression for groups of nearby passenger genes, but these expression changes would mostly be irrelevant to the treatment under study. To identify CNVs, DNA-seq data would be preferred over RNA-seq data, however, in many cases RNA-seq data could and probably should still be screened for signatures of large CNVs if it is the only data available from the samples. I am testing potential strategies to screen for large CNVs in RNA-seq data using existing and new tools for RNA-seq, in simulated and real data. This will help screen out expression changes due to passenger genes on CNVs, thus improving signal to noise ratio of biologically meaningful gene expression changes in response to treatment.

77 Performance of analysis software on TCGA exomes

Christoffer Flensburg and Ian Majewski

WEHI

DNA sequencing is an extremely powerful tool to understand cancer, but the analysis can often be challenging. There is a wide range of methods geared towards identifying different kinds of somatic mutations, each one with its own advantages and drawbacks. State of the art analysis of cancer exomes typically involves multiple tools that pipe output from one program to the next, but the upstream tools are not always optimised for the downstream use. This can result in severe quality issues of the final output of the pipe, despite each individual tool working as intended. We have developed superFreq, a tool that combines SNV and CNA calling with clonal tracking over multiple cancer samples from the same individual. Merging these three steps into one software allows for a more reliable analysis compared to current practice. To assess performance of state of the art tools for single nucleotide variants (SNVs), copy number alterations (CNAs) and clonal tracking, we used in silico mixing of exomes from The Cancer Genome Atlas to build a simulated control dataset. SuperFreq shows robust performance in the simulations and correctly handles difficult cases like low purity, low mutational burden or subclones more often than other tools.

78 The Monash Bioinformatics Platform (MBP) – Making bioinformatics accessible.

Nick Wong, Kirill Tsyganov, Adele Barugahare, Paul Harrison, Stuart Archer, Andrew Perry, Anup Shah, Haroon Naeem, Sarah Williams, Sonika Tyagi and David Powell

Monash University

The Monash Bioinformatics Platform (MBP) is one of many Research Technology Platforms available to the Monash research community. Our modus operandi is to empower the researcher in bioinformatic analyses. This is very different to a fee for service model. The team spans a diverse set of skills and expertise including biology, computer science and mathematics. We are located between the Monash Clayton and the Alfred campus and actively engage with the bioinformaticians and wet-lab researchers via various forums and online communication tools. We carry out collaborative research at Monash in the genomics and proteomics areas including but not limited to bulk and single cell RNAseq, Ribo-seq, ChIPseq, ATAC-seq, variant analysis and high throughput profiling of proteome and studying protein structures. We have developed several bioinformatics tools and applications tailored to particular research problems, some examples are: Degust, TCP-seq, Varistran, TopConfects. We offer tools for data analysis, visualisation and exploration. Degust is our flagship tool for differential expression analysis of RNA-Seq data. We also provide access to reproducible compute environments by developing bio-ansible containers to create a full installation of bioinformatics tools in re-usable modules on HPC, the cloud or a desktop environment. We promote community building through software carpentry styled training workshops and actively develop new training material for these workshops to teach modern techniques including tidyr and RShiny. We also run regular informal seminars from speakers across the community. As a platform, we facilitate and link researchers across the Monash campus with common research goals and interests.

79 Signature-based binning for metagenomic analysis

Timothy Chappell, James Hogan, Shlomo Geva, Dimitri Perrin and Lawrence Buckingham

Queensland University of Technology

Metagenomic analysis is highly sensitive to the coherence of the operational taxonomic units formed during the binning stage; however, with the increasingly large scale of metagenome samples rapidly outpacing the availability of computational resources for this task, more efficient approaches to the binning of reads and contigs are needed. In this work we present SigClust, an approach to binning that involves the transformation of sequence reads into read signatures that preserve pairwise similarity while markedly reducing the computational burden involved. With this technique we can rapidly and efficiently form highly coherent clusters, providing a good starting point for analysis. We illustrate the comparative performance of this method against state-of-the-art sequence binning tools on metagenomic samples drawn from a number of environments, and on a large set of closely related bacterial genomes. Furthermore, we present a practical use of this approach in the analysis of bacterial microbiota sampled in the context of a study of healing and non-healing wounds.

80 Building user friendly applications for biologists

Jarny Choi and Christine Wells

The University of Melbourne

A data analysis application aimed at the bench biologist is usually a package that comprises of relevant data and/or bioinformatics methods and visualisation tools. Such applications can be beneficial in many projects and are becoming increasingly popular, ranging from small Shiny apps to major data portals. Benefits include empowering the biologist who own the data for direct exploration (thus increasing the chances of new hypotheses being generated), enhancing collaborations and reducing burden on the bioinformatician. These applications can also be used to promote particular bioinformatics methods to a wider audience. It can however be time consuming to develop an application and one must weigh the benefits against the cost. I will discuss the key issues to consider when building such an application, and pass on some lessons learnt from years of doing such work across multiple projects: scoping, defining the target audience, choosing the right technologies and visualisation tools, etc. I will also use a newly developed standalone javascript application called iscandar (https://github.com/jarny/iscandar) to illustrate some of the points in context of single cell data analysis. Iscandar is an interactive reporting tool designed for single cell RNA-seq datasets, where the user can view PCA and TSNE plots and explore gene expression in these plots as well as define custom clusters.

81 The Effect of Binding on the Enantioselectivity of an Epoxide Hydrolase

Julian Zaugg, Yosephine Gumulya, Mikael Bodén, Alan Mark and Alpeshkumar Malde

School of Chemistry and Molecular Biosciences, The University of Queensland

Molecular dynamics simulations and free energy calculations have been used to investigate the effect of ligand binding on the enantioselectivity of an epoxide hydrolase (EH) from Aspergillus niger. Despite sharing a common mechanism, a wide range of alternative mechanisms have been proposed to explain the origin of enantiomeric selectivity in EHs. By comparing the interactions of (R)- and (S)-glycidyl phenyl ether (GPE) with both the wild type (WT, E = 3) and a mutant showing enhanced enantioselectivity to GPE (LW202, E = 193) we have examined whether enantioselectivity is due to differences in the binding pose, the affinity for the (R)- or (S)- enantiomers or a kinetic effect. The two enantiomers were easily accommodated within the binding pockets of the WT enzyme and LW202. Free energy calculations suggested that neither enzyme had a preference for a given enantiomer. The two substrates sampled a wide variety of conformations in the simulations with the sterically hindered and unhindered carbon atoms of the GPE epoxide ring both coming in close proximity to the nucleophilic aspartic acid residue. This suggests alternative pathways could lead to the formation of a (S)- and (R)-diol product. Together the calculations suggest that the enantioselectivity is due to kinetic rather than thermodynamic effects and that the assumption that one substrate results in one product when interpreting the available experimental data and deriving E-values may be inappropriate in the case of EHs.

82 SeqScrub: A web tool for automatic cleaning of FASTA file headers.

Gabriel Foley and Mikael Bodén

School of Chemistry and Molecular Biosciences, University of Queensland

Ensuring data consistency and minimising time spent on sanitising input data is crucial to bioinformatics workflows. Allowing collaborators to access tools that empower them to easily perform the same consistency checks makes data sharing across large-scale projects feasible. We developed SeqScrub as a web tool that streamlines the process of removing extraneous information from FASTA file headers while retaining a unique identifier and taxonomic information. SeqScrub uses identifiers to query external databases in order to ensure data remains consistent and species annotations are accurate. Headers that are standardised using this tool can then be parsed by a large range of bioinformatics tools, stay uniformly named between collaborators, and retain informative labels to aid with further research. SeqScrub is an example of how the process of responsive development can create tools that are suited to users' needs. We provide an easy to use method for performing a repetitive yet essential step that was built through consultation with its intended users. SeqScrub illustrates the importance of exposing even simple pipelines and techniques and making them easily accessible to collaborators.

