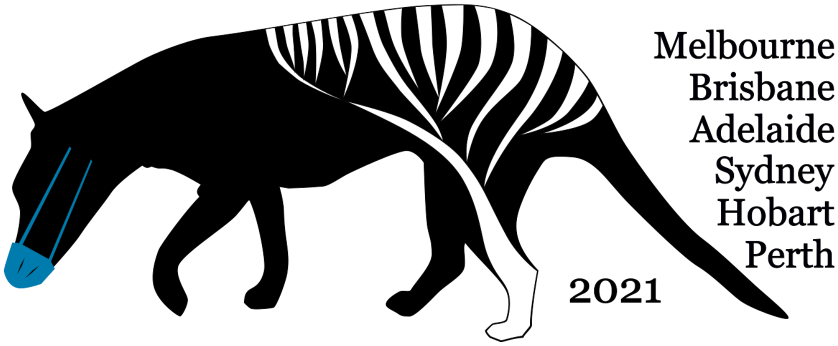
` t

COMBINE, ABACBS and Phylomania #13



ABACBS/COMBINE/Phylomania 2021 *Conference Program*

Organising Committee

Chair: Michael Charleston (UTas)

Jimmy Breen (UAdelaide)

Aaron Darling (UTS)

Melanie Smith (UAdelaide)

Jarny Choi (UMelb)

Denis Bienroth (MCRI)

A.J. Sethi (ANU)

Patricia Sullivan (CCIA, UNSW)

Sachintha Wijegunasekara (DeakinU)

Mirana Ramialison (ARMI/MCRI)

Ignatius Pang (CMRI)

Nicola Armstrong (CurtinU)

Program Committee

Mirana Ramialison (Coordinator, ARMI/MCRI)
Emily Wong (Co-coordinator, VCCRI)
A.J. Sethi (COMBINE)
Alicia Oshlack (Peter Mac)
Allan McRae (IMB)
David Powell (Monash U)
Eduardo Eyras (ANU)
Eleni Giannoulatou (VCCRI)
Ellis Patrick (USyd)
Fatemeh Vafaee (UNSW)
Jean Yang (USyd)
Jimmy Breen (UAdelaide)
John Ormerod (USyd)
Joshua Ho (HKU)
Lan Nguyen (Monash U)
Mike Charleston (UTas)
Pengyi Yang (USyd)
Sara Ballouz (Garvan)

Hub Coordinators

Adelaide: Michael Roach (Flinders U)

Brisbane: Ariane Mora, Apoorva Prabhu, Ebony Watson, Gabriel Foley and Sam Davis (UQ) and Anita Sathyanarayanan (QUT)

Hobart: Barbara Holland (UTas)

Melbourne: Fernando Rossello (UMelb)

Perth: Phillip Bayer (UWA)

Sydney: Mathieu Fourment (UTS)

Our Joint ABACBS/COMBINE/Phylomania 2021 Conference is proudly supported by

CSL™

Oxford NANOPORE™ Technologies

illumina

pawsey

INTERSECT RESEARCH FASTER

Melbourne Bioinformatics

iSCB INTERNATIONAL SOCIETY FOR COMPUTATIONAL BIOLOGY

Attendee information

Each ABACBS, COMBINE & Phylomania conference sessions will be chaired remotely on Zoom by separate locations around the country. Invited speakers and talks selected through abstract submissions will present via Zoom, or at a local hub site if available (Adelaide, Brisbane, Perth and Hobart). Additionally, attendees will have the opportunity to interact with each other via our conference interaction platform Gather.Town (https://www.gather.town/) and look at posters in each ABACBS/COMBINE/Phylomania room. Sponsors will also have the opportunity to interact with delegates interactively.

The information for the Zoom links (video conferencing) and Gather.town links (social interaction and posters presentation) for the ABACBS conference will be available in the following URL (https://conference2021.abacbs.org/abacbs/conf). Please authenticate with the ORCID that you have provided during registration and enter your ORCID password. You will be able to see the Zoom and Gather.Town links once you've logged in.

Poster/Talk Presentation Upload

For poster presentations, we will be uploading slides onto the Gather.town platform. Use only PNG image format as Gather.Town only supports PNG format currently. Images do need a min of 600px*1000px and max 3MB in size. Please upload the file using this CloudStor link (https://cloudstor.aarnet.edu.au/plus/s/NBSTiSm2ui9lmYn). Please setup your file name in the following format:
<Surname>_<First Name>_poster_<a random number between 1 to 10>_<a random alphabet letter>.PNG
If you want to replace a file just upload a file with the same file name again and the new file will be kept.
Please upload your slides before 5pm Friday 19th November

For short/lightning talk presentations, please send your slides to the session chairs by **5pm Friday 19th November**:

- COMBINE
  - Sachintha Wijegunasekara (sachinthajw@gmail.com)
  - AJ Sethi (aditya.sethi@anu.edu.au)
- ABACBS
  - Fernando Rossollo (fjrossello@gmail.com)
  - Philipp Bayer (philipp.bayer@uwa.edu.au)
  - Mathieu Fourment (mathieu.fourment@uts.edu.au)
  - Michael Roach (beardymcjohnface@gmail.com)
  - Ariane Mora (ariane.n.mora@gmail.com)
- Phylomania
  - Venta Terauds (venta.terauds@utas.edu.au)

