ABACBS 2020 Virtual Conference

Conference Program

Conference committee members:

Eduardo Eyras (Conference Convener)
Jimmy Breen
Jean Wen
Susan Wagner
Attila Horvath
Sonika Tyagi
Ignatius Pang
Aaron Darling
Gavin Huttley
Priyanka Surana
Melanie Smith
Fabio Zanini
Annette McGrath

ABACBS 2020 Virtual Conference is Proudly Supported by

Gold Conference Sponsor

CSL™

Australian National University

The John Curtin School of Medical Research

Australian BioCommons

Silver Conference Sponsor

UTS UNIVERSITY OF TECHNOLOGY SYDNEY

Google Cloud

i3 ithreeinstitute infection·immunity·innovation

Bronze Conference Sponsor

Oxford NANOPORE Technologies

DNAnexus®

illumina

UNSW Ramaciotti Centre for Genomics

ABACBS Day 1 (Tuesday 24th November)

Time (AEDT)	Parallel Session 1: Plant Genomics (Session Chair: Jen Taylor)	Parallel Session 2: Metagenomics (Session Chair: Aaron Darling)
	Canberra Cafe Remo Link	
12:30	**Invited Speaker: Sue Rhee** *Challenges and Opportunities for Bioinformatics and Computational Biology in Plant Science*	**Invited Speaker: Ami Bhatt** *Microproteins, Mobile Genetic elements and Strain-level resolution in the microbiome – a path to precision medicine*
12:50	Session Talk #38 **Stephanie Chen** *Unsupervised orthologous gene tree enrichment for cost-effective phylogenomic analysis and a test case on waratahs (Telopea spp.)*	Session Talk #35 **Feargal Ryan** *Intrapartum or Direct Antibiotic Exposure in Early Life Significantly Alters the Infant Microbiota and Whole-blood Transcriptional Responses to Immunisation*
13:00	Session Talk #112 **Charlotte Francois** *New insights into plant-microbe interactions through Quantitative Trait Locus (QTL) mapping*	Session Talk #127 **Luis Pedro Coelho** *The AMPSphere: antimicrobial peptides (AMPs) in the global microbiome*
13:10	Session Talk #110 **Chelsea Matthews** *Assessing PacBio long reads and de novo genome assembly tools for useability and suitability to applications where resources are limited*	Session Talk #8 **Daniela Gaio** *Over 25,000 metagenome assembled genomes reveal the development of the post-weaning pig gut microbial community*
13:20	**Invited Speaker: Bernice Waweru** *African-led genome sequencing of Lablab and African Yam bean orphan crop genomes*	**Invited Speaker: Rob Edwards** *Phage genome bioinformatics*
13:40	Networking Session Remo Link	
Time (AEDT)	Parallel Session 3: Regulation (Session Chair: Sonika Tyagi)	Parallel Session 4: Biomedical Informatics (Session Chair: Jimmy Breen)
15:00	**Invited Speaker: Paul Gardner** *Features of functional human genes*	**Invited Speaker: Jessica Mar** *Making sense of heterogeneity in gene expression data*
15:20	Session Talk #95 **Lixinyu Liu** *The landscape of alternative polyadenylation in CD8 T cells in single-cell transcriptome*	Session Talk #74 **Dhrithi Deshpande** *A comprehensive analysis of code and data availability in biomedical research*
15:30	Session Talk #9 **Qian Du** *DNA hypomethylation induces intrapopulation heterogeneity of DNA replication timing and 3D genome reorganisation*	Session Talk #99 **Gulrez Chahal** *CaraVaN: Prioritising Cardiac Variants in the Non-coding genome using boosting algorithm*
15:40	Session Talk #63 **Emma Gail** *A predictive model for commonly-repressed polycomb-target genes dissects DNA sequence from gene expression*	Session Talk #100 **Adria Closa** *Characterisation of a convergent malignant phenotype in B-cell acute lymphoblastic leukaemia*
15:50	Session Talk #136 **Chi Nam Ignatius Pang** *RNase III-CLASH of multi-drug resistant Staphylococcus aureus reveals a regulatory mRNA 3'UTR required for intermediate vancomycin resistance*	Session Talk #101 **David Goode** *An Evolutionary Approach to Network Analysis of Cancer Transcriptomes Reveals Common Indicators of Enhanced Malignancy Across a Range of Solid Tumours*
16:00	Session Talk #12 **Jiayue-Clara Jiang** *Integrated transcription factor profiling with transcriptome analysis: identification of L1PA2 transposons as global regulatory modulators in a breast cancer model*	Session Talk #120 **Nitika Kandhari** *Finding signatures of alternative polyadenylation as cancer biomarkers*
16:10	Short Break Remo Link	
16:40	ABACBS Senior award & talk	
17:40	ABACBS Day 1 Ends	

ABACBS Day 2 (Wednesday 25th November)

Time (AEDT)	Parallel Session 1: Phylodynamics & COVID19 (Session Chair: Fabio Zanini)	Parallel Session 2: Non-model organisms (Session Chair: Alexie Papanicolaou)
	Canberra Cafe Remo Link	
12:30	**Invited Speaker: Torsten Seemann** *How bioinformatics and genomics helped Australia's COVID response*	**Invited Speaker: Karine Le Roch** *Comparative 3D Genome Organization in Apicomplexan Parasites*
12:50	Session Talk #132 **Clare Slogett** *AusTrakka – Working towards integrated pathogen genomics for SARS-CoV-2*	Session Talk #28 **Aidan Tay** *Developing a computational safeguard to detect gene drive systems in wild populations*
13:00	Session Talk #70 **Harman Singh** *eMST, a scalable and interpretable method for Phylogenetic analysis of hundreds and thousands of SARS-CoV-2 genomes*	Session Talk #106 **Katarina Stuart** *Whole transcripts in genome assembly, annotation, and assessment: the draft genome assembly of the globally invasive common starling, Sturnus vulgaris*
13:10	**Invited Speaker: Rob Lanfear** *Confidence and truth in phylogenomics*	Session Talk #87 **Swapnil Tichkule** *Anthroponotic transmission and adaptive introgression underlies cryptic population structure of Cryptosporidium hominis in Africa*
13:30	Networking Session Remo Link	
	Parallel Session 3: Methods (Session Chair: Jean Yang)	Parallel Session 4: Transcriptomics / RNA (Session Chair: Jean Wen)
14:30	Invited Speaker: **Kim-Ahn Le Cao** *Navigating through 'omics data: a multivariate perspective*	Invited Speaker: **Yue Wan** *Direct RNA sequencing identifies isoform specific structures*
14:50	Session Talk #36 **Yingxin Lin** *Transfer learning for data integration of single-cell RNA-seq and ATAC-seq*	Session Talk #34 **Oak Hatzimanolis** *Implementing an integrated analysis to identify and validate circular RNAs using patient-derived neuronal stem cells.*
15:00	Session Talk #43 **Simon Sadedin** *Exploring Neural Network models for CNV detection from Exome Data*	Session Talk #125 **Vincent Corbin** *Moving beyond RNA sequence: uncovering the functional role of RNA structure*
15:10	Session Talk #82 **Ebony Watson** *Image-based Predictive Modelling for the Characterisation of Cellular Senescence*	Session Talk #133 **Sasdekumar Loganathan** *Application of mixture model to RNA-seq data to discover ageing regulators*
15:20	Session Talk #118 **Al J Abadi** *Integrating multi-modal single-cell studies with a latent component-based approach*	Session Talk #49 **Angel Liang** *RNA splicing is a hierarchical supernetwork that co-operates to drive osteoblast differentiation*
15:30	Session Talk #24 **Legana Fingerhut** *ampir: an R package for fast genome-wide prediction of antimicrobial peptides*	Session Talk #83 **Ruebena Dawes** *Features that determine 5' cryptic splice site selection in genetic disorders*
15:40	Session Talk #5 **Thomas Quinn** *Learning Distance-Dependent Motif Interactions: An Interpretable CNN Model of Genomic Events*	Session Talk #89 **Ke Ding** *Long short-term memory RNN for mirtron identification*
16:10	Short Break Remo Link	
16:40	ABACBS Awards session (MCR, ECR, ...)	
17:40	ABACBS Day 2 Ends	
18:30	Trivia Night Remo link	

ABACBS Day 3 (Thursday 26th November)

Time (AEDT)	Parallel Session 1: Indigenous Genomics (Session Chair: Bastien Llamas)	Parallel Session 2: Long reads (Session Chair: Lachlan Coin)
	Canberra Cafe Remo Link	
12:30	**Invited Speaker: Keolu Fox** *Creating accountability in human population genetics using base editing tools*	**Invited Speaker: Karen Miga** *Telomere-to-Telomere Chromosome Assemblies: New Insights Into Genome Biology and Structure*
12:50	Session Talk #58 **Irene Gallego Romero** *Characterising Diversity in Gene Regulation Across the Indonesian Archipelago*	Session Talk #4 **Luyi Tian** *Comprehensive characterization of single cell full-length isoforms in human and mouse with long-read sequencing*
13:00	Session Talk #123 **Sally Wasef** *How Ancient genomes can help Aboriginal Australian communities: lessons from the Cape York project*	Session Talk #33 **Nadia Davidson** *JAFFAL: Detecting fusion genes with long read transcriptome sequencing*
13:10	**Invited Speaker: Alex Brown** *Challenges In Aboriginal health in the genomics era*	Session Talk #46 **Zaka Yuen** *Systematic benchmarking of detection tools for CpG methylation from Nanopore sequencing*
13:20		Session Talk #93 **Stephen Kazakoff** *Mapping cancer transcriptomes with long-read sequencing*
13:30	Networking Session Remo Link	
	Parallel Session 3: Genomics (Session Chair: Sarah Kummerfeld)	Parallel Session 4: Single Cell (Session Chair: Ellis Patrick)
14:30	**Invited Speaker: Yu Lin** *Binning Metagenomic Sequences*	**Invited Speaker: Sara Ballouz** *Sex-specific co-expression: a baseline to explore disease*
14:50	Session Talk #41 **Daniel Cameron** *VIRUSBreakend: Viral Integration Recognition Using Single Breakends*	Session Talk #85 **Adam Chan** *An automated framework for elucidating hierarchical relationships in high dimensional cytometry data*
15:00	Session Talk #56 **Tinting Gong** *Structural variation signatures in primary prostate cancer*	Session Talk #109 **Elisabeth Roesch** *How does gene expression entropy change along developmental trajectories?*
15:10	Session Talk #76 **Jacob Bradford** *CRISPR, faster, better - The Crackling method for whole-genome target detection*	Session Talk #114 **Yue Cao** *Benchmarking single cell RNA-sequencing simulation methods*
15:20	Session Talk #98 **Anushi Shah** *Investigation of de novo mutations in human genomes using whole genome sequencing datasets*	Session Talk #117 **Belinda Phipson** *propeller: finding statistically significant differences in cell type proportions in single cell RNA-seq experiments*
15:30	Session Talk #103 **Varuni Sarwal** *A comprehensive benchmarking of WGS-based structural variant callers*	Session Talk #16 **Marie Trussart** *Removing unwanted variation with CytofRUV to integrate multiple CyTOF datasets*
15:40	Session Talk #90 **Loic Thibaut** *powerSFS: quantifying the intolerance of genes to mutation with a statistical model of the site frequency spectrum*	Session Talk #53 **Anna Trigos** *The next generation of biomarkers in cancer: single-cell spatial analysis of tumour and microenvironment cells*
15:50	**Invited Speaker: Phillippa Taberlay** *Recapitulation of a juvenile-like histone landscape in aged neurons*	**Invited Speaker: Fabio Zanini** *Hypothesis generation in the age of cell atlases*
16:10	Short Break Remo Link	
16:40	best Combine talk	
16:50	best Async Talk award	
17:00	ABACBS Day 3 Ends	

Workshops (Friday 27th November)

- **WORKSHOP 1: TIDY TRANSCRIPTOMICS FOR BULK AND SINGLE-CELL ANALYSES**
  - Instructor: Stefano Mangiola (mangiola.s@wehi.edu.au)
  - Time: 11 am - 2:30 pm (Sydney time, AEDT), including 30 mins break

- **WORKSHOP 2: USING CONTAINERS IN BIOINFORMATICS**
  - Instructor: Dr Marco De La Pierre, Pawsey Supercomputing Centre.
  - Please contact Christina Hall for queries on this workshop (christina@biocommons.org.au).
  - Time: 2:30 pm - 6:00 pm (Sydney time, AEDT), including 30 mins break

- **WORKSHOP 3: INTRODUCTION TO DEEP LEARNING AND TENSORFLOW**
  - Instructor: Titus Tang (titus.tang@monash.edu)
  - Time:
    - Part A 11 am - 2:30 pm (Sydney time, AEDT)
    - Part B 2:30 pm - 6:00 pm (AEDT)

- **WORKSHOP 4: PDB STRUCTURES: PROCESSING AND VISUALISATION**
  - Instructor: Carlos Miranda Rodrigues (cmiranda1@student.unimelb.edu.au)
  - Time: 2:30pm - 6pm (Sydney time, AEDT), including 30 mins break

Invited Speakers

Ami Bhatt
Stanford University

Ami Bhatt, MD, PhD is a physician scientist with a strong interest in microbial genomics, metagenomics, and global health. She received her MD and PhD from the University of California, San Francisco. She then carried out her residency and fellowship training at Harvard's Brigham and Women's Hospital and Dana-Farber Cancer Institute, and served as Chief Medical Resident from 2010-2011. She joined the faculty of the Departments of Medicine (Divisions of Hematology and Blood & Marrow Transplantation) and Genetics at Stanford University in 2014 after completing a post-doctoral fellowship focused on genomics at the Broad Institute of Harvard and MIT. Prof. Bhatt has received multiple awards for her academic scholarship including the Chen Award of Excellence from the Human Genome Organisation (HUGO).
Her lab develops molecular and computational methods to investigate the intestinal microbiome, with a strong focus on the use of next generation sequencing to define the microbiome and host genomic and transcriptomic features in patients with hematological diseases; developing custom computational tools for the identification of novel human commensals and pathogens in these immunosuppressed patient populations; and using statistical and functional biological methods to understand host and microbial factors that confer sensitivity to colonization or infection by certain bacteria, viruses and fungi.

Paul Gardner
University of Otago

My personal research goal is the complete functional classification of all functional RNAs. It is well established that RNAs such as tRNA and rRNA play a vital role in translation. The further discovery of more RNAs that build the translational machinery, such as snoRNAs, RNaseP and MRP along with spliceosomal RNAs, was not too great a shock. What has come as a surprise to RNA researchers in recent years is the importance of RNA in other areas such as regulation. Newly discovered RNAs and RNA-related processes involving regulation include: RNA interference (RNAi), microRNAs, siRNAs, riboswitches, thermosensors, bacterial antisense sRNAs, leader elements, IRES's and frame-shift elements. Further discoveries have shown that RNA is important for defending genomes from invasive elements such as transposons and bacteriophages, for deactivating entire chromosomes, and are essential for DNA replication. Furthermore, the discovery of RNA enzymes, or ribozymes, has lead to the suggestion that early life's genetic and numerous metabolic processes were RNA-based. Yet many groups are discovering thousands of RNAs with no known function. Clearly we are just beginning to appreciate the number of central roles that RNA mediated processes play in biology. What further unrealised roles does this "dark matter" of the cell play?

Karine Le Roch
University of California, Riverside

Dr. Le Roch is an professor at the University of California, Riverside (UCR). She obtained her master's degree in Parasitology at the University of Lille II and the University of Oxford, in 1997. She completed her Ph.D. in June 2001 at the University of Paris VI, working on the cell cycle regulation of the human malaria parasite, Plasmodium falciparum. In 2001, as a postdoctoral fellow, she joined the Scripps Research Institute, San Diego, California to carry out the functional analysis of the P. falciparum genome using microarray technologies. She joined the Genomics Institute of the Novartis Research Foundation (California) in January 2004 where she developed the malaria drug discovery program. Since April 2006 at UCR, Dr. Le Roch is using functional genomics approaches such as proteomics and high-throughput sequencing technologies to elucidate critical regulatory networks driving the malaria parasite life cycle progression and identify novel drug targets.

Karen Miga
University of California Santa Cruz

I am a satellite DNA biologist and the co-lead of the telomere-to-telomere (T2T) consortium. My research program combines innovative computational and experimental approaches to produce the high-resolution sequence maps of human centromeric and pericentromeric DNAs. In doing so, I am uncovering a new source of genetic and epigenetic variation in the human population, which is useful to investigate novel associations between genotype and phenotype of inherited traits and disease.

Seung Rhee

Carnegie Institution for Science, Washington DC, USA

Seung Yon (Sue) Rhee is a Senior Staff Member of Plant Biology Department at Carnegie Institution for Science. Her group strives to uncover the molecular mechanisms underlying adaptive traits in the face of heat, drought, nutrient limitation, and pests. Dr. Rhee's group studies a variety of plants including models, orphan crops, medicinal and desert plants. More recently their work has involved studying a model nematode C. elegans, fungal pathogens, and piezophilic bacteria. Her group employs computational modeling and targeted laboratory testing to study mechanisms of adaptation, functions of novel genes, organization and function of metabolic networks, and chemical and neuronal code of plant-animal interactions. Her group is also interested in developing translational research programs involving biomass maximization under drought in bioenergy crops. Dr. Rhee received her B.A. in biology from Swarthmore College and a Ph.D. in biology from Stanford University. She has been an investigator at the Plant Biology Department of Carnegie Institution for Science since 1999.

Yue Wan

Genome Institute of Singapore, Singapore

Yue Wan received her PhD degree in cancer biology from Stanford University, US, under the mentorship of Howard Y. Chang, during which she developed a high-throughput method for probing RNA structures genome-wide. She is currently a Principal Investigator at the Genome Institute of Singapore (GIS), a Society in Science-Branco Weiss Fellow, an EMBO Young Investigator and a CIFAR-Azrieli Global Scholar. She received the Young Scientist Award at the President's Science and Technology Awards in Singapore in 2015 and the L'Oreal-UNESCO for Women in Science, Singapore National Fellowship in 2016. Her research interests include functional RNA structures and understanding their roles in regulating cellular biology.

Keolu Fox
University of California San Diego

Dr. Fox earned his doctorate in Genome Sciences in 2016 at the University of Washington, Seattle. He then went on to serve as a postdoctoral fellow at UCSD since 2016, during which he was awarded the NIH, Institutional Research and Academic Career Development Award (2017) and the UC Chancellors' Postdoctoral Fellowship (2018). Dr. Fox's research program is multi-disciplinary in nature, reflecting his interdisciplinary research experience in anthropology, genomics, and computer science. His primary research focuses on questions of functionalizing genomics, which involves putting to the test theories of natural selection by editing genes and determining the function of the mutations. This unique approach of hypothesis testing through gene editing allows him to examine and test effects of genetic variants assumed to be under natural selection, such as "thrifty genes" in Polynesians, or Neanderthal variants in human cell lines. Dr. Fox is using the latest gene editing (CRISPR) technologies to ask anthropological questions about natural selection in humans and other closely related species that have never before been testable. Based on this work, he has been granted prestigious awards from Anthropological institutions including American Association for Physical Anthropology (Cobb Professional Development Grant) 2018 and the National Geographic Emerging Explorer (selected as one of fourteen 'world-changers'). His work has implications for understanding fundamental biological processes and diseases, and for these as they affect social groups. Dr. Fox connects biological anthropology with other subfields to address the relationship of genomics to society, the relationship of indigenous communities to science, questions of human health from a holistic biocultural perspective, and paleogenetics as a complement to archeological science.

Bernice Waweru
International Livestock Research Institute (ILRI), Nairobi, Kenya

Bernice Waweru has a background in plant breeding, biotechnology and bioinformatics through the Bioinformatics Community of Practice fellowship, conducted by the John Innes Center, Earlham Institute and BecA-ILRI hub, ILRI-Nairobi She has worked at KALRO studying resistance to stem rust of wheat in collaboration with CIMMYT. Bernice now works on genomics and bioinformatics at BecA-ILRI Hub. She and her colleagues are working to develop the first African led draft genome of the African Yam Bean, fully sequenced and analyzed in Africa.

Abstracts

National Keynote Speakers

Sara Ballouz

Garvan Institute of Medical Research, Sydney, NSW

I obtained my PhD from the University of New South Wales and the Victor Chang Cardiac Research Institute in 2013, working with Drs Merridee Wouters and Bruno Gaeta. I then moved to Cold Spring Harbor Laboratory for my postdoctoral training with Dr Jesse Gillis. In 2020, I started my own group at the Garvan-Weizmann Centre for Cellular Genomics at the Garvan Institute of Medical Research.
My central scientific interest has been to understand the genetic architecture of disease. With data from the genome, transcriptome, epigenome and proteome increasing exponentially, robust tools and practices need to be established to analyse this deluge, in particular if to be applied to personalized medicine.

Phillippa Taberlay

University of Tasmania

Associate Professor Phillippa Taberlay is a NHMRC Emerging Leadership Fellow and Senior Research Fellow in the Tasmanian School of Medicine, College of Health and Medicine. Her research centres on understanding two-dimensional and three-dimensional aspects of gene control, and uses cutting-edge methods to delineate mechanisms of epigenetic reprogramming in development, healthy ageing, cancer and neurodegenerative disorders. Phillippa attained her Bachelor of Science majoring in Biochemistry, Microbiology and Immunology from the University of Tasmania in 2002, and graduated with a Bachelor of Science (First Class Honours) in Biochemistry and Molecular Biology in 2003. She then joined the laboratory of Dr Adele Holloway where she sought to understand how leukaemic fusion proteins disrupt epigenetic mechanisms in Acute Myeloid Leukaemia.
Phillippa undertook her post-doctoral research training in the laboratory of Professor Peter Jones at the University of Southern California, USA (2008-2011). Her discovery that enhancer epigenetic states underpin cell reprogramming (Taberlay, Cell 2011) was an advance for the field that has shaped new theories of epigenetic regulation. Her early-career research has also received two prestigious Faculty of 1000 recommendations, and was named as one of the Top Clinical Advances 2012 (American Society of Clinical Oncology). Phillippa co-developed the NOMe-Seq technique, described as an “impressive” and “ingenious innovation” and named as a Top 10 Innovation of 2013. In 2012, Phillippa established her research group within the Epigenetics Research Program of Professor Susan Clark at the Garvan Institute of Medical Research, where she developed several new projects, including cutting-edge technologies to map higher-order (3D) genome structures inside cells.
Phillippa returned to the University of Tasmania as a National Health and Medical Research Council (NHMRC) Career Development Fellow in 2016. She established the ‘2D and 3D Epigenomic Remodelling’ Group under the umbrella of Medical Sciences in the Tasmanian School of Medicine and now mentors a number of research assistants, post-doctoral researchers, Ph.D. candidates and Honours students. She has published her work in prominent international journals including Cell, Genome Research, Cancer Cell and PNAS.

Jessica Mar

University of Queensland

Associate Professor Jessica Mar's research group focuses on the development of bioinformatics methods to understand how regulatory processes go awry in human diseases. Specifically, the groups is interested in modelling how variability of gene expression contributes to regulation of the transcriptome. This interest has led us very naturally into single cell biology where there is a great need to develop accurate statistical approaches for data arising from single cell sequencing. Elucidating heterogeneity and variability in gene expression in this context in important as this may uncover new cellular subtypes or identify stochasticity in the usage of key pathway or master regulators.
The explosive availability of big data sets, coupled with the speed at which sequencing technologies have advanced have created an exciting environment for the current state of computational biology research. The Mar group looks to modern tools in statistics, such as Bayesian methodologies and machine learning algorithms, to make sense of biology from big data.

Robert Lanfear

Australian National University

I grew up in England. My BSc and honors were in Ecology, my MSc was in Artificial Intelligence, and my PhD (at the University of Sussex in the UK) was in developmental biology. After a short postdoc attempting and failing to edit shrimp genomes at University College London, I switched gears to focus on my growing interest in molecular evolution and phylogenetics. In 2008 I moved the ANU as a postdoc, where I remained for 6 years. I then took up a permanent position as a Senior Lecturer in Genomics at Macquarie University in Sydney, and moved back to the ANU at the first opportunity in 2016. I now focus on a range of topics including molecular evolution, somatic mutation, phylogentics, comparative methods, and bioinformatics.

Fabio Zanini

UNSW Sydney

I am a group leader at UNSW in Sydney, Australia, focusing on integrating computer science, engineering, and biomedicine to understand viral infections, immunology, and basic biology.
I studied physics in Trento, Italy and Tübingen, Germany with a focus on mathematical physics, biophysics, and quantum optics. In my master thesis and related articles I studied the diffusional dynamics of proteins in solutions close to crystallization via light and neutron scattering. I was awarded a PhD at the Max Planck Institute for Developmental Biology in Tübingen, Germany with a thesis with Richard Neher on HIV evolution. I later spent 3 and a half years in Stephen Quake's lab at Stanford developing novel assays and data analytics to study viral infections at the single cell level. I started my lab in Sydney, Australia in September, 2019.

Kim-Anh Lê Cao

The University of Melbourne

Kim-Anh is committed to statistical education to instil best analytical practice. Kim-Anh has a mathematical engineering background and graduated with a Ph.D in statistics from the Université de Toulouse, France. She then moved to Australia to forge her own non-linear career path as a multi-disciplinary collaborator, both as a biostatistician consultant at QFAB Bioinformatics, and as a research group leader at the biomedical University of Queensland Diamantina Institute. She currently continues her strong research focus as a senior lecturer at the University of Melbourne. Kim-Anh has secured two consecutive NHMRC fellowships from 2014. In 2019 she received the Australian Academy of Science's Moran Medal for her contributions to Applied Statistics and was also awarded the Georgina Sweet Award from Prof Leann Tilley for Women in Quantitative Biomedical Science. Kim-Anh was selected to the international Homeward Bound program that aims to build a global collaboration of 1,000 women leaders in STEMM over ten years, culminating to a trip to Antarctica in 2019. Dr Lê Cao's main research focus is on variable selection for biological data (`omics' data) coming from different functional levels by the means of multivariate dimension reduction approaches. Since 2009, her team has been working on developing the statistical R toolkit mixOmics that is dedicated to the integrative analysis of `omics' data, to help researchers make sense of biological big data. She and her team regularly run statistical training workshops and short series seminars and mixOmics multi-day workshops.

Torsten Seemann

University of Melbourne

Associate Professor Torsten Seemann is a bioinformatician specialising in applied microbial genomics. He is lead bioinformatician at MDU PHL and Doherty Applied Microbial Genomics. His primary role is to manage the hardware and software analysis infrastructure needed to modernise public health microbiology services by replacing traditional assays with high resolution whole genome sequencing, and to build a national data sharing network to facilitate pathogen and antimicrobial resistance surveillance, and link with similar international networks.

Robert Edwards

Flinders University, Adelaide, SA

Rob received his BSc (Hons) from De Montfort University, Leicester, England, and then completed a PhD at the Nitrogen Fixation Laboratory at the University of Sussex, England exploring the regulation of nitrogen fixation in Klebsiella pneumoniae. He moved to the US as a PostDoc, first at the University of Pennsylvania in Philadelphia studying enterotoxigenic E. coli, and then at the University of Illinois, Urbana-Champaign studying the genomics and pathogenesis of Salmonella, work that he continued as an Assistant Professor at the University of Tennessee Health Sciences Center in Memphis, TN. Working with Argonne National Labs and the Fellowship for the Interpretation of Genomes (FIG) he developed the RAST and MG-RAST systems for bacterial genome and metagenome annotation.
In 2006, Rob took a position in the Departments of Computer Science and Biology at San Diego State University, and his work there led to breakthroughs in our understanding of how viruses interact with their hosts, and how viruses from around the world carry important genetic information. Rob has continued to push current sequencing and bioinformatics technologies, in 2013 took a next-generation sequencing machine to the remote Southern Line Islands to explore metagenomics of coral reefs in real-time. In 2014 Rob's team identified a virus that is present in the intestines of approximately half the people in the world, and in 2019 Rob assembled a consortium of 115 colleagues from every continent to demonstrate the global spread of the virus.
In addition to science and teaching Rob is also an advanced scientific SCUBA diver having led teams to study Coral Reefs all over the world. In his spare time, he is a cyclist, black-diamond skier and an avid international yachtsman, navigating in long-distance offshore races, including navigating the 2019 TransPac race from Los Angeles to Honolulu finishing 4th out of 89 boats.

Yu Lin

Australian National University (ANU), Canberra, ACT

Yu Lin joined the Research School of Computer Science in September 2016. Prior to this, he was a postdoctoral fellow at the Department of Computer Science and Engineering, University of California, San Diego. His research focuses on computational biology and bioinformatics, and he has been working on algorithms for genome assembly, the analysis of genome rearrangements, and phylogenetic reconstruction.
He received his PhD in Computer Science from École Polytechnique Fédérale de Lausanne (EPFL), Switzerland. He also holds a master's degree in Computer Science from Chinese Academy of Sciences and a bachelor's degree in Computer Science from the University of Science and Technology of China (USTC).
He has received two awards from the Swiss National Science Foundation. He is also the recipient of Chinese Government Award for Outstanding Self-Financed Students Abroad, Director's Award from Institute of Computing Technology, Chinese Academy of Sciences, Guo Moruo Presidential Award from University of Science and Technology of China.

Alex Brown

South Australian Health and Medical Research Institute

Professor Alex Brown is an Aboriginal medical doctor and researcher. He grew up on the south coast of New South Wales (NSW) with family connections to Nowra, Wreck Bay and Wallaga Lake on the far south coast of NSW. Alex trained in medicine at the University of Newcastle, before working in hospitals on the central coast of NSW. He subsequently travelled to Israel to complete a Master of Public Health and returned to Australia to begin work in Alice Springs, where he spent 14 years. Alex first managed the local Centre for Disease Control in Alice Springs, controlling outbreaks of disease, immunisation programs and the surveillance of disease, before starting in research for the Menzies School of Health Research. In 2007 he was appointed to set up a research program in Central Australia with Baker IDI Heart and Diabetes Institute, with a focus on heart disease and diabetes in Aboriginal people. During this time, Alex commenced and completed his PhD on depression and heart disease in Aboriginal men.

Over the last 20 years, Alex has established an extensive and unique research program focused on chronic disease in vulnerable communities, with a particular focus on outlining and overcoming health disparities. He leads projects encompassing epidemiology, psychosocial determinants of chronic disease, mixed methods health services research in Aboriginal primary care and hospital settings, and randomised controlled trials of pharmacological and non-pharmacological chronic disease interventions. Alex has been involved in policy since he commenced as a doctor. He has been heavily involved in engaging government and lead agencies in setting the agenda in Aboriginal cardiovascular disease management and control and chronic disease policy more broadly. He sits on a range of national committees,and co-chairs the Indigenous Research Health Fund through the MRFF.

In July 2012, Alex joined SAHMRI to lead Aboriginal health research, the same year that he was awarded the prestigious Viertel Senior Medical Research Fellowship to further his research into the impacts of psychosocial determinants on cardiovascular disease in Aboriginal communities. He also holds an NHMRC Research Fellowship.

Full Accepted Abstract List

Abstract #4

Luyi Tian (Walter and Eliza Hall Institute of Medical Research, Melbourne, VIC)
Comprehensive characterization of single cell full-length isoforms in human and mouse with long-read sequencing

Alternative splicing shapes the phenotype of cells in development and disease. Longread RNA-sequencing recovers full-length transcripts but has limited throughput at the single-cell level. Here we developed single-cell full-length transcript sequencing by sampling (FLT-seq), together with the computational pipeline FLAMES to overcome these issues and perform isoform discovery and quantification for single cells. With FLT-seq and FLAMES, we performed the first comprehensive characterization of the full-length isoform landscape in single cells of different cell types and species, identified thousands of unannotated isoforms. We found conserved functional modules that were enriched for alternative transcript usage in different cell populations, including ribosome biogenesis and mRNA splicing. Analysis at the transcript-level allowed data integration with scATAC-seq on individual promoters, improved correlation with protein expression data and linked mutations known to confer drug resistance to transcriptome heterogeneity. Our methods reveal previously unseen isoform complexity and provide a better framework for multi-omics data integration.

Abstract #5

Thomas Quinn (Applied Artificial Intelligence Institute (A2I2), Deakin University, Geelong, VIC)
Learning Distance-Dependent Motif Interactions: An Interpretable CNN Model of Genomic Events

In most biological studies, prediction is used primarily to validate the model; the real quest is to understand the underlying phenomenon. Therefore, interpretable deep models for biological studies are required. Here, we propose HyperXPair (the Hyper-parameter eXplainable Motif Pair framework) to model biological motifs and their distance-dependent context through explicitly interpretable parameters. This makes HyperXPair more than a decision-support tool; it is also a hypothesis-generating tool designed to advance knowledge in the field. We demonstrate the utility of our model by learning distance-dependent motif interactions for two biological problems: transcription initiation and RNA splicing.

Abstract #6

Enrique Zozaya-Valdes (Peter MacCallum Cancer Centre, Melbourne, VIC)
Accurate detection of cell free microbial DNA using a contaminant-controlled analysis framework

The human microbiome plays an important role in cancer. Accumulating evidence indicates that commensal microbiome-derived DNA may be represented in minute quantities in the cell-free DNA of human blood and could possibly be harnessed as a new cancer biomarker. However, there has been limited use of rigorous experimental controls to account for contamination, which invariably affects low-biomass microbiome studies. Here, we applied a combination of 16S-rRNA-gene sequencing and droplet digital PCR to determine if the specific detection of cell free microbial DNA (cfmDNA) was possible in 69 metastatic melanoma patients. Compared to matched stool and saliva samples, the absolute concentration of cfmDNA was low (mean of 2,719 gene copies/ml of plasma) but was significantly above the levels detected from negative controls (mean of 1,829 gene copies/ml). The microbial community of plasma was strongly influenced by laboratory and reagent contaminants introduced during the DNA extraction and sequencing processes. Through the application of an in silico decontamination strategy including the filtering and removal of Amplicon Sequence Variants (ASVs) with batch dependant abundances and those with a higher prevalence in negative controls, 31 high confidence plasma ASVs were identified. These included known gut commensal bacteria (e.g. Faecalibacterium, Bacteroides and Ruminococcus) but also other uncharacterised ASVs. Together, these observations indicate that plasma can harbour a low yet detectable level of cfmDNA. These results highlight the importance of accounting for contamination and provide an analytical decontamination framework to allow the accurate detection of cfmDNA for future biomarker studies in cancer and other diseases.