83 Small data bioinformatics: identifying leaderless secretory proteins in plant cell walls with limited sample data

Andrew Lonsdale, Melissa Davis, Monika Doblin and Tony Bacic

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

Leaderless secretory proteins (LSPs) are proteins that are secreted into the extracellular space yet lack the canonical N-terminal signal peptide sequence. The routes these proteins take are not fully understood, and it is an active area of research. Using proteomics to study LSPs in plant cells is further complicated by the open compartment nature of the cell surface, cell wall and apoplastic space. Typically a combination of destructive and non-destructive lab methods are used in the preparation stages in order to maximise the coverage of proteins. Both of these methods are thought to lead to the high number of potential LSP proteins found in plant cell wall proteomes. Finding sequences like these in a proteomic study poses a challenge. They could be genuine LSPS. They could be intracellular contamination. Knowing which is which requires either (a) follow up experiments using biochemical approaches such as immuno-localisation, or (b) comparison to known LSPS. Since (a) takes time and money, there are very few (b) to compare to. Bioinformatics techniques to predict likely LSPs from candidates would be a good solution, but how do we build a prediction tool without positive sample data? This work proposes using data from experimental observation, protein features relevant to the secretory environment, gene ontology terms and, protein interaction networks to distinguish likely candidate LSPs from contamination. Features associated with secretory and non-secretory proteins respectively are used classify potential LSPs and a create a reference set for further work in predicting leaderless secretory proteins in plants.

84 Using Singular Value Decomposition to alleviate batch effects in RNA sequencing data

Alexandra Garnham, Hannah Vanyai, Tim Thomas, Anne Voss and Gordon Smyth

The Walter and Eliza Hall Institute of Medical Research

Batch effects can be deleterious to an RNA sequencing (RNA-seq) analysis. They have the potential to mask true results, give false positives if not identified and appropriately corrected, leading to an incorrect biological interpretation of the results. A number of promising tools have been proposed for detecting unwanted and unexpected batch effects, but they can be unreliable or difficult to use for complex RNA-seq experiments with many experimental factors but only very small numbers of biological replicates. Here we use a linear modelling approach to RNA-seq data, and apply a singular value decomposition (SVD) to the residual space of the linear model to identify dimensions that can explain a large proportion of the variability. This approach is similar to surrogate variable analysis except that we transform to an orthogonal residual space before analysis. We present a case study on the role of the histone acetyltransferase gene Moz in mouse palate development. We show that SVD identifies surrogate variables that have biological interpretations. The SVD analysis allows gene and pathway targets of Moz to be identified with greater precision.

85 Learning Epistatic Interactions from Sequence-Activity Data to Predict Enantioselectivity

Julian Zaugg, Yosephine Gumulya, Alpesh Malde and Mikael Bodén

School of Chemistry and Molecular Biosciences, The University of Queensland

Enzymes with a high selectivity are desirable for improving economics of chemical synthesis of enantiopure compounds. To improve enzyme selectivity mutations are often introduced near the catalytic active site. In this compact environment epistatic interactions between residues, where contributions to selectivity are non-additive, play a significant role in determining the degree of selectivity. Using support vector machine (SVM) regression models we map mutations to the experimentally characterised enantioselectivities for a library of 136 variants of the epoxide hydrolase from the fungus Aspergillus niger (AnEH). We investigate whether the influence a mutation has on enzyme selectivity can be accurately predicted through linear models, and whether prediction accuracy can be improved using higher-order counterparts. Comparing linear and polynomial degree = 2 models, mean Pearson coefficients (r) from 50 × 5-fold cross-validation increase from 0.84 to 0.91 respectively. Equivalent models tested on interaction-minimised control sequences achieve values of r = 0.90 and r = 0.93. Testing linear and polynomial degree = 2 models on additional AnEH mutants, values increase from r = 0.51 to r = 0.87 respectively. The study demonstrates that linear models perform well, however the representation of epistatic interactions in predictive models improves identification of selectivity-enhancing mutations. The improvement is attributed to higher-order kernel functions that represent epistatic interactions between residues.

86 Development of computational methods to analyse single-cell high-dimensional mass spectrometry data

Marie Trussart, Charis Teh, Daniel Gray and Terry Speed

The Walter and Eliza Hall Institute of Medical Research Technological advances with cytometry time of flight mass spectrometry CyTOF (Bandura et al. 2009) has successfully enabled a comprehensive panel of surface and intracellular protein markers to unravel complex signalling networks and to delineate cell subsets in heterogeneous tissues such as blood, bone marrow and tumours (Bendall et al. 2011; Newell et al. 2012). Indeed, mass cytometers are able to analyse simultaneously more than 40 unique parameters per sample at the single cell level. Some markers are directly measuring a cellular process such as apoptosis (caspase 3 cleavage) or proliferation (cyclin expression) but also signalling (Behbehani et al. 2012). In the area of infectious disease and cancer, CyTOF is used to identify and quantify populations of immune cells, cancers cells or inflammatory cell states which allows the detection of clinical biomarkers and drug therapy resistance by comparing various patient groups (Amir el et al. 2013; Newell et al. 2012; Bruggner et al. 2014; Aghaeepour et al. 2013). In this quantitative technology, variation in instrument performance have been observed caused by both instrument calibration and fluctuations in signal strength. A multiplicative correction has been derived from control bead standards and is widely applied to normalize the raw mass cytometry data (Finck et al. 2013). Another pre-processing step is the implementation of the de-barcoding algorithm (Zunder et al. 2015). We are currently developing a computational approach with a different de-barcoding strategy and a new method for handling and removing such technical unwanted variation across runs in high-dimensional mass spectrometry data. Furthermore, we are also performing statistical analyses to understand the molecular changes induced by different treatments on cancer cell lines, clustering group of cells with similar markers profiles and investigate whether there are significant changes in the markers intensities and relative frequencies of the treated cell populations compared to the non-treated cells.

87 Calling variants from RNA-Seq data identifies significant eQTLs in a small cohort of asthma patients

Artika P. Nath, Milica Ng, Nick J. Wilson, Michael J. Wilson, Michael Inouye and Monther Alhamdoosh

The University of Melbourne

Deciphering the mechanism of action (MoA) of heterogeneous diseases, such as asthma, is essential for target and biomarker discovery. MoAs usually comprise gene regulatory mechanisms, genetic factors, and/or environmental factors. RNA-sequencing has been widely used to understand gene regulatory networks and genome-wide association studies (GWAS) have been utilized to map disease susceptibility loci. However, the identification of causal variants at these loci and their mechanistic effects remains a fundamental challenge. Traditionally, expression quantitative trait loci (eQTL) analysis has been used to identify genetic variants that affect gene expression levels by combining genotype data and transcriptomic data generated by RNA-seq or Microarrays. In this study, we show that RNA-seq data generated from airway epithelial cells can be solely utilized to identify significant eQTLs from a cohort of 57 asthmatic and 28 non-asthmatic adults. Firstly, raw reads were aligned to the human reference genome and then SNP variants were called and annotated for all the samples. Second, an imputation pipeline was used to infer non-coding variants using the 1000 genomes as a reference panel. Finally, joint eQTL association testing was performed on all the samples for each of the 14,406 transcripts and the 657,815 SNP genotypes that retained after pre-processing and quality control. Cis-eQTL effects were identified for 1,746 unique genes at a false discovery rate (FDR) cut-off threshold of 0.05. Of these unique genes, 150 genes were differentially expressed when asthmatic patients were compared with healthy individuals. Moreover, a large number of significant eQTLs were found in the GWAS catalogue and occurred at a regulatory locus on chromosome 17q12-21 that is known to be associated with asthma-related traits. In conclusion, this study demonstrates that the millions of sequences generated from RNA-seq experiments can be efficiently used to improve understanding of

MoA and biomarker discovery in complex disease setting.