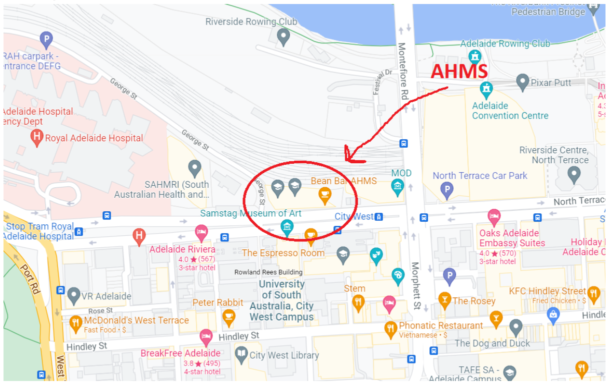
Social Networking

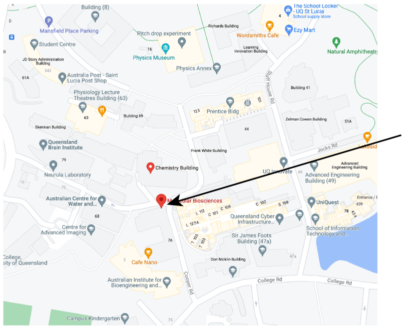
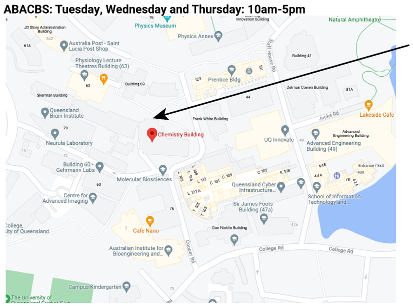
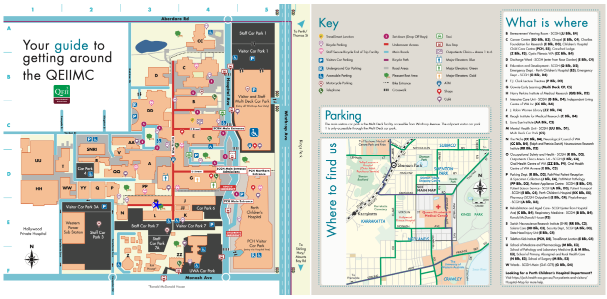
ABACBS Slack Channel: #conference2021_public https://abacbs.slack.com/archives/C02MKCJ45FY There is also a channel for each session of the conference. They are pinned in the main Slack channel or search for the channel using the string "conf2021_". We encourage users to continue discussions on Slack after the session is over.

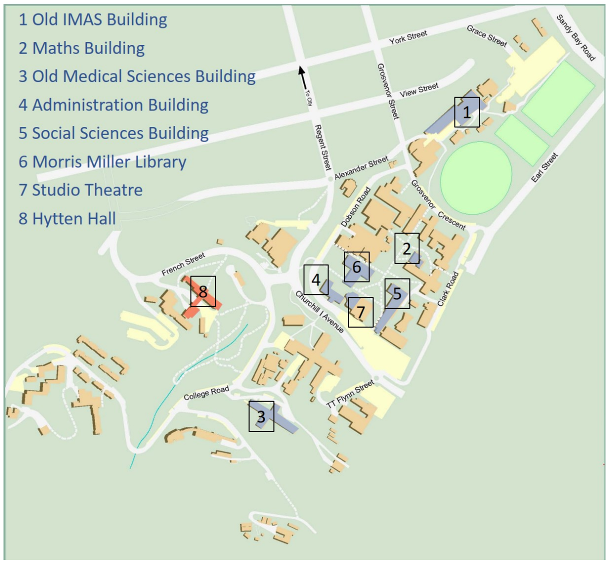
Twitter Hashtags: #ABACBS2021 as the official twitter hashtag. #COMBINE21 and #Phylomania can be used alongside the main one for the respective conferences.

In-person conference hubs

For states that are able to host events (Western Australia, South Australia, Queensland and Tasmania) we will be holding in-person conference hubs.

Hub City	Location	Additional information
Adelaide	Room G030 University of Adelaide Adelaide Health and Medical School (AHMS), 4 North Terrace, Adelaide (City West)	The Room G030 lecture theatre is on the bottom floor of AHMS. Monday and Wednesday will be catered with our local speaker Kat Pillman. Adelaide node dinner and in-person social networking event at the West Oak Hotel on Wednesday (24th Nov)  AHMS Riverside Rowing Club Adelaide Rowing Club Pedestrian Bridge Pixar Putt Adelaide Convention Centre Riverside Centre, North Terrace North Terrace Car Park North Terrace Montefiore Rd Festival Dr George St RAH carpark - Entrance DEFG Royal Adelaide Hospital SAHMRI (South Australian Health and... Bean Bar AHMS MOD Samstag Museum of Art City West Stop Tram Royal Adelaide Hospital Adelaide Riviera The Espresso Room Rowland Rees Building University of South Australia, City West Campus Stem Morphett St Oaks Adelaide Embassy Suites Holiday Inn Adelaide KFC Hindley Street The Rosey Phonatic Restaurant The Dog and Duck TAFE SA - Adelaide Campus VR Adelaide Rose St McDonald's West Terrace Peter Rabbit Hindley St BreakFree Adelaide City West Library Port Rd West Terrace
Brisbane	Monday (COMBINE): Molecular	Wednesday will be catered since we'll be