Abstract #7

Matthew DeMaere (ithree institute, University of Technology Sydney, Sydney, NSW)
Reference-free quality assessment for Hi-C sequencing data.

Hi-C is a sample preparation method that enables high-throughput sequencing to capture genome-wide proximity interactions between DNA molecules. The technique has been successfully applied to solve challenging problems such as 3D structural analysis of chromatin, scaffolding of large genome assemblies and more recently the accurate resolution of metagenome-assembled genomes (MAGs). Despite continued refinements, however, Hi-C library preparation remains a costly and complex laboratory protocol. QC options for Hi-C libraries are limited, with current wet-lab protocols only giving a very crude assay for the re-digest of ligation junctions in a Hi-C library. This QC approach does not provide a reliable estimate of the fraction of library templates containing a Hi-C junction, nor

is it available to all Hi-C protocols. QC via sequencing is another possible approach, but current tools require a reference genome to estimate quality metrics for the Hi-C library. We propose a new, reference-free approach for Hi-C library quality assessment that requires only a small amount of sequencing data from a library. Our tool, qc3C, implements an algorithm that estimates the fraction of reads in the library that contain Hi-C junctions, along with other quality metrics. The algorithm builds upon the observation that proximity ligation events are likely to create k-mers that would not naturally occur in the sample. The algorithm uses an empirical cumulative distribution of k-mer depths to compute the probability that a read containing a Hi-C junction sequence was generated by the proximity ligation reaction. This in turn enables the total fraction of reads containing Hi-C junctions to be estimated. We characterise the accuracy of the new algorithm on simulated and real datasets and compare it to reference-based methods. qc3C enables sequencing depth requirements to be estimated more precisely on a per-library and per-experiment basis, for chromosome conformation studies, for Hi-C scaffolding of assemblies, and for metagenomic Hi-C. Our qc3C software is an easy to use open-source tool that integrates with the multiQC framework. To our knowledge, qc3C is the first reference-free Hi-C quality assessment tool, enabling Hi-C to be more easily applied to non-model organisms and environmental samples. qc3C is available from https://github.com/cerebis/qc3C

Abstract #8

Daniela Gaio (University of Technology Sydney, Sydney, NSW)
Over 25,000 metagenome assembled genomes reveal the development of the post-weaning pig gut microbial community

Building on a new lower-cost metagenome sequencing technique developed by us, we have carried out the largest metagenomic analysis of the pig gut microbiome to-date. By processing over 800 time-series samples from 126 porcine hosts, we have generated over 8Tbp of metagenomic sequence data. From this data we reconstructed over 50,000 metagenome-assembled genomes (MAGs) of organisms resident in the porcine gut, 26,800 of which were above 70% complete with a <10% contamination, and 12,400 of which were nearly complete genomes. To do so we created co-assemblies for each individual host, pooling all the time-series samples available from each subject, thereby increasing the power to reconstruct genomes from low abundance microbes. We find that the microbiomes of post-weaning piglets appear to follow a highly structured developmental program in the weeks following weaning, and this development is robust to treatments including antibiotics and probiotics. The high resolution we obtained allowed us to identify specific taxonomic and genomic ‚Äúsignatures‚Äù characterizing the stages that the piglet gut microbiome shifts through during its development immediately after weaning (4 weeks) and up to the ninth week of life. These compositional shifts were also evident in analysis with an assembly-free phylogenetic profiling technique. Both the assembly-free method and the analysis of MAGs show strong associations between community composition and individual factors such as breed and mother, and can even resolve small differences in age (down to 3 days). Finally, by predicting proteins from the metagenomes and mapping them against the CAZy database, we described the carbohydrate repertoire of the post-weaning piglets. We tracked the shifts in abundance of these enzymes across time, and identified the species and higher level taxonomic groups that carry each of these enzymes in their metagenomes.

Abstract #9

Qian Du (Garvan Institute of Medical Research, Sydney, NSW)
DNA hypomethylation induces intrapopulation heterogeneity of DNA replication timing and 3D genome reorganisation

DNA replication timing is associated with epigenomic changes during human differentiation and between normal and cancer (Du et al. 2019 Nat. Commun.). A largely unaddressed question to date is whether alterations of the epigenome can ‚Äòdrive‚Äô changes in DNA replication timing? DNA methylation alterations in cancer, in particular hypomethylation, have been related to replication timing. For example, treatment of cancer patient lymphoblasts with 5‚Äô-azacytidine demethylating drug induces replication timing changes in key cancer genes such as RB1 (Dotan et al. 2008 BMC Cancer). To investigate whether alterations in DNA methylation can cause a change in DNA replication timing genome-wide, we profiled replication timing (Repli-Seq) in the colorectal cancer cell line HCT116, and the corresponding DNA methyltransferases DNMT1 and DNMT3B double knockout cell line (DKO1). Genome-wide, DKO1 cells showed a striking loss in precision of replication timing and loss of chromatin conformation integrity. Using single cell DNA sequencing, we found that DNA hypomethylation increased intra-population replication timing heterogeneity, indicating a possible loss of synchronicity of replication origins due to DNA methylation loss. Further, we found that discrete regions that undergo a large change in replication timing in the hypomethylated DKO1 cells show loss of allelic replication timing and shrinking of late-replicating partially methylated domains (PMDs). In contrast, conservation of replication timing after DNA hypomethylation at PMDs is associated with the acquisition of broad H3K9me3/H3K4me3 domains. We hypothesise that formation of these large bivalent domains serve to prevent ectopic transcription and loss of genome organisation, and provide an alterative ‚Äòrescue‚Äô pathway to maintain genome function in response to DNA methylation loss. In summary, global DNA methylation loss reduces the stability of replication timing and genome organisation. Further, we propose that the DKO1 cells may have adapted to protect against large changes in replication timing through changes in the chromatin landscape, in particular at PMDs.

Abstract #11

Khelina Fedorchuk (Swinburne University of Technology, Melbourne, VIC)
Machine Learning in Dynamic Microscopy

The tracking of individual, proliferating cells over time has long been more accurately done by eye than by algorithm. Fully automated, simultaneous tracking of multiple cells remains a daunting challenge, particularly for highly motile cells that grow and divide over days or weeks. Here we examine the use of deep convolutional neural networks to track several generations of T-lymphocytes through tens of thousands of time-lapse microscopy images. Neural networks are trained on short sequences of consecutive frames, where time is rendered in the 3rd dimension. This converts the usual 2D detection plus association problem into a single 3D problem. The network is required to identify cells (classification) and determine cell positions (regression). An additional neural network is used to perform high-quality segmentation of each cell in its cluttered environment. These networks work together to provide an automated solution for extracting cell lineage trees from movies of proliferating cells.

Abstract #12

Jiayue-Clara Jiang (University of Queensland, Brisbane, QLD)
Integrated transcription factor profiling with transcriptome analysis: identification of L1PA2 transposons as global regulatory modulators in a breast cancer model

LINE1 retrotransposons occupy approximately 17% of the human genome. Many LINE1 retrotransposons contain cis-regulatory sequences, such as transcription factor binding sites and promoters, which can be exapted to modulate human gene expression. While transposons are generally silenced in somatic tissues, many LINE1 retrotransposons can escape epigenetic repression in epithelial cancers, become transcriptionally active and contribute regulatory activity. We have developed a bioinformatic pipeline for the integrated analysis of transcription factor binding and transcriptomic data to identify transposon-derived promoters that are activated in specific diseases and developmental states. We applied this pipeline to a breast cancer model, and showed that L1PA2 retrotransposons, a primate-specific subfamily of LINE1, contribute abundant regulatory sequences to co-ordinated transcriptional regulation in breast cancer. Analysis of ChIP-seq data showed that 27% of L1PA2 retrotransposons were bound by transcription factors in MCF7 cells, with the majority of binding sites clustering in the 5' untranslated region. L1PA2 retrotransposons also facilitated binding site co-localisation of functionally interacting transcription factors, many of which are known to be involved in the transcriptional mis-regulation in cancer. In addition to being a rich reservoir of oncogenic transcription factor binding sites, L1PA2 retrotransposons also contributed transcription start sites to a number of transcripts in MCF7 cells, suggesting the activation of dormant promoters. These transcripts showed upregulated expression in MCF7 cells compared to the near-normal MCF10A cells, and their cancer-specific expression was supported by an active epigenetic profile. Our results demonstrate that L1PA2 retrotransposons are a prominent contributor of transcription factor binding sites in breast cancer, and drive the expression of cancer-specific transcripts through co-ordinated binding of oncogenic transcription factors. Understanding the regulatory impact of L1PA2 on breast cancer genomes may provide an insight into cancer genome regulation, and provide novel biomarkers for disease diagnosis and novel candidates for targeted therapy.

Abstract #13

Yue You (Walter and Eliza Hall Institute of Medical Research, Melbourne, VIC)
Longitudinal single-cell immune profiling revealed distinct innate immune response in asymptomatic COVID-19 patients

Recent studies have characterized the single-cell immune landscape of host immune response of coronavirus disease 2019 (COVID-19), specifically focus on the severe condition. However, the immune response in mild or even asymptomatic patients remains unclear. Here, we performed longitudinal single-cell transcriptome sequencing and T cell/B cell receptor sequencing on 3 healthy donors and 10 COVID-19 patients with asymptomatic, moderate, and severe conditions. We found asymptomatic patients displayed distinct innate immune responses, including increased CD56briCD16- NK subset, which was nearly missing in severe condition and enrichment of a new Th2-like cell type/state expressing a ciliated cell marker. Unlike that in moderate condition, asymptomatic patients lacked clonal expansion of effector CD8+ T cells but had a robust effector CD4+ T cell clonal expansion, coincide with previously detected SARS-CoV-2-reactive CD4+ T cells in unexposed individuals. Moreover, NK and effector T cells in asymptomatic patients have upregulated cytokine related genes, such as IFNG and XCL2. Our data suggest early innate immune response and type I immunity may contribute to the asymptomatic phenotype in COVID-19 disease, which could in turn deepen our understanding of severe COVID-19 and guide early prediction and therapeutics.

Abstract #14

Yiwen Wang (Univerisity of Melbourne, Melbourne, VIC)
A multivariate method to correct for batch effects in microbiome data

Microbial research has become critical to investigate the link between microbial composition and phenotypes, including human diseases. Microorganisms are highly dynamic and hence sensitive to variations in the environment. Microbiome, the genetic material of all microorganisms, is therefore vulnerable to batch effects. In this context, we define batch effects as sources of unwanted variation introduced by confounding factors that may originate from biological, technical or computational reasons. Batch effects are not related

to and obscure any factors of interest. Most existing methods that correct for batch effects were developed for gene microarray or RNA-seq data. They do not consider the data characteristics of microbiome, such as sparsity, overdispersion, skewed distribution and correlation between variables. To address these issues, we propose a new method based on partial least squares discriminant analysis for batch effect correction (PLSDA-batch). This method is non-parametric and thus can handle the skewed distribution caused by sparsity and overdispersion. As PLSDA-batch is also multivariate, we can account for the high correlation between microbial variables. Our method uses PLS-DA to first estimate treatment variation ,Äì thus preserving the biological variation of interest, then batch variation with latent components. The variation due to batch effects is then removed using matrix deflation. The resulting batch effect corrected data can then be input in any downstream statistical analysis. We developed two other variants: weighted PLSDA-batch for unbalanced batch x treatment design, and sparse PLSDA-batch to avoid overfitting in component estimation. We compared our approaches on simulated and three case studies, with existing batch correction methods removeBatchEffect and ComBat. For a balanced design, our methods (PLSDA-batch & sparse PLSDA-batch) led to competitive performance in removing batch variation while preserving treatment variation. For an unbalanced design, weighted PLSDA-batch led to a better performance than unweighed PLSDA-batch. When batch effects had high variability, sparse PLSDA-batch outperformed PLSDA-batch and the other two existing methods in both balanced and unbalanced designs. Our future work will investigate whether our approaches are suitable for other types of sequencing count data facing batch effects. Reproducible code and vignettes will soon be available on GitHub.

Abstract #15

Adrien Oliva (University Of Adelaide, Adelaide, SA)
Systematic Benchmark of aDNA Mapping Bias

Recent advances in molecular techniques, as well as the dramatic drop in the cost of DNA sequencing, gave a boost to the ancient DNA (aDNA) field. This resulted in population-level datasets and powered several important discoveries using aDNA, such as identifying and characterizing archaic hominin introgression, large-scale migrations, and replacements. However, mapping sequences, the cornerstone of ancient genomic discovery, is challenged by the shortness of these sequences, divergence from the reference sequence, and artefactual substitution resulting from the DNA molecules' degradation. Mapping vast amounts of short DNA sequences is a computationally challenging task that inevitably produces artifacts, including biases against alleles not found in the reference genome (reference bias). For nearly a decade, the gold-standard for aDNA mapping strategy remained unchanged while new software has emerged. Using simulated human aDNA sequences from different populations, we benchmarked 29 mapping strategies using four different mapping software ,Äì BWA-aln, BWA-mem, NovoAlign, and Bowtie2 ,Äì and quantified the impact of reference bias in downstream population genetic analyses. We show that after filtering out sequences with low mapping quality, in particular, resulted in high mapping precision with negligible levels of reference bias. Using an enhanced reference with population-level variation (i.e., IUPAC reference) with Novoalign results in very low bias levels and our study's best precision. Using the IUPAC reference demonstrates the benefit of incorporating population genetic information into the mapping process, echoing recent results based on graph genome representations.

Abstract #16

Marie Trussart (Walter and Eliza Hall Institute of Medical Research, Melbourne, VIC)
Removing unwanted variation with CytofRUV to integrate multiple CyTOF datasets

Mass cytometry (CyTOF) is a technology that has revolutionised single-cell biology. By detecting over 40 proteins on millions of single cells, CyTOF allows the characterisation of cell subpopulations in unprecedented detail. However, most CyTOF studies require the integration of data from multiple CyTOF batches usually acquired on different days and possibly at different sites. To date, the integration of CyTOF datasets remains a challenge due to technical differences arising in multiple batches. To overcome this limitation, we developed an approach called CytofRUV for analysing multiple CyTOF batches, which includes an R-Shiny application with diagnostic plots. CytofRUV can correct for batch effects and integrate data from large numbers of patients and conditions across batches, to confidently compare cellular changes and correlate these with clinically relevant outcomes.

Abstract #17

Emma Rath (Victor Chang Cardiac Research Institute, Sydney, NSW)
A survey of retrocopied genes in multiple cohorts

Retrocopied genes arise when spliced RNA of coding genes are reverse transcribed and integrated back into genomic DNA at a locus that is not that of the parent gene. Retrocopied genes not present in the reference genome are aligned to the parent gene and structural variant callers call ,Äúclean-intron-deletions,Äù where the entire intron appears to have been cleanly deleted, for all introns of the gene. At both UTR ends of the parent gene, a translocation is called to a different chromosome that is the true location of this retrocopied gene. We carried out a survey of clean-intron-deletions in various cohorts representing various diseases and controls. Genes involved included previously reported and retrocopied genes and genes not previously reported. We found a weak correlation per sample between retrocopy burden and mobile element insertion burden, and for the schizophrenia cohort this correlation was strong. We found

a weak correlation between retrocopy genes and mRNA stability as calculated by RNAfold. Samples taken from the same individual and sequenced at separate times resulted in the same retrocopies being reported. For the cohort sequenced from saliva, we observed the same clean-intron-deletions in the same genes that we observed for the cohorts sequenced from blood. The same putative insertion points in various individuals from out various cohorts were observed. For one of our cohorts, sequencing library prep involved two rounds of PCR amplification, and clean-intron-deletions were reported in more than double the number of different genes compared to other cohorts. Insertion points were not found for some suggesting that they may be the result of reverse transcription of RNA during library prep rather than representing retrocopies. The results of this survey permit us to more accurately interpret the results of our structural variants pipeline.

Abstract #18

Robert Lanfear (Australian National University, Canberra, ACT)
Confidence and truth in phylogenomics

How do you estimate a good phylogeny? How do you know if it's right? And how should you communicate your uncertainty? Phylogenies form the backbone of our understanding of the tree of life, and are crucial for understanding and tracking emerging diseases. But any phylogeny is just an estimate, and all estimates are associated with uncertainty. Accurately measuring and communicating uncertainty in these estimates is almost as important as building the phylogeny itself. Using the right measures of uncertainty can help avoid meaningless arguments, and in the case of emerging diseases can have important consequences for population health and national and international policy. In this talk I'll introduce a range of methods for measuring and communicating uncertainty in phylogenetics (bootstraps, concordance factors, and branch parsimony scores), and illustrate how and why each should be used with examples from estimating the tree of life to the genomic epidemiology of SARS-CoV-2.

Abstract #19

Tyrone Chen (Monash University, Melbourne, VIC)
Multi-omics data harmonisation for the discovery of COVID-19 drug targets

Background: The recent COVID-19 pandemic caused by SARS-COV2 virus has caused a high number of deaths globally. Despite the volume of experiments performed and data available, its biology is not yet fully understood. High throughput sequencing and mass spectrometry allow users to capture large quantities of high dimensional data. While it is possible to extract valuable information by analysing each omics data separately, integrating multiple modalities can yield information which would otherwise be hidden in individual omics analysis. This further enhances the robustness and reproducibility of the results. In a recent review we showed that existing multi-omics approaches require heavy data processing resulting in information loss. They usually are also restricted to specific data modalities or experimental set ups. Methods: In contrast to existing approaches, our pipeline unified proteome and translatome data at the lowest level features ensuring minimum information loss during integration. Using a multivariate approach, we accounted for the high dimensional structure of the data, and identified features in each omics data which most highly contribute to discriminating between the biological classes. The subset of features obtained here was further used to identify strong correlations within and between omics data types. From the strongly correlated features across the proteome and translatome, we identified potentially medically relevant biological targets for drug development. These findings were compared against publicly available datasets for association with the virus. Computational docking analyses were carried out for all identified drug-target combinations. To further narrow down this list, a pharmacokinetic analysis was used to assess drug efficiency and toxicity in humans and results were ranked. Results: We report 54 candidate drug targets, with 51 of these passing the pharmacokinetics criteria. One drug, etoposide phosphate that targets DNA topoisomerase2-alpha returns the highest molecular docking score. Thus, using a multi-omics integrative approach we report a high confidence list of computationally screened drugs for further biological validation. Some of our previous work on this project can be viewed here: https://doi.org/10.7490/f1000research.1118023.1

Abstract #21

Jacqueline Rehn (South Australian Health and Medical Research Institute (SAHMRI), Adelaide, SA)
Pseudo-alignment technique for accurate detection of disease-causing variants in acute lymphoblastic leukaemia

Background: Acute lymphoblastic leukaemia (ALL) is characterised by a diverse range of genomic alterations, some of which confer poor prognosis and increased risk of relapse. Timely and accurate detection of variants involved in disease pathogenesis is essential for patient risk stratification and precision medicine. Identifying gene fusions and single nucleotide variants (SNVs) from RNA sequencing (RNA-seq) data utilising established bioinformatics methods requires significant computational resources and produces a large number of variants, many of which are not clinically relevant. In this study, we assess the accuracy and efficiency of a targeted k-mer based pseudo-alignment approach for rapid variant detection, which requires a fraction of the computational power of current methods. Methods: Using a k-mer analysis program km, we tested 188 raw RNA-seq ALL samples against target sequences for 21 commonly observed ALL fusions as well as 15 pathogenic SNVs. Fusion target sequences were designed to represent commonly reported breakpoints or to detect mRNA transcripts present only as a result of a chromosomal rearrangement. Results were compared

to reported variants from popular alignment-based tools for detection of gene fusions and SNVs. Results: The km program accurately detected known fusions involving exon-exon breakpoints in 134 out of 135 patient samples (99%). This was completed in an average of 12 mins compared to >6 hrs for current alignment-based methods. In the sample where km failed, the gene fusion involved an alternative breakpoint junction and was subsequently detected after inclusion of an additional target sequence. Targets were also developed to accurately predict the presence of IGH-DUX4 and IGH-CRLF2 rearrangements, prognostically important lesions that are difficult to detect by cytogenetics and frequently missed or underreported by standard alignment-based tools. Our approach could detect these fusions in all (34) cases, showing 100% concordance with gene expression profiling or fluorescent in situ hybridization (FISH). Pathogenic SNVs were detected in 56 out of 63 patient samples (89%), including all 10 cases involving subtype defining ALL variants (PAX5 P80R and IKZF1 N159Y). Conclusions: We have demonstrated that a targeted pseudo-alignment approach to variant detection is not only highly accurate but also computationally efficient, enabling detection of multiple variant types with a single tool. Application of this algorithm in a clinical setting would enable rapid identification of patients with high-risk genomic alterations, ensuring appropriate treatment response. While this new approach cannot detect novel disease-causing variants, the repertoire of targets can easily be expanded as new clinically relevant alterations are identified.

Abstract #22

Yu Lin (Australian National University (ANU), Canberra, ACT)
Binning Metagenomic Sequences

Metagenomics studies have provided key insights into the composition and structure of microbial communities found in different environments. Among the techniques used to analyze metagenomic data, binning is considered as a crucial step in order to characterize the different species of microorganisms present in microbial communities. We propose two binning methods, GraphBin and MetaBCC-LR, to bin metagenomic sequences. GraphBin makes use of the assembly graph to refine binning results and to enable detecting shared sequences among multiple species. MetaBCC-LR is a new reference-free binning method which directly clusters long reads based on their k-mer coverage histograms and oligonucleotide composition.

Abstract #23

Fabio Zanini (University of New South Wales, Sydney, NSW)
Hypothesis generation in the age of cell atlases

While many laboratories are driven by computational method development or hypothesis-driven biomedical experimentation, we (http://fabilab.org) focus on data-driven hypothesis generation, especially single cell sequencing of viral infections, development, and cancer. We recently released northstar (Zanini et al. Scientific Reports 10, 15251 (2020)), a new algorithmic approach and software package to cluster/classify single cell transcriptomes guided by but not limited by a cell atlas. Northstar can annotate tens of thousands of cells from tumor samples into known cell types and unknown clusters within seconds on a laptop. We subsequently combined northstar with RNA velocity, ATAC-Seq, ChIP-Seq, knockdown, and overexpression to study gene regulation in acute myeloid leukemia (AML) and discovered an unknown level of heterogeneity within cancer cells. Single cell gene expression, chromatin state at enhancers, transcription factor binding motifs, and perturbation experiments highlighted an internal regulatory network that is modulated by the triad of transcription factors GATA2, TAL1, and ERG and shapes the phenotype of leukemic cells along trajectories that are distinct from but closely related to healthy haematopoiesis, suggesting new therapeutic approaches for leukemia. We will also illustrate our efforts on two distinct lines of research: (1) comparative single cell omics of viral infections and (2) constructing a cell atlas of the neonatal mammalian lung.

Abstract #24

Legana Fingerhut (James Cook University, Cairns, QLD)
ampir: an R package for fast genome-wide prediction of antimicrobial peptides

Antimicrobial peptides are natural antibiotics, part of the innate immune system, which help defend the host against pathogens and regulate the microbiome. Antimicrobial peptides occur in all life, are incredibly diverse, mostly quite small (< 200 amino acids), and typically only comprise a small proportion in a genome (~ 1%). This makes them very difficult to find. One way to discover antimicrobial peptides is by using statistical learning methods, but so far most attempts to do this have focussed on a subset of sequences that mostly include mature peptide sequences. This has limited utility in novel antimicrobial peptide discovery because gene predictions usually only provide a predicted precursor protein sequence within which the much shorter mature peptides is rarely known. We created a classification model (support vector machine with radial kernel) specifically trained for genome-wide scanning. The model was implemented in an R package, ampir. ampir was designed for high throughput and supports parallelisation. ampir was tested on multiple test sets (including complete proteomes) and performed with high precision. ampir can be used to narrow down the search space for novel antimicrobial peptides in genomes.

Abstract #25

Christina Azodi (St. Vincent's Institute of Medical Research, Melbourne, VIC)
Best practices for single-cell expression Quantitative Trait Locus (sc-eQTL) mapping studies from simulated data

Advances in single-cell (sc) technologies have made it possible to study omic variation at the cellular level, ultimately promising to improve our understanding of important phenotypes like disease risk and drug response. For example, with population-scale scRNA-sequencing data we can find cell-type and dynamic-state specific expression Quantitative Trait Loci (eQTL) that explain the cellular contexts in which stimuli or diseases have an effect. However, best practices for sc-eQTL mapping are yet to be established. Existing, pioneering efforts to study sc-eQTL have focused on novel sc-eQTL discovery using empirical data, where the ground truth is unknown. Here we share our best-practice recommendations based on eQTL mapping on simulated population-scale scRNA-seq data with known eQTL associations. To generate these simulations, we extend the functionality of Splatter (Zappia et al., 2017), an existing tool for simulating scRNA-seq data. Empirical data is used extensively to ensure the simulations reflect real data, including genotype data from the 1000 Genomes Project, empirical eQTL results from GTEx, and scRNA-seq data from HipSci. We compared how key preprocessing steps (e.g. normalization and data aggregation strategies) and study design considerations (e.g. batch effects and number of covariates) affect results and found that the use of best practices can improve eQTL mapping sensitivity by up to 13%. Our findings can inform the planning and execution of future eQTL studies. Finally, given the rapid rate of innovation in the generation and analysis of sc omics data, we made our simulated data, simulation framework, and workflow for testing sc-eQTL associations publicly available to facilitate future benchmarking.

Abstract #26

Johanna Wong (ithree Institute, University of Technology Sydney, Sydney, NSW)
Mining the wastewater microbiome with a metagenomic Hi-C approach to identify antimicrobial resistance risk in Australia

The rise of antimicrobial resistance (AMR) poses a severe risk to public health in Australia and globally. Overuse and misuse of antibiotics in humans and agriculture, international travel and water pollution are amongst the key factors contributing to the spread of AMR. The commensal bacteria of humans, animals and the environment can acquire AMR-associated mobile genetic elements, particularly plasmids, by horizontal gene transfer. Shotgun metagenomic sequencing from wastewater bacterial communities could help monitor the prevalence of AMR at a regional level, but shotgun metagenomics is unable to directly inform the linkages between AMR-associated plasmids and their bacterial host. As a consequence, our understanding of the origin and transmission of AMR-plasmids between different bacterial strains remains limited. To tackle this challenge, we are developing a metagenomic surveillance workflow that incorporates both long read and short read metagenomic sequencing with proximity-ligation (Hi-C) sequencing to discover AMR plasmids and associate them to bacterial hosts within wastewater microbial communities. We report on a pilot study using a total of 10 enrichment cultures representing 2 different wastewater influent samples obtained from different sites within New South Wales. Five different growth media, including two supplemented with antibiotics (Meropenem and Ciprofloxacin) were used to enrich for Gram-negative Enterobacteriaceae that are related to common human pathogens and resistant to antibiotics. Three types of sequencing data were generated: 10 Illumina shotgun samples (one per enrichment culture), 2 Oxford Nanopore Technology (ONT) samples (one per sampling site, combining 5 enrichments), and a Hi-C readset (combining all 10 enrichments). De novo long read assembly with MetaFlye produced approximately 4000 contigs per sample with a contig N50 > 90 kb. To resolve individual genomes from the mixture, the short-readsets were mapped back to the assembly contigs, metagenome binning was performed with MetaBAT2 and genome quality was assessed with CheckM. Near complete metagenome-assembled genomes (MAGs) of Proteus mirabilis, Pseudomonas aeruginosa and Comamonas kerstersii were recovered. Additionally, we have detected over 60 different AMR-genes and identified over 250 plasmids, with the majority reportedly associated with Escherichia coli and Klebsiella pneumoniae. Generally, these results suggested that microbes related to common human commensals and pathogens, along with their AMR elements, can be successfully detected with our workflow. Using the Hi-C data we have clustered contigs using bin3C, revealing a strong signal that associates the plasmids and bacteria to each other. Additionally, we observed a high frequency of single nucleotide polymorphisms (SNPs) in a number of contigs, suggesting a potentially high level of strain diversity in the microbial community. Aiming to differentiate closely-related strains within the mixture, we are currently developing a workflow that could deconvolute our metagenomes into individual, strain-specific genomes that may differ in AMR-plasmid association. Challenges of this dataset and our latest findings will be discussed in the presentation.

Abstract #27

Aaron Darling (University of Technology Sydney, Sydney, NSW)
Metagenomics Strain Resolution on Assembly Graphs

We introduce a novel bioinformatics pipeline, STrain Resolution ON assembly Graphs (STRONG), which identifies strains de novo, when multiple metagenome samples from the same community are available. STRONG performs coassembly, followed by binning into metagenome assembled genomes (MAGs), but uniquely it stores the coassembly graph prior to simplification of variants. This enables the subgraphs for individual single-copy core genes (SCGs) in each MAG to be extracted. It can then thread back reads from the samples to compute per sample coverages for the unitigs in these graphs. These graphs and their unitig coverages are then used in a

Bayesian algorithm, BayesPaths, that determines the number of strains present, their sequences or haplotypes on the SCGs and their abundances in each of the samples. Our approach both avoids the ambiguities of read mapping and allows more of the information on co-occurrence of variants in reads to be utilised than if variants were treated independently, whilst at the same time exploiting the correlation of variants across samples that occurs when they are linked in the same strain. We compare STRONG to the current state of the art on synthetic communities and demonstrate that we can recover more strains, more accurately, and with a realistic estimate of uncertainty deriving from the variational Bayesian algorithm employed for the strain resolution. On a real anaerobic digestor time series we obtained strain-resolved SCGs for over 300 MAGs that for abundant community members match those observed from long Nanopore reads.

Abstract #28

Aidan Tay (CSIRO, Canberra, ACT)
Developing a computational safeguard to detect gene drive systems in wild populations

Genome editing technologies such as CRISPR-Cas9, have made it possible to engineer gene drive systems for spreading desirable traits throughout wild populations. These systems rely on the fact that some genetic elements have a higher chance of being inherited, thereby allowing them to ‚Äòdrive‚Äô through a population over many generations. By releasing genetically modified individuals containing gene drive systems and allowing them to breed with wild individuals, desirable traits for managing wild populations such as invasive species or disease vectors, can be propagated. However, the use of gene drive systems for managing wild populations remains hampered by the potential risks associated with releasing genetically modified individuals containing these systems. To help minimise the risks associated with releasing genetically modified individuals into the wild and improve the traceability of gene drive systems, we developed a computational approach for detecting the presence of gene drive systems within a genome. This is done by analysing the characteristic frequency of oligonucleotide sequences (i.e., genomic signature). Different organisms display unique genetic signatures, which can be used to differentiate DNA originating from different species. By analysing the changes in genomic content of sequences at different locations, the native DNA sequence and that of a gene drive can be distinguished. We demonstrate how gene drive systems can be detected in whole genome sequencing data derived from experimental sequencing library for yeast, and a theoretical sequencing library for the Cas9 gene. Importantly, this approach requires no prior knowledge about the genomic sequence and requires no alignment to a reference sequence, meaning it can be readily applied to poorly characterized organisms.

Abstract #29

Bernard Pope (University of Melbourne, Melbourne, VIC)
Best practices for bioinformatics command-line software with Bionitio

The results-driven focus of bioinformatics means that shortcuts are often taken during software development for the sake of making something that works. Furthermore, many bioinformaticians are not trained in software engineering, and research-oriented projects have limited budgets for quality assurance. In response to this problem we have developed Bionitio, a tool that automates the process of starting new bioinformatics software projects following recommended best practices. With a single command, the user can create a new well-structured project in one of twelve programming languages. The resulting software is functional ‚Äî carrying out a prototypical bioinformatics task ‚Äî and thus serves as both a working example and a template for building new tools. Key features include command-line argument parsing, error handling, logging, defined exit status values, a test suite, a version number, standardised building and packaging, documentation, a standard open-source software license, revision control, and containerisation. For example, the following command creates a new Python 3 project called skynet using the BSD 3 Clause license and creates a remote repository on GitHub for username cyberdyne: bionitio-boot.sh -i python -n skynet -c BSD-3-Clause -g cyberdyne Bionitio serves as a learning aid for beginner-to-intermediate bioinformatics programmers and provides an excellent starting point for new projects. This helps developers adopt good programming practices from the beginning of a project and encourages high-quality tools to be developed more rapidly. Bionitio has been used in several workshops, providing a common codebase for coordination of workshop materials and an extensible platform for the delivery of hands-on practical activities. Additionally, by providing complete working examples in many different languages, Bionitio acts as a kind of Rosetta Stone and is therefore an excellent vehicle for comparative programming skills transfer. In this talk we will describe the design and implementation of Bionitio and demonstrate how it can be used to quickly start new open source bioinformatics projects. Project website: https://github.com/bionitio-team/bionitio Publication: Bionitio: demonstrating and facilitating best practices for bioinformatics command-line software, GigaScience, 2019, https://doi.org/10.1093/gigascience/giz109

Abstract #30

Marek Cmero (Peter MacCallum Cancer Centre, Melbourne, VIC)
Using equivalence classes for differential transcript usage and variant detection in RNA-seq data

RNA sequencing (RNA-seq) has enabled high-throughput and fine-grained quantitative analyses of the transcriptome, and has been utilised in distinct contexts such as differential expression and fusion detection. Traditionally, RNA-seq analyses have used alignment to the genome to make downstream inferences on differential expression profiles or to detect transcriptional variants such as fusions.