88 Identification of epigenetic complexes driving haematopoiesis

Yih-Chih Chan, Enid Lam, Brian Gloss, Jessica Morison, Marcel Dinger, Anthony Papenfuss and Mark Dawson

Peter MacCallum Cancer Centre

All blood cells have a finite life span, and haematopoiesis is a precise balance of self-renewal and differentiation of haematopoietic stem cells. Haematopoietic stem cells are a rare population of cells and the only haematopoietic cells that are able to self-renew, but they are also able to generate the one trillion blood cells that develop within an adult bone marrow each day. The disruption of this process results in various blood disorders including cancer. However, the molecular mechanism, such as the epigenetic and transcriptional programs, that govern the critical decisions of self-renewal and differentiation are not well understood. Polycomb (PcG) and Trithorax (TrxG) groups of proteins are evolutionarily conserved epigenetic modifiers that can control transcriptional repression and activation, respectively, of key genes in cellular differentiation and development. PcG or TrxG protein complexes comprise of core groups of proteins, some of which have 5-6 different members, as well as facultative proteins. Consequently it has been estimated that over 180 distinct PcG or TrxG complexes exists, each thought to have distinct functional roles. To unravel this complexity, we used in silico methods on RNA-Seq data to reconstruct the major PcG or TrxG complexes that dominate each main stage of haematopoietic development. Using a combination of expression profiling, clustering and rank based methods of ten different haematopoietic cell types, we identified the main PcG or TrxG complexes in haematopoietic stem cells, committed progenitors and terminally differentiated bloods cells. Importantly our in silico approach confirmed published experimental data demonstrating the importance of key PcG or TrxG complexes involved in self renewal and cell fate determination during haematopoiesis. The results suggest the PcG or TrxG complexes that play important roles in maintaining self-renewal and different lineage commitment, and provide a source of information to manipulate haematopoietic cells in vitro and potentially in vivo.

89 MicroRNA Regulatory Networks in Cancer Progression

Holly Whitfield, Melissa Davis and Joseph Cursons

Walter and Eliza Hall Institute

MicroRNAs are small, endogenous, non-coding RNAs which participate in gene regulation through the repression of mRNAs. MicroRNAs (miRNAs) play a fundamental role in regulating both normal cellular development and in the progression of disease. They can exert control over these phenotypes by the coordinated effects of multiple miRNA which includes the additive effects of multiple miRNA co-targeting individual mRNA, as well as a single miRNA targeting multiple mRNAs. For example, the mutual repression between the miR-200 family and transcription factors ZEB1 and ZEB2 form regulatory feedback loops which are known to contribute to cancer progression. It is the complex interactions between cell constituents, such as miRNA and mRNA, which drive cellular function rather than any individual molecule. Hence, by capturing these topological features in networks allows for a systems-level exploration of complex biological systems, as well as a powerful visualization tool. These regulatory networks can be constructed either from observed experimental data, or through computational inference. Here, these approaches will be integrated to construct miRNA regulatory networks for the identification of novel regulatory interactions. Networks will be constructed using information from both binding site prediction tools, and from TCGA breast cancer data. Using what is understood about miRNA:mRNA binding, databases such as TargetScan and DIANA-microT can predict putative relationships between miRNA and mRNA. The TCGA data will be used to establish associations between miRNA and mRNA, which when integrated with relationships from prediction databases will provide a framework for a tentative network. A revised network can then be permuted

with a condition or drug of interest to explore the regulatory system. Network analysis methods will be used to identify novel putative relationships, and regulatory mechanisms which drive the progression of cancer.

90 Reducing the impact of batch effects in single-cell RNA-Seq by imputing using Expectation Maximization algorithm

Alexander Hayes and Agus Salim

La Trobe University

Single cell RNA sequencing (scRNA-Seq) has the ability to analyse and study the transcriptome at a high resolution that was previously unobtainable. However, current technologies lead to a considerable amount of unwanted variation present within the data that requires sophisticated statistical methods to untangle from the experimental factors of interest. A pervasive cause of unwanted variation in scRNA-Seq experiments is batch effects, whereby replicates under the same condition but processed in different batches show variation. Also, as a result of the incredibly low amount of starting material and poor capture efficiency of current technologies, there is a high chance for transcripts of even moderate expression to fail to have their reads captured and counted in analysis. This phenomenon is referred to as a dropout event and results in excess of zero counts in the gene-count matrix, which can be easily confounded with true zero counts from unexpressed genes. By assuming that in the absence of dropout, observed counts can be reasonably modelled as Zero-Inflated Negative Binomial (ZINB) random variables, we use an EM algorithm approach to estimate the original numbers of molecules. In doing so, we find that imputing can reduce the impact of batch effects when the differences across batches are primarily caused by differences in capture rates. Imputation has the potential of improving signal-to-noise ratio when performing differential expression or clustering of scRNA-Seq data.

91 Integrating RNA, miRNA, DNA methylation and histone modification data uncovers biologically significant TGF-B-induced gene regulation

Haroon Naeem, Bo Wang, Guanyu Ji, Junwen Wang, Phillip Kantharidis, David Powell and Sharon D. Ricardo

Monash University

Background: Transforming growth factor-beta (TGF-B1) is recognised as an important regulator in the progression of renal fibrosis. Mesangial cells occupy the central position in the glomerulus and damage to the cells can lead to chronic kidney disease. To investigate the regulatory effects of TGF-B1 on mesangial cells at different molecular levels, mesangial cells were treated with TGF-B1 for 3 days after which genome-wide miRNA, RNA, DNA methylation and levels of H3K27Me3 expression were profiled using next generation sequencing. Results: We describe the first comprehensive study computationally integrating RNA-Seq, miRNA-Seq, and epigenomic (meDIP and histone H3K27Me3 ChIP-Seq) analyses across all genetic variations, confirming the existence of DNA methylation and H3K27Me3 in response to TGF-B1. Our data also provide support for epigenetic changes being associated with the expression of genes that are closely related to kidney diseases. Conclusions: These discoveries enhance our understanding of alternative mechanisms that contribute to TGF-B1-regulated gene expression that are involved in the pathogenesis of kidney injury. This study provides insight into the molecular underpinnings of TGF-1 stimulation in kidney cells and provides a robust platform for further target exploration.