	biosciences building: Learning theatre 76-228, the University of Queensland, St Lucia, QLD Tuesday-Thursday (ABACBS): Chemistry building, 68-415, the University of Queensland, St Lucia, QLD	having a live talk from Dr Ben Woodcroft, our invited local speaker. **COMBINE: Monday 10am-5pm**  Mansfield Place Parking Student Centre Pitch drop experiment Physics Museum Physics Annex Wordsmiths Cafe Ezy Mart Natural Amphitheatre Prentice Bldg Queensland Brain Institute Neurula Laboratory Chemistry Building Molecular Biosciences UQ Innovate Advanced Engineering Building (49) UniQuest Queensland Cyber Infrastructure... Sir James Foots Building (47a) Cafe Nano Centre for Advanced Imaging Australian Institute for Bioengineering and... Campus Kindergarten College Rd Cooper Rd Staff House Rd Jocks Rd Enter here and stay on the first floor walk through the building and you'll see ABACBS signs to the lecture theater :) **ABACBS: Tuesday, Wednesday and Thursday: 10am-5pm**  Physics Museum Physics Annex Natural Amphitheatre Prentice Bldg Queensland Brain Institute Neurula Laboratory Chemistry Building Molecular Biosciences UQ Innovate Lakeside Cafe Advanced Engineering Building (49) Queensland Cyber Infrastructure... Sir James Foots Building (47a) Cafe Nano Centre for Advanced Imaging Campus Kindergarten Enter here and go up to the 4th floor where you'll see ABACBS signs :)
Perth	Seminar room 1.18 M Block QE-II (UWA)	Directions: Go in the main entrance of M Block (QE2), up one flight of stairs and on your left is an old, brown corridor (no swipe needed). 1.18 is the first glass door on your left. See blue star below. Lunch will be provided on Monday and Tuesday.  Your guide to getting around the QEIIMC Aberdare Rd Winthrop Ave Monash Ave Key Parking Where to find us What is where
Hobart	Social Sciences Video Conferencing Room Sandy Bay Campus	Directions: From the central concourse of the Sandy Bay campus, head south up the hill; the Social Science building (5 on the map below) is to the left of Lazenby's Cafe. Follow the signs, the room numbers, or your nose to room 210.

		 1 Old IMAS Building 2 Maths Building 3 Old Medical Sciences Building 4 Administration Building 5 Social Sciences Building 6 Morris Miller Library 7 Studio Theatre 8 Hytten Hall To City York Street Grace Street Sandy Bay Road View Street Grosvenor Street Regent Street Alexander Street East Street Dobson Road Grosvenor Crescent French Street Churchill Avenue Clark Road College Road TT Flynn Street Dinner Thursday night at Botanica, Salamanca Place

Our in-person hubs are proudly supported by:

SAGC SOUTH AUSTRALIAN GENOMICS CENTRE

THE UNIVERSITY OF MELBOURNE

Centre for Cancer Research

Peter Mac
Peter MacCallum Cancer Centre
Victoria Australia

Oxford NANOPORE Technologies

SCRUM
Single Cell Research
User Group Meeting
go.unimelb.edu.au/s5gj

Supported by
illumina

Conference Program

COMBINE Student Symposium (Monday 22nd November)

Time (AEDT) (Melb/Syd/Canb/ Hob)	COMBINE Student Symposium
10:30	Symposium Start and Welcome
	Session 1: Genomics (Session Chair: Akari Komori)
10:50	**Abstract Talk:** Swapnil Tichkule *Genomic landscape of diversification, selective sweeps, and demographic history of an anthroponotic parasite*
11:05	**Abstract Talk:** Anushi Shah *An automated meta-caller to detect de novo mutations from whole genome trio data using cloud computing technology*
11:20	**Abstract Talk:** Nehleh-Fatemeh Kargarfard *PhiloBacteria: A new tool to infer phylogenetic trees from recombinant bacterial genomes*
11:35	**Abstract Talk:** Hugh Cottingham *Developing a novel CRISPR-Cas9 based method for characterisation of bacterial pathogens using Oxford Nanopore sequencing*
11:50	**Abstract Talk:** Ly Trong Nhan *AliSim: Phylogenetic Sequence Simulator in the Genomic Era*
12:05	**Abstract Talk:** Katherine Caley *Is the Human genome in mutation equilibrium?*
12:20	**Lightning Talks:** ● Chelsea Matthews ● Amarinder S. Thind ● Holly A. Withers ● Urwah Nawaz
12:40	Break Discussion session (Zoom)
1:10	Keynote Address **Prof. Alicia Oshlack** Peter Maccallum Cancer Center, Melbourne, Australia
	Session 2: Transcriptomics (Session Chair: Yunwei Zhang)
1:45	**Abstract Talk:** Wenjun Liu *RNA-seq regulatory network inference revealed an association between transcription factor SETX and neurogenerative pathways under prolonged autophagy induction*
2:00	**Abstract Talk:** Angelita Liang *A universal naming system for human alternative splicing*
2:15	**Abstract Talk:** Adrian (Abbas) Salavaty *Identification, classification, and prioritization of most influential players in normal biological processes and diseases*
2:30	**Abstract Talk:** Yue Cao *scFeatures: automatic feature generation for single-cell and spatial data*
2:45	**Abstract Talk:** Andy Tran *Using multimodal single cell data to predict regulatory gene relationships and to build a computational direct cell reprogramming model*
3:00	Break Discussion session (Zoom)
3:15	Session 3: Translational bioinformatics (Session Chair: Fred Jaya)
3:20	**Abstract Talk:** Anh Phuong Le (Phuong) Integrating genomic location and sequence contexts of point mutations to improve classifying cancer tissue of origin
3:35	**Abstract Talk:** Gulrez Chahal CaraVaN: Prioritising Cardiac Variants in the Non-coding genome using boosting algorithm
3:50	**Abstract Talk:** Keeley OGrady Genomic epidemiology of *Clostridioides difficile* in Australia, 2013-2018

4:05	**Abstract Talk:** Himal Shrestha *Landscape genetics to determine factors for ongoing transmission of onchocerciasis in the transition region of Ghana*
4:20	**Abstract Talk:** Rachel Bowen-James *disTIL: a turnkey approach to profile the immune landscape in cancer*
4:35	**Abstract Talk:** Alejandro Casar *Investigating how transcriptional plasticity drives drug resistance in cancer by combining computational modelling with single cell genomics*
4:50	**Lightning Talks:** ● Mikhail Dia ● Boris Ka Leong Wong ● Sanghyun Lee ● Ann Rann Wong ● Holly Martin
17:05	Break Discussion session (Zoom)
17:15	**Careers Panel** **(Chairs: Akarai Komori and Alice)** ● Sarah Beecroft ● Thom Quinn ● Sonika Tyagi ● Etsuko Uno
18:15	**Prizes and symposium close**

ABACBS Day 1 (Tuesday 23rd November)