Genome alignment, however, is computationally expensive, and can also lead to reference bias. Equivalence classes, which reflect the transcripts that a given read is compatible with, present an alignment-free alternative to isoform categorisation and quantification. Typically, equivalence classes are used as an intermediary unit to infer transcript abundance. We utilised equivalence class counts directly to perform differential transcript usage, which can elucidate the role of different transcript isoforms between experimental conditions, cell types or tissues. We find that equivalence class counts have similar sensitivity and false discovery rates as exon-level counts but can be generated in a fraction of the time through the use of pseudo-aligners. Equivalence classes can also be combined with de novo assembly to avoid reference bias, which may obscure variant isoforms. To demonstrate this, we present MINTIE, a catch-all variant finder that detects regular and irregular fusions, transcribed structural variants and splice variants by leveraging de novo assembly, equivalence classes and differential expression. We validated MINTIE on simulated and real data sets and compared it with eight other approaches for finding novel transcriptional variants. We found MINTIE was able to detect all defined variant classes at high rates (>70%) while no other method was able to achieve this. Applying the method to real cancer and rare disease data revealed several novel variants of potential clinical significance. We posit that equivalence classes are an efficient and flexible unit of quantification to perform diverse analyses on, such as differential transcript usage and in detection of novel structural and splice variants.

Abstract #31

Yatish Jain (CSIRO, Canberra, ACT)
sBeacon: cloud-native genomic data exchange

The Beacon protocol was published last year (Fiume M. et. al., Nat. Biotech. 2019), discussing secure and efficient genomic information exchange. However, the implementation described is costly to run, not extensible, and inefficient for large cohorts, jeopardizing the very premise of global data exchange. It excludes a large number of smaller and underprivileged providers from contributing their genomic data and results in less data diversity. At the other end of the spectrum, it also hampers large studies from expanding further. To overcome these issues, we fundamentally re-designed the approach using an economical yet highly scalable cloud-native architecture, resulting in the ‚Äúserverless‚Äù Beacon (sBeacon). sBeacon enables researchers and consortia to light their own beacon within their own secured infrastructure. sBeacon reduces the maintenance costs by 99% and improves dataset update time by a factor of 5000, which brings the investment cost to less than US $100/month and makes genomic data sharing more inclusive. It also allows query time to scale sub linearly with dataset sizes, catering for future population-scale cohorts. sBeacon can consolidate information across different providers, akin to the Beacon Network, but does so at the data level. This means individual providers can keep data in their own infrastructure and control access to individual files via consent codes (Dyke SO et. al., PLoS Genet. 2016), while still contributing to global summary statistics and other analytics. COVID-19 has highlighted the need to substantially improve disease preparedness to resolve this pandemic and avoid future ones. Bringing the advancement of genomic data sharing from the human space into other disciplines, we showcase how the Beacon protocol can be used for pathogen genomics. Our COVID sBeacon tracks the frequency of individual mutations over large, distributed COVID-19 datasets including INDICOV data, Genbank, and National Genomic Data Centre. With this, we can visualize the historic expansion of, for example, the D614G mutation through the different countries in each cohort. COVID sBeacon also demonstrates the extensible architecture of our implementation by enabling real-time calculation of minor allele frequencies and regional mutation rates.

Abstract #32

Caleb Lau (Swinburne University of Technology, Melbourne, VIC)
Fate specification and variability in cell lineage trees

Cell lineage trees provide a blueprint for understanding how cell fate is speci- fied. Sulston et al. (1983) famously showed how the lineage tree for C. elegans pinpoints when each cell type is specified during embryonic development. This success was possible largely because the invariant cell lineage tree for C. elegans allows trees from different founder cells to be inspected visually. For more complex organisms, however, cell specification is significantly more stochastic, making it impossible to use such simple visualisation techniques to interpret the lineage tree. To address this problem, Hicks et al (PLOS Comp Bio 2019) recently developed a new technique, called the lineage variability map, which provides the statistical framework for understanding lineage trees at the population, not just individual, level. This method showed how cell specification should be asso ated with the parts of the tree that are strong sources of variation. The original application of the method to studying the highly variable T-cell (CD8+) lineage tree was, however, limited by the large quantities of incomplete data. Here we show how to solve this problem using a Bayesian analysis that samples from the distribution of missing data as well as from the posterior distribution of the tree parameters. This enables us to extend our analysis of T cells into the regime of significant cell differentiation. .

Abstract #33

Nadia Davidson (Peter MacCallum Cancer Centre, Melbourne, VIC)
JAFFAL: Detecting fusion genes with long read transcriptome sequencing

Genomic rearrangements are common in the cancer landscape and have the potential to create novel oncogenes by fusing parts of two genes together. Massively parallel short read transcriptome sequencing has greatly expanded our knowledge of fusion genes across cancers with ~ 16% of cancers shown to have fusion gene events across a range of tumour types. These events are known to drive cancer and novel drugs have been developed to specifically target a number of these driver fusions. Short read sequencing requires RNA molecules to be reverse transcribed and fragmented. Therefore long range information about the structure of the fusion transcript away from the breakpoint is lost. Long read sequencing technologies, as offered by Oxford Nanopore Technologies (ONT), allow the full length of individual mRNA molecules to be sequenced. This can provide unprecedented opportunities to study splicing, RNA modifications and run rapid and remote diagnostics. However, the generated data has a very high rate of errors and fusion finding algorithms designed for short reads do not work. Here we present JAFFAL, a method to accurately identify fusion genes from long read transcriptomes. Our method is based on the JAFFA pipeline, which can call fusions from reads of any length provided they had a low error rate. To facilitate ONT transcriptomes for fusion finding we utilised the error tolerant aligner minimap2 with realignment of breakpoints to overcome insertion and deletion type errors. We also substantially improved computational efficiency to run approximately 10 million reads per hour using 8 threads. We validated JAFFAL using simulations and ONT data from cancer cell lines sequenced by our collaborators in the Singapore Nanopore-Expression Project (SG-NEx). We show that fusions can be detected in long read data with similar accuracy as short reads, and in addition, their splicing structure can be ascertained. Finally, by comparing ONT transcriptome sequencing protocols we show that numerous chimeric molecules are generated during cDNA library preparation that are absent when RNA is sequenced directly.

Abstract #34

Oak Hatzimanolis (Griffith University, Brisbane, QLD)
Implementing an integrated analysis to identify and validate circular RNAs using patient-derived neuronal stem cells.

In recent years it has been determined that a large proportion of non-coding RNAs have biochemical functionality (~80%), contrary to pre-existing consensus a few decades ago. Circular RNAs (circRNA) are an exciting addition to this class of RNA, and since 2012 a significant amount of research has accumulated around these biomolecules. They are regulatory non-coding RNAs playing a critical role in many cellular pathways with implications in cancers and neurological diseases. The most common function attributed to circRNAs is their ability to act as a microRNA sponge. Advances in the fields of computational biology and molecular biology have allowed for widespread study of this class of RNA, with several specific detection tools that have arisen since the start of the last decade. This study aims to take advantage of these recent advances in transcriptomics and bioinformatics to identify circRNAs in schizophrenia patients and healthy controls from total RNA sequencing data from olfactory neuronal stem (ONS) cells. We will use these patient derived ONS cells as a model for neurological diseases as they have previously been used in several other multi-omics studies describing molecular and cellular differences between schizophrenia individuals and healthy controls. A large amount of circRNA was identified from the total RNA sequencing data and a subsequent differential expression analysis was conducted on these molecules. Over 500 circRNAs were differentially expressed between schizophrenia and controls, with CADM3 being the most downregulated in Schizophrenia and GPR133 being the most upregulated. Potential MicroRNA regulation was also examined, wherein we found potential sponging occurring from microRNAs found with predicted strong binding to the identified circRNAs. Interestingly we also found microRNAs implicated in schizophrenia such as miR-137 and miR-2682 which were found to putatively bind to some of the identified circRNAs. Several circRNAs such as DOCK1, COL4A2, PTPN13, PRKCA, FAT4, LAMA3 and ITGA3 had a tremendous amount of potential microRNA binding sites suggesting they play an important role as decoyer for microRNAs which effectively decrease the functional amount of microRNA in the cells. Future study into circRNAs could prove to be fruitful in discovering links between diseases states and genetic dysregulation, as well has helping to build the framework for observing neurological diseases as the result of the shared dysregulation of multiple pathways such as the interaction networks of circRNA-miRNA-mRNA-protein.

Abstract #35

Feargal Ryan (South Australian Health and Medical Research Institute (SAHMRI), Adelaide, SA)
Intrapartum or Direct Antibiotic Exposure in Early Life Significantly Alters the Infant Microbiota and Whole-blood Transcriptional Responses to Immunisation

Antibody-mediated responses play a critical role in vaccine-mediated immunity, however, for reasons that are poorly understood, these responses are highly variable between individuals and vaccine immunogenicity is frequently impaired in low income countries. Previously, we have demonstrated that antibiotic-driven dysregulation of the gut microbiota in infant mice leads to impaired antibody responses to five vaccines that are routinely administered to infants worldwide. Antibiotics are frequently prescribed to both mother and baby during the pre- and postnatal periods but their impact on immune responses to vaccination in human infants is currently unknown. To investigate this, we recruited a cohort of 226 vaginally born, healthy, term infants that were either unexposed or exposed to direct or intrapartum antibiotics. We used high-resolution shotgun metagenomics to profile the composition and encoded functional capacity of the gut microbiota in these infants at ~7 days of life and again at ~6 weeks of age when they received the majority of their first routine immunisations. Antibiotic exposure, both direct and intrapartum, resulted in a significant reduction in the relative abundance of the beneficial commensal bacterial family Bifidobacteriaceae which has previously been associated with vaccine responses in infants. Antibiotic treatment was also predicted to significantly alter the functional capacity of the gut microbiota, including its ability to produce

multiple amino acids and fatty acids. The largest source of variation in these data is an axis defined by an inverse relationship between Escherichia coli and Bifidobacterium spp. Bloods were also collected from the infants at baseline (median 45 days old, SD 3 days), ~1 week post-immunisation and at 7 and 15 months for RNAseq, serology and multi-parameter flow cytometry. Antibody response data is currently pending all infants completing their 15 month visits. RNA-Seq was used to profile the blood transcriptome of 182 infants at baseline and at ~1-week post-immunisation. Samples were sequenced on an Illumina NovaSeq to an average of 25.3 million 2x100bp reads. These data provide the most detailed assessment of the impact of immunisation in early life on the blood transcriptome to date. We found that post-immunisation there was a significant up-regulation of an array of immune gene expression signatures including interferon signalling, antigen presentation, T and B cell pathways, together with a strong up-regulation of genes involved in transcription and the cell cycle. Interestingly, this signature was altered in antibiotic exposed infants who had decreased T and B cell related gene expression signatures relative to non-exposed infants. Integrating the metagenomics and gene expression data revealed significant associations between metagenome-encoded functions, such as tryptophan biosynthesis, and immunisation induced gene expression signatures. These data suggest that antibiotic-driven changes in the infant gut microbiota alter immune responses to vaccination and may explain why vaccine responses are highly variable between different individuals and populations.

Abstract #36

Yingxin Lin (University of Sydney, Sydney, NSW)
Transfer learning for data integration of single-cell RNA-seq and ATAC-seq

Single-cell transcriptomics profiling with single-cell RNA-seq (scRNA-seq) has provided unprecedented resolutions in charatersing cell identities, cell functions across diverse tissues and conditions. Recent advances in measuring multiple modalities of single cells, such as single-cell ATAC sequencing (scATAC-seq), further enable characterisation of cells from different aspects. While scATAC-seq data provides the epigenomics profiling of cells, its extreme sparsity leads to its lack of the power of cell type identification. Therefore, integration of scRNA-seq and scATAC-seq allows not only cell type label transferring but also a better understanding of the cellular phenotypes. Here, we present a new end-to-end semi-supervised transfer learning algorithm, scJoint, to integrate heterogeneous collections of scRNA-seq and scATAC-seq data. By building an integrative method with neural network based dimension reduction and semi-supervised cell type prediction model, our algorithm is able to transfer labels from scRNA-seq to scATAC-seq data and construct a joint embedding for the two modalities. We illustrate the performance of our algorithm with unpaired scRNA-seq and scATAC data collections, including unpaired large mouse cell atlas data (177,577 cells, 82 cell types) and multimodal data coupled with protein level profiles. Our algorithm outperforms the existing methods by a large margin in both joint visualisation of two modalities and cell type prediction, with accuracy rate improved by 7~14%. Using paired transcriptomic and epigenomic data as ground truth, we have further verified the label transfer performance of our algorithm.

Abstract #37

Cameron Hosking (CSIRO, Canberra, ACT)
Detection of recombination amongst SARS-CoV-2 strains

The emergence and spread of the SARS-CoV-2 virus has been accompanied by a significant increase in the genetic diversity of the virus, with multiple phylogenetic clusters emerging. Contributing to this diversity is the potential for recombination events to occur between distinct viral strains. When two strains coinfect the same host, they may exchange genetic information during replication, creating new strains with characteristics of both parents. While a recombination event between zoonotic coronavirus strains has been postulated as the origin of the current pandemic virus, so far there has been little research into whether recombination is occurring among the human-specific SARS-CoV-2 strains currently circulating. We sought to undertake a thorough analysis of all viral genomes currently available (>100,000 strains) to investigate whether the virus is undergoing active recombination. Examining existing phylogenetic trees, we find that many strains are poorly explained by just one parent. To examine whether these can be better explained be recombination events we developed an algorithm to find possible recombination events. We first used phylogenetic analysis to calculate the average mutation frequency distribution of the virus. Using this we simulated both natural evolution of the virus as well as recombination, varying the frequency and extent of the events. Based on these simulated datasets, we developed and benchmarked an algorithm for detecting recombination between SARS-CoV-2 strains which we then applied to viral sequences downloaded from GISAID. Using this approach, we identified a number of recombination events amongst circulating strains, characterizing their distribution in terms of both time and geography. Based on our analysis, we believe that there is evidence SARS-CoV-2 is undergoing recombination and that this is contributing to the genetic diversity being observed.

Abstract #38

Stephanie Chen (University of New South Wales & Royal Botanic Garden Sydney, Sydney, NSW)
Unsupervised orthologous gene tree enrichment for cost-effective phylogenomic analysis and a test case on waratahs (Telopea spp.)
Whole-genome shotgun sequencing is becoming increasingly common in phylogenetic research due to the falling cost of whole genome sequencing compared to traditional methods which target subsets of genomes. However, there are few existing packages for assembling putatively orthologous loci from evolutionarily diverged samples and making alignments for phylogenetic analysis from

these data. Additionally, short-read Illumina sequencing data are highly accurate but at low coverages, it can be difficult to draw out meaningful phylogenomic inferences, especially for non-model organisms for which there is no reference genome available. We have developed a scalable method of rapidly generating species trees from short-read data without the need for a reference genome. The workflow involves (1) de novo genome assembly with ABySS at a range of k values (2) extracting the most complete BUSCO (Benchmarking Universal Single-Copy Orthologs) genes from each set of assemblies with the BUSCO Compiler and Comparison tool (BUSCOMP) (3) generating gene trees, and (4) constructing a species tree. The workflow has been applied to a whole genome shotgun sequencing waratah (Telopea spp.) dataset of five species, comprising of two samples from each of the seven lineages; there are three lineages of T. speciosissima (New South Wales waratah) ‚Äì coastal, upland, and southern. We have also generated a reference genome for T. speciosissima, and examine the robustness of the workflow by comparison to a reference-based approach. It is anticipated that the workflow will maximise the recovery of informative data from genomic datasets for reproducible phylogenomic studies and be especially useful for non-model organisms.

Abstract #39

Hieu Nim (Australian Regenerative Medicine Institute, Monash University, Melbourne, VIC)
Single-cell and network analyses reveal organ-specific transcriptomic identity of adult fibroblasts

BACKGROUNDS. Organ fibroblasts are essential components of homeostatic and diseased tissues, as they participate in sculpting the extracellular matrix, sensing the microenvironment and communicating with other resident cells. They are also involved in pathological remodeling and fibrosis in response to injury or disease, caused by excessive and disordered deposition of connective tissue, which impairs organ function. Organ fibroblasts display a unique organ-specific identity, suggesting important roles in normal and pathological organ remodeling. METHODS. Liver, heart, lung, kidney, tail, kidney, gonad and ventral skin of adult mice and E16.5 embryos were mechanically and enzymatically digested to obtain single cell suspensions. Interstitial cells isolated from the different tissues were cultured for 5 days in the same conditions and FACS-sorted to enrich for fibroblasts (CD45-CD31-Thy1+). Single-cell and bulk transcriptomic profiles were obtained from independent platforms: microarray, bulk RNA-sequencing and single-cell RNA-sequencing. Data extraction and pre-processing were performed using the EdgeR package. Gene ontology analysis, bioinformatics analyses and visualisation were performed using MeV. Differentially expressed genes showing more than 10-fold change in any given organ were retrieved and an interaction file listing in which organs these genes were enriched was constructed. The interaction file was used as input for Cytoscape in order to reconstruct the network of genes shared by two or more organs, or only specifically enriched in one organ. The network layout was constructed using a spring embedded layout in Cytoscape. Single-cell data were analysed using Seurat v3: raw data were natural-log normalised and scaled using the top-2000 most variables features in the raw data; Principal component analysis (PCA) dimensionality reduction was calculated on 50 principal components; the Uniform Manifold Approximation and Projection (UMAP) dimensional reduction was calculated on 24 dimensions; cluster determination was performed using shared nearest neighbor (SNN) at a 0.5 resolution. Clusters markers genes were identified with the FindAllMarkers function, using the default Wilcoxon Rank Sum test, at a threshold of 0.25 and a minimum difference in the fraction of detection (min.diff.pct) of 0.3. Pairwise comparison was done using the FindMarkers function, with MAST assay and only testing genes that are detected in 25% of cells in either of the two populations (min.pct=0.25). RESULTS. Two independent single-cell datasets were obtained: stromal cell data from the Mouse Cell Atlas and in-house single-cell fibroblast data set. Out of the original Mouse Cell Atlas aggregate containing 21 samples and 4830 cells, 5 populations of interested were identified: Lung, Testis, Kidney, Liver, NeonatalHeart, corresponding to 682 cells. From the in-house single-cell data set, 7 fibroblast populations were isolated: ‚ÄúHeart‚Äù, ‚ÄúLung‚Äù, ‚ÄúKidney‚Äù, ‚ÄúGonad‚Äù, ‚ÄúLiver‚Äù, ‚ÄúSkin‚Äù and ‚ÄúTail‚Äù. Both datasets confirmed that fibroblasts isolated from the adult mouse organs each display positional and organ-specific transcriptome signatures that reflect their embryonic origins. This fibroblast positional code is presumably conserved to maintain and recapitulate adult organ morphology and function, and opens novel opportunities for the treatment of fibrotic diseases in a more precise, organ-specific manner. CONCLUSIONS. Systems analysis coupled with human-driven data exploration can be a powerful tool for understanding the development and diseases associated with fibroblasts.

Abstract #40

James Ferguson (Garvan Institute of Medical Research, Sydney, NSW)
Molecular barcoding of native RNAs using nanopore sequencing and deep learning

Nanopore sequencing enables direct measurement of RNA molecules without conversion to cDNA, thus opening the gates to a new era for RNA biology. However, the lack of molecular barcoding of direct RNA nanopore sequencing data sets severely affects the applicability of this technology to biological samples, where RNA availability is often limited. Here, we provide the first experimental protocol and associated algorithm to barcode and demultiplex direct RNA nanopore sequencing data sets. Specifically, we present a novel and robust approach to accurately classify raw nanopore signal data by transforming current intensities into images or arrays of pixels, followed by classification using a deep learning algorithm. We demonstrate the power of this strategy by developing the first experimental protocol for barcoding and demultiplexing direct RNA sequencing libraries. Our method, DeePlexiCon, can classify 93% of reads with 95.1% accuracy or 60% of reads with 99.9% accuracy. The availability of an efficient and simple multiplexing strategy for native RNA sequencing will improve the cost-effectiveness of this technology, as well as facilitate the analysis of lower-input biological

samples. Overall, our work exemplifies the power, simplicity, and robustness of signal-to-image conversion for nanopore data analysis using deep learning.

Abstract #41

Daniel Cameron (Walter and Eliza Hall Institute of Medical Research, Melbourne, VIC)
VIRUSBreakend: Viral Integration Recognition Using Single Breakends

An important cause of disease is the integration of viruses into the human genome. Recent studies have shown that the position of viral integration is a critical determinant of HIV expression and latency, as well as the oncogenic effects of HPV and HBV infections. Although several tools exist to detect these viral integrations from sequencing data, there are some key limitations preventing the widespread uptake of existing viral integration detection software. Firstly, these tools require a priori knowledge of the virus, or family of viruses to be detected. Secondly, these programs are computationally expensive with runtimes ranging from several hours to several days. Finally, and most critically, these tools rely on the identification of read pair and/or split read alignments that have uniquely mappable alignments to both the host and viral genomes. This prevents the detection of viral integrations in regions of the genome such as the centromeres. Here I present VIRUSBreakend: Viral Integration Recognition Using Single Breakends. VIRUSBreakend is a high-speed viral integration detection tool designed to be incorporated in the whole genome sequence pipelines with minimal additional cost. VIRUSBreakend identifies viral reads, aligns them to the most abundant host-infecting virus, and detects viral integration sites using single breakend variants. Single breakend variants are breakpoints in which only one side can be unambiguously placed. By detecting single breakends in this viral genome, then annotating the single breakend sequences with the potential host integration site(s), viral integrations can be identified in regions of the host genome that are not uniquely mappable. Benchmarking VIRUSBreakend on 13 HCC cell lines shows that VIRUSBreakend is faster than and outperforms existing viral integration detection software. Running VIRUSBreakend on 3,782 samples in the Hartwig Medical Foundation cohort, I show that VIRUSBreakend can uncover novel biological insights into the centromeric viral integration behaviours of HBV and HPV.

Abstract #42

Joel Robertson (Royal Melbourne Institute of Technology (RMIT), Melbourne, VIC)
Revealing interactions between coding and non-coding transcripts in plants using heterogeneous networks

Once thought to be junk genetic material, non-coding RNAs are increasingly recognised as playing an important role in transcription regulation, RNA splicing, chromatin modification and translation control. Dysregulation of non-coding transcripts has also been associated with a number of human diseases. In plants, however, their role is less understood. High-throughput sequencing allows us to examine levels of co-expression between transcripts, and co-expression across a sample-set is likely to indicate involvement in similar cellular processes. A useful technique for analysing this co-expression information is to view it as a network. Network science is the study of complex, non-random systems represented abstractly as a set of nodes along with a set of links that signify an interaction between a node pair. The benefit of framing expression data in this way is that it allows the use of a range of measures that focus on global and local structures in the network to infer relationships between nodes. Community detection methods are commonly used to analyse the global coexpression network topology, grouping transcripts into modules that allow functional annotation of rare or unknown transcripts. Less-utilised information is also observed on a local level, and graphlet counting offers a way to capture this. Graphlets are small-scale subnetworks that can repeat many times throughout a larger network. If a particular graphlet is significantly overrepresented in a network then it is designated as a network motif. Similarly, a network can be characterised by its graphlet profile, allowing comparison of different networks based on this higher-order information. In the biological context, graphlet counting has mainly been deployed on directed networks (e.g. gene regulatory networks or protein-protein interaction networks), where more graphlet types are available to characterise and differentiate networks. This increased granularity can also be obtained in undirected co-expression networks if they are constructed as coding/non-coding heterogeneous networks. Graphlet counting then also provides a method to examine the relationships between mRNA protein-coding transcripts and the less understood families of non-coding RNA. Our research employs graphlet counting techniques to identify significant patterns of interaction between coding and non-coding transcripts in plants. Raw plant sequencing data obtained from _Cicer arietinum L._ (chickpea) samples are assembled into a _de novo_ transcriptome with each transcript type determined via a filtering process to determine coding or non-coding status. After quantification of transcript expression counts, whole-transcriptome heterogeneous co-expression networks are constructed, and a typed graphlet counting algorithm is applied to characterise the network by its higher-order structure. Significant patterns between coding and non-coding transcripts reveal information about regulatory interactions and ultimately identify a set of candidate non-coding transcripts to be investigated experimentally. A comparison of networks across different experimental conditions is also used to indicate which coding/non-coding interactions are particularly important at different stages of the plant's lifecycle. The longer-term goal of the project is to integrate graphlet counting processes with widely-used module detection workflows to facilitate a richer network analysis toolset for biologists.

Abstract #43

Simon Sadedin (Murdoch Childrens Research Institute, Melbourne, VIC)

Exploring Neural Network models for CNV detection from Exome Data

In recent years many methods have been developed to detect copy number variants (CNVs) from exome and targeted sequencing data. These methods typically rely exclusively on read depth to ascertain CNVs as other signals are diluted or missing due to the sparsity of targeted regions in a typical exome capture. Despite some significant success, accurate CNV detection remains challenging, especially for small events that span few exons or target regions of the exome capture. A striking feature of this problem is that humans can often easily recognise CNV false positives through visual inspection of the read depth signal based on the precise shape and local features of the surrounding signal variation. This phenomenon raises the question of whether approaches used in other domains for image classification could be applicable to the problem of CNV detection from exome data. In this presentation, I explore applicability of various neural network models to detect CNVs, including combinations of convolutional neural networks (CNNs), Recurrent Neural Networks (RNNs) and Long-Short-Term-Memory networks (LSTM). Through the use of small scale models and simulated data I will show that there is strong potential for these models to improve on accuracy of existing CNV detection methods.

Abstract #44

Divon Lan (University of Adelaide, Adelaide, SA)
genozip: an advanced universal compressor for genomic data files

genozip is a new universal lossless compressor for genomic data files. It supports all common genomic data formats, such as FASTQ, SAM/BAM, VCF, FASTA. genozip is designed as a drop-in replacement for gzip for data archival and transfer. It compresses 3 to 5 times better than gzip, and significantly faster. genozip provides significant gains in time and cost related to CPU usage and data storage and transfer. genozip also offers some advanced features: data are encrypted with AES-256 to ensure security and privacy (with --password), MD5 is calculated on-the-fly (with --md5), and multiple files can be bound into a single genozip file for efficient delivery and storage. While lossless by default, genozip also offers the --optimize option, which modifies the data in ways that are usually harmless for downstream analysis, but significantly improve the compression. genozip is open source and available on github, conda and DockerHub.

Abstract #45

Sara Ballouz (Garvan Institute of Medical Research, Sydney, NSW)
Sex-specific co-expression: a baseline to explore disease

Biological states such as cell-type, cell-state, tissue, sex, disease, or age, are all encoded in a sample‚Äôs transcriptional profile, as measured by assaying mRNA levels. Transcriptomic data has therefore been used as proxy for functional phenotypes. Yet, no gene acts alone, with expression levels of one gene influencing the expression levels of others. Co-variation of gene expression profiles (i.e., co-expression) is used to extract information on co-transcription, co-regulation and co-functionality. Therefore, we can also expect that sex differences in the structure of gene co-expression networks can generate sexual dimorphism in downstream phenotypes. However, this still remains relatively unknown. Here, we explore and characterize sex- and tissue- specific derived co-expression networks in order to provide a baseline for disease specific applications. We find very subtle differences between the sexes, suggesting that sex differences in expression may be relevant to very specific functions and pathways, necessitating careful exploration.

Abstract #46

Zaka Yuen (Australian National University, Canberra, ACT)
Systematic benchmarking of detection tools for CpG methylation from Nanopore sequencing

Nanopore sequencing can access native DNA at single-molecule resolution and detect base modifications from the Nanopore signal patterns. In recent years, multiple computational tools for detecting methylation using Nanopore sequencing data have been developed. However, the lack of comprehensive benchmarking of tools presents a challenge for users in selecting the right method and assessing the reliability of their predictions. We performed a systematic benchmarking of six tools (Nanopolish, DeepSignal, Tombo, Megalodon, Guppy and DeepMod) for cytosine methylation detection in DNA from Nanopore sequencing using individual reads, controlled mixtures of methylated, and unmethylated reads and whole genome bisulfite sequencing data. Our analyses indicated that although most tools show high correlation with the expected percentage methylation, some present a high proportion of false positives, whereas others present a high proportion of false negatives. All tools showed an overall concordance with the bisulfite data, but some presented significant local variations with respect to the bisulfite data and the other tools. Finally, we proposed multiple strategies to improve the accuracy of methylation calls, including a novel consensus method called METEORE that combines the outputs from two or more tools to achieve higher accuracy at both single-read level and per CpG site. We provide Snakemake pipelines to run all these tools in a standardized workflow for the systematic characterisation of CG methylation from Nanopore sequencing https://github.com/comprna/METEORE.

Abstract #47

Matt Field (James Cook University, Cairns, QLD)
Recurrent miscalling of missense variation from short-read genome sequence data

Short-read resequencing of genomes produces abundant information of the genetic variation of individuals. Due to their numerous nature, these variants are rarely exhaustively validated. Furthermore, low levels of undetected variant miscalling will have a systematic and disproportionate impact on the interpretation of individual genome sequence information, especially should these also be carried through into in reference databases of genomic variation. We find that sequence variation from short-read sequence data is subject to recurrent-yet-intermittent miscalling that occurs in a sequence intrinsic manner and is very sensitive to sequence read length. The miscalls arise from difficulties aligning short reads to redundant genomic regions, where the rate of sequencing error approaches the sequence diversity between redundant regions. We find the resultant miscalled variants to be sensitive to small sequence variations between genomes, and thereby are often intrinsic to an individual, pedigree, strain or human ethnic group. In human exome sequences, we identify 2,Äì300 recurrent false positive variants per individual, almost all of which are present in public databases of human genomic variation. From the exomes of non-reference strains of inbred mice, we identify 3,Äì5000 recurrent false positive variants per mouse ,Äì the number of which increasing with greater distance between an individual mouse strain and the reference C57BL6 mouse genome. We show that recurrently miscalled variants may be reproduced for a given genome from repeated simulation rounds of read resampling, realignment and recalling. As such, it is possible to identify more than two-thirds of false positive variation from only ten rounds of simulation. Identification and removal of recurrent false positive variants from specific individual variant sets will improve overall data quality. Variant miscalls arising are highly sequence intrinsic and are often specific to an individual, pedigree or ethnicity. Further, read length is a strong determinant of whether given false variants will be called for any given genome ,Äì which has profound significance for cohort studies that pool datasets collected and sequenced at different points in time.

Abstract #48

Nicholas Darci-Maher (University of California Los Angeles, California, USA)
Secondary analysis of publicly available omics data across almost 3 million publications

As today,Äôs high throughput sequencing techniques become increasingly affordable and accurate, the number of publicly available omics datasets is rapidly accumulating. Bioinformatics methods provide unprecedented opportunities for analysis of omics datasets in quantitative biological research. Traditionally, such research has included primary analysis of novel omics data developed as part of the study. However, this data has the potential to be reused, and is often valuable beyond the scope of the study that introduced it. Data-driven research by secondary analysis on existing datasets is becoming more important. Increased availability of public omics data represents an opportunity to find novel insights and discoveries across different datasets. This study presents a quantitative analysis of the reusability of omics datasets in two online repositories, the Sequence Read Archive (SRA) and the Gene Expression Omnibus (GEO). We downloaded over 2.5 million publications from the PubMed Central Open Access corpus, and identified those that referenced SRA or GEO datasets. We used these papers to examine reusability based on various factors, including journal, repository, sequencing technology, and species. We find that most datasets are never reused,Äîthese datasets are mentioned once in the study that introduced them, but then never referenced again. In recent years, however, data reuse is rising. We aim to shed light on the landscape of data sharing in the quantitative biology research community, and illuminate the benefits of secondary analysis of omics data.

Abstract #49

Angel Liang (University of New South Wales, Sydney, NSW)
RNA splicing is a hierarchical supernetwork that co-operates to drive osteoblast differentiation

Splicing is essential for the proper expression of mature messenger RNA (mRNA) transcripts in eukaryotes. A growing body of evidence shows alternative splicing of transcripts to be widespread and important for normal cell development and differentiation, where transitions from immature isoforms into their mature counterparts have been demonstrated in literature. Our study sought to understand the functions of alternative splicing in human mesenchymal stem cells (hMSCs), a type of multipotent adult stem cells with great therapeutic promise but severely unrealised due to an incomplete understanding of their stem cell maintenance and differentiation processes. Using Illumina short-read RNA-Seq, we explored the extent to which the landscape of mRNA splicing is changing and regulated as hMSC-TERT4 cells differentiate into osteoblasts, the cells which build bone. By integrating proteome, phosphoproteome and single-cell RNA-Seq data, we discovered that alternative isoforms not only have functional consequences in both the transcriptome and proteome, but can also demarcate homogenous cell sub-populations. Our findings represent a major advance in the knowledge of developmentally regulated splicing, with implications for the selection criteria of the ‚Äúbest‚Äù type of hMSCs to use in cell-based therapies for wound healing and bone regeneration.

Abstract #50

Dominic Maderazo (University of Melbourne, Melbourne, VIC)

Detection and identification of cis-regulatory elements using change-point and classification algorithms

Transcriptional regulation is primarily mediated by the binding of factors to non-coding regions in DNA. Identification of these binding regions enhances understanding of tissue formation and potentially facilitates the development of gene therapies. However, successful identification of binding regions is made difficult by the lack of a universal biological code for their characterisation. We extend an alignment-based method, changept, and identify clusters of biological significance, through ontology and de novo motif analysis. Further, we apply a Bayesian method to estimate and combine binary classifiers on the clusters we identify to produce a better performing composite. The analysis we describe provides a computational method for identification of conserved binding sites.