92 Galaxy Training Network - Training Material Repository

Simon Gladman and Maria Doyle

Melbourne Bioinformatics

Over the last 7 years there has been a lot of training material produced for and with Galaxy, especially in the bioinformatics space. However, it can be hard to find Galaxy training material that is complete, up to date and maintained. The Galaxy Training Network was born from a birds of a feather session at the 2014 Galaxy conference. This community of Galaxy trainers and training material developers has recently held 3 hackathons during which we created a repository and web page (http://training.galaxy.org) for Galaxy training materials. Training materials for users and administrators of Galaxy, and for bioinformatics analyses in Galaxy, were both produced and collected and appropriate materials collated including slides, tutorial instructions, training datasets and Galaxy tours. Each tutorial also has a docker container associated with it with a pre-installed Galaxy server and the required tools. Maintainers and collators for the materials are also listed. The slides and tutorial documents are all written in markdown with version control, and there are contribution guidelines and procedures. Training material authors are encouraged to store training datasets in a repository such as Zenodo where they can receive a DOI so the authors of the datasets receive attribution.. There are currently over 40 tutorials for data analysis and 20 for Galaxy server administration and development. This repository has been produced and is maintained by the international Galaxy community. Contributions are welcome!

93 Resolving cardiac stromal-cell diversification through single-cell transcription profiling

Ralph Patrick, Nona Farbehi, Munira Xaymardan, Robert Nordon and Richard Harvey

Victor Chang Cardiac Research Institute

The potential for employing stem cells in treating cardiac disease has garnered much interest in research, but remains a controversial area. The discovery of an adult cardiac progenitor cell population - named cardiac colony forming unit fibroblasts (cCFU-Fs) - has opened up the possibility of activating regeneration potential as a means for cardiac repair. The cCFU-Fs are housed in a fraction of cardiac stromal cells that contain both the cCFU-Fs and their progeny, and become activated after injury. Understanding the specific sub-lineages that exist in this stromal population, and how these are altered in response to injury, will be critical for developing strategies to manipulate their capacity. In this study, we have employed single-cell RNA-seq (scRNA-seq) to generate a large-scale (~16,000 cells) transcriptional data-set investigating the impact of myocardial infarction (MI) on the cCFU-F housing population of stromal cells in the adult murine heart. Through clustering procedures we identify novel cell populations representing responses to MI across two time-points, as well as cell populations in homeostasis that are maintained after injury. Applying trajectory analysis to order cells along differentiation pathways reveals distinct injury response lineages corresponding to time-dependent myofibroblast populations. This study provides an unbiased examination of the heterogeneity of cardiac stromal cells, and reveals a high-resolution response to injury.

94 Computational Prediction of Serotonin Distribution in the Human and Rodent Colon

Helen Dockrell, Phil Dinning, Damien Keating and Lukasz Wiklendt

Flinders University

Approximately 95% of serotonin is produced in the intestines by enterochromaffin cells where it has been shown to modulate muscular contraction pattern and contraction force. Perturbation of intestinal muscle contraction and of serotonin concentrations underlie intestinal pathologies including Irritable Bowel Syndrome and Crohn's Disease, where changes in total digestion time, gastrointestinal emptying and gastrointestinal motility are observed. Serotonin concentration fluctuations are studied in relation to muscle contraction using ex vivo human and animal colonic samples. However, it has proved difficult to identify the complex relationship between these factors. This project collates the experimental data in a three-dimensional computational model of the intestines. This allows potential interactions between serotonin and muscle contraction to be explored and the physics constricting possible interactions to be applied consistently to inform hypothesis development. Our model replicates a transverse segment of human colon ex vivo study containing a buffer solution in the lumenal space. All parameters applied to the model are experimentally determined, and are physiologically present. We have found that the addition of involutions in the mucosal epithelium, termed crypts, results in the formation of distinct serotonin concentration patterns in the mucosal pool with respect to the mucus and buffer pool concentrations ($P < 2.2e-16$, Man-Whitney-Wilcoxon test). The mucus sitting in the crypts acts as a well of serotonin, stabilising serotonin release from the mucus into the lumenal buffer solution. As a result, mucus and lumenal serotonin concentrations remain comparatively steady during muscle contraction, where serotonin concentrations in the mucosa change rapidly. This correlates with ex vivo experimental findings as well as neural signalling theory, which together predict that serotonin concentrations remain high in the lumenal space, but fluctuate at lower concentrations within the mucosa in order to prevent neural desensitisation and communicate complex information to the nerves affecting muscle contraction.

95 Data Visualisation for Clinical diagnosis

David Ma

Peter MacCallum Cancer Centre

Data Visualisation for Clinical diagnosis is different to Data Visualisation for research and exploratory work. In the Molecular Pathology department at the Peter MacCallum Cancer Centre, Data Visualisation is used for quality control of samples before curation is done. And Data Visualisation is also used for the curation of variants. In both of these use cases, it is very important that the visual representation of data is as unambiguous as possible and gives a clear result every time, no matter who is doing the interpreting. Data Visualisations in research are often exploratory in nature, and often one-off creations. In research we can often find new and interesting things through data visualisation. Data Visualisation in the Clinical diagnosis comes with different priorities to Data Visualisation in the research setting, and I think it would be interesting and insightful to give a short ABACBS talk highlighting how these different priorities manifest.

96 SWATH-MS Spectral Reference Library Species Conversion with the R Package dialects

Madeleine J Otway, Peter G Hains and Phillip J Robinson

Children's Medical Research Institute

Mass spectrometry (MS) based proteomics is a methodology used to measure the relative abundance of proteins in biological samples. Proteins are extracted from tissue, enzymatically cleaved into peptides and sequenced by the fragmentation of selected peptides. Shotgun mass spectrometry (data-dependent acquisition; DDA) approaches produce biased results, where only the most abundant peptides are measured. SWATH-MS is a relatively new approach that measures all theoretical peptide fragments. The resulting file is extremely complex and identification of peptides relies on a separately generated spectral reference library (SRL). This is a table of peptide fragments, defined though a series of DDA-MS runs, which are used to search the SWATH-MS data. Creation of an SRL is a lengthy process with large computational requirements. Re-searching of MS files with a species other than that of the original tissue is often avoided. This precludes the use of a large well characterised SRL designed in one species from being applied to SWATH-MS data generated in another species. To overcome this, we developed an R package named dialects (Data Independent Acquisition Library Editing to Convert The Species) for the conversion between species in an SRL. This package has five core functions:

1. Import a UniProt formatted protein sequence database (fasta file)

2. Import a PeakView/OneOmics or OpenSWATH formatted SRL

3. Perform an in silico trypsin digestion on the proteins from the UniProt database

4. Swap the species of the SRL to that of the digested protein sequence database, only for peptides with full sequence homology occurs

5. Export the newly created SRL in either PeakView/OneOmics or OpenSWATH format

This package aids the creation of comprehensive SRLs, without the need for repeated MS runs and/or reprocessing of MS searches. This reduces the time to convert between species of a SRL and expands utility of large well characterised SRLs.