Time (AEDT)	Session 1: Transcriptomics (Melbourne) (Session Chair: Fernando Rossello)
8:30-9:05	International Invited Speaker **Roser Vento-Tormo** Wellcome Sanger Institute, UK *Mapping tissues* in vivo *and* in vitro
10:50	**ABACBS Conference opening**
11:00	International Invited Speaker **Smita Krishnaswamy** Yale School of Medicine, USA
11:35	National Invited Speaker **Matthew Lewsey** La Trobe University *Gene regulatory dynamics of plant embryos from bulk tissue to single-cell resolution*
12:10	**Abstract Talk:** Patricia Sullivan *Breaking the black box: Dissecting deep neural network models to reveal splicing motifs*
12:25	**Abstract Talk:** Jeffrey Pullin *Benchmarking methods for selecting marker genes in single cell RNA sequencing data*
12:40	**Abstract Talk:** Yingxin Lin *Large scale single-cell multi-sample multi-condition data integration using scMerge2*
12:55	**Lightning Talks:** ● Natalie Charitakis, *Benchmarking methods for the identification of spatially variable genes in spatial transcriptomics datasets.* ● Andy Tran, *Using multimodal single cell data to predict regulatory gene relationships and to build a computational direct cell reprogramming model* ● Alejandro Casar, *Investigating how transcriptional plasticity drives drug resistance in cancer by combining computational modelling with single cell genomics* ● David Humphreys, *Overlapping transcripts within gene models can influence bioinformatic analysis.*
13:10	Break Poster session (GatherTown)
Time (AEDT)	Session 2: Genomics (Perth) (Session Chair: Philipp Bayer)
13:30	International Invited Speaker **Isobel Parkin** Agri-Food, Canada *Towards fully assembled plant genomes and the promise of pan-genomes*
14:05	**Abstract Talk:** Ulf Schmitz *Multi-omics data analysis identifies epigenetic regulators of alternative splicing*
14:20	**Abstract Talk:** Isobel J Beasley *Predicting the cross-population portability of human expression quantitative trait loci (eQTLs)*
14:35	**Abstract Talk:** Dan Andrews *Most disease phenotypes are actually very subtle: Deep learning with a massive, mouse, mutant cellular dataset to detect sub-clinical phenotypic variation*
14:50	National Invited Speaker **Parwinder Kaur** University of Western Australia *Breaking Barriers: From conventional to next generation genetics*
15:25	**Lightning Talks:** ● Aidan P. Tay, *INSIDER: alignment-free detection of foreign DNA sequences* ● Yunwei Zhang, *HISP: cohort heterogeneity identification and risk factors detection for survival prediction* ● Jiru Han, *Population-level genome-wide STR typing in Plasmodium species reveals higher resolution population structure and genetic diversity relative to SNP typing* ● Hieu T. Nim, *Finding needles in a haystack: identifying cardiac disease-causing genes from*

	cis-regulatory elements
15:40	Break Poster session (GatherTown)
Time (AEDT)	Session 3: Metagenomics, Long Reads & Proteomics (Sydney) (Session Chair: Mathieu Fourment)
16:15	International Invited Speaker **Rohan Williams** Singapore Centre for Environmental Life Sciences Engineering (SCELSE), Singapore *Long (and short) read metagenome analysis of complex microbial communities: progress, problems and prospects*
16:50	**Abstract Talk:** Rebecca C Poulos *A pan-cancer proteomic map of 949 human cell lines*
17:05	**Abstract Talk:** Pablo Acera-Mateos *Detection of m6A and m5C RNA modifications at single-molecule resolution using Nanopore sequencing*
17:20	**Abstract Talk:** Aaron Chuah *Genome-wide estimation of the change in protein stability due to missense mutation*
17:35-18:30	Main Poster Session (GatherTown)

ABACBS Day 2 (Wednesday 24th November)

Time (AEDT)	Session 3: Metagenomics, Long Reads & Proteomics (Sydney) - Continued (Session Chair: Mathieu Fourment)
11:00	International Invited Speaker **Joep de Ligt** ESR, New Zealand *Putting single molecule sequencing to use for health*
11:35	**Lightning Talks:** ● James M. Ferguson, *InterARTIC: an interactive web application for whole-genome nanopore sequencing analysis of SARS-CoV-2 and other viruses* ● Richard J Edwards, *DepthSizer and DepthKopy: genome size and copy number prediction using single-copy long-read depth profiles* ● Kaitao Lai, *Shotgun microbial profiling reveals geographic disparities in aggressive prostate cancer* ● Hasindu Gamaarachchi, *SLOW5: a new file format enables massive acceleration of nanopore sequencing data analysis*
11:50	Break Poster session (GatherTown)
Time (AEDT)	Session 4: Biomedical (Adelaide) (Session Chair: Michael Roach)
12:00	International Invited Speaker **Jill Moore** University of Massachusetts Chan Medical School, USA *The ENCODE Registry of candidate Cis-Regulatory Elements: a summary of curation and applications*
12:35	National Invited Speaker **Katherine Pillman** Centre for Cancer Biology, University of South Australia *Uncovering the functions of splicing and transcription factors in Breast Cancer*
13:10	**Abstract Talk:** Feng Yan *DNA Methylation in pre-leukemic stem cells in a T-cell acute lymphoblastic leukemia mouse model*
13:25	**Abstract Talk:** Feargal Ryan *Long-term perturbation of the peripheral immune system months after SARS-CoV-2 infection*
13:40	**Abstract Talk:** Bennet McComish *Using signatures of natural selection to focus the search for causal genetic variants in multiple sclerosis.*
13:55	**Lightning Talks:** ● Andreas Zankl, *Ontoclick: a web browser extension to facilitate biomedical knowledge curation*