Abstract #51

Malindrie Dharmaratne (University of Queensland, Brisbane, QLD)
A statistical approach for modelling differential distributions in single-cell transcriptomic data

Single cell RNA sequencing (scRNA-seq) allows the sequencing of the whole transcriptome at the resolution of a single cell, with the ability to unveil complex and rare cell populations. However, scRNA-seq data can be driven by a large number of outliers, over-dispersion and dropouts, posing challenges for identifying genes showing genuine heterogeneity. Although statistical methods developed for the analysis of scRNA-seq data addresses some of these issues, all these methods assume that all the genes in the transcriptome follow a single distribution. We argue that expression profiles of all the genes in the transcriptome cannot be summarised using a common statistical distribution and thus propose a statistical framework for identifying distributional shapes of transcriptomic data. The UMI counts for a given gene are modelled using a generalized linear model with cellular sequencing depth used as an offset to normalize for sequencing depth differences. First, a Kolmogorov-Smirnov (KS) test is performed to select genes that belong to the family of Zero Inflated Negative Binomial distributions (ZINB). Read counts of genes with significant p-values for the KS test are next modelled using the error distributions Poisson, Negative Binomial, Zero Inflated Poisson and ZINB with log link function, independently under each biological condition. The model with the least Bayesian Information Criterion value is selected as the best model. Additional model adequacy tests are conducted using the deviance statistic and the likelihood ratio test to ensure the best distribution is selected for each gene. Furthermore, differential gene expression analysis can also be performed under the same framework, using a likelihood ratio test to compare populations across biological conditions. We demonstrate our framework using mouse ageing gene expression measurements on adipose and muscle tissues, to identify genes that change distributions across biological conditions. Whilst some of the genes and pathways discovered through our framework overlap with those identified through traditional analysis of transcriptomic data, importantly our method is able to identify additional genes and pathways both at tissue level and cell specific level which have been linked to aging. This suggests that modelling the distributional shape of gene expression level independently for each gene, provides an opportunity to extract precise genes and pathways related to the underlying biological condition not necessarily put forth by traditional methods of differential expression. Here, we provide a novel framework for quantifying gene expression heterogeneity by modelling gene expression levels under the differential distribution pipeline. Compared to existing methods, the framework has higher power to detect age-associated genes and pathways, supporting the potential of our approach. This method is also able to model biological conditions that involve change which is either subtle or heterogeneous, such as ageing. Moreover, this framework has the flexibility of adjusting for covariates and for performing multiple comparisons across biological conditions.

Abstract #52

Kerui Peng (University of Southern California, California, USA)
pyTCR: a comprehensive cloud-based platform for TCR-Seq data analysis using interactive notebooks to facilitate reproducibility and rigor of immunogenomics research

T cells are crucial components of the adaptive immune system as they are activated after being exposed to antigens. During the activation, V (variable), D (diversity), J (joining) segments in the T cells receptor loci undergo VDJ recombination to create diverse repertoires for recognizing and binding to the epitopes of the antigens presented by major histocompatibility complex (MHC). With the development of high throughput sequencing, TCR-seq provides the opportunities to understand adaptive immune responses, further helps with diagnosis, prognosis prediction, treatment outcome prediction in a variety of diseases including cancer, autoimmune disease, infectious disease, and allergies. Due to the diversity and complicity of the TCR repertoire, computational methods are needed are important in understanding the features. Existing tools have promoted the advancement in TCR analysis. However, the existing tools fail to provide easy to use interface for biomedical researchers with no or limited background. They don‚Äôt offer integrative analysis as they provide disjoined commands instead. Moreover, the analysis is not comprehensive as other tools are usually needed in order to finish the analysis. Furthermore, existing tools have limited options to customize the analysis and visualization. An alternative solution is urgently needed in this field. pyTCR is a comprehensive platform with a rich set of functionalities of TCR repertoire analysis for biomedical researchers. Our cloud-based easy to use platform is based on the interactive notebook with the enhancement of reproducibility and transparency, by providing comprehensive and integrative functions, and customizable manipulations. The platform that pyTCR utilizes is interactive notebooks which code and results are all available to the users. pyTCR provides basic sample statistics such as number of reads, number of clonotypes, and convergence, clonality analysis, overlap analysis, segment usage

analysis, diversity analysis, motif analysis. In each analysis type, metrics, visualization, and statistical analysis are provided, which offers a comprehensive solution to TCR analysis. The existing gap between traditional biomedical research and bioinformatics provides a substantial barrier for biomedical researchers to utilize computational tools to analyze high throughput data. Our tool will illustrate the capacities of cloud-based notebooks as the solution to bridge the gap, where users with no to limited bioinformatics background or experience would be able to use notebooks to analyze the data with transparent analysis and reproducible results.

Abstract #53

Anna Trigos (Peter MacCallum Cancer Centre, Melbourne, VIC)
The next generation of biomarkers in cancer: single-cell spatial analysis of tumour and microenvironment cells

Spatial technologies that query the location of cells in tissues at single-cell resolution are gaining popularity and are likely to become commonplace. The resulting data includes the X, Y coordinates of millions of cells, cell phenotypes and marker or gene expression levels. However, to date, the tools for the analysis of this data are largely underdeveloped, making us severely underpowered in our ability to extract quantifiable information. We have developed SPIAT (Spatial Image Analysis of Tissues), an R package with a suite of data processing, quality control, visualization, data handling and data analysis tools. SPIAT includes our novel algorithms for the identification of cell clusters, tumour margins, distance to tumour margins, cell gradients, and the calculation of neighbourhood proportions. SPIAT also includes speedy implementations of the calculation of cell distances and detection of cell communities. This version of SPIAT is directly compatible with Opal multiplex immunohistochemistry images analysed through the HALO and InForm analysis software, but its intuitive implementation allows use with a diversity of platforms. We expect SPIAT to become a user-friendly and speedy go-to package for the spatial analysis of cells in tissues.

Abstract #54

Feng Yan (Monash University, Melbourne, VIC)
New Insights Of Cancer DNA Methylation By Studies Of Pre-Leukemic Stem Cells In A Mouse Model Of T-Cell Acute Lymphoblastic Leukemia

The role of DNA methylation in the initiation and clonal evolution of cancer remains poorly understood, in part due to lack of studies of the early pre-malignant state. Recent studies showed that variably methylated regions are associated with multiple cancers, but how it regulates gene expression remains unknown due to sample heterogeneity. To address this, we have analysed three stages of leukemogenesis using a Lmo2 transgenic mouse model of T-cell acute lymphoblastic leukemia (T-ALL). FACS purified pre-leukemic stem cells (Pre-LSCs), LSCs, bulk T-ALL and wild-type controls were profiled with enhanced reduced representation bisulfite sequencing (ERRBS) for DNA methylation and RNA-seq for gene expression. Hierarchical clustering for DNA methylation showed the greatest change occurred between pre-LSCs to LSCs. Hypermethylation predominated in pre-LSCs and LSCs, with hypomethylation predominantly in T-ALL. The genomic location of differentially methylations cytosines (DMCs, compared with earlier stages) in pre-LSCs and LSCs were distinct. In pre-LSCs, DMCs occurred most frequently in CpG Open Seas and ChIP-Atlas analysis showed enrichment for histone marks of active enhancers (H3K27Ac/H3K4me2), the demethylase Kdm1a and the transcription factor RUNX1. In contrast, DMCs in LSCs were more typical of those reported in multiple cancers and ageing, which were most frequently seen at GpG islands and ChIP-Atlas analysis of these sites identified enrichment for the repressive H3K27me3 and activating H3K4me2 modifications. Finally, the transition from LSCs to T-ALL saw new regions of hypomethylated DMCs in CpG open sea. Integrative analysis of ERRBS and RNA-seq data showed that the differentially methylated regions (DMRs) in pre-LSCs were not associated with altered gene expression. However, hypermethylated promoter DMRs of LSCs were correlated with down regulation of 201 genes, including multiple transcription factors, growth factors and signal transduction molecules. Although these genes were lowly expressed in wild-type and pre-LSCs compared to background genes, they were further down regulated in LSCs. Because pre-LSCs are the earliest stage of TALL development, we asked whether there are already changes of methylation at a clonal level. We calculated an epi-polymorphism score, measured by the heterogeneity of 4 adjacent CpG sites (epialleles). The epi-polymorphism increased in pre-LSCs and further in LSCs but decreased in T-ALL. Most differentially heterogenous epialleles (DHE) did not overlap with DMRs at the same stage, but half of all DHEs in pre-LSCs was sites of DMRs in LSCs, suggesting that heterogeneity of DNA methylation is a pre-seeding event. Pathways analysis of LSC-specific DHEs showed enrichment for Wnt signaling and pluripotency pathways, exemplified by Wnt3a and Wnt10a. In conclusion, we have used mouse model of T-ALL to describe the DNA methylation and associated gene expression changes in leukemic stem cells during leukemogenesis. We show for the first time that the well-recognised promoter hypermethylation at bivalently marked sites is preceded by clonal heterogeneity and hypermethylation of Open Seas at active enhancer regions. We propose that these early changes establish a platform for clonal selection of gene expression changes that promote leukemogenesis. Further study of these early changes will provide new insights into the mechanism and role of DNA methylation in cancer development.

Abstract #55

Charlie Higgs (University of Melbourne, Melbourne, VIC)
Optimising genomic approaches for detection of vancomycin resistant Enterococcus faecium transmission in the hospital environment

Objectives: Vancomycin-resistant Enterococcus faecium (VREfm) is a leading cause of nosocomial infections and globally significant public health threat. In an effort to better understand the transmission of VREfm within the hospital environment, whole genome sequencing (WGS) is increasingly being used. While the flexibility and plasticity of the VREfm genome enables its adaptation to the hospital environment, it limits our ability to identify transmission events. Currently multiple genomic analysis approaches are being used worldwide without consensus. This project aims to determine the optimal method for analysing WGS data to identify transmission events, using epidemiological data for comparison. The optimal method would be stable over time as new isolates are added to the analysis, be standardised to allow for comparison across sites and require minimal interpretation. Methods: This study combined WGS data from VREfm samples (n= 305) with comprehensive patient bed movement data, collected during a 15-month prospective study across four Victorian hospital networks (Controlling Superbugs study). To determine the most reliable method for identifying possible transmission events, multiple genetic comparison methods were used, including: core genome single nucleotide polymorphism (SNP) count grouped by multi locus sequence type (MLST), core genome MLST (cgMLST), k-mer based comparison method (SKA) and pairwise mapping following de novo assembly. This was compared with the patients,Äô bed movement data to infer likelihood of transmission based on temporal and spatial overlap. Current literature suggests an isolate pairwise distance threshold of ,â§25 SNPs indicates putative transmission of VREfm. Results: Genomic comparison methods that compare isolates in a pairwise manner (SKA and de novo pairwise comparison) were stable over time and can easily be standardised. When set at the same relatedness thresholds, cgMLST displayed a lower concordance with the epidemiological data compared to the core genome alignments but could be more easily standardised. Methods that maximise the amount of genomic diversity captured (SKA and de novo pairwise comparison) showed greater discriminatory power in identifying transmission clusters that were supported by ward-move data. Nested clusters were able to be separated within groups of isolates that were indistinguishable using core genome alignment methods. We also showed that although a SNP threshold of ,â§25 may be effective for some data sets, it cannot be uniformly applied across all ST backgrounds using core genome alignments. Conclusion: Of the genomic comparison methods tested, those that maximise the genomic diversity captured, such as SKA and the de novo pairwise comparison, showed better concordance with the epidemiological data. These methods could increase the sensitivity of VREfm transmission analyses and more discriminately identify outbreaks.

Abstract #56

Tingting Gong (Garvan Institute of Medical Research, Sydney, NSW)
Structural variation signatures in primary prostate cancer

Prostate cancer has a predominance of large complex genomic rearrangements, known collectively as structural variations (SVs). Through deep whole-genome sequencing analysis (90X tumour/46X normal coverage) of 180 primary prostate cancer samples from African and European patients, including 138 identified as high-risk, we comprehensively studied somatic SVs and identified their signatures in SV types and genomic positions. Using Manta and GRIDSS for high-confidence somatic SV calling, we found a large variability in the number of SVs among samples (ranging from 0 to 754), including 6 hyper-duplicated and 6 hyper-deleted samples. Additionally, we identified loci most frequently targeted by SVs and further correlated the presence of SV hotspots with different SV types and ethnic groups. TMPRSS2 and ERG gene regions were found as SV hotspots and their fusion has previously been identified as common gene fusion. We therefore identified that 33 samples are TMPRSS2-ERG fusion positive, in which around 50% of them involved multiple SV events with different SV types. Adding findings from our previous study on TMPRSS2-ERG gene transcripts (Blackburn et al., 2019), we confirmed that a single genomic fusion event can result in multiple fusion transcript isoforms. Copy number variation data was also used to validate duplications (DUPs) and further provide an estimate of the number of copies for DUPs in hyper-duplicated samples. Oxford Nanopore long-read Sequencing data was used to validate the presence of SV and precision of SV breakpoints in one of the hyper-duplicated sample. This study provides an invaluable resource for discovering SV signatures and insights into the different mechanisms underlying SV types in primary prostate cancer.

Abstract #57

Frederick Jaya (University of Technology Sydney, Sydney, NSW)
Evaluation of recombination detection methods for viral sequence analysis

To accurately infer the evolutionary history of viral genomes, the process of recombination needs to be accounted for and addressed appropriately. A vast choice of recombination detection methods have been developed over the past 20 years, but their ability to address the needs presented by high-throughput sequencing of viral data is unclear. Here, we present the key considerations for selecting a suitable method for viral analyses. We assess five published methods used to detect recombination in nucleotide sequences - PhiPack (Profile), 3SEQ, GENECONV, UCHIME and gmos. The performance of methods were evaluated with analysis of within-host hepatitis C virus populations, simulated across a wide range of mutation and recombination rates. Scalability was assessed by recording the CPU time required to analyse datasets with n = 500, n = 1000 and n = 5000 sequences per alignment (1680 nt). In addition, empirical datasets of two bovine RNA viruses were analysed by each method and compared with simulation findings. We find critical trade-offs between the methods, where the most scalable methods (PhiPack (Profile), UCHIME and gmos) may not be suitable for analysis of high coverage, within-host sequencing. Analysis of highly similar sequences (mean pairwise diversity < 1%) produced a high rate of positive detections in PhiPack (Profile), whereas 3SEQ and GENECONV are unable to process these. Overall, the five

evaluated methods are inadequate for a rapid and reliable analysis of recombination in large viral datasets, presenting a severe unmet need for the development of scalable and accurate viral recombination detection methods.

Abstract #58

Irene Gallego Romero (University of Melbourne, Melbourne, VIC)
Characterising Diversity in Gene Regulation Across the Indonesian Archipelago

Title: Genetic Drivers Of Gene Expression And DNA Methylation Across The Indonesian Archipelago Lack of diversity in human genomics limits our understanding of the genetic underpinnings of complex traits, hinders precision medicine, and contributes to health disparities. Island Southeast Asia ‚Äì a region that includes Indonesia, the Philippines, Papua New Guinea and other smaller island states ‚Äì accounts for nearly 7% of the world‚Äôs population and remains dramatically underrepresented in modern human genetics studies. Genetically, the region is home to a wealth of genetic diversity not present anywhere else in the world, including signals of Denisovan introgression that account for up to 5% of the genome of present-day Papuans and Indigenous Australians. I will discuss ongoing research, in partnership with local researchers, to address two fundamental questions about the peopling and settling of the region: characterising the legacy of DNA from archaic Denisovans in present-day Papuans, and contextualising Indonesia ‚Äì the world's 4th largest country by population ‚Äì in the global human genetics landscape. By examining the function of archaic hominin alleles within 72 genomes from individuals of Papuan genetic ancestry we find that introgressed SNPs are often located within cis-regulatory elements, suggesting that they are actively involved in a wide range of cellular regulatory processes. Our analyses identify 39,269 high-confidence Denisovan variants that have the potential to alter the affinity of multiple transcription factors to their cognate DNA motifs, and point towards a consistent signal across Denisovan variants of strong involvement in immune-related processes. In parallel, we have generated whole genome sequencing, CpG methylation, and gene expression data in over 100 Indonesian individuals. We identify substantial differences in both methylation and expression, some of which are directly associated with varying ancestry proportions. Genes identified in these analyses are enriched in pathways involved in immunity, hinting at possible adaptation to the local environment. We identify nearly 1,900 expression QTLs and over 48,000 methylation QTLs, many of which are not shared between Indonesian and European populations, and contribute to hematological traits. Altogether, our research highlights the importance of diverse sampling and inclusion in human genetics if it is to deliver benefits to all.

Abstract #59

Yashpal Ramakrishnaiah (Monash University, Melbourne, VIC)
linc2function: Predicting function of lncRNA transcripts using an Artificial Neural Network (ANN) Model

Long Noncoding RNAs (LncRNA) modulate cellular processes by interacting with various biomolecules and processes affecting gene regulation. Recent studies have shown that lncRNAs play a major regulatory role in a number of genetic conditions including complex diseases. Thus, the study of lncRNA is gaining momentum and efforts are underway to create comprehensive annotation of these transcripts using experimental and in silico techniques. The later approaches in general employed machine learning techniques and escalated the efforts of annotation of lncRNA to a large extent. A number of lncRNA reference data repositories are currently available, some sourced reliably while others are crowdsourced, however, there is not much overlap amongst them (Ramakrishnaiah et al. 2020). Further, other than a few well-studied species, there is very little to no such annotations available for the rest of sequenced genomes. Therefore, there is a need for new methods that can be used to identify and annotate lncRNA reliably and in a species agnostic manner. We have built a model that predicts if a given sequence is a lncRNA, and also identifies its functional motifs using Artificial Neural Networks (ANN). Data used in developing the model consisted of consensus transcripts from major data repositories curated by us (Ramakrishnaiah et al. 2020). Sequence, secondary structure, and interactome related features were extracted from it, and the ones with the highest discriminative power are chosen by using the recursive feature elimination (RFE) technique. We achieved accuracies over 99.5% consistently, both on training and test datasets. linc2function can identify lncRNAs with a high degree of accuracy in a species agnostic manner. In addition to accurately annotating the transcripts from existing data repositories, this method would also assist in the efforts to extend the annotations for non-model species. Source code will be made available via GitLab https://gitlab.com/tyagilab/linc2function under MIT license. References: Ramakrishnaiah, Y.; Kuhlmann, L.; Tyagi, S. Computational Approaches to Functionally Annotate Long Noncoding RNA (lncRNA). Preprints 2020, 2020060116 (doi: 10.20944/preprints202006.0116.v1).

Abstract #60

Ivan de la Rubia (Pompeu Fabra University, Barcelona, Spain)
Reference-free reconstruction and quantification of transcriptomes from Nanopore long-read sequencing

Single-molecule long-read sequencing with Nanopore provides an unprecedented opportunity to measure transcriptomes from any sample. However, current analysis methods rely on the comparison with a reference genome or transcriptome, or the use of multiple sequencing technologies, thereby precluding cost-effective studies in species with no genome assembly available, in individuals underrepresented in the existing reference, and for the discovery of disease-specific transcripts not readily identifiable from a reference

genome. Methods for DNA assembly cannot be directly transferred to transcriptomes since their consensus sequences lack the required interpretability for genes with multiple transcript isoforms. To address these challenges, we have developed RATTLE (https://github.com/comprna/RATTLE), the first tool to perform reference-free reconstruction and quantification of transcripts from Nanopore long reads. Using simulated data, isoform spike-ins, and sequencing data from tissues and cell lines, we demonstrate that RATTLE accurately determines transcript sequence and abundance, is comparable to reference-based methods, and shows saturation in the number of predicted transcripts with increasing number of input reads.

Abstract #61

Yunwei Zhang (University of Sydney, Sydney, NSW)
Risk prediction survival model utilising both omics data and clinical data

Utilisation of omics data such as gene expression data for disease diagnosis, e.g. cancer stage prediction has been well established in recent years. However, across the literature, the survival information has been mostly put aside. It is common to define the study-specific patients,Äô outcome by truncating the raw time information at some pre-defined time point. Also, patients,Äô clinical features are not included in most of the studies. Therefore, we aim to utilise both the survival time information and clinical information to build risk prediction models as well as identifying potential significant clinical variables. We use a Melanoma data set in our study. Survival times are included via establishing survival models: random survival forest (RSF) and penalised Cox model are used in our study, instead of binary classification models. Pre-validated methods are used to pre-validate the omics data information to be combined with the clinical information. We compare the survival model performances and the clinical variables identified corresponding to each model. Our preliminary results show that with a considerable concordance index which gives the goodness-of-fit of the survival models, some clinical variables are significant in predicting patient,Äôs outcome. As a comparison between RSF and penalised cox model, RSF slightly performs better. In conclusion, clinical and survival time information should be used simultaneously and additionally with omics data to enhance patients,Äô risk prediction.

Abstract #62

Eduardo Eyras (Australian National University, Canberra, ACT)
ISOTOPE: ISOform-guided prediction of epiTOPEs in cancer

Immunotherapies provide effective treatments for previously untreatable tumors, but the molecular determinants of response remain to be elucidated. Here, we describe a pipeline, ISOTOPE (ISOformguided prediction of epiTOPEs In Cancer), for the comprehensive identification of cancer-specific splicing-derived epitopes. Using RNA sequencing and mass spectrometry for MHC-I associated proteins, ISOTOPE identified neoepitopes from cancer-specific splicing event types that are potentially presented by MHC-I complexes. We found that, in general, cancer-specific splicing alterations led more frequently to the depletion of potential self-antigens compared to the generation of neoepitopes. The potential loss of native epitopes was validated using MHC-I associated mass spectrometry from normal cells. Furthermore, analysis of two cohorts of melanoma patients with ISOTOPE identified that splicing-derived neoepitopes with higher MHC-I binding affinity associate with positive response to immune checkpoint blockade therapy. Additionally, we found a more frequent depletion of native epitopes in non-responders, suggesting a new mechanism of immune escape. Our analyses highlight the diversity of the immunogenic impacts of cancer-specific splicing alterations and the importance of studying splicing alterations to fully characterize the determinants of response to immunotherapies. ISOTOPE is available at https://github.com/comprna/ISOTOPE

Abstract #63

Emma Gail (Biomedicine Discovery Institute, Monash University, Melbourne, VIC)
A predictive model for commonly-repressed polycomb-target genes dissects DNA sequence from gene expression

BACKGROUND: Histone H3 lysine 27 trimethylation (H3K27me3) is a hallmark of facultative heterochromatin, deposited exclusively by the polycomb repressive complex 2 (PRC2) and marks genes for repression. PRC2 has an important role in normal development and is dysregulated in cancer and congenital disorders. Previous works indicate that the recruitment of PRC2 to its target genes is dictated by various molecular cues, including chromatin marks, transcriptional state, coding and non-coding RNAs and the interplay with repressive transcription factors, insulators and other chromatin modifiers. DNA sequence composition determines chromatin occupancy of PRC2 in human cells, with low complexity RNA and DNA sequence elements were identified as determinants, including CpG islands, DNA shape and G-tracts. Yet, it is unknown to what extent DNA sequence composition cooperates with transcription to dictate H3K27me3 deposition in a lineage-specific manner. METHOD: We used a Random Forest algorithm and predicted with high accuracy genes that are likely to be marked with the H3K27me3 repressive mark, deposited by PRC2. This model specifically examines the relationship between low complexity DNA sequence signatures, transcription level and the H3K27me3 mark in humans, spanning 24 different cell types and tissues. RESULTS: In agreement with previous studies, gene expression level is the best single predictor for the presence or absence of the H3K27me3 mark on genes. Yet, the models with the best predictability obtained by combining 4 to 6 different features of low complexity DNA sequence elements. Adding gene expression level to sequence-driven models did not improve the predictability of genes marked with H3K27me3 and in some cases even reduced it. Hence, gene expression plays two contrasting roles in predicting

H3K27me3 deposition: good predictability by itself but poor predictability when combined with low complexity DNA sequence elements. CONCLUSION: The results indicate that gene expression data is dispensable for the accurate prediction of polycomb-target genes that are repressed in a wide range of cell types. Our analysis implies that H3K27me3 deposition at commonly-repressed genes is largely driven by low complexity DNA sequence elements while lineage-specific repressed genes are likely selected by other determinants.

Abstract #64

Amy Longmuir (Deakin University, Geelong, VIC)
A high-quality reference genome for the plant pathogen, Phytophthora cinnamomi

The plant pathogen, Phytophthora cinnamomi results in noteworthy crop losses globally, including the avocado and macadamia industries in Australia, while also posing a significant threat to the biodiversity of Australia‚Äôs native flora. To date, attempts to sequence and assemble the genome of P. cinnamomi using first and second generation sequencing techniques have resulted in hugely fragmented genomes, with large gaps of missing information. Specifically, significant sections of the genome containing abundant repeat regions and virulence genes are not adequately resolved, limiting the utility of such genomic resources. Utilising recent advances in long-read sequencing by Oxford Nanopore Technologies we sequenced ~5.5 gigabases of an Australian P. cinnamomi isolate (~50x coverage). By optimising the assembly process to include error-correction and repeat graphs we have produced the first genome assembly under 1,000 contigs for any Phytophthora sp.; a 85 Mb assembly in 597 contigs with an N50 of 322,750. The presented genome is a vast improvement on the current available reference genome which consists of 9,537 contigs in 1,314 scaffolds, but importantly enables us to understand the importance of large scale structural variations in the genome. Here, we present the results of our assembly and annotation work, and highlight how this improved contiguous assembly enables us to investigate and answer fundamental evolutionary questions regards Phytophthora genomics and pathogenicity. I will then go on to detail our future work (enabled by this assembly) where we use whole-genome sequencing to study the mechanisms and evolution of P. cinnamomi infection in important native flora.

Abstract #65

Ingrid Tarr (Victor Chang Cardiac Research Institute, Sydney, NSW)
Genomic analysis of sporadic spontaneous coronary artery dissection

Spontaneous coronary artery dissection (SCAD) is a major cause of heart attack in women and may cause up to 4% of all acute coronary syndromes. However, until recently SCAD was considered to be very rare. SCAD occurs when bleeding within the layers of the coronary artery wall results in separation of these layers. This leads to complete or partial obstruction of the artery, which can lead to myocardial infarction or death. Recent efforts to understand the mechanisms underlying SCAD have identified clinical risk factors for disease, such as migraines, pregnancy, physical or emotional stress, and connective tissue or vascular disorders. Notably absent for SCAD patients are many of the traditional cardiovascular disease risk factors. However, while a genetic contribution to disease is accepted and multi-generational SCAD pedigrees are known, the genetics underlying SCAD remain poorly understood. We investigated common and rare variation in a whole genome sequencing cohort of 91 sporadic SCAD patients, with the Medical Genome Reference Bank (n = 1127) used as a control cohort. Rare variation was assessed a) in 90 genes associated with connective tissue disorders or vasculopathies; b) in three genes previously reported in SCAD cohorts; c) genome-wide, via novel loss of function variants, to identify new SCAD-associated genes; and d) in the potential enrichment of genes or sets of genes for carriers of rare, potentially pathogenic variants. The role of common variation was investigated with the creation and application of a SCAD genomic risk score. Rare splice altering and pathogenic coding variants in connective tissue or vascular disorder genes were identified in 16.5% of our cohort, despite zero diagnosed disorders in these patients. Novel loss of function variants were identified in 8 genes intolerant to this class of variant and previously not associated with SCAD. No rare, pathogenic variants were identified in cases in TSR1, TLN1, and F11R, the three genes previously reported. One truncating variant was found in each of TLN1 and TSR1 in controls, however. No single gene was found to have a greater proportion of cases carrying rare and potentially pathogenic variants than controls, yet significantly more cases had variants in our connective tissue disorders and vasculopathy gene list than did controls. Finally, our SCAD cohort showed significantly higher genomic risk scores than either our controls or a (non-vascular) cardiomyopathy control group. There was no relationship between genomic risk score and carrying rare, pathogenic variant in connective tissue or vascular disorder genes. We have provided further evidence to support a link between connective tissue and vascular disorders and SCAD, as well as identifying a set of genes worthy of further investigation. We were unable to confirm genes previously reported to be associated with SCAD and found evidence for a common genetic contribution to disease. Taken together, this suggests that the genetic contribution to SCAD pathophysiology is likely complex and highly heterogeneous.

Abstract #66

Mikhail Gudkov (Victor Chang Cardiac Research Institute, Sydney, NSW)
Quantifying negative selection on synonymous variants

Relatively little is still known about the full complexity of disease-causing genetic variants. Most studies focus primarily on potential loss-of-function (LoF) variants, such as stop-gain and frameshift variants, at the expense of other classes of mutations. In particular, synonymous genetic variants, that is, those single-nucleotide variants (SNVs) that do not alter the produced amino acid sequence, are routinely considered to be non-deleterious. However, the role of these so-called 'silent mutations' could be more important than was previously thought. For instance, synonymous variants may create nonoptimal codons, thus affecting the stability of the produced mRNA and the overall translational efficiency, which may have implications in terms of Mendelian disease. Therefore, it is unsurprising that synonymous codons have even been dubbed 'a secondary genetic code' for their supposedly more subtle yet powerful mechanism of preserving genetic information. It has also been shown that synonymous SNVs reducing codon optimality undergo purifying selection, the extent of which, nonetheless, remains unknown. The latter presents a significant limitation for variant prioritisation and, consequently, for finding the true causes of genetic disorders. Indeed, SNVs with unknown, unquantified deleteriousness are generally more likely to be overlooked or even excluded from further analysis. To fully understand the potential role of synonymous variants in human disease, a quantitative framework is needed to assess their deleteriousness in comparison with other, better studied classes of mutations. Here we quantify the intensity of the negative selection acting on each possible synonymous amino acid change. We applied MAPS, a recently developed metric of deleteriousness, to a QC compliant subset of synonymous variants affecting codon optimality from 125,748 gnomAD exome sequences (release 2.1.1). MAPS, or the Mutability-Adjusted Proportion of Singletons, which was developed as a 'corrected' version of the originally proposed Proportion Singleton metric, known to be biased towards non-CpG variants, shows the intensity of natural selection and is highly correlated with CADD and PolyPhen predictions. MAPS has already proved to be extremely useful for quantitatively predicting the effect of missense and LoF variants, and in this work we demonstrate how its principles can be applied to differentiate between truly benign synonymous SNVs and harmful ones. The results of our work contribute to further understanding of the potential role of synonymous variants in Mendelian disease.

Abstract #67

Yidi Deng (Melbourne Integrative Genomics, University of Melbourne, Melbourne, VIC)
Benchmarking single cell transcriptomes with bulk transcriptional atlases.

Single cell RNA sequencing (scRNA-seq) allows for the study of cell-specific variation and cell population heterogeneity at an unprecedent resolution. One statistical challenge when analyzing scRNA-seq data is to comprehensively characterize each cell‚Äôs molecular identity defined by their cell types and cell states. To address this challenge, building a reliable referencing cell atlas of human tissues on which results from different studies can be projected and benchmarked is gaining popularity from the global research community, as exemplified by the Human Cell Atlas consortium. Well curated and annotated single cell atlases, nonetheless, are still lacking. Here, we propose to leverage on our previously built bulk transcriptional atlases (Angel et al., 2020) to gain insights into single cell biology The major challenge we face is the difference in data distribution and scale between the bulk and the scRNA-seq data which can greatly obstruct our ability to capture biologically relevant variation in the single cells when we consider the bulk atlas as reference. Therefore, we have developed a new computational framework for projecting query scRNA-seq data onto a reference bulk atlas, whilst preventing batch differences between technologies that affect the projection results. Our approach is based on either aggregating the cells to create pseudo-bulk samples (aggregation) or imputing zeros to allow for library normalization (imputation). These techniques aim to transform scRNA-seq data so that their distribution resembles bulk. In particular, we have developed a modified version of MAGIC (van Dijk et al., 2018), a popular data smoothing based single cell imputation method, by changing its affinity calculation to construct graph of cells which emphasize on each cell‚Äôs local connectivity. Furthermore, we propose to use Random Forest to profile continuum identities of query single cells and quantify the uncertainty of each particular cell mapping based on their locations in the lower-dimensional expressional space of the atlas. We show on four scRNA-seq datasets projected on two bulk transcriptional atlases that our approach projects cells into the correct biological niches of bulk atlas, showing high agreement with the biology described in the query study. By comparing the projection results of scRNA-seq data imputed by different methods, we further demonstrate the value of a transcriptional atlas for computational method evaluation. The projection result show that our imputation method leads to improved performance compared to MAGIC when the query dataset is highly imbalanced in cell population composition. Besides enabling accurate data projection, our framework can also help unlock other bulk based statistical toolkits (i.e. models for DE analysis. cell type predictors) for analyzing scRNA-seq data.

Abstract #68

Steven Morgan (St. Vincent's Institute of Medical Research, Melbourne, VIC)
Unravelling the genetic mystery of lipoedema with whole exome sequencing

Lipoedema is a chronic, progressive adipose deposition disorder, which becomes exacerbated during periods of hormonal change. It has been suggested that there is an inherited component to the condition. The distribution of abnormal lipoedema adipose tissue, joint hypermobility and commonly, lymphedema in the lower limbs, leads to impaired mobility, decreased quality of life, and secondary health consequences. Despite the severity of this disease, the molecular and genetic basis of lipoedema remains poorly understood. We sought to conduct whole exome sequencing and tissue analysis to determine what genetic changes may contribute to the phenotype and pathophysiology of lipoedema. We used a custom pipeline to call variants in the exomes of lipoedema patients and their unaffected family members. We identified several genetic variants in important pathways related to the integrity of the extracellular matrix (ECM).

To validate this finding, transcriptomic and proteomic data were obtained from tissue and adipose-derived stem cells isolated from lipoedema patients. This revealed differentially expressed proteins and transcripts in ECM related pathways. Histological analysis of lipoedema tissue confirmed differences in ECM components. Our study highlights the importance of the ECM in lipoedema and explain some of the clinical findings. It is hoped that an understanding of the genetics of lipoedema may allow more accurate diagnoses and provide a basis for further therapeutic and diagnostic approaches.