97 A blood-based signature of cerebral spinal fluid AB_1-42 status

Benjamin Goudey, Bowen Fun, Christine Schreiber and Noel Faux IBM It is increasingly recognized that Alzheimer's disease (AD) exists before dementia is present and that shifts in amyloid beta occur long before clinical symptoms can be detected. Early detection of these molecular changes is a key aspect for the success of interventions aimed at slowing down rates of cognitive decline. Recent evidence indicates that of the two established methods for measuring amyloid, decreases in cerebral spinal fluid (CSF) amyloid (AB_1-42) levels may be an earlier indicator of Alzheimer's disease risk than measures of amyloid obtained from Positron Emission Topography (PET). However, CSF collection is highly invasive and expensive. In contrast, blood collection is routinely performed, minimally invasive and cheap. In this work, we develop a blood-based signature that can provide a cheap and minimally invasive estimation of an individual's CSF amyloid status. We show that a Random Forest model derived from plasma analytes can accurately predict subjects as having abnormal (low) CSF AB_1-42 levels indicative of AD risk (0.84 AUC, 0.73 sensitivity, and 0.76 specificity). Post-hoc analysis indicates that only six analytes are required to achieve these high levels of accuracy. Furthermore, we show across an independent validation cohort that individuals with predicted abnormal CSF AB_1-42 levels transitioned to an AD diagnosis over 120 months significantly faster than those predicted with normal CSF AB_1-42 levels. This is the first study to show that a plasma protein signature, together with age and APOE4 genotype, is able to predict CSF AB_1-42 status, the earliest risk indicator for AD, with high accuracy. Biomarkers in plasma have previously been shown to be predictive of PET amyloid levels. This work further highlights the potential for developing a blood-based signature for improved AD screening, critical for drug and intervention trials.

98 Precision Medicine: A clinical perspective on genome data

Gulrez Chahal, Sonika Tyagi and Mirana Ramialison

Australian Regenerative Medicine Institute, Monash University Clayton VIC and SBI Australia.

Precision (personalized) medicine is integrating traditional medicine with genomic profiling to make therapeutic, prognostic and preventive decisions in various human diseases, including cancer, heart diseases and inherited syndromes. While the research labs aim at identifying these genomic markers, there is still limited understanding of how is genetic testing actually implemented at the clinic. Cancer genomics being one of the largest collaborations between the clinic and genomic research, offers a good example to understand what are the factors which affect the utility, efficiency and scalability of these tests in the clinic. Clinical cancer genomics generates several gigabytes of data, which is trickled down to a few candidate pathogenic variants/markers by using several computational analysis, visualization and interpretation tools. However, it is important to understand, how do clinicians integrate this information with traditional methods to take therapy decisions for the patient? Are there enough clinical guidelines to integrate this information? How effective are these decisions and what factors govern their efficiency? What are other socio-economic factors which affect the therapy decisions? Here we present a bird's eye view of precision medicine which integrates the view at the lab, health industry and the clinic.

99 Integrative Analysis of Lipid Metabolic Pathways in Prostate Cancer Reveals DECR1 as a Key Cancer-Related Gene that Promotes Tumour Cell Survival

Chui Yan Mah, Max Moldovan, Zeyad Nassar, David Lynn and Lisa Butler

The University of Adelaide

Prostate cancer (PCa) is the most commonly diagnosed malignancy and the second leading cause of cancer-related deaths in Australian men, with advanced metastatic PCa remaining a lethal disease. Altered lipid metabolism is one of the hallmarks of PCa, which commonly overexpresses lipogenic enzymes, including those involved in lipid uptake, binding, transport and metabolism. These changes are observed early during transformation of normal prostate epithelial cells to malignant PCa cells, which results in increased dependency on lipids as the major energy source rather than glucose. Here, we analysed differential expression of the major lipid metabolic genes, in order to explore which lipid-related pathways are characteristic of PCa compared to benign or normal tissues. We selected five gene expression datasets from online open repositories (Gene Expression Omnibus and GDC Data Portal) consisting of individual microarray and transcriptome profiles. The z-scores of lipid metabolism genes of individual datasets were extracted for meta-analysis. Our results showed significant dysregulation of multiple genes involved in lipid metabolism and identified 'DECR1' as the most commonly overexpressed gene in PCa cells. DECR1 catalyses the rate limiting step of polyunsaturated fatty acid (PUFA) oxidation, an important source of cellular energy. Further analysis revealed a significant correlation between DECR1 expression and shorter biochemical recurrence-free and overall PCa patient survival. To validate these in silico findings, we detected overexpression of DECR1 protein levels in PCa cell lines compared to non-transformed prostate cells. Consistent with the vital role of DECR1 in PUFA metabolism, we showed that DECR1 down-regulation decreased cellular ATP levels and PCa cell proliferation. Our results suggest that DECR1 represents an exciting new therapeutic target for PCa. Further network analysis will focus on defining the interactions of DECR1 with other key metabolic pathways involved in PCa progression.

100 HIV-1 RNA structure heterogeneity in cells

Vincent Corbin, Phillip Tomezko, Lachlan Mcintosh, Paromita Gupta, Margalit Glasgow, Sitara Persad, Anthony Papenfuss and Silvia Rouskin

Walter and Eliza Hall Institute

RNA is unique among all biomolecules as it can be both information-storing and enzymatic. These features are tightly linked to its structure, in which base-pairing interactions give rise to a highly folded macromolecule. Indeed, in addition to the genetic code, which specifies the composition of proteins, we now know there is a secondary layer of information encoded in every transcript in the form of RNA structure that can regulate processes as diverse as splicing, localization, and translation efficiency. We have developed and validated a unique method to detect variability in RNA structure using next generation sequencing data. Our method probes RNA structure at single molecule and single nucleotide level, and is capable of detecting alternative RNA structures which form from the same underlying sequence, both in vitro and ex vivo. Application of this approach to Human Immunodeficiency Virus-1 (HIV-1) reveals HIV-1 genomic RNA structure heterogeneity with novel functional implications.

101 Usage of VLSCI Compute Resources by the Life Sciences Research Community

Clare Sloggett, Andrew Isaac and Edmund Lau

Melbourne Bioinformatics

Life Science research is becoming increasingly dependent on data and computation. Here we explore the diversity of computer and storage needs over 8 years of hosting high performance resources for the life sciences community. The Victorian Life Sciences Computation Initiative (VLSCI), now Melbourne Bioinformatics, was established in 2009 to provide high-performance compute resources and expertise to the life sciences community. The initiative was established with IBM Blue Gene hardware, as well as more traditional x86 clusters. In addition to the focus on fast compute resources, VLSCI hardware was designed to accommodate the needs of life sciences research by providing high-speed disk to handle I/O bound processes, large-memory nodes, and sufficient quantity of disk space to handle -omics data volumes. Here we explore the historical record of jobs run across our clusters, and report on the observed usage patterns of compute resources by the life sciences research community. We look at the demands placed upon our compute resources in practice, and discuss the real-life resource requirements of life sciences researchers across different domains. In particular, we consider issues such as: is a desktop computer enough? can I move to the cloud? should our lab buy a server? The breadth of activities available to life science researchers on a high performance cluster has implications for institutional and national support for life science research.

102 scPipe: a flexible data preprocessing pipeline for single-cell RNA-sequencing data

Luyi Tian, Shian Su and Matthew Ritchie

Walter and Eliza Hall Institute of Medical Research

Single-cell RNA sequencing (scRNA-seq) technology allows researchers to profile the transcriptomes of thousands of cells simultaneously. Protocols that incorporate both designed and random barcodes to label individual cells and molecules have greatly increased the throughput of scRNA-seq, but give rise to a more complex data structure. There is a need for new tools that can handle the various barcodes used by different protocols and exploit this information for quality assessment at the sample-level and provide effective visualization of these results in preparation for higher-level analyses. We developed scPipe, an R package that allows barcode demultiplexing, read alignment, gene-level quantification and quality control of raw sequencing data generated by multiple 3 prime end sequencing protocols that include CEL-seq, MARS-seq, Chromium 10x and Drop-seq. scPipe produces a count matrix that is essential for downstream analysis along with an HTML report that summarises data quality. These results can be used as input for downstream analyses including normalization, visualization and statistical testing. scPipe performs this processing in a few simple R commands, promoting reproducible analysis of single-cell data that is compatible with the emerging suite of scRNA-seq analysis tools available in R/Bioconductor.The scPipe R package is available from Bioconductor.