	● Chelsea Mayoh, *Utilising the transcriptome to functionally resolve molecular aberrations in precision medicine* ● Andreas Halman, *STRipy: a graphical application for detecting short tandem repeat expansions in all known pathogenic loci* ● Letitia Sng, *Novel Coronary Artery Disease Variants and Epistatic Interactions Identified with Machine Learning Pipeline using VariantSpark and BitEpi*
14:10	Break Poster session (GatherTown)
Time (AEDT)	Session 5: Stats & Methods (Brisbane) (Session Chair: Ariane Mora)
14:30	National Invited Speaker **Anna Trigos** Peter MacCallum Cancer Centre *Single-cell and spatial transcriptomics to dissect modulators of radionuclide therapy response in prostate cancer*
15:05	National Invited Speaker **Ben Woodcroft** Queensland University of Technology *Scalable community profiling of shotgun metagenomes*
15:40	**Abstract Talk:** Nicolas P. Canete *spicyR: Spatial analysis of in situ cytometry data in R*
15:55	**Abstract Talk:** Yuzhou Feng *SPIAT: Novel Computational and Statistical Tools for the Simulation and Analysis of Spatial Data*
16:10	**Abstract Talk:** Sabrina Yan *Helium: Automatic variant prioritization for surfacing cancer drivers*
16:25	**Lightning Talks:** ● Dillon Hammill, CytoExploreR: Next-Generation Open-Source Software for Cytometry Data Analysis ● Mikhail Gudkov, Quantifying negative selection on synonymous variants ● Daniel Cameron, Single breakends: a new paradigm for structural variant calling ● Angelita Liang, A universal naming system for human alternative splicing
16:40	Break
16:45	ABACBS Awards

ABACBS Day 3 / Phylomania Day 1 (Thursday 25th November)

Time (AEDT)	ABACBS Session 6 / Phylomania Session 1: Phylogenetics, Evolution & BigData (Hobart) (Session Chair: Michael Charleston)
11:00	National Invited Speaker: **Barbara Holland** University of Tasmania *Some possibly bad ideas for detecting convergent selection*
11:20	**Abstract Talk:** Zhaoxiang Cai *Integrating multi-omics data with biological knowledge by Transformer-based deep learning*
11:35	**Abstract Talk:** Paola Cornejo-Páramo *Transcription Factor-binding motif composition of cis-regulatory regions in a sea sponge species suggest the existence of some constraints in metazoan regulatory grammar*
11:50	Nicholas Fountain-Jones *Hunting alters viral transmission and evolution in a large carnivore*
12:05	**Lightning Talks:** ● Tyrone Chen, *Identifying the complete functional and regulatory signatures driving antimicrobial resistance in sepsis.* ● Ariane Mora, *Integrated gene landscapes uncover multi-layered roles of repressive histone marks during mouse CNS development* ● Nehleh-Fatemeh Kargarfard, *PhiloBacteria: A new tool to infer phylogenetic trees from recombinant bacterial genomes* ● Jarny Choi, *Creating interactive visualisation of gene expression data online using latest web*

	technologies
12:20	Best COMBINE talk
12:40-13:00	ABACBS Wrap Up
	Break **ABACBS ends - but Phylomania continues!**
Time (AEDT)	Phylomania Session 2: applications (Session Chair: Jeremy Sumner)
13:45	Folagbade Abitogun: Identification and Analysis of Novel Putative Drug Targets in Hypervirulent Klebsiella pneumoniae Proteomes
14:00	Abhinay Thakur: *In silico* based approach of FDA Approved Drugs for targeting against nsp10 (6W75) of 2019-nCoV (novel coronavirus)
14:15	Eike Steinig: Signatures of epidemic growth in the emergence of community-associated MRSA
14:30	Swapnil Tichkule : Genomic landscape of diversification, selective sweeps, and demographic history of an anthroponotic parasite
14:45	Amarinder Thind : Potential non-coding regulations in cSCC Metastasis cancer
15:00	Break
15:30	Joshua Stevenson: Rearrangement Events on Circular Genomes
15:45	Venta Terauds: Genome algebras in action
16:00	Jiahao Diao: A stochastic model of evolution to improve the prediction of independent or dependent gain and loss of genes
16:15	Jonathan Mitchell: Consistent inference of species networks from gene trees under the multispecies coalescent using a generalized AIC
16:30	Break
17:00-17:40	International Invited Speaker: ~~**Arndt von Haeseler**~~ **Sebastian Burgstaller-Mühlbacher** Center for Integrative Bioinformatics Vienna, Vienna University and Medical University, Vienna, Austria
19:00	(Hobart) conference dinner - Botanica in Salamanca Place

Phylomania Day 2 (Friday 26th November)

Time (AEDT)	Phylomania Session 3: methods comparing trees (Session Chair: Barbara Holland)
09:00 (Earlier start!)	Martin Smith: Improving consensus trees by detecting rogue taxa
09:15	Caitlin Cherryh: Does Filtering Recombinant Loci Improve the Estimation of Species Trees?
09:30	Ana Serra Silva: From islands of trees to clumps: can we generate informative clusters of partially overlapping trees?
09:45	Sarah Alver: Improvement to GLASS/Maximum Tree Method of Species Tree Inference from Estimated Gene Trees Using Measurement Error Modified Single Linkage Clustering
10:00	Lena Collienne: Distances between phylogenetic time trees
10:15	Gleb Zhelezov: Trying out a million genes to find the perfect pair with MTrip
10:30	Break
Time (AEDT)	Phylomania Session 4: statistical models (Session Chair: Venta Terauds)
11:00	Ben Kaehler: Phylogenetics if you stop ignoring non-stationary evolution
11:15	Cassius Manuel Perez: A test for phylogenetic saturation