Abstract #69

Beth Signal (University of Technology Sydney, Sydney, NSW)
Quick determination of RNA-Seq strandedness with how_are_we_stranded_here

RNA-Seq is a commonly performed method to analyse RNA transcripts on a global scale - usually either for differential expression or transcript assembly. Library preparation and sequencing of samples can give either single-end or paired-end, and stranded or unstranded reads. While paired and single end data can be inferred by the number of files produced for each sample, inferring strand-specificity generally requires mapping of RNA-Seq reads, and there is currently no way to quickly check specificity before running your analysis pipeline. Strand-specificity impacts differential expression, mapping, and assembly - and using incorrect parameters in downstream tools results in lower accuracy. Moreover, we found that nearly half of publications with RNA-Seq data did not report strand-specificity of their data - either explicitly, or by reporting library preparation methods. In addition, the vast majority (94%) did not report the strandedness parameters of downstream software. To address these issues, we developed a Python package that can quickly infer strand-specificity of raw fastq files. how_are_we_stranded_here uses kallisto psuedoalignment, which takes less than 30 seconds per sample, and then counts the proportion of reads that are explained by a stranded (RF/FR) or unstranded layout. When testing on published data, we found a number of publications reported the incorrect layout, again highlighting the need for this tool as a part of quality control. We found that use of how_are_we_stranded_here can also point towards potential sample contamination with other nucleic acids, such as genomic DNA and small RNA. We present this software as an essential quality control check performed on RNA-Seq samples prior to analysis which can then inform the correct processing of samples.

Abstract #70

Harman Singh (Indian Institute of Technology Delhi, New Dehli, India)
eMST, a scalable and interpretable method for Phylogenetic analysis of hundreds and thousands of SARS-CoV-2 genomes

A novel coronavirus, known as SARS-CoV-2, was identified as the cause of an outbreak of pneumonia in Wuhan, China, in December 2019. Travel-associated cases of coronavirus disease 2019 (COVID-19) were reported outside of China as early as January 13, 2020, and the virus has subsequently spread to nearly all nations. Sequencing and phylogenetic analysis of viral genomes is essential for tracking the transmission of SARS-CoV-2. Previous attempts to provide a phylogenetic analysis for studying the transmission of SARS-CoV-2 in the United States, are not scalable to large datasets, provide limited information about the network connectivity and lack user friendly visualization.We present a new network analysis method called eMST(epsilon Minimum Spanning Tree). This method can be used to create a graph, with genetic samples as nodes, connected by edges with weights corresponding to the hamming distance between the nodes. Given a value of epsilon (e), the eMST is then constructed by considering the union of all possible MST,Äôs with one edge of weight w replaced by another edge of weight less than w(1+e). The output of the eMST is in the form of an edge list, which is visualized using Gephi. We validate the results derived from our phylogenetic analysis using eMST, with the results obtained from NextStrain on the data from a previous study (Fauver, Joseph R., et al.Cell (2020)). We then extend our analysis to a larger number of strains with emphasis on specific states, namely California, New York and Washington. For each of these cases, we observe that Nextstrain and eMST results are in agreement with each other, and eMST provides a better visualization and a negligible running time. We finally plan to scale up our analysis to large genomic datasets of size more than 80k, to create a global network which proves the scalability of our approach. Phylogenetic clustering of SARS-CoV-2 genomes is an important first step in studying the coast to coast spread of SARS-CoV-2 during the early epidemic in the United States. eMST will be of broad interest to all scientists engaged in such research as this method improves user visualization, provides detailed information about network connectivity and is scalable to large datasets, thus allowing scientists to draw inferences from the spread of SARS-CoV-2 in the United States.

Abstract #71

Tarun Bonu (Monash University, Melbourne, VIC)
A Deep Learning Approach to Recover Combination of Biologically Significant Motifs

DNA or RNA motifs are short (5-20 bp) recurring patterns that represent binding sites for regulatory proteins such as Transcription Factors (TF) or RNA binding proteins (RBP). Searching for these small patterns in large genomic data (up to billions bp) is very challenging. Further, these motifs may work in collaboration with one or more other motifs, and it is a computationally expensive task to find various permutations with a biological function. We built an Artificial Neural Network (ANN) model to predict the co-occurrence of motifs from given instances of a TF binding motif. One training data consisted of labelled co-occurring and non-co-occurring motif pairs bound by TF in the gene promoter region. Then we grouped closely located motif instances into clusters, also known as co-regulatory

motifs (CRM). Multiple features based on DNA shape, DNA composition, and protein-protein interaction (PPI) were calculated to study the biophysical characteristics of co-occurring motifs. Extensive feature engineering was performed to rank and select features and to train the ANN model. The motifs involved in binding, DNA shapes such as OC2, EP, Opening and Stagger, and DNA binding domains and PPI score of TFs occupying these motifs were found to be the most distinguishing features. We have applied this model to locate collaborative TF binding sites in the data generated through ChIP-seq experiments with 85% accuracy. We curated the UniBind database containing ChIP-seq data from over 6300 samples and 213 TFs. We predicted 4461 potential motif pairs arranged in 433 CRMs. These predicted CRMs are output in BED format and can be viewed in browsers such as IGV. Future work would involve the extension of the same workflow to RBP sites on RNA molecules.

Abstract #73

Sarah Williams (Monash University, Melbourne, VIC)
Exploring mechanisms of nephron maturation using scRNAseq

Three paired kidney structures arise sequentially during mammalian development: the pronephros, the mesonephros and the metanephros. Only the metanephric kidney is maintained into adulthood but each kidney structure contains immature nephrons - precursors of the filtration units that underwrite kidney function. Stem cell models of the human brain recapitulate the timing and order of developmental processes, taking months to reach maximal size and maturity. In contrast, stem cell-derived kidney organoids form immature nephrons in 10-14 days, roughly half the time expected for metanephric nephrons. The timing and maturity of nephron formation in kidney organoids raises the possibility that current protocols may generate tissue closer to pronephric or mesonephric kidneys. However, it is unclear whether there are molecular differences between early nephrons in each context, and why only metanephric nephrons reach functional maturity. In this work we compare single cell RNAseq data from the developing mouse mesonephric and metanephric kidney. Published and new datasets were combined with the Seurat integration method. Cluster labels were assigned via label transfer and identification of known markers. Differential expression was calculated between datasets and within individual cell types via a pseudobulk approach with limma. Initial results have recapitulated known expression patterns in these tissues, lending confidence to analyses of novel marker genes and cell communication driving nephron maturation. This work will facilitate further comparative analysis of embryonic kidney structures and kidney organoids. It may also lead to new approaches to drive nephron maturation, which are required to model disease states that manifest in adulthood

Abstract #74

Dhrithi Deshpande (University of Southern California Los Angeles, California, USA)
A comprehensive analysis of code and data availability in biomedical research

In biomedical research, it is not only imperative to publish a detailed description of the study design, methodology, results and interpretation, but, there is a pressing need to make all the biomedical data and code used for scientific analyses sharable, well documented and reproducible. Analytical code and data availability is consequential for ensuring scientific transparency and reproducibility. However, raw data is not sufficient to make scientific analyses reproducible. We have reviewed the code and data availability in 11 different prominent biomedical journals published between 2016-2020 and our current results indicate that while the majority of articles comply with the data sharing policies of journals, most of them are not accompanied with code. 98.5% of the research papers have data availability whereas only a meagre 40.9% of the research papers have code available. A majority, 59.1% of the studies do not share their code. Code sharing can warrant for reproducibility of the scientific analyses and transparency. For those research papers which do share code, we further intend to verify whether the code is usable and reproducible. We also plan to extend our survey to corroborate if every figure in the article is backed up by code and attempt to run the code to evaluate its reproducibility and the language used for data analysis. We hope our results will abet the researchers and journals in adoption of best practices to ensure scientific transparency and reproducibility.

Abstract #75

Jieun Kim (University of Sydney, Sydney, NSW)
PhosR enables processing and functional analysis of phosphoproteomic data

Mass spectrometry (MS)-based phosphoproteomics has revolutionised our ability to profile phosphorylation-based signalling in cells and tissues on a global scale. To infer the action of kinases and signalling pathways in phosphoproteomic experiments, we present PhosR, a set of tools and methodologies implemented in a suite of R packages facilitating comprehensive analysis of phosphoproteomic data. By applying PhosR to both published and new phosphoproteomic datasets, we demonstrate capabilities in data imputation and normalisation using a novel set of ‚Äòstably phosphorylated sites‚Äô, and in functional analysis for inferring active kinases and signalling pathways. In particular, we introduce a ‚Äòsignalome‚Äô construction method for identifying a collection of signalling modules to summarise and visualise the interaction of kinases and their collective actions on signal transduction. Together, our data and findings demonstrate the utility of PhosR in processing and generating novel biological knowledge from MS-based phosphoproteomic data.

Abstract #76

Jacob Bradford (Queensland University of Technology, Brisbane, QLD)
CRISPR, faster, better - The Crackling method for whole-genome target detection

CRISPR-Cas9 systems have become a leading tool for gene editing. However, the design of the guide RNAs used to target specific regions is not trivial. Design tools need to identify target sequences that will maximise the likelihood of obtaining the desired cut, and minimise the risk of off-target modifications. Achieving this across entire genomes is computationally challenging, with some existing methods already attempting this, however they lack the accuracy and performance required for whole-genome analysis. There is a clear need for a tool that can meet both objectives while remaining practical to use on large genomes. Here, we present Crackling, a new method for whole-genome identification of suitable CRISPR targets. We test its performance on 12 genomes, of length 375 to 9965 megabases, and on data from validation studies. The method maximises the efficiency of the guides by combining the results of multiple scoring approaches, including: inhibition of gRNA expression due to Polymerase-III terminators, poor site binding due to GC-content, poor hairpin formation, the presence of an indel-causing guanine in position 20, and via machine learnt bias derived from an existing model. The results, that are validated on experimental data, show the consensus approach selects guides of higher efficacy (with precision of up to 92%) than those selected by existing tools. Following efficacy checks, guide specificity is considered only for guides that pass. For this, we employ an approach based on Inverted Signature Slice Lists (ISSL) - a locality-sensitive, signature-based search method for large-scale data. ISSL provides a gain of an order of magnitude in speed when calculating a position-specific off-target risk score, all whilst preserving the same level of accuracy. Overall, this makes Crackling a faster and better method to design guide RNAs at scale. Crackling can be installed locally, with the source code and license at https://github.com/bmds-lab/Crackling. We further improve the convenience and availability of Crackling by adapting it for a serverless architecture. This enables rapid scaling for extremely large sized inputs at minimal cost and outperforms traditional server-based approaches that are often limited by a lack of compute resources.

Abstract #77

Karishma Chhugani (University of Southern California, California, USA)
Comprehensive analysis of usability and archival stability of RNA-seq tools

As technology has advanced, RNA-seq methods have become increasingly popular and has become an exemplar technology for transcriptome analysis, revolutionizing modern biology and clinical applications over the past decade. It has gained immense momentum driven by continuous efforts of the bioinformatics community to develop accurate and scalable computational tools. RNA-seq data analyzed by computational tools can be used to effectively tackle important biological problems such as estimating gene expression profiles across various phenotypes and conditions or detecting novel alternative splicing on specific exons. We have surveyed 235 computational tools developed from 2008 to 2020 across 15 varying domains of RNA-seq analysis. The average annual growth rate of computational tools developed for RNA-seq analysis was 114.4% from 2008 to 2014, but the rate of new tool development slowed after 2015; the average annual growth rate in tools from 2015 to 2020 was 8.97%. On an average, across the domains, there have been 18 tools developed each year between 2008-2020. Additionally, we also assessed the usability and archival stability of the computational tools designed for various types of RNA-seq analysis.Maintaining the archival stability of bioinformatics tools is increasingly important in preserving scientific transparency and reproducibility. We accessed the archival stability of the tools present in our survey and the majority of these tools are stored on archivally stable repositories (e.g, GitHub) and other tools are hosted on personal or academic webpages, which often have limited archival stability. We have also accessed the computational expertise required to install and use RNA-seq tools. A vast majority of tools require the user to operate the command line interface and only 8.09% of tools were web-based. We have also compared the availability of package managers across RNA-seq tools and majority of RNA-seq tools lack a package manager implementation. For the tools with available package manager implementation, Anaconda was the most commonly used package manager platform. The second most popular platforms were Bioconductor and CRAN. Tools that are available with package managers exhibit increased citations per year ($p=1.85 \times 10^{-7}$) compared with the tools that are not available as package managers ($p=1.43 \times 10^{-7}$). According to our survey, only 41.4% of the tools are available as package managers. Lastly, we evaluated the effect of usability on the popularity of RNA-seq tools. We found that tools that are available as package managers had significantly more citations per year compared with tools which are not available as package managers. In addition to information about usability and archival stability of the tools, we plan to create a resource which will engage the biomedical community through sharing their feedback on utilizing these tools. We hope our resources will help researchers make a more informed decision when selecting a tool for a specific type of data and research question

Abstract #78

Gunjan Dixit (Australian National University, Canberra, ACT)
Single-cell RNA-seq Analysis To Explore Bone-marrow Immune Landscape

Bone marrow (BM) contains multiple immune cell subsets with critical functions and is considered an immune regulatory organ. It contains osteoclasts and immune cells fundamentally involved in physiological and pathological bone remodelling. In autoimmune diseases, inflammation can impair the BM niche, disturb hematopoietic and immune development, and induce osteoporosis. Specific cytokines exhibit pleiotropic effects on the immune system, and their discovery in the regulation of survival, differentiation and propagation of activated T cells paved the path for its direct clinical implications in immunotherapy. This project addresses the fundamental question of how low-dose of CytokineX modulates BM immune landscape by comprehensively mapping the therapy-induced changes using single-cell technologies. I analyzed the scRNA-seq data obtained from BM of CD45 cells of mice to identify different immune cell types and compared their expression across four experimental conditions- a control (sham), control treated with CytokineX (Sham+Treatment), ovariectomy-induced osteoporosis (OVX) and OVX treated with CytokineX (OVX+Treatment). The study reveals cellular heterogeneity in different experimental conditions and identifies rare cell type ILC2. Gene expression profiles of diseased and treated samples show a significant decrease in differentially expressed genes. Pathway analysis highlights important mechanisms like osteoclast differentiation being upregulated in the diseased samples whereas downregulated after the treatment with CytokineX explaining a potential role in inhibiting bone resorption during Osteoporosis.

Abstract #79

Ning Liu (South Australian Health and Medical Research Institute (SAHMRI) & University of Adelaide, Adelaide, SA)
Abstract #3DFAACTS-SNP: Using regulatory T cell-specific epigenomics data to uncover candidate mechanisms of Type-1 Diabetes (T1D) risk

Background Genome-wide association studies (GWAS) have enabled the discovery of single nucleotide polymorphisms (SNPs) that are significantly associated with many autoimmune diseases including type 1 diabetes (T1D). However, many of the identified variants lie in non-coding regions, limiting the identification of mechanisms that contribute to autoimmune disease progression. To address this problem, we developed a variant filtering workflow called 3DFAACTS-SNP using cell type-specific data integration to link genetic variants that are associated with T1D to the loss of immune tolerance in regulatory T cells (Treg). Results Using 3DFAACTS-SNP we identified 36 SNPs with plausible Treg-specific mechanisms of action contributing to T1D from 1,228 T1D fine-mapped variants, identifying 119 novel interacting regions resulting in the identification of 51 candidate target genes. We further demonstrated the utility of the workflow by applying it to three other meta-analysed SNP autoimmune datasets, identifying 17 Treg-centric candidate variants and 35 interacting genes. Finally, we demonstrate the broad utility of 3DFAACTS-SNP for functional annotation of all known common (>10% allele frequency) variants from the Genome Aggregation Database (gnomAD). We identified 7,900 candidate variants and 3,245 candidate target genes, generating a list of potential sites for future T1D or autoimmune research. Conclusions We demonstrate that it is possible to further prioritise variants that contribute to T1D based on regulatory function and illustrate the power of using cell type specific multi-omics datasets to determine disease mechanisms. Our workflow can be customised to any cell type for which the individual datasets for functional annotation have been generated, giving broad applicability and utility.

Abstract #80

Dan Andrews (Australian National University, Canberra, ACT)
Predictive functional classification of pharmacogenetic variation

Pharmacogenetic variation is important to drug responses through diverse and complex mechanisms. To predict the functional impact of missense pharmacogenetic variants, predictive tools are employed that primarily rely on the degree of sequence conservation between species as a primary discriminator. However, off-target drug-variant interactions commonly involve effects that are unrelated to their intended biological function, which likely violates the assumptions of the predictive methodology. We have exhaustively assessed the effectiveness of missense mutation functional inference tools on the pharmacogenetic missense variants contained in the Pharmacogenomics Knowledgebase (PharmGKB) repository. We categorize PharmGKB entries into sub-classes to catalog likely off-target interactions , such that we may compare predictions across different variant annotations. Functional inference tools perform poorly on the complete set of PharmGKB variants, with large numbers of variants incorrectly classified as ‚Äòbenign‚Äô. We find substantial differences amongst PharmGKB variant sub-classes, particularly in variants known to cause off-target, type B adverse drug reactions, that are largely unrelated to the main pharmacological action of the drug. Specifically, variants associated with off-target effects (hence referred to as off-target variants) were most often incorrectly classified as ‚Äòbenign‚Äô. These results highlight the importance of understanding the underlying mechanism of pharmacogenetic variants and how variants associated with off-target effects will ultimately require new predictive algorithms. We describe how to identify variants associated with off-target effects within PharmGKB in order to generate a training set of variants that is needed to develop new algorithms specifically for this class of variant. Development of such tools will lead to more accurate functional predictions and pave the way for the increased wide-spread adoption of pharmacogenetics in clinical practice.

Abstract #81

Arash Bayat (CSIRO, Canberra, ACT)
Understanding Polygenic Disease with BitEpi and EpiExplorer

Polygenic diseases are driven by a large number of Single Nucleotide Variations (SNVs) and many of these interact in complex ways. Identifying these interactions is difficult due to computational complexity, especially in the case of higher-order interactions where more than two SNVs are involved. Here we introduce BitEpi, a fast and accurate method to test all SNVs and combinations of up to four SNVs. BitEpi introduces a novel bitwise algorithm that is 2.1 and 56 times faster than a 3-SNV search with MPI3SNP and 4-SNV search with MDR respectively. Prior to the development of BitEpi, MPI3SNP was the fastest exhaustive 3-SNV search tool and MDR was the only software to perform an exhaustive 4-SNV search. BitEpi uses a novel test to identify statistically relevant SNVs and interactions. Our method is 44% more accurate than BOOST and MPI3SNP when identifying interactive SNVs. BitEpi is compatible with standard genomic format and offers p-value-based significance testing. To aid in the visualisation of statistically significant SNVs from BitEpi, our novel tool, EpiExplorer, utilizes an interactive Cytoscape graph. EpiExplorer uses various visual elements to facilitate the discovery of the underlying biology in a complex polygenic environment. For example, it is possible to layout the graph to separate genomic regions with different functionality or highlight part of the graph based on a query. The combination of BitEpi and EpiExplorer forms a tool set that empowers researchers to identify novel genomic variants and further investigate their functionality. Furthermore, with the use of VariantSpark the pipeline can be expanded to process large-scale and whole-genome datasets.

Abstract #82

Ebony Watson (Australian Institute for Bioengineering and Nanotechnology, The University of Queensland, Brisbane, QLD)
Image-based Predictive Modelling for the Characterisation of Cellular Senescence

Senescence is a cellular stress response characterised by a state of irreversible cell-cycle arrest, and the persistent secretion of a large variety of pro-inflammatory factors. This stress response plays an essential role in embryonic development, wound healing and the prevention of tumorigenesis. Despite this, the secretions of senescent cells also give rise to chronic inflammation of the tissue environment and is believed to be the underlying driver of age-associated disease. Clearance of senescent cells has shown promising results for the prevention, delay and alleviation of these pathologies, and research into a range of clinical treatments is underway. However, the senescence response involves extensive changes to the cells morphology, chromatin organisation and methylation, metabolism, and transcription, which are further influenced by the cell-type, tissue, the stressor that induced senescence, and time since the induction of senescence. As a result of this dynamic and heterogeneous nature of the senescence phenotype, a specific and universal biomarker remains unidentified. To address this challenge, I present the development of an image-based predictive model for senescence. High-content imaging is the most informative tool for capturing associations and interactions between multiple cellular elements at high resolution, making it the ideal data-type for comprehensively characterising the multifactorial nature of the senescence response. Mesenchymal Stem Cells (MSCs), and subsequently derived osteocytes, adipocytes and chondrocytes, underwent high-content imaging at multiple time points throughout the ageing and differentiation processes. Bright field and multi-channel fluorescence imaging was conducted to visualise morphology, function and protein expression of the cells. To prepare the images for the model, an end to end image processing pipeline has been developed in Python 3. The pipeline takes the raw, grayscale images of the cells taken in 5 different channels, and outputs pixel-registered, segmented and labelled images of cells in each channel, that have been corrected for illumination, noise and background fluorescence. These processed images are now being used to train a supervised convolutional neural network to learn the complex relationships between a range of different cell types, morphological features, biomarker signatures, and stages of senescence. The pixel-registered pairing of bright field and fluorescence microscopy will enable the trained model to predict senescence in new, unseen cells from bright field images alone, allowing us to determine the quality of cell populations without fixing or staining cells. Through the application of post-hoc interpretability and visualisation methods, such as feature importance and activation maximisation, the final model will also provide a more comprehensive characterisation of the phenotypes of various sub-types and transitional states of cellular senescence. This will facilitate the identification of novel, robust biomarkers for improved targeting of therapeutics.

Abstract #83

Ruebena Dawes (Kids Neuroscience Centre, Sydney, NSW)
Features that determine 5‚Äô cryptic splice site selection in genetic disorders

Background: Splicing variants are a common cause of genetic disorders, though are challenging to interpret. Variants affecting a consensus 5' splice site motif (5'SS) often result in the spliceosome inappropriately utilising a cryptic-5'SS that may be in-frame or out-of-frame. Despite advances in the in-silico analysis of splicing variants, predicting specific mis-splicing events triggered by a genetic variant remains difficult. Aim: To define features that determine cryptic-5'SS selection in the context of a variant affecting the authentic-5'SS, or a variant that creates or modifies a cryptic 5'SS (with or without affecting the authentic-5'SS). Methods: A) Analyse features of confirmed cryptic-5'SS used by the spliceosome in the context of a 5'SS variant (4,416 cryptic-5'SS in 2,919 genes). B) Determine the abundance, location and features of decoy-5'SS which are tolerated (not utilised by the spliceosome) in human exons and introns. C) Use binary features determined in A-B to derive an algorithmic model to predict cryptic 5'SS selection. Results:

Abstract #84

Ryan Wick (Monash University, Melbourne, VIC)
Trycycler: working towards the perfect bacterial genome

Long-read assembly has progressed rapidly in recent years, with many different tools now available for assembling bacterial genomes from Oxford Nanopore or PacBio reads. However, none of these assemblers are perfect. They often fail to circularise bacterial sequences, either duplicating or omitting sequence at the start/end of a contig. They sometimes produce spurious contigs, e.g. assembling a repetitive part of the chromosome into a separate contig. They sometimes omit entire replicons, e.g. failing to include a plasmid. They sometimes create significant indel errors, e.g. deleting 50 bp from the genome. And they occasionally create large-scale misassemblies, e.g. a major structural rearrangement. Trycycler is a new tool that takes as input multiple separate long-read assemblies of the same bacterial genome and produces a consensus long-read assembly. The result is a trustworthy assembly that is free from the kinds of problems listed above. Using Trycycler, researchers can assemble bacterial genomes with more confidence and accuracy than was previously possible. In this study, we outline how Trycycler works and demonstrate its effectiveness using both simulated and real datasets.

Abstract #85

Adam Chan (University of Sydney, Sydney, NSW)
An automated framework for elucidating hierarchical relationships in high dimensional cytometry data

With the progressive influx of high dimensional cytometry data as instruments become capable of measuring up to fifty parameters for millions of cells, the challenge of finding meaningful relationships in the data continues to remain a lofty challenge. Analysts have traditionally analysed high dimensional cytometry data using a method called manual gating - where 2D scatterplots are drawn using certain markers to identify cell populations in a sequential manner. Scientists then measure the absolute proportion of the identified cell types, as well as the proportion of the cell types relative to a parent subpopulation previously gated. These proportions are then used to discover biological relationships between different patient conditions. The main drawback of this method however is the infeasibility to analyse all the cells using each of the markers, a consequence of the curse of dimensionality. Newer automated methods developed to handle the high dimensionality of modern cytometry data ‚Äì such as FlowSOM, Citrus, and Phenograph ‚Äì have been successful in handling larger amounts of data with much greater efficiency than manual gating, however these methods have overlooked the importance of measuring proportions of cells to a parent subpopulation, which can potentially lead to overlooking important relationships in the data. We present a framework which leverages hierarchies implicit in automated clustering methods to recapitulate the detection of significant cell subpopulations in a manner similar to traditional gating workflows, whilst overcoming its lengthy manual process. In particular, the framework applies a hierarchy to FlowSOM clustering and measures the proportions of these cell subpopulations relative to its parent population ancestor in the hierarchy, as opposed to exclusively measuring the proportion of FlowSOM clusters relative to all cells, to avoid missing significant relationships present in the data and emulate the traditional manual gating workflow. In addition, we analyse several high dimensional single cell datasets to: highlight the importance of child-parent proportions through comparing significance testing results using absolute proportions and proportions relative to parent; and demonstrate the use of proportions relative to parent as stratifying features in improving multivariate classification of patient outcomes.

Abstract #86

Emma Macdonald (University of Auckland, Auckland, NZ)
Is it a Mammal Thing? Determining When the Change in rDNA Unit Size Occurred in Amniotes

The ribosomal RNA genes in eukaryotes are organised into long arrays of tandem repeats, collectively termed ribosomal DNA (rDNA). Each rDNA repeat unit comprises a rRNA coding region and an intergenic spacer. Eukaryotes exhibit remarkably little variation in rDNA unit size, with the majority ranging between 9 - 15 kb in length. Mammals, though, present a striking exception to this rule, with rDNA unit sizes of ~45 kb. However, rDNA unit size is poorly characterized in amniotes, of which mammals are a member, partly due to the repetitive nature of the rDNA limiting short-read assemblies of this region. Thus, it is unclear where in amniote evolution this rDNA unit size increase occurred, or why. Here, we utilised Oxford Nanopore datasets to determine if long rDNA unit sizes are limited to mammals or are widespread among members of the amniote clade. We used BLAST to identify reads with at least two rDNA units and calculated their size through the average distance between rDNA units. Analysing over 100 reads per species revealed that the bird, snake, lizard, and turtle lineages, as well as the tuatara, all have short rDNA unit sizes. In contrast, we show that marsupials have a long rDNA unit size. These results demonstrate that expansion of the rDNA unit size occurred after the divergence of mammals and marsupials from the amniotes. We are now attempting to obtain monotreme long-read datasets to determine whether this increase in rDNA unit size occurred before or after monotremes split from the remaining mammalian lineages. Our results illustrate the value of long-read sequencing for investigating regions of the genome that are typically refractory to analysis. Pinpointing where in amniote evolution the rDNA size underwent a dramatic increase will allow us to then ask how and why this increase might have occurred.

Abstract #87

Swapnil Tichkule (Walter and Eliza Hall Institute of Medical Research, Melbourne, VIC)

Anthroponotic transmission and adaptive introgression underlies cryptic population structure of Cryptosporidium hominis in Africa

INTRODUCTION: Cryptosporidiosis is a major cause of diarrhoeal illness among African children, and it is associated with increased childhood mortality, malnutrition, cognitive development and growth retardation. Cryptosporidium hominis (C. hominis) is the dominant pathogen for this disease in Africa. Genotyping at the glycoprotein 60 (gp60) gene has revealed a complex distribution of different subtypes across Africa. However, a comprehensive exploration of the metapopulation structure and evolution based on whole genome data has yet to be performed. METHODS: In this study, we sequenced and analyzed the genomes of 26 C. hominis isolates, representing different gp60 subtypes, collected at rural sites in Gabon, Ghana, Madagascar and Tanzania. These whole genomes were subjected to variant calling with GATK pipeline, followed by PCA, Phylogeny, STRUCTURE and network based analysis to infer cryptic population structure. HybridCheck and RDP4 programs were employed to investigate introgression and recombination among the isolates, respectively. Polymorphic genes were then evaluated with population genetic metrices and tested their significance towards transmission and evolution. RESULTS: Phylogenetic and cluster analyses based on Single Nucleotide Polymorphisms showed that isolates predominantly clustered by their country of origin, irrespective of their gp60 subtype. We found a significant isolation-by-distance signature that shows the importance of local transmission, but we also detected evidence of hybridization between isolates of different geographic regions. We identified 37 outlier genes with exceptionally high nucleotide diversity, and this group is significantly enriched for genes encoding extracellular proteins and signal peptides. Furthermore, these genes are found more often than expected in recombinant regions, and they show a distinct signature of positive- or balancing selection. CONCLUSION: Our study showed that: 1) the metapopulation structure of C. hominis can only be accurately captured by whole genome analyses; 2) local anthroponotic transmission underpins the spread of this pathogen in Africa; 3) hybridization occurs between distinct geographical lineages; and 4) genetic introgression provides novel substrate for positive- or balancing selection in genes involved in host-parasite coevolution.

Abstract #88

Dan Andrews (John Curtin School of Medical Research, Australian National University, Canberra, ACT)
Machine Learning Prediction of End-stage Kidney Disease for Clinical Decision Support

Chronic kidney disease (CKD) is a major source of morbidity and mortality globally. Whilst CKD may ultimately culminate in end stage kidney disease (ESKD), CKD may progress at highly variable rates or not progress at all. ESKD is associated with a marked increase in mortality and morbidity and without renal replacement therapy (RRT) in the form of haemodialysis, peritoneal dialysis, or kidney transplantation is a terminal condition. As ESKD approaches, clinicians are required to make difficult decisions. RRT requires either the formation of permanent dialysis access or evaluation of suitability for transplantation. Preparation for RRT is associated with significant cost and side effects such as post-operative infection and bleeding. The capacity of physicians to correctly identify which, and when, patients will require RRT is poor. Therefore, any method which will improve the ability to correctly identify patients who will require RRT is highly desirable. Data-driven predictive modelling is a rapidly advancing field and has been employed in a range of clinical scenarios. Recent advances have demonstrated the capacity of predictive modelling as an adjunct in clinical care in kidney disease. In this work, we will present a machine learning algorithm capable of predicting which patients will progress to ESKD and the time when they will reach ESKD. This data-driven approach, based on thousands of patient records from the Canberra Hospital, collected over nearly a decade, will assist treating physicians in planning if and when an individual patient should be prepared for RRT.

Abstract #89

Ke Ding (John Curtin School of Medical Research, Australian National University, Canberra, ACT)
Long short-term memory RNN for mirtron identification

MicroRNAs are small regulatory RNAs mediate extensive networks of post-transcriptional regulation and are implicated in a variety of diseases. Unlike most microRNAs generated by the canonical pathway involving Drosha and Dicer, we discovered a new subclass of microRNAs which bypass Drosha cleavage and generate functional miRNAs via splicing, termed Mirtrons. As mirtrons are lowly expressed, their detection using high-throughput sequencing is difficult. In this study, we use a recurrent neural network (RNN) known as long short-term memory (LSTM) to build a model for mirtron identification. Unlike classical machine learning methods which require pre-selected features to build the models, deep neural network models can extract relevant features without previously defining them. Comparing to other deep neural network models such as the convolutional neural network (CNN), LSTM networks are capable of capturing the long-range dependency in sequential data and are famous for solving the vanishing gradient problem in traditional RNNs when dealing with long sequence data. We showed that our model achieves a satisfactory prediction result with the area under the ROC curve (AUC) of 0.857 by only taking mirtron sequence data as input. Comparing with other models in pre-miRNA classification, we showed that our model outperforms it all by achieving F-1 score 95.3 (3.6 point absolute improvement). By training and testing our model on mammalian species, we verified mirtrons in humans and mice share similar sequence and structural features. By concatenating small RNA sequencing coverage data with mirtron sequence data as input, we further improved the performance of our model, reaching an AUC to 0.896. We also experiment with the Attention Mechanism and visualize the critical segment in sequences for identifying mirtrons. Lastly, we run our model genome-wide and predicted the novel mirtrons for future study.

Abstract #90

Loic Thibaut (Victor Chang Cardiac Research Institute, Sydney, NSW)
powerSFS: quantifying the intolerance of genes to mutation with a statistical model of the site frequency spectrum

Scores of genic intolerance such as RVIS, GeVIR and LOEUF help to prioritize candidate disease-causing genes and facilitate the interpretation of patient genomes. Existing methods considered the deficit of common functional variation in a gene (RVIS), the deficit of potential loss of function variants (LOEUF) and the distribution pattern of variants within a gene (GeVIR) to define metrics of intolerance. Although the site frequency spectrum (SFS) ‚Äî the distribution of the allele frequencies in a gene ‚Äî is routinely used by classical tests of natural selection, such as Tajima‚Äôs D, no genic intolerance method so far uses the whole SFS for quantifying intolerance. Here we develop a gene-level intolerance metric that explicitly models the signature that purifying selection leaves in the site frequency spectrum of a gene. Specifically, we assume that the proportion of variants observed at a given allele frequency in a population decreases with the allele frequency according to a power law. As expected from population genetics theory, we further assume that the magnitude of the slope of the power law increases with the intensity of purifying selection. Using data from 125,748 exome sequences released by the gnomAD v2 project, we compiled the distribution of allele frequencies of single nucleotide functional variants for every known protein-coding gene. We then estimated the slope of the power law for each gene in a maximum likelihood framework. Our results show that powerSFS can prioritize dominant and recessive disease genes and is comparable to existing constraint metrics but performs better for small genes and for genes intolerant to missense mutations. Furthermore, we identify and characterize a set of genes that powerSFS predicts as significantly intolerant but that are not prioritized by other methods. Our findings show that powerSFS is complementary to other scores of intolerance and can help to identify genes causing Mendelian human diseases.