103 Identification and characterization of new cell populations using droplet-based single-cell RNA-seq: an example on breast cancer T cell infiltrate

Chengzhong Ye, Agus Salim, Peter Savas, Sherene Loi and Terence Speed

The Walter and Eliza Hall Institute of Medical Research

Single-cell RNA-seq (sc-RNAseq) has provided researchers with an unprecedented opportunity to investigate the heterogeneities of cell populations. Recently popularized droplet-based technologies greatly increased the number of cells that can be profiled in each experiment. However, this type of data suffers from even higher technical noise and dropout rate, presenting new computational and analytical challenges. Here we try to identify and characterize a particular group of cells in a breast cancer infiltrating T cell population sequenced with 10X Genomics droplet-based scRNA-seq platform. We undertook a combined analysis including clustering, differential expression and trajectory analysis. Particularly we made vital use of imputation to overcome the high dropout rate of the data. A newly developed model, DECENT, was used for performing differential expression analysis on the imputed pre-dropout data.

104 A Systems Biology Approach to Investigate SRSF7 Functions in RNA Regulation of Autism Spectrum Disorder

Monika Mohenska, Mirana Ramialison and Minna-Liisa Anko

ARMI, Monash University

There are more than 150,000 diagnosed cases of autism spectrum disorder (ASD), in Australia alone, every year. Although ASD affects the lives of many Australians, there is no known cause for the disease. In general, many neuronal diseases are heavily characterised by erroneous RNA processing, such as the splicing of mRNA. A family of proteins which are greatly involved in RNA regulation are the serine-arginine rich splicing factors. One of these proteins is serine-arginine rich splicing factor 7 (SRSF7). A previous study identified an Australian patient with a point mutation in the SRSF7 gene, which is predicted to encode a truncated SRSF7 protein. With increasing evidence that SRSF7 is involved in neuronal function, this mutation could provide insights into the role of SRSF7 in the RNA regulation of ASD. Here we show that the truncated form of SRSF7 has a link to ASD based on the novel findings of Next Generation RNA Sequencing (RNA-seq) analyses. Differential gene expression and differential

splice variant analyses in a murine Srsf7-mutant in vitro model have identified potential targets of importance during neuronal development. Interestingly, by combining publically available datasets, such as neuronal RNA-seq and cross-linking immunoprecipitation sequencing (CLIP-seq) data, we uncovered potential SRSF7-regulated pathways that contribute to ASD. These findings of the role of SRSF7 in neuronal RNA regulation will aid our understanding of the underlying molecular cause of ASD.

105 Reference-free methods for genomic prediction and selection

Kevin Murray and Justin Borevitz

ANU

Genomic prediction uses knowledge of the population and family genetic relatedness to explain and predict variation underlying complex quantitative traits. This process has accelerated the breeding of crops and livestock as selection can occur on generally more accurate predicted phenotypes and can be extended to predictions on unobserved individuals. Genomic prediction using gBLUP currently relies on relatedness data as determined from SNP genotypes mapped to a reference genome. This can induce bias, and precludes its use in non-model species, where reference genomes are either missing or poor. Additionally, incorporating the predictive power of the microbiomes associated with crops or livestock is arduous at best, requiring assembly of complex metagenomes. We propose to capture this missing predictive power using a new method, which incorporates the weighted covariance between k-mer counts into traditional gBLUP-based genomic prediction approaches.

106 Systems biology framework as an integrative approach in psychiatric research

Liliana G Ciobanu, Micah Cearns and Bernhard T Baune

University of Adelaide

We live in the era of 'big data', with consortiums and repositories providing unprecedented opportunities to explore new avenues of psychiatric research. In an attempt to answer critical questions regarding the biological underpinnings of psychiatric conditions, millions of data points at various levels of biological abstraction are being generated every year. However, even with the advent of 'big data', deriving clinically translatable insights that progress the field beyond symptomatic diagnosis remains a challenge. This posits a problem for classical statistical approaches looking at associations between individual molecules and clinically relevant phenotypes. For example, the capacity to detect multiple small effects from gDNA-mRNA and other biological factors places large sample size requirements on individual studies. Achieving breakthroughs with such methodologies is proving both costly and unsustainable. We advocate that a systems biology framework that allows for dimensionality reduction in genome-wide data is an effective approach to produce biologically meaningful results in a cost-effective manner. This framework is centred on the co-expression network, the main aim of which is to cluster numerous individual players into fewer co-regulated modules; compute a principal component, i.e. eigenegene of each module and use the eigenegenes to assess the relationship between gene expression and other elements of complex biological interactions - genotypes, epigenetic markers and metabolomics perturbations associated with a disease. Using this framework we have been able to identify several pathways and biological substrates involved in the pathophysiology of depression using transcriptome data for a moderate sample size of 465 participants. The main results will be presented and discussed. Our findings are consistent with recently conducted GWASs in depression and have provided meaningful functional interpretation for nine susceptibility genetic loci. Further investigation following this proposed integrative approach will lead to greater understanding of the pathophysiology of depression with the potential to identify translatable drug targets.

108 UMID-dedup: A flexible tool for marking duplicate reads in NGS data using unique molecular identifiers

Paul Wang, Wendy Parker, Joel Geoghegan and Andreas Schreiber

ACRF Cancer Genomics Facility, Centre for Cancer Biology, SA Pathology and University of South Australia, Australia

Due to the rapidly decreasing cost of next generation sequencing (NGS) it has become commonplace to increase sequencing depth so as to detect ultra-low level variants or rare transcripts. However, there are some challenges involved in high coverage depth data. First, the classical methods of marking PCR duplicates, using just read positions, will erroneously remove growing number of fragments that are actually unique. Second, detection of very low allele frequency variants is hampered by the increased noise and errors associated with very high coverage depth, especially in amplicon data. To circumvent these problems, many sequencing facilities and biotech companies have started to incorporate unique molecular identifiers (UMID), which are fixed-length random barcodes attached to fragments before PCR amplification. De-duplication using UMIDs together with fragment positions provides a means to significantly increase the accuracy of transcript or fragment counts and decrease the noise due to sequencing errors. Here we present UMID-dedup, a Python tool for marking duplicate reads in BAM files. Besides its ability to mark duplicates using the UMIDs, it also provides more flexibility in data input/output and processing. For example, unlike most other UMID tools, it can handle multiple UMID encoding methods. It can also handle different NGS data types: DNA-seq, RNA-seq, and amplicon data, and both single and paired-end sequencing. More significantly, given a group of duplicate reads, it has the ability to either select a representative (like the classical methods) or to generate a more accurate consensus read. We demonstrate that UMID-dedup can retain significantly more reads than classical methods, especially in high depth data, while significantly reducing random errors (using consensus reads), which in turn improves detection of ultra-low frequency variants. UMID-dedup is designed to be a replacement for classical dedup tools (e.g. Picard MarkDuplicates) and should be easy to deploy in standard NGS pipelines.