11:30	Qin Liu: Is AIC An Appropriate Metric For Model Selection In Phylogenetics?
11:45	Samuel Davis: Inference of biochemical and biophysical parameters for ancestral proteins
12:00	Conrad Burden: Stationary Distributions of Neutral Multi-type Branching Diffusions
12:15	Break (Lunch in Hobart)
Time (AEDT)	Phylomania Session 5: statistics, now with geometry (Session Chair: Jimmy Breen)
13:15	Yao-ban Chan: The effect of copy number hemiplasy on gene family evolution
13:30	Albert Soewongsono: The Shape of Phylogenies Under Phase-Type Distributed Times to Speciation and Extinction
13:45	Matthew Macaulay: Hyperbolic Tree Embeddings for Bayesian Phylogenetic Inference
14:00	Ruriko Yoshida: Tree Topologies along a Tropical Line Segment
14:15	Ben Wilson: Learning phylogenetic trees as hyperbolic point configurations
14:30	Michael Charleston: Landscapes of split space part 2: the quest for utility.
14:45	
15:00	Break
Time (AEDT)	Phylomania Session 6: Eclectica (Session Chair: Jonathan Mitchell!)
15:30	Luke Cooper: Models of biomolecular network evolution
15:45	Andrew Francis: The perfect picnic project: Sandwiches, eggs, and trees.
16:00	Mareike Fischer: Phylogenetic Diversity Rankings in the Face of Extinctions: the Robustness of the Fair Proportion Index
16:15	**Phylomania ends** **Brief announcements and slightly whimsical awards**
16:16	Beer o'clock

Workshops

PATHWAY ANALYSIS OF MULTI-OMICS DATA USING THE REACTOME DATABASE (Thursday, 25 November)

- Ignatius Pang (main contact) ipang@cmri.org.au
- Nader Aryamanesh
- Pablo Galaviz

Start time: AEDT 14:15

Run Time: 2 hours 45 mins

Overview

Transcriptomics and proteomics experiments often produce lists of genes or proteins that show statistically significant differential expression when samples from different experimental conditions, genotypes, or tissue types were compared. To investigate whether significantly differentially expressed genes or proteins were overrepresented in specific biological pathways and provide insights on their functions, curated databases that map genes and proteins to biological pathways were often used. One such curated database is the Reactome

(Jassal *et al*. 2020 NAR, 48(D1):2115-2125), which is a database of manually curated biological pathway maps (e.g. metabolic, signalling and regulatory pathways). The Reactome analysis engine also enables quantitative -omics data as input, using tools such as CAMERA (Wu and Smyth 2012 Mol. Cell Proteomics, 19(12):e133) and PADOG (Tarca et al. 2012 BMC Bioinformatics 3:136), to increase statistical power of identifying relevant pathways as compared to overrepresentation analysis with gene lists. The interactive web interface enables users to visually locate their protein of interest in the curated pathway map, while the Bioconductor/R interface enables analysis to be automated for larger or more complex analyses. This workshop will include an introduction to the Reactome database and hands-on tutorials, including:

- Analysis of a publicly available proteomics dataset using the interactive web interface,
- Comparison of proteomics and transcriptomics data using the ReactomeGSA Bioconductor/R package, and
- Analysis of a publicly available single-cell expression atlas dataset using the ReactomeGSA (Griss et al. 2020 NAR, 48(D1):D498-D503) Bioconductor/R interface.

The workshop is tailored for researchers of all levels of experience. The workshop will include step-by-step instructions on how to perform analysis on the web-based tool and also to run the R command lines on Google Colab. Some programming experiences (e.g. R in particular) is advantageous but is not essential to fully participate in this course.

The online environment for the participants to run the R command lines during the hands-on part of the tutorial will be Google Colab's R notebook (https://colab.to/r). Details will be delivered to people who have registered by email about one week prior to the workshop.

VIRAL METAGENOMICS WITH HECATOMB (Friday, 26th November)

Organisers

- Michael Roach (michael.roach@flinders.edu.au)
- Rob Edwards
- Scott Handley
- Sarah Giles

Start time: AEDT 10:30-16:30

Run Time: 6 hours

Overview

One particular virus has been at the forefront of everyone's mind recently. However, the impact from the pandemic is a drop in the ocean compared to the impact that viruses (including bacteriophages) have on our daily lives. Viruses are a core component of every microbiome and they influence every environment including human health and disease. While some are associated with disease states, most viruses are harmless or even beneficial. For instance, there is a renewed interest in the use of phages for combating the rise of antimicrobial resistance.

Viral metagenomics is fraught with challenges. Viruses are both highly diverse and poorly represented in reference databases. Viruses share large portions of sequence homology with other domains of life. Viral metagenomics studies are hence typically populated with false-positive hits while novel viruses fly under the radar. We've developed Hecatomb to address these issues and make accurate viral metagenomics accessible to anyone.

This workshop with cover an overview of the challenges associated with viral metagenomics; methods and approaches for viral enrichment and DNA extraction in the lab; a detailed overview of installing and running the Hecatomb pipeline for different platforms and usage cases; and an extensive hands-on component for visualizing and statistically interrogating your data. By the end of this workshop, you will have everything you need to perform your own viral metagenomics analysis.

Running sheet (Adelaide time)

1hr 09:00 - 10:00 - OPTIONAL - Installation of required software (Michael Roach & Rob Edwards)

1hr 10:00 - 11:00 - Challenges associated with viral metagenomics (Scott Handley)

45min 11:00 - 11:45 - Viral enrichment and DNA extraction methods (Sarah Giles)

45min 11:45 - 12:30 - Running Hecatomb demonstration (Michael Roach or Rob Edwards)

1hr 12:30 - 13:30 - Lunch break

1.25hr 13:30 - 14:45 - **Hands-on part 1**: Loading files and filtering alignments (Michael Roach or Rob Edwards)

1.25hr 14:45 - 16:00 - **Hands-on part 2**: Visualisation and statistical interrogation (Michael Roach or Rob Edwards)

30min 16:00 - 16:30 - OPTIONAL - Hands-on extra help and troubleshooting (Michael Roach & Rob Edwards)