Abstract #91

Renzo Balboa (Australian National University, Canberra, ACT)
PACIFIC: A lightweight deep-learning classifier of SARS-CoV-2 and co-infecting RNA viruses

Viral co-infections occur in COVID-19 patients, potentially impacting disease progression and severity. However, there is currently no dedicated method to identify viral co-infections in patient RNA-seq data. We developed PACIFIC, a deep-learning algorithm that accurately detects SARS-CoV-2 and other common RNA respiratory viruses from RNA-seq data. Using in silico data, PACIFIC recovers the presence and relative concentrations of viruses with >99% precision and recall. PACIFIC accurately detects SARS-CoV-2 and other viral infections in 63 independent in vitro cell culture and patient datasets. PACIFIC is an end-to-end tool that enables the systematic monitoring of viral infections in the current global pandemic. Biorxiv: https://doi.org/10.1101/2020.07.24.219097

Abstract #92

Konstantinos Bogias (University of Adelaide, Adelaide, SA)
Characterisation of transcript expression in placenta across early gestation reveals variable transcript usage

The human placenta is a rapidly developing, highly specialised organ that plays a central role in pregnancy and is essential to the development of the fetus. The myriad functions of the placenta include exchange of nutrients and wastes between mother and fetus, mediation of maternal immune system activity, regulation of maternal insulin sensitivity and protection of the fetus against xenobiotics. In early gestation, the placenta develops in a physiological but low oxygen environment until approximately 10 weeks‚Äô gestation at which time maternal blood flow into the placenta is initiated and oxygenation occurs. Recent work in our lab has shown differential gene expression (DGE) in placenta from 6-10 weeks‚Äô versus 11-23 weeks‚Äô gestation, where striking changes in expression of immune system genes were evident. Identification of individual isoform expression associated with developmental processes in early gestation, and with pregnancy complications, have been previously reported in the literature. To assess the impact of individual transcript isoforms at a genome-wide scale, we initially used selective alignment implemented in Salmon to quantify transcript abundances from 96 placenta samples throughout 6-23 weeks‚Äô gestation. Comparing 6-10 weeks‚Äô and 11-23 weeks‚Äô gestation placenta, differential transcript expression (DTE) was performed using edgeR and differential transcript usage (DTU) analysis using DRIMSeq. Subsequent validation and increased FDR control of DTU results were performed using StageR and gene ontology (GO) enrichment analysis using goseq. Both the DTE and DTU analyses were subsequently compared with each other, along with DGE data. A total of 1,005 DE transcripts were identified from 853 genes, with transcripts more likely to be upregulated at 11-23 weeks‚Äô gestation. In the DTE results, 501 genes were also detected in DGE analysis while 352 genes had observable differential expression in transcripts without any detectable changes in gene-level expression. In DTU analysis, 1441 transcripts from 611 genes were found to have changing proportions from early to mid-gestation. Of the DTU genes, 55 also showed DTE in the absence of any detectable change in gene expression. GO enrichment for DTE genes revealed enrichment in cell surface receptor signalling, cellular migration and inflammatory response, while DTU genes were enriched for intracellular protein transport, positive regulation of splicing, and cadherin binding. The most significant DTU genes (FDR < 1.0e-23), ADAM10, VMP1, MTUS1, ASAH1, and GPR126, exhibited variable usage of transcripts coding for functional protein domains associated with processes including autophagy, embryogenesis, angiogenesis and apoptotic resistance. Overall, this study presents the most comprehensive genome-wide profile of transcript usage in early to mid-gestation

placenta to date. It highlights the diversity and plasticity of transcript expression during early placental development. Future studies will examine functional implications.

Abstract #93

Stephen Kazakoff (QIMR Berghofer Medical Research Institute, Brisbane, QLD)
Mapping cancer transcriptomes with long-read sequencing

Whilst tissue-specific transcript expression is highly regulated during the stages of development, dysregulation in cancer cells can lead to aberrant transcriptional signaling and increased variation of transcripts. The variation may lead to a selective advantage to the cancer cells, contribute to the hallmarks of cancer and affect tumour treatment response. Although short-read RNA-Seq has been widely used in cancer genomics to quantify gene expression, emerging long-read technologies can now be used to sequence full-length transcripts. This ability enables the profiling of cancer transcriptomes and identification of potential disease biomarkers and treatment targets. We used the PacBio RS II and Sequel II platforms to generate long reads from RNA extracted from colorectal cancer (CRC) samples from four patients. The samples from each patient comprised primary CRC tumour, liver metastases and matched normal colon tissue. The data was processed using the PacBio IsoSeq3 analysis pipeline to generate high-quality full-length mRNA transcripts. Transcript variation was assessed by comparison of splice junctions in cancer and normal datasets aligned using minimap2 (v2.17). Matched RNA-Seq was used to verify splice junctions and publicly available RNA-Seq provided independent frequency data of splice junction used in pan-cancer cohorts. An average of 31,000 and 5,000 full-length transcripts were identified per sample using the Sequel II and RS II platforms, respectively. By comparing the RS II data to the GENCODE v31 gene model an average of 73% of transcripts matched annotated splice junctions and approximately 25% had unannotated splice boundaries. A novel spliced transcript was also identified and was verified using matched short-read RNA-Seq. The sequence could be uniquely placed in the genome and BLAST searches of nucleotide and protein databases did not return significant hits across 44 vertebrate species (including human). We also could not find any evidence of expression at the genomic locus in human GTEx data. To identify if this transcript was expressed in other cancers short-read RNA-Seq from eleven TCGA projects was accessed totaling 4357 samples. Expression of the unannotated gene was detectable in >10% of cases from TCGA CRC, breast, endometrial and oesophageal samples. Analysis of data from the more recent Sequel II platform provided detection of transcripts at a higher resolution allowing determination of allele-specific and somatic driver mutation isoform expression. We show that long-read sequencing of cancer transcriptomes can be used to identify patterns of cancer-specific splicing events, cancer driver mutation expression, allele-specific expression and novel cancer-specific transcripts. We also show that publicly available RNA-Seq data can be used to verify and identify the frequency of these events in a given cohort. Further work could support somatic transcript splice junctions as disease biomarkers.

Abstract #94

Mark Pinese (Children's Cancer Institute, Sydney, NSW)
Nimpress brings polygenic scores to the sequencing era

Polygenic scores enable the quantitative prediction of phenotype from genotype, and the application of polygenic scores to emerging genomic cohorts promises to revolutionise the prediction of complex traits such as disease risk. However, existing software for the calculation of polygenic scores are poorly-suited to modern genomic data, requiring cumbersome and lossy conversion steps. These steps demand significant computational and storage resources, and their lossy nature can impact the accuracy of the resulting polygenic scores. In all, current software for polygenic score calculation are a poor match to modern genomics data, and an impediment to the integration of polygenic scores into genomics research and clinical genomics pipelines. Here we present nimpress, a lightweight and portable tool for exact polygenic score calculation, direct from VCF or BCF. nimpress has rich inbuilt imputation options tailored to modern sequencing data, and computes polygenic scores more than 10 times faster than existing software, while using a few percent of the memory. nimpress also includes helper scripts to create polygenic scores from published summary statistics. These scripts automatically handle identifier conversion, consistency checking, and tag SNP imputation, to dramatically reduce the effort and error involved in the definition of new polygenic scores. nimpress is available as open source code, static binary, and Docker images, to enable the rapid and simple integration of polygenic scores into existing genomics workflows.

Abstract #95

Lixinyu Liu (Australian National University, Canberra, ACT)
The landscape of alternative polyadenylation in CD8 T cells in single-cell transcriptome

Alternative cleavage and polyadenylation (APA) enables the production of different isoforms of mRNA and affect more than 70% eukaryotic genes. APA is a critical process as it diversifies the transcriptome in cells, and can dynamically affect transcript stability, cellular localisation, nuclear export, and translation efficiency. APA is cell-type specific and has been found to be abundant in immune cells. In particular, specific APA events have been detected in several pathological conditions, including immunological and autoimmune diseases. However, a comprehensive exploration of APA in immune cells at the single cell level has not been conducted previously. In this study, we profiled mouse CD8+ T cells, known as cytotoxic T cells acting as intermediaries of adaptive immunity,

from eight tissues by single-cell sequencing (scRNA-seq) to uncover cell type and developmental trajectories of differential APA varying across cell types, developmental states and tissues. We first comprehensively maps the APA sites in 13 cell types. At the cell-type level, na√Øve CD8+ cells show more lengthening isoforms than other cell types, and the comparatively shorter isoforms are associated with drug metabolism, regulation of lymphocyte-mediated immunity, and positive regulation of immune effectors. Effector CD8+ cells prefer proximal sites compared to others, and the shorter isoforms are associated with biosynthesis, endocytosis, cell to cell recognition, and upregulation of immune defences. By comparing differential APA usage at the single cell level, we detect cell type-specific APA markers. For example, Id2 prefers its proximal polyadenylation site in effector memory CD8+ cells and prefers its distal polyadenylation site in Na√Øve CD8+ cells. To correlate APA usage and gene expression, we further use a deep learning method, Autoencoder, to compress APA usage and expression features at a single-cell level. With the autoencoder model, the single-cell map of APA usage can be used to improve the expression-based clustering map. Finally, we develop a visualisation method to visualise the APA usage of CD8+ cells at single-cell resolution.

Abstract #97

Malathi S.I. Dona (Baker Heart and Diabetes Institute, Melbourne, VIC)
Single-cell Transcriptional Profiling Reveals Novel Cellular and Molecular Drivers of Cardiovascular Fibrosis

Fibrosis is a leading factor in the development of many cardiovascular diseases, which are major causes of morbidity and mortality worldwide. Recent research has revealed that cardiovascular tissues,Äîsuch as the heart and aorta,Äîare complex cellular networks comprising diverse arrays of cell types. However, little is known of how these networks change during chronic physiological stress and how specific cell populations contribute to fibrosis. To address these gaps in our knowledge, we utilised single-cell RNA-sequencing to characterise changes in the cellular landscapes of mouse hearts and aortas in the context of chronic physiological stress. Our research has identified previously undescribed cell types, global cellular changes in gene expression, and shifts in intercellular communication networks that drive tissue fibrosis. Our research has also highlighted sex-specific characteristics of the homeostatic and stressed tissue that have been previously unappreciated. Our high-resolution analyses of the heart and aorta provide new insights and challenge many long-held paradigms of how fibrosis develops and contributes to cardiovascular disease.

Abstract #98

Anushi Shah (Victor Chang Cardiac Research Institute, Sydney, NSW)
Investigation of de novo mutations in human genomes using whole genome sequencing datasets

De novo mutations (DNMs) are genetic alterations occurring for the first time in a family member, which could be germline or somatic. DNMs have been shown to be a major cause of severe developmental genetic disorders. With the advent of next generation sequencing technologies, accurately detecting DNMs is crucial. A number of de novo variant callers to call DNMs from whole genome sequencing(WGS) data have been developed that differ in algorithms, filtering strategies and output. However, there is no study which has systematically evaluated these tools. We evaluated four DNM calling tools TrioDenovo, Phasebytransmission, DenovoGear and VarScan2 with regards to their concordance and accuracy in calling DNMs. Validation gold standard dataset consists of Illumina Hiseq WGS data of one CEU trio from 1000 Genomes Project. We also performed evaluation using simulated trio WGS datasets spiked-in with known DNMs for which we independently developed ‚ÄúTrioSim‚Äù, an automated framework to generate simulated genomic datasets for trios. Our analysis on CEU 1000G trio dataset shows 8.7% DNM concordance amongst 4 DNM callers, while up to 36.2% of DNMs were called as unique to each caller. Our analysis on simulated trio dataset spiked-in with 100 DNMs show 1.9% concordance while 0.6% to 66.9% DNM calls were unique to each caller. This shows large false positives detected by these tools and stringent post-filtering is required to obtain high confidence DNMs. Our meta-caller approach of utilizing consensus of all 4 callers shows comparable sensitivity and specificity of 97%.

Abstract #99

Gulrez Chahal (Monash University, Melbourne, VIC)
CaraVaN: Prioritising Cardiac Variants in the Non-coding genome using boosting algorithm

Heart diseases have been one of the leading causes of death worldwide. There are several gene candidates that have been identified to diagnose and treat heart diseases. However, in many cases the genetic cause still remains unknown. This can be attributed to the fact that protein-coding genes contribute only ~2% of the genome, while the non-coding genome (~98%) comprises of functional regions, such as cis-regulatory elements, that are involved in the regulation of the expression of the genes. Recent evidence indicates that variations in these regulatory regions such as enhancers and insulators impact gene expression and result in heart disease. However, there is no method to investigate/predict non-coding variants in cardiac disease yet. Finding disease-causing variants in the non-coding genome poses a challenge, as they do not follow a genetic code. In recent years, several conservation and machine learning-based tools have been developed to prioritise these variants in the non-coding genome, however, they are not disease-specific. Given the unique complexity of heart development and its associated defects, we present a cardiac-specific model which annotates and prioritises potentially pathogenic variants in the non-coding genome pertaining to heart disease, using

decision-tree based ensemble learning boosting algorithm. For training the machine learning model on cardiac-specific human functional, epigenomic and structural consequence, we have identified 4 major feature categories comprising of: cardiac-specific histone marks (n=18), human cardiac transcription factor binding sites (n=98), cardiac-specific 3D chromatin organisation (n=4) and deleteriousness scores from existing non-coding genome variant assessment tools, to capture variants with role in regulation of cardiac development and disease. These cardiac-specific features are used to prioritise variants with a potential pathogenic role in heart disease.

Abstract #100

Adria Closa (John Curtin School of Medical Research, Australian National University, Canberra, ACT)
Characterisation of a convergent malignant phenotype in B-cell acute lymphoblastic leukaemia

Acute lymphoblastic leukaemia (ALL) is the most common form of cancer in children worldwide. Although combination chemotherapy provides in general an effective treatment, resulting in an overall survival of >90%, subtypes of paediatric ALL affecting children in the first year of life or carrying rearrangements of the mixed lineage leukaemia (MLL) gene remain with a dismal prognosis. These poor outcomes highlight the unmet need for a better understanding of the molecular mechanisms of acute leukaemia and motivate the search for new therapeutic strategies for high- risk paediatric acute leukaemia. Genome sequencing studies of ALL patients have shown a very low frequency of somatic mutations, indicating that MLL-r may not require additional alterations to induce full transformation. However, the mechanisms of how gene fusions relate to disease transformation remain to be fully explained. To uncover new molecular mechanisms potentially linked to the observed poor outcome, we performed an exhaustive multicohort analysis of gene-fusions and RNA processing alterations in 428 B-ALL patients. We identified 84 fusions with significant allele frequency across patients, 6 of them novel and 19 known from other blood and solid tumours but which had not been observed before in ALL. We have analysed and uncovered the similarities in their potential functional impacts. Furthermore, using MLL-r and ETV6-r as proxies for high and low risk, respectively, we found an expression signature involving MYC target genes and regulators of RNA processing in association with MLL-r patients. Moreover, this signature has a predictive power related with risk in an independent set of patients with other or no fusions, demonstrated by a Random Forest model of survival and a Cox-regression test. This risk signature includes the upregulation of the splicing factor SRRM1, which we show that through the interaction with other splicing factors potentially impact a set of alternative splicing events associated with high risk. Our findings provide evidence for a convergent mechanism of aberrant RNA processing that sustains a malignant phenotype in a subset of gene-fusion-driven B-ALL cases. This convergent phenotype can complement the risk diagnosis currently based on gene-fusion detection.

Abstract #101

David Goode (Peter MacCallum Cancer Centre, Melbourne, VIC)
An Evolutionary Approach to Network Analysis of Cancer Transcriptomes Reveals Common Indicators of Enhanced Malignancy Across a Range of Solid Tumours

Despite diverse origins and significant genomic heterogeneity, all types of cancers display common molecular features, including uncontrolled proliferation, altered metabolism and dedifferentiation. These features resemble properties of unicellular organisms such as bacteria, yeast and protozoa, suggesting cancer is driven by the breakdown of gene regulatory networks (GRNs) that evolved in the earliest metazoans (multicellular animals) to control basic processes such as cell division and differentiation. To study this hypothesis, we developed Evolutionary Network Analysis (ENA). Our approach combines phylogenetics with network biology and cancer transcriptome data to investigate how GRNs are disrupted and rewired in cancer and how this relates to the evolutionary origins of the genes involved. Applying ENA to data from The Cancer Genome Atlas showed strong enrichment of somatic mutations in master regulators that control communication between unicellular and multicellular genes in cancer, resulting in increased expression of unicellular genes across multiple tumour types. To better understand how disrupted communication between unicellular and multicellular genes drives tumour progression, we identified gene co-expression modules in over 30 type of cancers, and matched normal tissues where available. Using protein sequence conservation, modules were classified as having predominantly unicellular genes, predominantly multicellular genes or a mix of unicellular and multicellular genes (mixed UC-MC modules). Mixed UC-MC module showed the greatest degree of difference in their topology and expression between tumours and normal tissue. Thus, changes in co-expression of unicellular and multicellular genes are common features of tumours. These mixed UC-MC module were frequently associated with copy-number alterations, particularly amplifications. Comparing high-grade to low-grade tumours showed progression to higher grades involved further disruption and rewiring of the links between genes in mixed UC-MC modules, suggesting continued alterations to connections between unicellular and multicellular genes enhance tumour malignancy. To examine the potential prognostic implications of this finding, we built Random Forest classifiers based on the expression of unicellular, multicellular or mixed UC-MC modules and tested their ability to distinguish low-grade from high-grade brain tumours. Classification based on mixed UC-MC modules performed the best for discriminating low-grade glioma tumours from high-grade glioblastoma tumours. Furthermore, survival in the low-grade glioma cohort was linked to activity of selected mixed UC-MC modules. Evolutionary Network Analysis demonstrates the power and utility of incorporating the evolutionary origins of genes into the study of cancer transcriptomes. This opens the door to development of new classes of evolutionary-informed prognostic gene expression signatures potentially applicable to any type of cancer.

Abstract #102

Renzo Balboa (Australian National University, Canberra, ACT)
Alu Repeat Diversity in the Human Genome

More than half of the human genome is comprised of repeat sequences. Alu elements are primate-specific ~300bp repeat sequences considered to be the most successful transposable elements in the human genome. In any individual, it is believed that there are ~1.1 million copies of Alu repeats comprising ~11% of the human genome. Their higher activity (~1:20 births) and high polymorphism rates contribute to human diversity. Alu elements have also been implicated in genome regulation where for example, Alus contribute up to ~30% of methylation sites in the human genome, and due to its activity, have been implicated in genetic instability and disease. Therefore, a comprehensive understanding of Alu polymorphisms is required for understanding their contribution in genome regulation and in health and disease. We have developed an assembly-based approach to comprehensively annotate Alu variations at an individual and population level using whole genome sequence data from 273 individuals from the Simons Genome Diversity project encompassing 128 global populations. We find that Alu elements form the largest proportion of annotated structural variation (~15% of all structural variants) in the human genome. We detect ~1.3 million Alu across all individuals, where ~99% of all detected Alu elements overlap with the reference genome, and ~93% of all reference Alu elements are detected in our analyses. Our approach can confidently describe polymorphisms for ~700 non-reference Alu regions per individual; these elements are largely individual and population-specific. A greater understanding of Alu polymorphisms in humans will reveal wider patterns of human diversity and will pave the way for understanding their roles in DNA regulation and in health and disease.

Abstract #103

Varuni Sarwal (University of California Los Angeles, California, USA)
A comprehensive benchmarking of WGS-based structural variant callers

Structural variants (SVs) are genomic regions that contain an altered DNA sequence due to deletion, duplication, insertion, or inversion, and have varying pathogenicity of disease. Dissecting SVs from whole genome sequencing (WGS) data presents a number of challenges and a plethora of SV-detection methods have been developed. Currently, there is a paucity of evidence which investigators can use to select appropriate SV-detection tools. We evaluated the performance of 15 SV-detection tools based on their ability to detect deletions from aligned WGS reads using a comprehensive PCR-confirmed gold standard set of SVs to find methods with a good balance between sensitivity and precision. While the number of true deletions is 3710, the number of deletions detected by the tools ranged from 899 to 82,225. 53% of the methods reported fewer deletions than are known to be present in the sample. The length distribution of detected deletions varied across tools and was substantially different from the distribution of true deletions. 53% of tools underestimate the true size of SVs and deletions detected by BreakDancer were the closest to the true median deletion length. We allowed deviation in the coordinates of the detected deletions and compared deviations to the coordinates of the true deletions from 0 to 10,000 bp. Manta achieved the highest f-score for all thresholds. Methods with high specificity rates tend to also have significantly higher f-score and precision rates. CLEVER was able to achieve the highest sensitivity while the most precise method was PopDel. We assessed the performance of SV callers at coverages from 32x to 0.1x generated by down-sampling the original WGS data. DELLY showed the highest F-score for coverage below 4x while Manta was the best performing tool from 8x to 32x. We assessed the effect of deletion length on the accuracy of detection. Manta and CREST were the only tools with high specificity for deletions shorter than 500bp. LUMPY was the only method able to deliver an F-score above 30% across all categories. Manta and LUMPY were the best performing tools for general applications. Our recommendations can help researchers choose the best SV detection software, as well as inform the developer community of the challenges of SV detection.

Abstract #104

Ying Zheng (Australian National University, Canberra, ACT)
Discovery of Tissue-specific Gene Expression Patterns in CD8 T Cells by Single-cell RNA-seq

The current most effective clinical cancer treatment, immune checkpoint blockade (ICB), has been argued to be the basis of next-generation immunotherapy. Nevertheless, although the efficacy of ICB has been attributed to CD8 T cells, the underlying precise mechanism remains poorly understood. In recent years, with the clinical application of ICB, the distinct effects of ICB on the treatment of different cancer types and the tissue-specific complications of ICB (immune-related adverse events, irAEs) have been widely reported. In allusion to the widely reported association between tissues and ICB effects, in this research, we used single-cell technologies to profile murine tissue-infiltrating CD8 T cells obtained from 8 different tissues, including the small intestine, kidney, liver, lymph nodes, lung, PBMCs, spinal cord, and spleen. We comprehensively compared and characterised tissue-infiltrating CD8 T cells to explore tissue-specific gene expression patterns. Our research proves cellular heterogeneity of tissue-infiltrating CD8 T cells, profiles their abundant tissue-specific gene expression patterns, and most importantly, identifies several tissue-specific-infiltrating CD8 T cell subpopulations in the liver, kidney, and small intestine. The trajectory analysis reveals the distinct differentiation statuses of infiltrating CD8 T cells across tissues and identified subpopulations. The pathway enrichment analysis further highlights the immune-related

mechanisms involving detected tissue-specific gene expression patterns. These discoveries have important implications for illustrating the underlying association between tissues and ICB, including the relationship between tissue-specific irAEs and ICB effects in different cancer types.

Abstract #105

Huiwen Zheng (Australian Institute for Bioengineering and Nanotechnology, The University of Queensland, Brisbane, QLD)
Systematic evaluation for metrics of gene expression variability in single-cell RNA sequencing data

During ageing, transcriptional noise has been shown to increase in multiple organs and tissues. Transcriptional noise is defined as the variability of gene expression, and this property reflects the heterogeneity that results from stochastic cell to cell variation. Although the concept of transcriptional noise is not new, different metrics are being used to measure this and it is unclear what the optimal approach is. With the advent of single cell sequencing techniques, it is now becoming possible to quantify how noise is distributed through the genome. The project focuses on understanding how to accurately model transcriptional noise as a regulatory property of the genome and its contribution to the fundamental feature of ageing. To conduct a systematic evaluation, we selected 12 different metrics that commonly used in scRNA-seq studies. Performance of these metrics is tested with simulated and experimentally-derived datasets. We investigated the performance of these metrics against different data structures, stably expressed genes,and other properties. Using a publicly available scRNA-seq datasets with multiple tissues and age groups for mice, we intend to investigate how transcriptional noise changes during ageing and between cell types. Through these analysis, the goal is to understand how transcriptional noise impacts the regulatory processes that underlie ageing.

Abstract #106

Katarina Stuart (University of New South Wales, Sydney, NSW)
Whole transcripts in genome assembly, annotation, and assessment: the draft genome assembly of the globally invasive common starling, Sturnus vulgaris

Native to the Palearctic, the common starling (Sturnus vulgaris) is a near-globally invasive passerine that has now colonised every continent barring Antarctica. Ecological interest in the species is two-fold ,Äì they are considered a conservation risk and crop pest within the invasive ranges, while recent decades have brought with them a worrying decline in starling numbers within historical native ranges. Despite the global interest in this species, there are still fundamental knowledge gaps in our understanding of the genetics and population differences of this species across their native and invasive range. We present the Australian S. vulgaris draft genome and transcriptome to be used as a reference for further investigation into evolutionary characterisation of this ecologically significant species. An initial 10x Genomics linked-read assembly was scaffolded and gap-filled with low coverage nanopore sequencing, complemented by PacBio Isoseq full-length transcript data. Isoseq data was incorporated into assembly scaffolding, annotation, and assembly assessment to inform workflow decisions. We produced a draft assembly with a scaffold N50 size of 72.5 Mb, and assess this alongside a North American S. vulgaris draft genome, previously assembled from Illumina data. Lastly, we use these different reference genomes, alongside a non-scaffolded version of the Australian S. vulgaris genome to assess how choice of reference genome affects common population genetic downstream analysis using a global whole genome resequencing data set.

Abstract #107

Christoffer Flensburg (Walter and Eliza Hall Institute of Medical Research, Melbourne, VIC)
clonal tracking as quality control for mutational signature analysis

High throughput sequencing of cancer samples is an incredibly sensitive assay that can reveal deep information about the dynamics of a cancer. Unfortunately, the high sensitivity to detect somatic mutations comes with a high sensitivity to noise. Sequencing data is frequently discarded due to quality and some issues can lead to false conclusions if not identified. Quality control is often the most important and difficult part of an analysis but rarely takes the spotlight in publications or presentation. However, good quality control can identify low quality samples to avoid false signals, or even remove noise to extract a robust signal from precious samples. I will show that clonal tracking can be a valuable quality control tool for bulk sequencing data, even when phylogeny or clonal structure is not a primary research question. We analysed a dataset of multiple polyps from an MBD4 deficient individual and used clonal tracking to overcome FFPE artefacts as well as contamination from a different individual. We found a consistent pattern of mutation in APC, AMER1, BFXW7 in clones consistent with a SBS1 signature, as well as *RAS mutations in clones with a novel CA>AA signature.

Abstract #108

Qianhui Wan (Flinders University, Adelaide, SA)
DNA methylation differences in human placenta from female and male bearing pregnancies

Sex differences in fetal growth and development are well known. Males are generally larger than females at birth and sex-specific pregnancy outcomes are associated with maternal diet and health status, though the underlying mechanisms of sex differences are not fully understood. DNA methylation differences between female and male placentas could be one of the reasons that explain the sex-specific disease risks. In this study, we profiled placental DNA methylation across early to mid-gestation according to fetal sex using Illumina Infinium¬Æ MethylationEPIC BeadChips (EPIC arrays) on 131 human placenta samples (63 female and 68 male) from 6-23 weeks,Äô gestation. ENmix was used to correct the dye bias and background of array data, failed probes and samples were identified and removed by quality control methods using minfi, and BMIQ method implemented in ChAMP was used for data normalisation. Differential methylation analysis was performed using DMRcate for female and male placenta samples with DMRs identified between 6-10 and 11-23 weeks,Äô gestation groups. Weighted linear regression models were used to deal with heteroskedasticity such that samples with more variances were weighted less in the models to reduce the effects of non-constant variations. Batch effects were considered and adjusted in the models for identifying DMRs. When comparing 6-10 and 11-23 weeks,Äô gestation groups for female and male placentas, the majority of DMRs were identified on autosomes except for 3 DMRs on the X chromosome (overlapped SMARCA1, ACE2 and FIRRE). 99 DMRs increased in methylation and 4 DMRs with decreased methylation in 11-23 compared to 6-10 weeks,Äô gestation groups were found in both female and male placentas, which was in line with previous reports that global DNA methylation increased across gestational weeks. Between 11-23 versus 6-10 weeks,Äô gestation groups, we also found increased DNA methylation of 348 and 4 DMRs and decreased DNA methylation of 3 and 2 DMRs were identified only in female and male placentas, respectively. Upon functional enrichment of 313 genes that overlapped 348 DMRs with increased methylation using methylGSA, we identified genes regulating transmembrane transport. Overall, our study is the first to show DNA methylation differences associated with fetal sex in human placenta during early to mid-gestation with a large sample size, which could provide novel insights to understand sex differences in neonatal and early programming of adult disease. In future, different omics data from these placenta samples will be integrated to investigate sex-specific features in placental development.

Abstract #109

Elisabeth Roesch (University of Melbourne, Melbourne, VIC)
How does gene expression entropy change along developmental trajectories?

The differentiation of stem or progenitor cells is accompanied by profound changes in gene expression. In addition to overall changes in expression profiles there is also widespread evidence that also the variability of gene expression varies along the developmental trajectory [1]. Here we investigate this in the context of simple mathematical models. These models are inspired by, and share important features with Waddington,Äôs epigenetic landscape [2], and they allow us to disentangle the effects of developmental dynamics and molecular noise on patterns of gene expression variability [3], and to investigate the role of molecular noise in the cell fate decision making process. In particular we consider dynamical systems that exhibit the types of qualitative change ‚Äì ‚Äúbifurcations‚Äù in the language of dynamical systems ‚Äì that occur in e.g. stem cell differentiation and developmental patterning processes. Using simulations we construct the corresponding epigenetic landscapes and we show how noise can affect, and even mask some types of bifurcations. Noise, we show, affects the shape of the epigenetic landscape. Indeed, landscapes for the same developmental system, but with different levels of molecular noise can give rise to qualitatively different behaviour and outcomes. This analysis complements ongoing efforts by developmental biologists and the single cell analysis community to distill features of cell differentiation dynamics from data directly. We conclude by pointing out, how these modelling studies can help in the analysis of single cell transcriptomic data. [1] Angelique Richard, Lois Boullu, Ulysse Herbach, Arnaud Bonnafoux, Valerie Morin, Elodie Vallin, Anissa Guillemin, Nan Papili Gao, Rudiyanto Gunawan, Jeremie Cosette, Ophelie Arnaud, Jean-Jacques Kupiec, Thibault Espinasse, Sandrine Gonin-Giraud, and Olivier Gandrillon. Single-Cell Based Analysis Highlights a Surge in Cell-to-Cell Molecular Variability Preceding Irreversible Commitment in a Differentiation Process. PLOS Biology, 2016. [2] Conrad Hal Waddington. The strategy of the genes. Allen & Unwin, 1957. [3] Naomi Moris, Cristina Pina, and Alfonso Martinez Arias. Transition states and cell fate decisions in epigenetic landscapes. Nature Reviews Genetics, 2016.

Abstract #110

Chelsea Matthews (University of Adelaide, Adelaide, SA)
Assessing PacBio long reads and de novo genome assembly tools for useability and suitability to applications where resources are limited.

Due to the further development of long read sequencing techniques, reductions in the cost of sequencing per kbp, and increases in sequencing throughput, a growing number of laboratories are undertaking long read de novo assembly projects. Making informed decisions regarding sequencing technology, read depth, and assembly tools is imperative for managing project costs, timelines, and expectations. Where projects are mismanaged, costs and timelines can blow out and the risk of an assembly never eventuating due to staff/students moving to new positions increases. Unfortunately, for a laboratory with limited experience generating de novo assemblies, informing these decisions can be very challenging. To investigate some of the issues surrounding the project management of a de novo genome assembly project, de novo assembly and quality assessment workflows for PacBio continuous long reads (CLRs) and PacBio HiFi reads were implemented using a range of tools (Flye, Wtdbg2, Raven, and Canu for CLRs and Flye, Canu, HiCanu, and Hifiasm for HiFi reads). Using the rice strain MH63, assemblies were generated at a range of coverages for each tool. Assembly

resource usage and quality were measured, the difficulty of implementing a workflow from scratch was assessed, and the costs of sequencing were estimated based on quotes and estimated throughput. As a result of this research, a number of recommendations and guidelines were able to be made that may assist with project planning and costing a de novo assembly project. In particular: ,Ä¢ Paying an experienced person at a higher rate may result in overall project savings ,Ä¢ Unless the genome is very large, the main component of project expense is generally labour ,Ä¢ Accurately estimating the cost of sequencing can often only be achieved by actually doing some sequencing ,Ä¢ HiFi read assemblies are more contiguous than CLR assemblies ,Ä¢ Flye is an accurate, easy to use assembler of both CLRs and HiFi reads with large RAM requirements ,Ä¢ Where RAM is limited, Canu (CLR or HiFi reads) or Hifiasm (for HiFi reads) perform well

Abstract #111

Barbara Brito Rodriguez (University of Technology Sydney, Sydney, NSW)
Viral RNA metagenomics reveals the Australian bovine respiratory virome

Metagenomic next-generation sequencing is transforming public health and has the potential to do the same with the agricultural industries and national biosecurity. Beef production contributes to 1.5% of Australia,Äôs key industry GDP. The most impactful infectious disease in intensive beef cattle production is bovine respiratory disease (BRD). Traditionally a limited number of bacteria and viruses were thought to be associated with respiratory disease. Recent metagenomic studies have unveiled a number of additional viruses in the respiratory tract of bovines, potentially associated with respiratory disease. We collected pharyngeal swabs from two feedlots in NSW to characterise the virome of beef cattle in Australia. The samples were collected in viral transport media and transported on dry ice to the Bioscience laboratory at the University of Technology Sydney. The RNA was extracted using Qiagen RNeasy Micro kit. Samples were pooled to obtain a minimum of 3 ug of RNA. Library preparation was done using the Illumina TruSeq¬Æ Stranded (human, mouse, rat) kit for library preparation and sequencing was run in an Illumina NovaSeq S1 Lane, 300 cycles (150bp paired) at AGRF. A total of 353 Gb of data was generated from 15 pooled libraries. The Bos taurus genome assembly (GCA_002263795.2) was used to filter out host reads using bwa mem and samtools. The reads were assembled and taxonomically classified using the Genome Detective online platform. Briefly, low-quality reads are filtered out using trimommatic. Then, the reads are classified using DIAMOND and the Swissprot Uniref90 protein database. The reads are sorted into groups (bins). Each bin contains reads of one viral species. De novo assembly was done in SPAdes for each bin. The contigs were then classified using NCBI RefSeq virus database. Contigs were joined and consolidated using the Advanced Genome Aligner. The reads were also queried to investigate the presence of antimicrobial resistance genes from the metatranscriptome obtained using ResFinder. Using an NGS RNA metagenomics approach we obtained near whole-genome sequences of multiple viruses, including Bovine Nidovirus, Bovine Rhinitis A, Bovine Viral Diarrhea Virus-1, Bovine Betacoronavirus, Enterovirus E, as well as the partial genome of Bovine Ungulate tetraparvovirus 1, and a novel paramyxovirus. We also found smaller contigs (sequences) that suggest the presence of segments of Influenza D, Bovine Rhinitis B, Ungulate bocaparvovirus 6, Influenza D, Bovine Respiratory Syncytial virus and Parainfluenza 3. Surprisingly, the most abundant virus in feedlot cattle was Bovine Nidovirus, a virus that was only recently discovered in the US (2015) and that has only been reported in two other countries to date. One antimicrobial resistance gene was consistently identified in several pools: Beta-lactam bla-TEM-116. Although none of the viruses found in this study are reportable, they are part of an unexplored respiratory complex in Australian cattle.