109 Exploring The role of Vitamin D receptor in THP-1 cells

Ming Lu

Menzies Institute for Medical Research

Background: Vitamin D receptor, as the exclusive target of Vitamin D in cells, has pervasive regulatory effects and large cistrome region, which partly explained the anti-inflammatory effect and deep involvement of Vitamin D in autoimmune diseases. Monocytes express VDR, and the expression will be upregulated after differentiation to dendritic cells and macrophages. SE (super-enhancer) is a stretch of enhancer, which can determine cell fate and is enriched for disease genetic risk. Method: We reanalysed the VDR ChIP-seq (chromatin immunoprecipitation sequencing) data and FAIRE-seq data (open chromatin region) in THP-1 (immortalized monocyte) cells from NCBI GEO database with Samtools and BEDtools. I used HOMER to identify associated SE region and its enriched motifs. Result: I found only a small part of VDR SE overlapped with FAIRE-seq region at 24h after Vitamin D stimulation, indicating their different mechanism and function. By integrating them, we found SEs enriched for MS risk SNPs and are near genes involving immune regulation. Classic VDR DR3 motif is only enriched for early VDR binding regions at 2h, but not at 24h. Further research about the SE region and VDR genomic regulatory region will help to find potential biomarkers for autoimmune diseases.

110 Learning representations for sequence comparison

Dhananjay Kimothi, Akshay Soni, Pravesh Biyani, James Hogan and Wayne Kelly

QUT Brisbane

Grouping molecular sequences based on similarity is an important task in bioinformatics. Comparisons of this nature may be based directly on local or global alignments, on alignment based heuristics such as BLAST, or on k-mer based alignment free approaches. Functional annotation is usually inferred based on these relationships, propagating existing annotations. This general approach has been enormously successful, although each method has its limitations: alignment based methods may prove slow for longer sequences and are not robust when faced with structural re-arrangements; alignment free methods rely on careful normalisation, and may lose precision rapidly as sequences diverge. In this work we look to combine both sequence similarity and annotations in a single vector representation. By adapting methods from text processing, we are able to embed each sequence within a vector space so that sequences with similar k-mer content are mapped close together. By including class information – protein family, functional annotations – when learning the representation, we can tailor the embedding to represent better a range of relationships not always apparent from scored alignments or other similarity measure. We apply our method initially to the problem of retrieving functionally similar protein sequences, demonstrating its value through evaluation and comparison of precision-recall values with known alignment and alignment free methods. We also report preliminary work on its application to broader functional annotation problems.

111 An improved geometric measure for comparison of biological sequences

Lawrence Buckingham, Timothy Chappell, James Hogan and Shlomo Geva

Queensland University of Technology

Sequence comparison is a fundamental task in computational biology, traditionally dominated by alignment-based methods such as the Smith-Waterman, Needleman-Wunsch, or BLAST algorithms. In the era of Next Generation Sequencing a need for improved scalability of sequence comparison and database search is driving a focus on alignment-free alternatives to these approaches. While some alignment-free algorithms have proven successful for particular tasks, many continue to exhibit a marked decline in sensitivity as closely related sequences diverge. Avoiding the alignment step yields improved scalability, but at the cost of decreased search effectiveness. In this paper we describe an enhanced version of the Similarity Projection algorithm, an alignment-free sequence comparison algorithm based on variants of the Hausdorff set distance, which offers sensitivity comparable to alignment based methods while retaining the scalability characteristic of alignment-free approaches. The algorithm decomposes sequences into fragments – collections of overlapping sequential k-mers. Inter-fragment similarity is obtained via approximate k-mer matching, and fragment similarities are combined to produce an overall measure of sequence similarity which reflects those components which match well while lessening the impact of those which do not. Formally, the algorithm generates a large mutual similarity matrix between sequence pairs based on their component fragments; successive reduction steps yield a final score over the sequences. However, only a small fraction of these underlying comparisons need be performed, and by use of an approximate scheme based on vector quantization, we are able to achieve an order of magnitude improvement in execution time over the naive approach. The present work describes an enhanced approximate nearest neighbour search algorithm which offers improved sensitivity while retaining favourable scaling properties. We evaluate the approach on two large protein collections obtained from UniProtKB, showing that Similarity Projection achieves accuracy rivalling, and at times clearly exceeding, that of BLAST, while exhibiting markedly superior execution speed.

112 TRNDiff: Large-scale Visualisation for Transcriptional Regulation

Samuel Smith, James Hogan, Lawrence Buckingham, Xin-Yi Chua, Margot Brereton, Daniel Johnson and Markus Rittenbruch

Queensland University of Technology

This poster reports recent extensions to TRNDiff Regulon Explorer, a visualisation tool which displays

transcriptional regulatory information from the curated RegPrecise database and user supplied datasets. TRNDiff is intended to provide an efficient way to analyse large regulatory datasets on both small and large scale devices. The application displays Transcriptional Regulatory Networks (TRNs) as a series of radial network diagrams, or wagon-wheels, with a Transcription Factor (TF) hub radially connected to (potentially) regulated Target Genes (TGs). Users may interact with these diagrams in several ways in order to explore a dataset. TRNs may be positioned on the screen or organised into distinct groups via drag and drop. Groups may also be created automatically through k-means clustering based on the topological similarity between the wagon wheels. Two or more networks can be selected and the comparison view used to show automatically the differences in target genes and interactions between them. While the RegPrecise database is comprehensive, TRNDiff allows import and export of datasets using a simple CSV-based file format. This allows users to save groupings of the data, to add and preserve annotations, and to use the tool to display their own data from experiments or other sources. An on-line version of the tool is available at http://trndiff.org/ and the source code is available at https://bitbucket.org/biovisml/trndiff.2015

113 VDJPuzzle: A computational method for BCR and TCR reconstruction from single-cell sequencing data

David Koppstein, Simone Rizzetto and Fabio Luciani

UNSW

Single-cell methods have revolutionized the field of immunology by providing an unparalleled view into the heterogeneity of lymphocyte transcriptional states. Although established protocols impute expression levels for all transcripts for a given cell, most existing methods fail to capture the full-length receptor, a critical parameter for investigating behavior of individual clonotypes and for mining receptor sequences from clinically-relevant subsets of cells for novel therapeutics. Here we present VDJPuzzle, a computational method for reconstructing both the BCR and TCR receptors from short paired-end reads derived from whole-transcript single-cell sequencing methods such as SMART-Seq2. We validate our method using Sanger sequencing and show that our BCR reconstruction method outperforms existing methods. Further, we demonstrate an integrated workflow coupling transcriptomics to index data derived from flow cytometry and apply this method to B cells derived from patients with Sjogren's syndrome and tetramer-positive T cells from patients chronically infected with Hepatitis C virus.

114 Discovering a mechanism of drug-resistance in *P. falciparum*

Jocelyn Penington, Jennifer Thompson, Anthony Papenfuss and Alan Cowman

The Walter and Eliza Hall Institute of Medical Research

Malaria is caused by single-celled parasites of the genus Plasmodium, which invades and remodels red blood cells. The World Health Organization estimates that malaria caused 212 million clinical episodes and 429,000 deaths in 2015—the majority due to the species Plasmodium falciparum. Resistance to therapies targeting blood-stage plasmodium is a major barrier to treating and curing malaria. To understand the mechanisms that underlie the emergence of resistance to new therapeutics currently under development, the genomes of laboratory strains of Plasmodium falciparum that were sensitive to these new drugs and strains with induced resistance were sequenced. We developed a pipeline for analysing matched mutant and wildtype strains for single nucleotide polymorphisms, indels, copy numbers and structural variation. This included the use of a variant caller for haploid genomes (SNVer), as well as somatic calling algorithms from cancer genomics (varscan, mutect, QDNAseq and GRIDSS). The analysis did not identify any recurrent SNVs or indels across different resistant strains. However, both copy-number analysis and GRIDSS structural-variant calling found amplifications covering a gene thought to be relevant to the mechanism of action of the molecule.