International Invited Speakers

Roser Vento-Tormo

Wellcome Sanger Institute

Vento-Tormo's research interest is to understand the influence of cellular microenvironments on individual cellular identities and responses, in the context of immunity and development. Her team (https://ventolab.org/) employs single-cell and spatial transcriptomics methods to deconstruct the cell signals in human organs and tissues, and utilise this information to inform the reconstruction of novel in vitro models. Essential for this work, is the novel computational tools her team develops to build cell–cell interactions networks from transcriptomics data. Her training in genomics and bioinformatics puts her in a unique position to lead multidisciplinary projects. In her predoctoral research, she studied the interplay between cell signalling and epigenetic machinery key to regulating cellular fate decisions in myeloid cells. She pursued her postdoctoral studies in the Teichmann laboratory as an EMBO / HFSP fellow, where she developed CellPhoneDB.org, a computational tool to study cell-cell communication from single-cell transcriptomics data, and use it to disentangle the complex communication between maternal and fetal cells in the uterine-placental interface during early human pregnancy. Vento-Tormo work has been funded by many recognised international agencies (H2020, MRC, CZI, Wellcome-LEAP), and she has recently obtained the Early Career Research Award from the Biochemistry Society (2021).

Smita Krishnaswarmy

Yale School of Medicine

Smita Krishnaswamy is an Associate professor in Genetics and Computer Science. She is affiliated with the applied math program, computational biology program, Yale Center for Biomedical Data Science and Yale Cancer Center. Her lab works on the development of machine learning techniques to analyze high dimensional high throughput biomedical data. Her focus is on unsupervised machine learning methods, specifically manifold learning and deep learning techniques for detecting structure and patterns in data. She has developed algorithms for non-linear dimensionality reduction and visualization, learning data geometry, denoising, imputation, inference of multi-granular structure, and inference of feature networks from big data. Her group has applied these techniques to many data types such as single cell RNA-sequencing, mass cytometry, electronic health record, and connectomic data from a variety of systems. Specific application areas include immunology, immunotherapy, cancer, neuroscience, developmental biology and health outcomes. Smita has a Ph.D. in Computer Science and Engineering from the University of Michigan.

Isobel Parkin
University of Saskatchewan

Isobel Parkin has been a Research Scientist with Agriculture and Agri-Food Canada (AAFC) since 1999 and an adjunct Professor at the University of Saskatchewan since 2004. Her research interests include Brassica genomics, comparative genome organisation, global gene expression analysis, and abiotic stress responses. Together with Prof. Andrew Sharpe, she co-led the Canadian Canola Genome Sequencing (CanSeq) industry consortium project that developed a high quality genome sequence as part of the multinational Brassica Genome Sequencing initiative. Isobel's current research funding sources include the Global Institute for Food Security (GIFS), Saskatchewan Agriculture Development Fund, AAFC Crop Genomics Initiative, and canola producer groups.

Rohan Williams
Singapore Centre for Environmental Life Sciences (SCELSE)

Rohan Williams is an Australian computational biologist based at the Singapore Centre for Environmental Life Sciences Engineering (SCELSE), a Research Centre of Excellence co-hosted by the National University of Singapore and Nanyang Technological University. At SCELSE he is Head of the Integrative Analysis Unit, which combines research, service and training activity in bioinformatics and computational biology. His research interests lie in metagenomics and multi-omics of microbiomes and complex microbial communities, mostly in the context of wastewater bioprocess systems. Previously he was a research group leader at the Australian National University (2077-2011) and an NHMRC Peter Doherty Fellow at UNSW (2003-2007). He holds a PhD (UNSW) in Medicine and a BAppSc (UTS) in Physics. Williams is cofounder of the ASEAN Industrial Wastewater Genomics Consortium, a new initiative targeting microbial communities underpinning the agri-industrial systems in South East Asia and a scientific confounder of BluMaiden Biosciences Pte Ltd, a Singapore headquartered biotechnology platform company that is developing new small molecule therapeutics derived from human microbiomes.

Jill Moore

University of Massachusetts, Amherst

Dr. Moore's research focuses on the large scale integration of multi-omic datasets to better understand gene regulation, particularly in the context of human disease. She received her Bachelors of Science in Mathematics with minors in Biology and Chemistry as a member of the Commonwealth Honors College from the University of Massachusetts Amherst. She received her PhD in 2017 from the University of Massachusetts Medical School under Prof. Zhiping Weng where she was an active member of the Encyclopedia of DNA Elements (ENCODE) project. Since completing her graduate training, Dr. Moore has continued working with the ENCODE consortium as the Project Manager for the Data Analysis Center. In this role, she is very interested in improving ways to make multi-omic data and computational analyses accessible to the wider biological community.

Arndt von Haeseler

Vienna University and Medical University, Vienna, Austria

Professor Arndt von Haeseler is the scientific director of the Max F. Perutz Laboratories at the Vienna Biocenter and a professor of bioinformatics at the University of Vienna and the Medical University of Vienna. Arndt obtained his doctorate in mathematics from the University of Bielefeld in 1988. He habilitated in 1994 at the Department of Zoology of the University of Munich, where he remained as a lecturer until 1998. From 1998 until 2001, von Haeseler was a group leader at the Max Planck Institute for Evolutionary Anthropology in Leipzig. Subsequently, he was professor of bioinformatics at the University of Düsseldorf and, in 2005, he joined the Max F. Perutz Laboratories (MFPL), where he leads the Center for Integrative Bioinformatics Vienna (CIBIV). At the University of Vienna, he is the dean of the Center for Molecular Biology and at the Medical University of Vienna, he is the head of the Department for Medical Biochemistry.

Arndt's research focuses on developing computational methods for the reconstruction of phylogenetic trees. He co-authored the phylogenetics software packages TREEFINDER, TREE-PUZZLE, and its successor, IQ-TREE. He sits on the editorial boards of Molecular Biology and Evolution and BMC Evolutionary Biology. In 2015, von Haeseler was elected as a corresponding member of the mathematics and science class of the Austrian Academy of Sciences. Since 1999, he holds an honorary professorship in theoretical biology at the University of Leipzig.