Abstract #112

Charlotte Francois (La Trobe University, Melbourne, VIC)
New insights into plant-microbe interactions through Quantitative Trait Locus (QTL) mapping

Like humans, plants are colonised by bacteria on virtually every surface. Surrounding roots is a carefully-regulated region known as the rhizosphere, which supports a special subset of bacteria known as beneficial rhizobacteria. Plants regulate rhizosphere structure and function through the secretion of root exudates that contain sugars, secondary metabolites and phenolics suspended in an exopolysaccharide mucilage. The activity and makeup of the rhizosphere-associated bacterial community is similarly regulated by the plant host. Beneficial rhizobacteria provide numerous benefits to their hosts, including but not limited to improved nutritional status and enhanced stress and pathogen tolerance. In exchange, beneficial bacteria receive sugars and other metabolites. This ancient relationship is representative of a long coevolutionary process and selective pressure on both partners has been applied. The magnitude of the benefit conferred is partially dependent on the host plant genotype and many studies have demonstrated that there is genetic variation for this trait. Studies using Arabidopsis thaliana have shown that varieties within a plant species respond differently to a single beneficial rhizobacterial species, which suggests that there is host genetic variation for this trait. It is therefore expected that rhizosphere-associated traits in plants are adaptive traits, and that an enhanced ability to gain from beneficial rhizobacteria may be evolutionarily advantageous, having consequences that affect plant fitness. The genes underlying this variation are currently unknown, but they represent attractive targets in emerging biotechnologies that seek to exploit the interaction between plants and beneficial bacteria, such as phytoremediation. Using a small Arabidopsis recombinant inbred line (RIL) mapping population combined with Quantitative Trait Locus mapping and candidate gene identification in R, a computational genomic analysis was conducted to understand the complex interaction between plants and their beneficial rhizobacteria, elucidate underlying genes and provide new insights into this variation.

Abstract #113

Ellis Patrick (University of Sydney, Sydney, NSW)
Spatial analysis of in situ cytometry data

Understanding the interplay between different types of cells and their immediate environment is critical for understanding the mechanisms of cells themselves and their function in the context of human diseases. Recent advances in high-parameter in situ cytometry technologies have fundamentally revolutionized our ability to observe these complex cellular relationships providing an unprecedented characterisation of cellular heterogeneity in a tissue environment. We will introduce an analytical framework, spicyR, for analysing data from high-parameter in situ cytometry assays including CODEX, CycIF, IMC and High Definition Spatial Transcriptomics. Ultimately, this framework is applicable to any assay that has undergone single-cell segmentation and produces information on a cells location in x-y space and the abundance of various molecular markers on the cell. Using point process models, our framework includes methodology for testing if the colocalisation of two cell-types has changed between two groups of subjects. While not necessary, this methodology can benefit from situations when multiple measurements are performed for each subject. We will also demonstrate how point process models can be exploited to identify consistent spatial organisation of multiple cell-types in an unsupervised way. This can be used to enable the characterization of interactions between multiple cell-types simultaneously and can facilitate the identification of distinct tissue compartments or identification of complex cellular microenvironments.

Abstract #114

Yue Cao (University of Sydney, Sydney, NSW)
Benchmarking single cell RNA-sequencing simulation methods

Single cell RNA-sequencing (scRNA-seq) is a powerful technique for profiling the transcriptomic at the single cell resolution and has gained tremendous popularity since its emergence in 2009. In recent years, there has been an increasing number of simulation tools designed specifically for simulating scRNA-seq data. For simulation data to be useful to aid in the development of analytical algorithms, simulation methods must generate faithful and biologically meaningful representation of the scRNA-seq data. Using a systematic framework, the aim of our study is thus to evaluate each method at capturing the underlying biological structure of scRNA-seq datasets. We survey the literature and include a total of 12 simulation methods in the evaluation framework. We select over 40 datasets across a variety of tissue types, biological conditions and sequencing platforms to ensure diversity in the framework. The evaluation framework considers a range of criteria, including both marginal distribution and joint distribution. Marginal distribution concerns simple parameter estimates such as mean and variance distribution of gene expression. Joint distribution considers the relationship between parameters, such as the mean-variance relationship. Using multiple metrics, we found that there exists discrepancies in methods performance, where some methods performed consistently better than others by a large margin. In addition, some more recent methods published within these two years do not necessarily outperform the methods that were published in earlier years. We also discovered some parameters such as gene correlation are harder to be captured by current methods modelling than other parameters such as cell correlation. Discrepancies in time consumption was also observed, with some methods able to model 8000 cells within 3 hours and some failed to model beyond 1500 cells. Overall, we‚Äôve identified recommended methods for users and areas that could benefit from further improvement for methods developers.

Abstract #115

Heroen Verbruggen (University of Melbourne, Melbourne, VIC)
A workflow for the detection and phylogenetic placement of eukaryotes from metagenomes

Microbial eukaryotes (protists) play key roles in global element cycles, ecosystem functioning, food production and disease, but relatively little genome data are available for microbial eukaryotes, limiting our understanding of their biodiversity, evolution and roles in the environment. As has been the case for prokaryotes, metagenomic data could shed some light on this unknown diversity via metagenome-assembled genomes, but most existing tools are geared towards bacteria and do not function well for protists. We have designed a workflow for automated download of metagenome reads, data cleaning, assembly, computation of contig statistics (k-mer frequencies, read coverage) and metagenome binning. The workflow is implemented in WDL and dependencies packaged in a Singularity container, allowing us to deploy it on a wide variety of platforms. We have also implemented several new methods to extract eukaryote organelle genomes from the assembled data and to place them in the eukaryote tree of life. A large library of publicly available metagenomes were processed, resulting in the detection and phylogenetic placement of a large diversity of eukaryotes, many of which unknown to science. This improves our knowledge of eukaryote evolution and delivers insights into the occurrence of these eukaryotes in natural environments.

Abstract #116

Nicolas Canete (University of Sydney, Sydney, NSW)

SpicyR - Spatial analysis of in situ cytometry data

Highly multiplexed imaging techniques such as cyclic immunofluorescence (CycIF), as well as the mass cytometry based techniques imaging mass cytometry (IMC) and multiplexed ion beam imaging (MIBI), has allowed many antibody parameters to be visualised within the same image. These technologies enable a variety of distinct cell types to be analysed concurrently in their native microenvironment. Indeed, there is increasing evidence that cell microenvironments are programmed not just by cell ontogeny, but by signals from the microenvironment. Highly multiplexed imaging hence provides the necessary spatial data required to interrogate the role of the microenvironment and identify any interdependencies between complex cell subsets in health and disease. Standard image processing, cell segmentation, and cell classification, the phenotype of a single cell and its position within an image can be identified. However, the analysis of the spatial relationships between these cells can be difficult. Here, we present the statistical use of marked point process models to identify the spatial relationship between different cell types. Specifically, these models can be used to identify specific localisations and avoidances between pairwise cell types. This approach can provide insights for immune responses in cancer models, autoimmune disorders such as type 1 diabetes and multiple sclerosis, and in infectious disease such as HIV. Such models can prove to be useful for applications in other high parameter disease models.

Abstract #117

Belinda Phipson (Peter MacCallum Cancer Centre, Melbourne, VIC)
propeller: finding statistically significant differences in cell type proportions in single cell RNA-seq experiments

Single cell RNA Sequencing (scRNA-seq) has rapidly gained popularity over the last few years for profiling the transcriptomes of thousands of single cells, enhancing our understanding of complex tissues, and enabling the discovery of novel cell types. However, scRNA-seq data is not without significant analytical challenges. The low amount of starting RNA, combined with shallow sequencing, leads to data that is noisy and sparse. Technical variation, such as batch effects, can be larger than the biological signal. To date, a large number of single cell-specific software tools have been developed, focusing on visualisation, clustering and trajectory analysis with the aim to discover and define new cell types. With the maturation of the technology, there is a shift towards applications comparing cell type composition and gene expression of samples between experimental conditions. In the context of bulk RNA-seq, the main goal of analysis between multiple groups is to find significant differentially expressed genes. It has not been possible to deconvolve the difference between genes that are lowly expressed across the majority of cells making up a sample and genes that are highly expressed in a small proportion of cells. With scRNA-seq data we can not only compare expression levels between cell types but also the cell type composition of samples between conditions. Variation in cell type proportions in a sample can be due to differences in capture efficiency and dissociation protocols. In addition, in a designed experiment with multiple samples, there is additional variability due to biological (sample-to-sample) variation. These sources of variability need to be taken into account when testing for differences in cell type proportions between conditions. Here I will demonstrate our new method for finding statistically significant changes in cell type proportions between experimental conditions. I will focus on a relatively large (>54,000) single nuclei heart dataset that profiles fetal, young and adult human heart samples, with three biological replicates in each group. The samples were obtained from heart biopsies and are highly heterogeneous. When analysing differences in proportions with a classic chi-square test, we found that every cell type was highly significantly different between developmental time points. In contrast, when applying our new method, propeller, which accounts for the biological variability between samples, we found 4 of the 8 broad cell types significantly different between fetal, young and adult groups. Using simulations we show that methods that do not account for additional biological variability, such as chi-square tests, do not adequately control the false discovery rate. The propeller test models transformed proportions based on empirical Bayes linear modelling and is thus highly flexible and robust. The propeller method is available in the speckle R package (https://github.com/Oshlack/speckle).

Abstract #118

Al J Abadi (Melbourne Integrative Genomics, University of Melbourne, Melbourne, VIC)
Integrating multi-modal single-cell studies with a latent component-based approach

Single-cell multi-modal sequencing offers unprecedented insights into complex cellular processes such as embryonic development and complex disease. Integrative methods which leverage the multitude of modalities have helped to gain novel molecular insights which were not possible using single-omic approaches. These insights include epigenetic mechanisms governing cell fate decisions in mouse and deciphering chromatin accessibility heterogeneity in interneuron subpopulations. However, these methods often focus on integration of data from independent cell populations which are presumably related. These approaches are also limited by the common features across various modes to anchor the linked molecular layers (e.g. Seurat and LIGER). By contrast, for the same cell populations, MOFA+ uses non-negative matrix factorisation to characterise cellular heterogeneity across modes. However, it ignores genomic interactions by assuming independence between features. We apply a multivariate approach based on projection to latent structures (multiblock PLS, Tenenhaus et al. 2014. Biostatistics 15) to integrate multiple single cell datasets and to characterise the coordinated variation among modalities. The method makes no assumptions on data distribution or statistical independence among features and enables the selection of highly correlated variables across different datasets. In particular, by building sets of latent components which maximise the sum of covariance between transcriptome and the other epigenetic modalities, we can identify most

relevant biomarkers to resolve the linked cellular heterogeneity. Using this framework, we integrate 13 modalities from multiple stages of early mouse embryo during gastrulation (Argelaguet et al. 2019. Nature 576). The measurements from 826 matching cells include the transcriptome, DNA methylation in gene bodies, promoters, enhancers (P300 and CTCF), CpG islands, and DNase Hypersensitive sites, as well as chromatin accessibility for the same regions. Using the available phenotypic data as well as gene set enrichment analysis, we demonstrate the relevance of our approach to identify biomarkers and potentially reveal novel genomic interactions. We are currently extending multiblock PLS to account for uncertainty in epigenomic measurements, and a supervised analysis to characterise the stages of gastrulation using our statistical method.

Abstract #119

Jack Clarke (University of New South Wales, Sydney, NSW)
The role of gene duplication in the evolution of snake venoms

Snakes are one of the most venomous animals on the planet, using their venom for defence and the capturing of prey. Snake venoms have evolved independently of other venoms in other vertebrates, and there is considerable variation between species in their proteomic composition. One of the primary mechanisms through which snake venoms are thought to evolve is the duplication, recruitment and specialisation of proteins from other tissues. In some cases, this evolution is known to involve the tandem duplication of genes resulting in chromosomal clusters of venom genes in some gene families. We have recently sequenced and assembled the genomes of two highly venomous Australian snakes: Notechis scutatus (mainland tiger snake) and Pseudonaja textilis (eastern brown snake). In conjunction with publicly available proteomes from 10 other venomous snakes and 2 non-venomous snakes, these genomes provide an excellent opportunity to examine the role that duplication and neofunctionalisation has played in snake venom evolution. We have analysed 43 protein families known to play a role in snake venom and examined their pattern of duplication in snakes, compared to high quality reference genomes of other reptiles and non-venomous vertebrates. We find evidence for extensive duplications across some of these families, but no clear enrichment for duplication in the evolution of venom specifically. Instead, we identify a trend where numerous duplications specific to venomous snakes occur in proteins that seem predisposed to evolve by duplication and specialisation, even in non-venomous vertebrates. A subset of high-quality snake genomes was then used to further explore the nature of duplications. While tandem gene duplication is evident in some larger families, it remains absent in many. The snake venom metalloproteinase (SVMP) family provides an excellent case study, with multiple duplication events throughout its evolutionary history in vertebrates. Part of the broader ADAM (‚Äúa disintegrin and metalloproteinase‚Äù) family of single-pass transmembrane and secreted zinc proteases, SVMP appears to have expanded by independent tandem duplications in different snake lineages. We also identify a second ADAM subfamily, ADAM20, with an abundance of venomous snake-specific duplications. Ongoing work in exploring the possible role of ADAM20 proteins in snake venoms and the role that genome assembly quality has played in our ability to robustly detect the presence or absence of gene duplication events.

Abstract #120

Nitika Kandhari (Monash University, Melbourne, VIC)
Finding signatures of alternative polyadenylation as cancer biomarkers

Alternative transcript cleavage and Polyadenylation (APA) is linked with cancer cell transformation and proliferation. APA gene biomarkers are strong predictors of clinical outcomes. If incorporated to standard prognostic measures such as gene-expression and clinical parameters, these could reform cancer prognostic testing and therapy. Nearly 70% of mammalian genes harbour APA sites resulting in distinct transcripts with variable 3‚ÄôUTR length and regulatory content. Short 3‚ÄôUTRs are generally associated with de-differentiated proliferative cells (eg, stem cells) whereas longer 3‚Äô-UTRs associate with more complex regulation and cellular specialisation. ~91% APA genes switch to shorter mRNA isoforms in tumour cells. Our study aims to detect signatures of APA changes that are specific to triple-negative breast cancer (TNBC) that could be applied as a novel prognostic biomarker in early-stage breast cancer. Using bioinformatic analyses of 3‚Äô-focused RNA-seq approaches we studied the landscape of transcription and APA in three cancer cell lines in response to loss of PCF11, a core regulator of 3‚Äô-end formation. This shows a conservation of an expression and processing response to loss of 3‚Äô-end processing machinery. In addition to gene expression changes, we identify 3‚Äô-end ‚Äúshifted‚Äù signature genes that are common to all 3 cell lines. These signature genes highlight a different biology than gene expression. We will present our current work around the idea that systematic lengthening of 3‚ÄôUTR interferes with key signalling pathways and sensitizes cells to PTEN/AKT inhibitor.

Abstract #121

Mathieu Fourment (University of Technology Sydney, Sydney, NSW)
Fast variational Bayesian inference in phylogenetics

Markov chain Monte Carlo algorithms have been the workhorse of Bayesian inference in phylogenetics for almost two decades. Although these algorithms have been successfully used in a wide range of applications they do not scale well to large numbers of sequences. Recent advances in statistical machine learning techniques have led to the creation of probabilistic programming

frameworks. These frameworks enable probabilistic models to be rapidly prototyped and fit to data using scalable approximation methods such as variational inference. In this talk I will present some work on phylogenetic variational inference using phylostan, a Stan-based tool. Because trees are unusual statistical objects, phylostan tends to lack the flexibility required for accurately modeling molecular evolution. We can meet this challenge by using more general modeling frameworks such as Pytorch. This framework brings in an efficient machine learning toolbox in a popular programming language that can be easily extended. I will emphasize the current challenges in the new field of phylogenetic variational inference.

Abstract #122

James Hogan (Queensland University of Technology, Brisbane, QLD)
Metagenomic Geolocation with Read Signature Clustering

Metagenomic sequencing produces large quantities of reads to characterise environmental samples. These reads can be binned and assembled, or simply fed into a fast sequence classification tool such as Kraken, but only at enormous computational expense. We present a novel approach that takes an entire metagenomic sample and reduces it to a small number of vectors that characterise the underlying sample effectively enough that they can be used to predict its geographic origin. In this presentation we will introduce an approach wherein we compute read signatures using random orthonormal k-mer vectors and cluster them down to a small number of centroids that retain much of the expressivity of the original samples. We will then show, using ground truth from the CAMDA Metagenomic Location Challenge, that we are able to use the resulting read signatures in conjunction with a nearest-neighbour classifier to predict the geographic locations of many of the metagenomic samples.

Abstract #123

Sally Wasef (Griffith University, Brisbane, QLD)
How Ancient genomes can help Aboriginal Australian communities: lessons from the Cape York project

Paleogenetics is a relatively new and promising field that has the potential to provide new information about past Indigenous social systems, including insights into the complexity of burial practices. Yet, DNA evidence alone would not be enough. We present results from the first interdisciplinary study in Australia to incorporate modern and ancient genomics, isotopes, bioarchaeology and new archaeological excavations in order to provide important insights into the question of the population history, repatriation of Indigenous remains and possibility of late contact with Melanesian populations in Cape York. We reveal information that provides insights into gene flow from the north but note that the pattern is complex. Equally complex is the mortuary record from Flinders Island as revealed through the isotope and mitogenomic data, and we compare this to other groups in Cape York and Queensland. The significant expansion in contemporary and ancient genomic research over the last 5 years has unlocked powerful new data to investigate the question of possible genomic admixture events between Australian and Papuan populations. However, there have been several new studies that have provided contemporary papuan genomic data, greatly expanding the potential to reconstruct later phases of the region‚Äôs population history. Unfortunately, parallel Aboriginal Australian datasets do not currently exist for genomic datasets. From a genomics perspective, Australia remains relatively understudied and for some locations restricted to mitochondrial data only. Further limiting the opportunity for comparison is the issue of access to these datasets, particularly the Aboriginal Australian whole modern genomic data, which is one of the largest existing datasets and currently is not publicly shared. The restrictions on who can use these data is a point of considerable concern for many geneticists and other researchers Recently, we completed an Australian Research Council-funded project that focused on human remains from the Cape York Peninsula of Queensland, in collaboration with several local Aboriginal communities. What we found suggests no single method such as DNA testing or using geological clues will be enough to reliably determine the origin of remains ‚Äì an interdisciplinary approach using all available evidence will be required to repatriate Aboriginal remains. Interdisciplinary approaches employing a range of techniques commonly applied in archaeology elsewhere in the world can only be possible in Australia when undertaken in close collaboration and partnership with Traditional Owners. By reproducing the past by incorporating the stories from the people themselves, and not just the artefacts and material culture they left behind, we have the potential to provide very important and new insights into the stories from Aboriginal Australia.

Abstract #124

Daniel Russell (Griffith Institute for Drug Discovery, Griffith University, Brisbane, QLD)
Developing a computational analysis to identify differentially allelic expressed loci in patient-derived stem cells

Epigenetics is the inherited and acquired modifications in the genome that affect the gene activity without changing the DNA sequence. Instead the epigenome is determined by chemical modifications within the genome such as the modification of histone proteins and the methylation of DNA which in turn change gene regulation. Epigenetics plays a vital role in the modulation of gene activity, especially during embryonic development and in brain function. Recent research has found that it also plays a role in disease, particularly neurodegenerative and neuropsychiatric diseases such as Alzheimer‚Äôs disease and schizophrenia. An intriguing type of epigenetic regulation is genomic imprinting, where one of the two alleles of a particular gene are silenced permanently depending on its parent of origin. Other instances of allele-specific expression are of great interest as when the expression of a heterozygous gene deviates

towards one allele, it becomes functionally homozygous. In the case of risk alleles, this level of deviation may increase its impact in the onset of polygenic diseases. This study aims to develop a computation analysis to identify protein-coding genes with allele-specific expression by integrating whole genome sequencing (WGS) data and RNA sequencing (RNA-seq) data from the same patient-derived stem cells. We implemented a computational pipeline to determine differential allele expression based on read counts from RNA-Seq data of all heterozygous loci identified in WGS data from the same patient-derived stem cells. Heterozygous variants that showed statistically significant biased expression were validated using a gold standard data set for known imprinted genes and available epigenome maps. Our pipeline may be useful for finding epigenetic regulated loci associated with diseases using patient-derived stem cells. As a proof-of-concept we used olfactory neuronal stem cells from a well-studied schizophrenia cohort. Many of these epigenetically regulated loci have been previously identified to be associated with schizophrenia. We propose heterozygous risk alleles located in these loci may have very different ranges of functional effects based on their allelic-biased expression which is defined by epigenomic regulation.

Abstract #125

Vincent Corbin (Peter MacCallum Cancer Center, Melbourne, VIC)
Moving beyond RNA sequence: uncovering the functional role of RNA structure

RNA is unique among biomolecules in its possession of both information storing and enzymatic capacities. In addition to the genetic code itself, the conformation of RNA secondary structure can regulate processes as varied as splicing, localisation, translation efficiency and protein binding. Existing assays based on RNA chemical probing (DMS, SHAPE) have attempted to explore RNA structure in the cell, however to date they have assumed structural homogeneity and derived a single average estimated structure, leading to a potentially false structural interpretation and an inability to untangle the multiple conformations of RNA inside a single cell. We developed a method capable of detecting alternative RNA structures which form from the same underlying sequence, both in vitro and ex vivo. We applied this to the RNA retrovirus human immunodeficiency virus-1 (HIV-1), and successfully revealed genomic structure heterogeneity with novel functional implications. HIV-1 has a 10-kb single-stranded RNA genome. It expresses all gene products from the same primary transcript, which undergoes alternative splicing to produce diverse protein products which include structural proteins and regulatory factors. Despite the critical role of alternative splicing, the mechanisms driving splice-site choice are poorly understood, as HIV-1 does not encode any of its own splicing factors. We used standard DMS-MaPseq to probe the structure of HIV-1 RNA in cells, and developed a novel algorithm called Detection of RNA folding Ensembles using Expectation-Maximization (DREEM), which successfully revealed alternative conformations assumed by the same RNA sequence [1]. DMS-MaPseq highlights nucleotides within unpaired sections of RNA structures by mutating them. DREEM models the mutational profile of the reads using a Mixture of Multivariate Bernoulli distributions, and uses an Expectation-Maximization algorithm to cluster the reads into groups corresponding to distinct RNA conformations. Contrary to previous models which analyzed population averages, our results revealed the widespread heterogeneous nature of HIV-1 RNA structure. In addition to confirming in vitro characterized alternative structures for HIV-1 exists in cells, we discovered alternative conformations at critical splice sites which influence the ratio of transcript isoforms. Our simultaneous measurement of splicing and intracellular RNA structure provides the first evidence for the long-standing hypothesis that RNA folding regulates splice site usage, and indicates a major role for RNA conformation heterogeneity in regulating RNA gene expression. [1] Corbin VDA*, Tomezsko P*, Gupta P et al. Determination of RNA structural diversity and its role in HIV-1 RNA splicing. Nature 582, 438-442 (2020)

Abstract #126

Gayathri Thillaiyampalam (Griffith Institute for Drug Discovery, Griffith University, Brisbane, QLD)
Developing a systems-based analysis of miRNA target networks

microRNAs (miRNAs) are regulatory RNAs (19 ,Äì 24 nt) critical for the control of gene expression via direct binding to 3,ÄôUTR of messenger RNAs (mRNAs) to suppress the protein synthesis. One of the major challenges in miRNA studies is identifying their gene targets. Even though there are enormous number of tools developed to identify miRNA targets based on the studies elucidating miRNA biology this remains a challenging task for researchers due the complexity in miRNA functions. Computational algorithms often resulted in high fraction false positive results (~70% accuracy is acquired) in predicting miRNA targets. In recent years there has been significant progress in understanding the role of microRNAs and their potential involvement in several biological processes and diseases through high-throughput technologies such as high-throughput sequencing of RNA isolated by crosslinking immunoprecipitation (HITS-CLIP) transcriptomic profiling of miRNA targets. Here we are proposing a systems-based analysis of miRNA ,Äì gene network which can be modulated to gain/loss functions of biological interest. We developed an enumerated motif-based miRNA target prediction model based on the sequences of enriched mRNAs from HITS-CLIP pulldown. We identified the enriched small motifs which are enumerated list of sequences generated from the mature miRNA sequence and trained an SVM model on training data set (70% of HITS-CLIP data) using the enriched motif counts. The trained model showed ~80% accuracy while tested on the remaining 30% of test data set. Target genes were predicted based on the motif enrichment and the model reported different weights for each motif in separating the targets and non-targets. To generalise the application of our model we applied the same approach to an independent HITS-CLIP pull down data and the model performed well with ~84% accuracy in predicting the gene targets. Taking together our enumeration method based on enriched small motifs yields higher accuracy in predicting miRNA targets

which can be adapted for a better miRNA target prediction tool and further characterisation of miRNA ‚Äì gene network and biological pathway inference.

Abstract #127

Luis Pedro Coelho (Fudan University, Fudan, China)
The AMPSphere: antimicrobial peptides (AMPs) in the global microbiome

Antimicrobial peptides (AMPs) are short peptides that inhibit the growth of microbial organisms Particularly as antimicrobial resistance becomes a public-health crisis, AMPs can be valuable for use in both clinical and industrial applications. At the same time, it is important to recall that these molecules did not evolve for human benefit, but they must play a role in the natural habitats and we expect that they have an impact in structuring microbial communities. Due to their small size, standard gene mining approaches are not directly applicable to short peptides. To address this problem, we previously developed macrel, a machine-learning based pipeline to find AMPs in (meta)genomes. Having validated macrel on benchmark and test datasets, we then applied it to ProGenomes2, which includes 86 thousand high-quality genomes, as well as a dataset of over 35,000 publicly available metagenomes. After redundancy removal, we produced a collection of AMPs from the global microbiome, which we termed the AMPsphere. This first version of the AMPsphere contains 317,790 distinct amino acid sequences (97.1% of which were found exclusively in metagenomes). These sequences were further clustered by sequence similarity into 4,705 AMP families by sequence similarity. While macrel finds peptides that are present as short genes (as opposed to being formed by fragmentation of larger proteins), 9% of the AMPsphere peptides could be matched to larger proteins in the eggNOG databases. These can represent cryptic peptides that originated as fragments and have evolved to become independent genes. Current work aims at elucidating the impact of AMPs on microbial communities and quantifying the extent to which they play a role in structuring the community.

Abstract #128

Mark Read (University of Sydney, Sydney, NSW)
Using single-cell cytometry to illustrate the generalisable unbiased evaluation of clustering algorithms using Pareto fronts

Clustering is widely used in biological fields such as microbial ecology, genomics, and cytometry to partition cells on basis of similarity. Many automated gating algorithms now exist to cluster cytometry and single cell sequencing data into discrete populations. Comparative algorithm evaluations on benchmark datasets rely either on a single performance metric, or a few metrics considered independently of one another. However, single metrics are biased as they emphasise different aspects of clustering performance and hence differ in how clustering solutions are ranked. This undermines the translatability of results onto other non-benchmark datasets, and underlies the lack of consensus regarding optimal clustering algorithms in the field. We propose the Pareto fronts framework as an integrative evaluation protocol, wherein individually biased metrics are instead leveraged as complementary perspectives. Judged superior are algorithms that provide the best trade-off between the multiple metrics considered simultaneously. This yields a more comprehensive and complete view of clustering performance. Moreover, by broadly and systematically sampling algorithm parameter values using the Latin hypercube sampling method, our protocol discounts (un)fortunate parameter value selections as confounding factors. Furthermore, it reveals how meticulously each algorithm must be tuned in order to obtain good results, vital knowledge for users with novel data. We exemplify the protocol by conducting a comparative study between two clustering algorithms using four common performance metrics applied across four cytometry benchmark datasets. To our knowledge, this is the first time Pareto fronts have been used to evaluate the performance of clustering algorithms in any application domain.

Abstract #129

Xiangnan Xu (University of Sydney, Sydney, NSW)
A multi-step model for micobiome data with application to Parkinson's disease prediction

Parkinson's disease (PD) is one of the most common neurodegenerative diseases and increasingly studies highlight that imbalances in the composition of the gut microbiome may play important roles in the occurrence and progression of PD. Statistical and machine learning methods such as lasso, support vector machine, and random forest have been used to predict the occurrence of PD using microbiome composition. However, extensive modulating factors such as dietary intake have great impact on microbiome composition, which is a major source of heterogeneity in datasets and poses particular challenge on the model‚Äôs ability to predict well. Here, we propose a multi-step model to predict PD, incorporating both nutritional information and microbiome composition. The model first builds classifiers using microbiome composition and a cross validation procedure is used to determine if an individual can be reliably classified. Then, a decision tree using nutritional information is built to explain the outcome of the microbiome classifier. Next, a decision tree is constructed to divide the heterogeneous samples into several sub-groups. Finally, we build classifiers within each sub-group to predict PD state. When a new sample comes in, the decision tree will first determine which sub-group it belongs to, then the classifier in this sub-group will predict whether it is PD. We apply our model on a study consisting of 103 PD patients and 81 healthy controls. In this study, gut microbiome profiles were characterised using high-throughput sequencing, targeting the 16S rRNA gene and the matched nutritional information were derived from questionnaires. Cross validation results show that when splitting the samples

using carbohydrate intake, the AUC of the classifier can be improved from 0.66 to 0.75. This demonstrates that our multi-step model can count for the heterogeneity in dataset and betters prediction.

Abstract #130

Givanna H. Putri (University of Sydney, Sydney, NSW)
TrackSOM: immunopathogenic temporal mapping through clustering time-series cytometry data

Disease, and our immune response to it, are dynamic in time. Under challenge, the immune response deviates from homeostasis to create many immune cell populations that enact effector function in a coordinated fashion. The process's fallibility is attested to by autoimmunity, non-communicable diseases and lethal infections. Effective intervention requires holistic characterisation of these immune and resident tissue cell populations, and their spatio-temporal dynamics. Single-cell cytometry and, more recently, RNA sequencing methods are a mainstay technique for profiling such cell populations at single time-points. Interest in profiling evolving immunopathogeneis through a time-series of such assays is growing. Yet, whilst computational algorithms to support the automated identification of cellular phenotypes within such data have emerged, technologies to annotate temporal dynamics of cell populations are scarce. We present here TrackSOM, an automated clustering and temporal tracking algorithm tailored to operating over a time-series of cytometry datasets. It can articulate cellular infiltration, differentiation and functional dynamics as they evolve during disease resolution development, remission and relapse. TrackSOM can capture moving, growing, shrinking, splitting, merging and transient clusters. With TrackSOM we map out the evolving immune response in the bone marrow and brains of West Nile virus-infected mice. We detect infiltrating macrophages in brains at day 2-post infection. Interestingly, as soon as day 1-post infection, CNS-resident macrophages (microglia) commence a split into two phenotypes, with one merging with that of infiltrating macrophages, the other remaining largely invariant. TrackSOM is built upon the popular and computationally expeditious FlowSOM algorithm, extensively used within cytometry, thus facilitating ready adoption by cytometrists and bioinformaticians.