115 Direct transcriptional regulation by microRNAs

Klay Saunders, Cameron Bracken, Greg Goodall and Katherine Pillman

University of South Australia

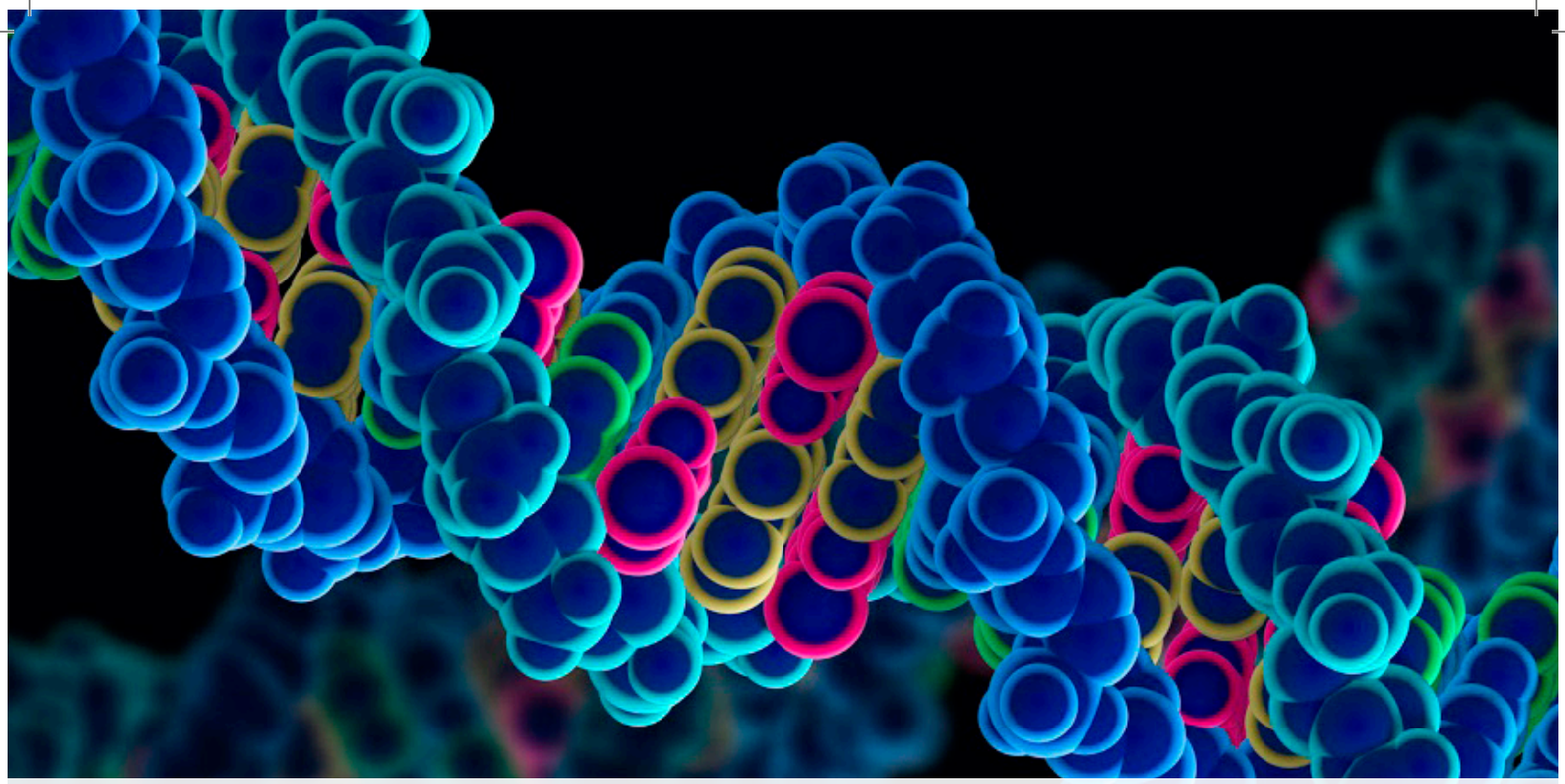
MiRNAs are almost always regarded as cytoplasmic, negative post-transcriptional regulators of gene expression. Recent studies however reveal an abundance of mature miRNAs (and accessory proteins) within the nucleus and several reports now postulate direct gene regulatory roles for miRNAs (both positive and negative) at the level of transcription. We have strong evidence that the contribution of these nuclear roles to miRNA activity has thus far been vastly under-appreciated. Combining high throughput RNAseq datasets with a custom miRNA target prediction database we can identify genes which are being directly regulated by miRNAs. From here we can begin to investigate the effects various targeting types and context specific roles in regulation. Working initially with miR-200c, a key promoter of mesenchymal-epithelial transition (MET), we have found that miRNA binding sites within gene promoters are strongly associated with transcriptional regulation. Further, we found that endogenous miRNA:Argonaute (AGO) complexes are present proximal to the transcription start sites of many genes.

116 Integrative genomic assessment of advanced CML reveals widespread genomic instability mediated by the recombination activation gene (RAG) pathway

Thomson D[1], Wang P[1], Schreiber A[1] and Branford S[1]

[1]Centre for Cancer Biology, SA Pathology, Adelaide, Australia.

The survival rate of patients diagnosed with Chronic Myeloid Leukaemia (CML) has increased dramatically over the last decade with testing for the oncogenic BCR-ABL fusion-gene, which directs treatment with tyrosine kinase inhibitors (TKIs). However, a subset of patients gain resistance to treatment drastically affecting their prognosis. In an effort to study genomic variation that leads to therapy-resistance, we have accumulated a cohort (n=186) of CML patient blood samples. Samples were taken at diagnosis and blast crisis (advanced stage), and were assessed with RNAseq and whole exome sequencing (WES). Gene expression analysis and unsupervised hierarchical clustering revealed aberrant expression of the recombination activation gene (RAG) pathway, specifically in patients that had undergone lymphoid blast crisis. RAG1 is an endonuclease with an essential role in acquired immunity; In developing lymphocytes, RAG forms DNA-breaks at immunoglobulin genes and facilitates recombination to produce a diverse array of antibodies. Through analysis of copy number variation (CNV), we observe loss of immunoglobulin loci in CML patients specifically at blast crisis, indicative of RAG activity. This is supported by breakpoint analysis showing DNA breaks occurring at known RAG target sites. We investigate the potential that aberrant RAG activity in advanced CML promotes genomic instability through the formation of DNA breaks and fusion genes outside of its canonical role of processing immunoglobulin genes. We identify hundreds of putative fusion genes that occur since diagnosis, from which unbiased motif enrichment searches reveal are associated with the recombination signal sequence (RSS) motifs targeted by RAG1. We identify, and PCR validate a number of novel fusion genes, which in many cases include mobilised immunoglobulin genes. Disease progression of CML and resistance to treatment by TKI's is associated with accumulation of genomic lesions, this work provides evidence that transposition by RAG1 is a major driving force behind genomic instability in advanced CML.



#abacbs17
@abacbs
@combine_au

www.abacbs.org
www.combine.org.au