Abstract #131

A.J. Sethi (John Curtin School of Medical Research, Australian National University, Canberra, ACT)
A machine learning model to predict splice factor expression directly from transcriptome-wide splicing patterns

Dysregulated splicing is a major driver of cancer and inherited genetic disease, of which the underlying mechanisms are poorly understood. Variation in splicing outcomes (isoform usage) can arise from several sources, including the differential expression of splice factors (SF). These proteins regulate spliceosome assembly, and may function both synergistically and redundantly to coordinate splice-site selection. Although numerous studies have characterised transcriptome-wide differential splicing patterns following the depletion or overexpression of individual SFs, the observed differential splicing phenotypes are always a cumulative effect of the experimental treatment in combination with the compensatory expression of other, interdependently-regulated splice factors. As such, these conventional methods fail to link individual splice factors directly to the alternative splicing events which they regulate. Our study aims to quantify the impact of the siRNA depletion of 56 individual splicing-related proteins on the global expression of splice factors, and to further to measure the resultant alternative splicing patterns in a publicly available dataset in drosophila (Brooks et al. 2015). Using this data, we aim to develop a statistical learning model to understand the complex relationships between transcriptome-wide splicing patterns (i.e. exon percentage spliced-in) and the underlying splice-factor transcript expression levels. This model will allow for the prediction of SF expression levels directly from exon inclusion values. Using this information, we will gain further insight into the roles of individual SF in controlling alternative splicing. Furthermore, we will gain insight in how we may be able to modulate disease-relevant alternative splicing evets directly at the splice factor level. In contrast to previous approaches to understand the roles of splice factors, our model will directly quantify the proportion of variation in isoform usage that is driven by differential splice factor expression, clearly delineating the effects of SF from the other mechanisms which regulate pre-mRNA splicing. References: Brooks AN, Duff MO, May G, et al. Regulation of alternative splicing in Drosophila by 56 RNA binding proteins. Genome Res. 2015;25(11):1771-1780. doi:10.1101/gr.192518.115 ,ÄÉ

Abstract #132

Clare Sloggett (Peter Doherty Institute for Infection and Immunity & University of Melbourne, Melbourne, VIC)
AusTrakka ,Äì Working towards integrated pathogen genomics for SARS-CoV-2

The COVID-19 pandemic has highlighted the need for SARS-CoV-2 genomics sequencing and analysis to track transmission and identify emerging clusters and outbreaks. Simultaneously, the need to share genomic data between jurisdictions has become apparent to highlight inter-jurisdictional spread of SARS-CoV-2. Here we describe the development and early implementation of the AusTrakka genomic data sharing platform and describe its utility for nationally integrated SARS-CoV-2 genomics. The development of the AusTrakka platform by the Communicable Diseases Genomics Network (CDGN) was accelerated to address the urgent need for timely SARS-CoV-2 genomic data sharing between public health laboratories to contribute meaningfully to public health intervention. Over six months, AusTrakka has built in capacity for public health laboratories to upload SARS-CoV-2 consensus genome sequences that are analysed by bioinformatic pipelines and tools integrated into the platform, as well as capability to generate and visualise national

phylogenetic trees that identify and notify relevant jurisdictional public health laboratories of genomic ‚Äòmatches‚Äô that may infer interstate transmission or emerging clusters for further investigation. Since its endorsement as the national SARS-CoV-2 genomic data sharing platform, over 13,000 SARS-CoV-2 genomic sequences have been uploaded to the platform from every state and territory, as well as New Zealand and has been used for national reporting against the Australian National Disease Surveillance Plan for COVID-19. The next phase of its development sees the integration of genomic and epidemiological data, required for SARS-CoV-2 to enable optimal high-resolution genomic clustering and determination of putative transmission events. The development of AusTrakka continues to be a highly consultative process, leveraging on Australia‚Äôs genomic and bioinformatics expertise to remove the barrier of entry into SARS-CoV-2 genomics for public health laboratories by bridging the gap between raw genomic sequence data, bioinformatic analysis and epidemiology for real-world application of public health pathogen genomics.

Abstract #133

Sasdekumar Loganathan (Australian Institute for Bioengineering and Nanotechnology, The University of Queensland, Brisbane, QLD)
Application of mixture model to RNA-seq data to discover ageing regulators

Ageing is a complex process where the combined effects of environmental and genetic factors make it challenging to isolate specific regulators of ageing. Moreover, both the variation in ageing between individuals and the variation between different tissues make the task of identifying regulators even more of a challenge. Previous studies that have modelled gene expression have been successful in identifying regulatory information by detecting novel genes and pathways. Mixture models can model variability through the detection of multimodality at the gene level from RNA-sequencing (RNA-seq) data. Therefore, we used mixture models to interrogate this variability to study the regulatory programs involved in ageing. We applied mixture models using three different distribution-based assumption to transcriptome-wide RNA-seq data for four human tissues (subcutaneous adipose, skeletal muscle, skin and whole blood) from the Genotype-Tissue Expression (GTEx) cohort. We identified lists of candidate genes that clustered according to multimodal distributions with donors that showed significant changes in age. A large percentage of these genes (71 ‚Äì 87%) have also been detected by standard differential gene expression (DE). However, genes that are unique to mixture models (MMUG) (ie. genes that are only detected by mixture models), have a similar percentage overlap to known known ageing databases (15.8%) as genes discovered by DE (15.2%). Furthermore, pathway over representation analysis of these MMUG on Hallmark, Reactome and Gene ontology (GO) databases have detected pathways not detected by DE genes (164) such as, TNFA signalling via NFKB in skeletal muscle. 75 pathways were common between MMUG and DE genes, indicating that mixture models potentially can detect different parts of the pathway such as TNFA signalling via NFKB for both blood and skin. The results indicate that modelling gene expression variability using mixture models can help uncover regulators that have a potential role in understanding human ageing. The application of unsupervised clustering from mixture models identified genes that had significantly altered ages between donors that corresponded to different modes. Showcasing the importance of not assuming a single mode when analysing RNA-seq data to model a gene‚Äôs expression distribution in a cohort. Using resources like the Digital Ageing Atlas and GenAge we confirmed that some of these genes have been previously implicated in ageing. Most genes not in these resources have also been implicated in ageing. Lastly, the results also indicate that mixture models detecting different genes in a pathway, which might lead to potential drug targets.

Abstract #134

David Chisanga (Olivia Newton-John Cancer Research Institute, Melbourne, VIC)
Integrated transcriptional and chromatin accessibility profiling uncovers sex-specific adipose tissue imprinting of regulatory T cells

Integrated transcriptional and chromatin accessibility profiling reveals sex-specific adipose tissue imprinting of regulatory T cells Adipose tissue, commonly referred to as body fat is a dynamic endocrine organ whose main function is as an energy repository together with cushioning and insulating the body. Especially, visceral adipose tissue (VAT) is critical for the regulation of systemic energy homeostasis by acting as a caloric reservoir. Therefore, the impairment of VAT function like in obesity is associated with insulin resistance and type 2 diabetes. Regulatory T (Treg) cells that express the transcription factor FOXP3 are critical for limiting immune responses and suppressing tissue inflammation, including in the VAT. Differences between sexes with respect to the physiology and organismal metabolism in adipose tissue are well documented across species. Here, we integrated transcriptome (RNA-seq) and chromatin accessibility (ATAC-seq) datasets to uncover distinct sexual dimorphism in Treg cells in the VAT. Our results showed that VAT in males was enriched for Treg cells in comparison to VAT in females, and Treg cells from VAT in males were strikingly different from their female counterparts in phenotype, transcriptional landscape and chromatin accessibility. Furthermore, increased inflammation in the VAT in males facilitated the recruitment of Treg cells via the CCL2,ÄìCCR2 axis. Androgen regulated the differentiation of a unique IL-33-producing stromal cell population specific to the male VAT, which paralleled the local expansion of Treg cells. Sex hormones also regulated VAT inflammation, which shaped the transcriptional landscape of VAT-resident Treg cells in a BLIMP1 transcription factor-dependent manner. Overall, we find that sex-specific differences in Treg cells from VAT are determined by the tissue niche in a sex-hormone-dependent manner to limit adipose tissue inflammation. References Vasanthakumar, A., Chisanga, D., Blume, J. et al. Sex-specific adipose tissue imprinting of regulatory T cells. Nature 579, 581‚Äì585 (2020). https://doi.org/10.1038/s41586-020-2040-3

Abstract #135

Rohan Williams (Singapore Centre for Environmental Life Sciences Engineering, Nanyang Technological University, Singapore)
Recovery of complete genomes and non-chromosomal replicons from activated sludge enrichment microbial communities using Nanopore long read metagenome sequencing

New long read sequencing technologies offer huge potential for effective recovery of complete, closed genomes from complex microbial communities. Using long read data (ONT MinION) obtained from an ensemble of activated sludge enrichment bioreactors, we 1) describe new methods for validating long read assembled genomes using their counterpart short read metagenome assembled genomes; 2) assess the influence of different correction procedures on genome quality and predicted gene quality and 3) contribute 21 closed or complete genomes of community members, including several species known to play key functional roles in wastewater bioprocesses: specifically microbes known to exhibit the polyphosphate-- and glycogen--accumulating organism phenotypes (namely Accumulibacter and Dechloromonas, and Micropruina and Defluviicoccus, respectively), and filamentous bacteria (Thiothrix) associated with the formation and stability of activated sludge flocs, and 4) demonstrate the recovery of close to 100 non--chromosomal replicons and a small microbial genome from order Saccharimonadales, that are present in these communities. Our findings further establish the feasibility of long read metagenome--assembled genome recovery, and demonstrate the utility of parallel sampling of moderately complex enrichment communities for recovery of genomes of key functional species relevant for the study of complex wastewater treatment bioprocesses.

Abstract #136

Chi Nam Ignatius Pang (University of New South Wales, Sydney, NSW)
RNase III-CLASH of multi-drug resistant Staphylococcus aureus reveals a regulatory mRNA 3,ÄôUTR required for intermediate vancomycin resistance

Treatment of methicillin-resistant Staphylococcus aureus (MRSA) infections is dependant on the efficacy of last-line antibiotics like vancomycin. Vancomycin treatment failure is most commonly linked to the emergence of vancomycin-intermediate resistance in clinical isolates (termed VISA). These isolates have not acquired resistance genes but appear to accumulate a heterogenous collection of single nucleotide polymorphism that collectively alter the physiology of the cell to increase vancomycin tolerance. Cell wall thickening is common among VISA isolates and is thought to decrease vancomycin permeability. Changes in regulatory sRNA expression have been correlated with antibiotic stress responses in VISA isolates however the functions of the vast majority of these RNA regulators is unknown. The 5,Äô and 3,Äô untranslated regions (UTRs) of mRNAs are often the site of regulatory RNA interactions. Therefore, we generated a highly detailed transcriptome architecture of methicillin-resistant Staphylococcus aureus JKD6009. This include using RNA-seq to identify expressed transcripts, differential RNA-sequencing (Sharma and Vogel, 2014, Curr Opin in Microbiol, 19:97,Äì105) to identify transcription start sites, and the use of Term-Seq to identify transcripts termination site (Dar et al., 2016, Science, 352:aad9822). The ANNOgesic pipeline (Yu et al., 2018, GigaScience, 7(9):giy096) was used to analyse the transcriptome data to generate the detailed transcriptome map, which resulted in the identification of 2867 coding sequences, 1156 5' UTRs, 1031 3' UTRs, and 499 sRNAs. We have used the endoribonuclease RNase III to capture RNA-RNA interactions using an RNA proximity-dependant ligation technique termed CLASH. From seven independent RNase III-CLASH experiments, 256 sRNA-mRNA interactions were observed in vivo allowing functional characterisation of many sRNAs for the first time. Surprisingly, we found that an mRNA encoding an unusually long 3,ÄôUTR (here termed vigR) functions as a regulatory ,Äòhub,Äô within our RNA-RNA interaction network. We present evidence that vigR promotes expression of the cell wall lytic transglycosylase encoded by isaA through a direct mRNA-mRNA interaction. Further, we found that the vigR mRNA 3,ÄôUTR is required for cell wall thickening and that deletion of the vigR 3,ÄôUTR re-sensitises VISA to vancomycin. Our results demonstrate the utility of RNase III-CLASH for identifying new regulatory RNA functions and indicate that S. aureus may use mRNA-mRNA interactions to co-ordinate gene expression much more widely than previously appreciated.

Abstract #137

Jaqueline Brito (University of Southern California, California, USA)
Tampa: interpretable analysis and visualization of metagenomics-based taxon abundance profiles

Taxonomic metagenome profiling aims to predict the identity and relative abundances of taxa in a given whole genome shotgun (WGS) metagenomic sample. A recent surge in computational methods that aim to accomplish this, called taxonomic profilers, has motivated community-driven efforts to create standardized benchmarking datasets, standardized taxonomic profile formats, as well as a benchmarking platform to assess tool performance. However, existing tools are either integrated into a single taxonomic profiling method or lack the flexibility and interpretability to analyze and visualize multiple taxonomic profiles. We address this lack of a metagenomic taxonomic profile analysis and visualization platform by proposing a software package Tampa (Taxonomic metagenome profiling evaluation). Tampa allows any user to effectively analyze one or more taxonomic profiles produced by numerous taxonomic profiling methods. Additionally, Tampa can leverage a wide range of formats, including the widely utilized and community developed BIOM and CAMI profiling formats. We demonstrate Tampa's ability by illuminating the critical biological differences between samples and conditions otherwise missed by commonly utilized metrics. Additionally, we show that Tampa can enable biologists to effectively

choose the most appropriate profiling method to use on their real data when a ground truth taxonomic profile does not exist; the most frequently encountered roadblock when a biologist asks, Which tool should I use to analyze my data? When ground truth taxonomic profiles are available (such as with simulated or mock metagenomic communities), we show how Tampa can augment existing benchmarking platforms such as OPAL. Tampa will be provided in a platform-independent fashion (utilizing Bioconda) and integrated into the Galaxy Toolshed for easy point and click analysis for less computationally inclined users. Tampa will allow scientists to quickly contextualize, assess, and extract insight from taxonomic profiles instead of relying primarily on statistical summaries or manual manipulation.

Abstract #138

Diksha Sharma (University of Auckland, Auckland, New Zealand)
Population dynamics of rDNA copy number in yeast and humans

In most eukaryotes the ribosomal RNA genes (rDNA) are organized as tandem repeat arrays, with each species having a characteristic number of copies. Despite this, there is a high degree of variability in copy number between individuals, and the copy number is thought to return to the ‚Äúhomeostatic‚Äù number when perturbed. However, it is unclear whether homeostatic rDNA copy number is a species-level property, or whether different populations have different homeostatic rDNA copy numbers. To address this question, we developed a novel bioinformatics approach to measure rDNA copy number from whole genome sequence data using the most frequent (modal) coverage. We validated this method with Saccharomyces cerevisiae strains having known, different rDNA copy numbers, and show this approach is robust and reliable, even with low coverage datasets. We then applied our pipeline to investigate variation in rDNA copy number between different S. cerevisiae and human populations. Our results using 1002 S. cerevisiae Genome project data suggest that different populations have different homeostatic rDNA copy numbers. We validated this result using a molecular biology approach, with yeast isolates from different populations showing no recovery to a common rDNA copy number. In contrast, investigation of 430 individuals coming from 14 populations from around the world showed no evidence for population-level differences in rDNA copy number, except for the Papuan population that might harbor a higher homeostatic rDNA copy number than other populations. Together, this work establishes a robust and a simple platform to determine rDNA copy number using whole genome sequence data. This pipeline provides evidence for population-level differences in S. cerevisiae, but little evidence for such differences in most human populations. This discrepancy might result from S. cerevisiae populations have deeper genetic divergence, and/or differences in the population dynamics of these species.

Abstract #139

Kim-Anh L√™ Cao (University of Melbourne, Melbourne, VIC)
Navigating through ‚Äòomics data: a multivariate perspective

Technological improvements have allowed for the collection of data from different molecular compartments (e.g. gene expression, protein abundance) resulting in multiple ‚Äòomics data from the same set of biospecimens or individuals (e.g. transcriptomics, proteomics). We propose to adopt a systems biology holistic approach by statistically integrating data from these multi-omics. Such an approach provides improved biological insights compared with traditional single omics analyses, as it allows to take into account interactions between omics layers. Integrating data include numerous challenges ‚Äì data are complex and large, each with few samples (< 50) and many molecules (> 10,000), and generated using different technologies. In this talk I will give a broad overview of the different methods we have developed for multivariate statistical data integration. I will illustrate these approaches to explore and integrate different multi-omics studies, from bulk to single cell resolution. Across all these studies, our main goal is to identify a signature composed of biological markers of different types to characterise a specific phenotype or disease status, and thus better understand the underlying molecular mechanisms of a biological system. Our methods are based on dimension reduction matrix factorisation techniques and implemented in the R package mixOmics.

Abstract #140

Seung Rhee (Carnegie Institution for Science, Washington DC, USA)
Challenges and Opportunities for Bioinformatics and Computational Biology in Plant Science

Plants make up the biggest biotic component of the biosphere and play essential roles in all ecosystems. Our survival and well-being depend on plants and this dependence will increase as the climate changes rapidly. To improve how we obtain food, energy, and materials from plants and steward the health of our environment for future generations, we need to understand how plants work at multiple scales from molecules to cells to ecosystems. A major challenge to achieving this goal is a limited understanding of functions of plant genes. The majority of genes in plant genomes are uncharacterized and many of them are found only in plant lineages. Traditional sequence-similarity based biochemical function inference cannot address this challenge. Another aspect of gene function that is critical but generally lacking is the spatial and temporal context under which gene products operate. These challenges have, in part, driven the spectacular advances and inventions in genomics, imaging, mass spectrometry and we are now capable of high-throughput, high-content, and high-resolution measurements of gene and protein function parameters. Along with these technologies and emerging

datasets, we need advances in computational biology and bioinformatics tools, concepts, and methods. In this talk, I will describe these challenges and some of the efforts we are making in addressing them.

Abstract #141

Locedie Mansueto (Southern Cross Plant Science - Southern Cross University, Lismore, NSW)
An Open Platform for Cannabis Genomics Research

Recent global appreciation of the therapeutic properties of Cannabis has led to revised national legislations under the Single Convention on Narcotic Drugs (https://www.unodc.org/unodc/en/treaties/single-convention.html), leading to rapidly growing hempseed and medicinal cannabis industries. Part of government regulation is concerned on limiting the THC content, although some countries still prohibit possession. The high heterozygosity of the Cannabis genome results in wide morphological and metabolite variation, including cannabinoids (THC, CBD and up to 60 others) and terpene content. It is a challenge for growers to produce legally compliant products while maintaining other desired properties. While there are publicly available molecular data for Cannabis that could aid in crop improvement, these are scarce and scattered. Looking at these datasets, we see that Cannabis is an ideal model system for multi-omics study since chemovar or metabolite concentration is the primary phenotype of interest, which may readily be associated with genome through the transcriptome and proteome. In this project, we aim to build a bioinformatics portal to enable community-inclusive cannabis ‚Äòomics and genetics research. We adopted Tripal (tripal.info), a well-established toolkit with numerous modules for storage and visualization of biological datasets. Publicly available omics-type data were obtained, re-analysed, and updated to add value to the data beyond the original publication, and harmonized to interoperate across the Tripal modules. Public genome data includes CBDRx cs10 genome and its NCBI RefSeq annotation and the Purple Kush genome that we annotated using GeneMark EP pipeline; these data can be accessed in a JBrowse genome browser. Genome variant calls were generated from 5 genome sequencing projects and these could be mined using the embedded SNP-Seek interface. Various gene expression data from GEO and publications can be visualized over biological pathways using the MapManJS interface through the curated pathways. More public data and Tripal modules are in active development. The use of Tripal as underpinning infrastructure allows for contribution of relevant contents by the cannabis research community, enabling users to register, be assigned to roles and groups with different levels of access to the site such as view, edit or create certain contents, and post comments about the contents to encourage discussion. We hope this portal will bring together the cannabis research community and encourage sharing of relevant data and information for the advancement of cannabis research.

Abstract #146

Karine Le Roch (University of California Riverside, California, USA)
Comparative 3D Genome Organization in Apicomplexan Parasites

The malaria parasite, Plasmodium falciparum, is a major cause of mortality in young children and pregnant women living in endemic areas. The spread of drug-resistant parasites is alarming and calls for the development of novel antimalarial drugs. However, the development of novel antimalarials is hampered by our lack of understanding about how the parasite controls its development and gene expression profiles through the different stages of its life cycle. Increasing amounts of evidence points towards a role for chromatin structure and three-dimensional (3D) nuclear organization to regulate gene expression throughout the parasite life cycle. In eukaryotic cells, the positioning of chromosomes in the nucleus is highly organized and has been showed to have a complex and dynamic relationship with gene expression. In the human malaria parasite Plasmodium falciparum, the clustering of genes involved in virulence and pathogenicity correlates with their coordinated silencing and has a strong influence on the overall genome organization. To identify conserved and species-specific principles of genome organization, we performed Hi-C experiments and generated 3D genome models for five Plasmodium species and two related apicomplexan parasites. Plasmodium species mainly showed clustering of centromeres, telomeres, and virulence genes. In P. falciparum, the heterochromatic virulence gene cluster had a strong repressive effect on the surrounding nuclear space, while this was less pronounced in Plasmodium vivax and Plasmodium berghei, and absent in Plasmodium yoelii. In Plasmodium knowlesi, telomeres and virulence genes while still interacting were more dispersed throughout the nucleus, and its 3D genome showed a strong correlation with gene expression. The Babesia microti genome showed a classical Rabl organization with colocalization of subtelomeric virulence genes, while the Toxoplasma gondii genome was dominated by clustering of the centromeres and lacked virulence gene clustering. Collectively, our results demonstrate that spatial genome organization in most Plasmodium species is constrained by the colocalization of virulence genes. P. falciparum and P. knowlesi, the only two Plasmodium species with gene families involved in antigenic variation, are unique in the effect of these genes on chromosome folding, indicating a potential link between genome organization and gene expression in more virulent pathogens.

Abstract #147

Paul Gardner (University of Otago, Otago, New Zealand)
Features of functional human genes

Proteins and non-coding RNAs are functional products of the genome that carry out the bulk of crucial cellular processes. With recent technological advances, researchers can sequence genomes in the thousands as well as probe for specific genomic activities of multiple species and conditions. These studies have identified thousands of potential proteins, RNAs and associated activities, however there are conflicting conclusions on the functional implications depending upon the burden of evidence researchers use, leading to diverse interpretations of which regions of the genome are ‚Äúfunctional‚Äù. Here we investigate the association between gene functionality and genomic features, by comparing established functional protein coding and non-coding genes to non-genic regions of the genome. We find that the strongest and most consistent association between functional genes and any genomic feature is evolutionary conservation and transcriptional activity for protein-coding and ncRNA sequences. Other strongly associated features include sequence alignment statistics, such as between-site covariation and protein substitution scores such as synonymous variation. In sum, our results demonstrate the importance of evolutionary conservation and transcriptional activity for sequence functionality, which should both be taken into consideration when differentiating between truly functional sequences and noise.

Abstract #148

Ami Bhatt (Stanford University, California, USA)
Microproteins, Mobile Genetic elements and Strain-level resolution in the microbiome ‚Äì a path to precision medicine

From climate change to agriculture, and human health to the oceanic food chain, microbes are at the base of every major system in the earth and life sciences. Far from being passive passengers, these organisms strongly interact with the environment ‚Äì be it the ocean floor or the human body. Yet, for all of this interaction, the dynamics between human hosts and bacteria (microbiome) has only been explored in earnest for the last twenty or so years, and even then, most studies have collapsed spectacular strain heterogeneity (indeed, millions of different strains) into monolithic ‚Äúspecies‚Äù. Positing that ‚Äústrains matter‚Äù, as do the genes that specific organisms encode ‚Äì our lab has developed and applied molecular and computational methods that help us link microbes to specific biological phenomena. Recently, our laboratory discovered >4,500 new small protein families encoded in microbial genomes. These genes had laid ‚Äòhidden in plain sight‚Äô due to the computational and experimental challenges in identifying them. We have validated that many of these predicted genes are transcribed and translated, and have enabled their rapid annotation using a deep learning approach. Moving forward, we hope to wield the precise microbiome measurement tools we have developed to systematically dissect the role of genomic plasticity in microbial adaptation (including antibiotic resistance), leverage mobile genetic elements to revolutionize genome engineering and gene therapy, and decode microbial microprotein communication to enable a breakthrough in microbe-inspired drug discovery. In so doing, we hope to revolutionize how we investigate microbiomes.

Abstract #149

A.J. Sethi (John Curtin School of Medical Research, Australian National University, Canberra, ACT)
An integrated approach for interrogating the dynamics of co-transcriptional splicing with unparalleled depth and minimal bias

Although dysregulated alternative splicing underlies 15% of human genetic diseases and underlies the initiation of numerous human neoplastic malignancies, the in vivo regulatory mechanisms remain poorly understood. Deconvoluting the regulation of alternative splicing is particularly arduous due to the manifold, highly interdependent mechanisms at play, including local epigenetic structure, transcriptional dynamics, splice factor expression levels and more. In order to understand how perturbations to any of these factors result in alternative isoform usage, the current gold standard is to sequence and analyse pre-mRNAs, the results of which reveals the speed, order, coordination and efficiency of intron removal. However, current approaches suffer either from read-depth, obscuring the dynamics of co-transcriptional processes, from read-length, obscuring the coordination of co-transcriptional splicing, or from methodologically induced bias, leading to biologically unrepresentative findings. Here, we present an integrated experimental and bioinformatic method to quantify the speed, order, and yield of intron removal from transcriptionally active pre-mRNAs. Our experimental protocol involves the biochemical fractionation of cellular RNA into transcriptionally active RNAPII-bound pre-mRNAs (nascent transcripts) and cytoplasmic polyadenylated mRNAs. Nascent transcripts are then subject to deep, short-read paired-end sequencing using a broad range of insert sizes, capturing a broad array of exon-defining units (ExoDUs). ExoDUs are cDNA fragments which span internal exons and capture the presence or absence of upstream and downstream splicing events, without vulnerability to splicing-dependant length bias. Paired-end alignments are then analysed using our novel bioinformatic pipeline, ClaiRO (https://git.nci.org.au/as7425/ClaiRO) which quantitates the presence of ExoDUs read-pairs. By studying the abundance of ExoDUs in different conditions, we can infer the order, speed, and yield of co-transcriptional splicing. Using an existing differential-splicing package (Whippet), we then identify alternative splicing events in polyadenylated cytoplasmic transcripts and understand the link between altered pre-mRNA splicing dynamics and alternative isoform usage in poly(a)+ cytoplasmic mRNA populations. ClaiRO provides a novel, low-bias method for studying the dynamics of co-transcriptional splicing and understanding the mechanisms which underly alternative isoform usage throughout various disease contexts.

Abstract #156

Phillippa Taberlay (University of Tasmania, Hobart, Tasmania)
Recapitulation of a juvenile-like histone landscape in aged neurons

The greatest risk factor for dementia is increasing age. During healthy aging the activity of neurons underlie a range of cognitive trajectories from unimpaired to significant decline. The epigenome is the interface between our genes and the environment and comprises a highly interactive network of chemical moieties (including histone modifications) unique to each cell type. We characterised H3K27ac and H3K4me3 histone modifications using ChIP-seq in forebrain neurons from 3, 6, 12, and 24 month (m) old C57/Bl6 mice. H3K27ac and H3K4me3 marking was enriched at promoters and enhancers in neurons from juvenile (3m) and aged (24m) mice compared to neurons from adult mice (6m and 12m). Differentially marked regions annotated to nucleosome organisation, protein folding and immune responses associated with H3K4me3, and protein localisation, RNA regulation and synaptic plasticity through modulation of H3K27ac. These process remained dynamic across life. Interestingly, a change in chromatin assembly pathways were unique to adult neurons, while differences in apoptosis signalling was characteristic of aged neurons. Interestingly, variability was primarily driven by the H3K27ac dataset. Surprisingly, we detected a partial recapitulation of a juvenile-like histone landscape in aged neurons; H3K27ac and H3K4me3 differentially marked sites were shared between juvenile and aged neurons and the majority of these shared sites were consistently enriched in both juvenile and aged neurons. This work reveals epigenetic alterations that impact neurons across aging.

Abstract #213

Karen Miga (University of California Santa Cruz, California, USA)
Telomere-to-Telomere Chromosome Assemblies: New Insights Into Genome Biology and Structure

We are entering into an exciting era of genomics where truly complete, high-quality assemblies of human chromosomes are available end-to-end, or from ‚Äòtelomere-to-telomere‚Äô (T2T). Recently, the Telomere-to-Telomere (T2T) consortium announced our v1.0 assembly that includes more than 100 Mbp of novel sequence compared to GRCh38, achieves near-perfect sequence accuracy, and unlocks the most complex regions of the genome to functional study. This technological advance, crediting the confluence of new assembly methods with long read sequencing technologies, offers a new opportunity to comprehensively the genomic structure and epigenetic organization in the most repeat-dense regions of our chromosomes. In particular, I will focus on the release of initial genetic and epigenetic reference of all human centromeric regions. High-resolution study of the pericentromeric sequence content and organization reveals new satellite families, sites of transposable element insertion, segmental duplications, and pericentromeric gene predictions. Using unique markers (marker-assisted method) to anchor ultra-long nanopore reads to human centromeric regions regions we report hypomethylated dips at every centromeric region, as previously described for the T2TX centromere. These sites are shown to coincide with regions enriched in centromere protein A (CENP-A) and may provide a signature of sites of kinetochore assembly genome-wide.

Abstract #214

Keolu Fox (University of California San Diego, California, USA)
Creating accountability in human population genetics using base editing tools

Appropriate empirical-based evidence and detailed theoretical considerations should be used for evolutionary explanations of phenotypic variation observed in the field of human population genetics (especially Indigenous populations). Investigators within the population genetics community frequently overlook the importance of these criteria when associating observed phenotypic variation with evolutionary explanations. A functional investigation of population-specific variation using cutting-edge genome editing tools has the potential to empower the population genetics community by holding ‚Äújust-so‚Äù evolutionary explanations accountable. In this lecture I will detail currently available precision genome editing tools and methods, with a particular emphasis on base editing, that can be applied to functionally investigate population-specific point mutations.

Abstract #215

Alex Brown (South Australian Health and Medical Research Institute (SAHMRI), Adelaide, SA)
Challenges In Aboriginal health in the genomics era

Engagement in the genomic era poses significant concerns for Indigenous people. As it currently stands, genomic research has excluded Indigenous peoples from its focus and remit. This leaves up to 370 million of the world‚Äôs population unrepresented in existing datasets. The reasons for this marginalisation are complex, but are embedded in a legacy of conflict, depopulation, distrust, institutionalised racism, social disadvantage and estrangement from health and political institutions within society. The scientific community has started to consider the importance of harnessing a more complete understanding of human genetic variation. Increasing focus has been applied to ‚Äòunlocking‚Äô Indigenous and diverse ethnic populations across the globe. But on whose terms and at what cost? The starting point for engaging Indigenous peoples in genomic research has been framed around two strategies. ‚ÄòSelling the Dream‚Äô which fundamentally speaks to the unfulfilled promise of genomics to cure the ills of Indigenous peoples; and ‚Äò Don‚Äôt miss out‚Äô which levers fear of widening disparities between Indigenous people and non-Indigenous populations. As yet, it seems the hype has outstripped the reality ‚Äì but that is unlikely to be the case forever. There is a clear and present need for

Indigenous populations to be involved in any genomics future. In part this must be to increase representation. But what is rarely discussed is what else Indigenous peoples offer the world in relation to the conduct, stewardship, ethics, integration, communication and interpretation of genomics research and clinical care. Our particular strengths lie in our ability to understand, communicate and re-interpret context and the broader determinants of health, within which genomics may offer critical insights into our past, present and future. But this will not be possible until Indigenous peoples are afforded our rightful place at the head of decisions that influence our own health and well-being.

Abstract #216

Yue Wan (Genome Institute of Singapore, Singapore)
Direct RNA sequencing identifies isoform specific structures

The ability to correctly assign structure information to an individual transcript in a continuous and phased manner is critical to understanding RNA function. RNA structure play important roles in every step of an RNA‚Äôs lifecycle, however current short-read high throughput RNA structure mapping strategies are long, complex and cannot assign unique structures to individual gene-linked isoforms in shared sequences. To address these limitations, we present an approach that combines structure probing with SHAPE-like compound NAI-N3, nanopore direct RNA sequencing, and one-class support vector machines to detect secondary structures on near full-length RNAs (PORE-cupine). PORE-cupine provides rapid, direct, accurate and robust structure information along known RNAs and recapitulates global structural features in human embryonic stem cells. The majority of gene-linked isoforms showed structural differences in shared sequences both local and distal to the alternative splice site, highlighting the importance of long-read sequencing for phasing of structures. Structural differences between gene-linked isoforms are associated with differential translation efficiencies globally, highlighting the role of structure as a pervasive mechanism for regulating isoform-specific gene expression inside cells.

Abstract #217

Jessica Mar (University of Queensland, Brisbane, QLD)
Making sense of heterogeneity in gene expression data

Any model that we apply makes assumptions. But how often do we step back and question the validity of these assumptions? This talk includes studies from my group where a focus on the shape of gene expression distributions, specifically shape diversity, has revealed new insights into biology. Using topics from cancer and ageing, we have found that studying the shape of a gene‚Äôs expression distribution can represent a powerful means to modelling heterogeneity. Collectively, this work raises new questions and opportunities to understand how heterogeneity is gene expression contributes to regulation in genomics.

Abstract #218

Robert Edwards (Flinders University, Adelaide, SA)
Phage genome bioinformatics

Phages, viruses that infect bacteria, experience unusual selective pressures on their genomes. Phage genomes need to be packaged to move from host to host, and phages often carry additional genes in case they are beneficial in a new environment. But phage genes need to be expressed in their new host and create new phages. Therefore, phages have compact genomes with reduced gene length and very little non-essential DNA. We have used the unique features of phages to build novel phage-gene identification algorithms and to build machine learning tools to predict the function of phage proteins. New phage genome bioinformatics tools are helping us illuminate the true meaning of the viral dark matter.

Abstract #219

Bernice Waweru (International Livestock Research Institute (ILRI), Nairobi, Kenya)
African-led genome sequencing of Lablab and African Yam bean orphan crop genomes

The African Orphan Crops Consortium (AOCC) is a global partnership promoting strategic, genome-enabled improvement of under-researched crops for biodiversity-based nutritious food solutions in Africa. We present current status, opportunities and examples of successes of AOCC. Orphan crops like Lablab (Lablab purpureus; ~452 Mbp), African Yam bean (Sphenostylis stenocarpa; ~800 Mbp) and Moringa (Moringa oleifera; ~315 Mbp), are often of high nutritive value and are climate resilient. The Oxford Nanopore MinION was used to generate long reads of all three crops, to complement and generate contiguous draft genome assemblies. Work is under way to improve these assemblies to chromosome-scale using Hi-C mapping.

Abstract #220

Torsten Seemann (University of Melbourne, Melbourne, VIC)

How bioinformatics and genomics helped Australia's COVID response

Complete genomes of microbial pathogens are essential for the phylogenomic analyses that increasingly underpin core public health lab activities. Here, we present complete genomes of pathogen strains of regional importance to the Southwest Pacific and Australia. These enrich the catalogue of globally available complete genomes for public health while providing valuable strains to regional public health labs. Whole-genome sequence (WGS) data is increasingly important in public health microbiology. The data can be used to replicate many of the basic bacterial sub-typing approaches, as well as support epidemiological investigations, such as surveillance and outbreak investigation. The appeal of WGS data comes from the promise of a single workflow to process all microbial pathogens that can provide easily portable data that promotes deeper integration of surveillance and investigation efforts across jurisdictions. This promise is leading to a concerted effort to move microbial public health to a primarily genome-based workflow at numerous jurisdictions, including Australia.