National Invited Speakers

Matthew Lewsey

La Trobe University, Melbourne University

I am a hands-on lab biologist turned data-crunching genome scientist. My lab studies how plants perceive the world around them and interact with their environments by regulation of their genomes. We apply this work with commercial partners who grow a range of agricultural crops including cannabis, opium poppies, barley, oats and peas. I completed my PhD in Molecular Virology at the University of Cambridge with Prof John Carr. I next conducted postdoctoral research at the Salk Institute (La Jolla, USA) with Prof Joe Ecker and at the Centro Nacional de Biotecnología (Madrid, Spain) with Prof Roberto Solano. My projects at that time focused on applying functional genomics to understand plant hormone biology and small RNA signalling between organs. This work was performed in both labs as part of my Marie Curie Fellowship. I joined La Trobe University in April 2016 to further my interests in genomics, systems biology and plant science. In 2017 I became Director of the La Trobe Genomics Platform, and in 2019 Co-Deputy Director of the ARC Research Hub for Medicinal Agriculture.

Parwinder Kaur

The University of Western Australia

Assoc. Prof. Parwinder Kaur leads an innovative Translational Genomics research program that aims to translate fundamental science into ready-to-use solutions across the agricultural and medical sectors. Her DNA Lab team enables research to span the spectrum of scientific activities beyond the traditional ‘Lab-to-Landscape’ model, using new age technologies such as CRISPR, single-cell and 3D genomics. With DNA Zoo Australia she is on a mission to provide genomic empowerment to unique Australian biodiversity facilitating conservation efforts for the threatened and endangered species She has made substantial contributions to the field of biotechnology and was appointed as UWA “Be Inspired” for Agricultural & Environmental Biotechnology in 2019. Her studies tracking genomic variation to breed low methanogenic forages in Australia provided a new paradigm for reducing the environmental footprint of ruminants. She has been honored by the prestigious “Science and Innovation Award” for Young People in Agriculture, Fisheries and Forestry by the Australian Academy of Sciences in 2013. For DNA Zoo innovative work developments won the Microsoft's AI for Earth award for 2019-20. She is an active mentor for gender equity and GirlsXTech international ambassador working to close the gender gap in technology.

Katherine Pillman

Centre for Cancer Biology, University of South Australia

Dr Katherine Pillman is a bioinformatician and RAH Florey Fellow at the Centre for Cancer Biology in South Australia. She uses a variety of transcriptomic methods and next-gen sequencing data types to dissect gene regulation through analysis of microRNA biology and targeting, circular RNAs, alternative splicing, epigenetic modifications, gene regulatory networks and expression. She obtained her PhD in Molecular Biology from the University of Adelaide in 2009 working on transcriptional regulation in barley plants. Her postdoctoral research at Oregon State University involved RNA-seq analysis of stress-responsive gene regulatory networks in potato plants, which fuelled her interest and transition into the field of Bioinformatics. In 2012, she returned to Australia to take up her current role as Lead Bioinformatician in the Gene Regulatory Section at the Centre for Cancer Biology, working in collaboration with Prof. Greg Goodall. Career highlights include a seminal paper in Cell which identified the first protein known to control the formation of circular RNAs, and the subsequent discovery that this protein, Quaking, is a key emerging player in circular splicing and linear alternative splicing in cancer (published in EMBO J). Other work has identified genome-wide microRNA regulatory networks controlling cell invasion (also in EMBO J) and uncovered the cooperative role of microRNAs in cancer (published in Nucleic Acids Research and Cell Systems). In addition to her fellowship, she is a Co-Investigator on NHMRC Ideas and MRFF grants which fund her current work.

Anna Trigos

Peter MacCallum Cancer Centre

Dr. Anna Trigos is a postdoctoral researcher at the Peter MacCallum Cancer Centre in the team of A/Prof. Shahneen Sandhu in Prof. Rick Pearson’s lab. She was recently awarded two NHMRC Ideas grants and a CASS Foundation grant to investigate prostate cancer evolution and heterogeneity and develop novel spatial analysis methods to study the tumour immune microenvironment. In 2019 she was awarded the Joseph Sambrook Prize for Research Excellence and the Peter Mac Postgraduate Research Medal, and in 2020 she received the Lea Medal Award recognising emerging female researchers.

Ben Woodcroft

School of Biomedical Sciences, Queensland University of Technology

Starting from a computational background at the University of Queensland, Ben's broad interest in biological systems was sparked by an undergraduate project in protein structure with Dr. Nicholas Hamilton, then an honours project in Prof. Bernie Degnan's marine biology laboratory studying the genome structure of the most basal animals, sponges. Then he moved south to the University of Melbourne for his PhD, under the guidance of A/Prof. Stuart Ralph and Prof. Terry Speed, concentrating on the development of bioinformatic tools to understand the malaria parasite's complex cell biology. During his post-doctoral studies with Prof. Gene Tyson at the Australian Centre for Ecogenomics, UQ, and continuing until the present time as microbial informatics team leader within the Centre for Microbiome Research at QUT, he has studied the microbial world using informatic techniques. Direct sequencing of DNA and RNA derived from natural microbial systems has great scientific and applied potential, and the wealth of sequencing data bring many bioinformatic opportunities and challenges, including the recovery and annotation of genomes with strain-level specificity, exploring large public datasets and linking microbial communities with their function. A primary study site he has been involved with is Stordalen Mire, in northern Sweden, a permafrost thaw gradient home to complex microbial ecosystems. This system serves as a model for understanding how climate change is affecting microbial communities in thawing permafrost, which are in turn generating the potent greenhouse gas methane, exacerbating global warming. Primary contributions at this site involved the recovery of >1000 high quality genomes from the site, discovery of novel methanogenic lineages and linking specific community members to the isotopic composition of released methane, with implications for global climate modeling.

Barbara Holland

University of Tasmania

Barbara completed a PhD in Mathematical Biology at Massey University in New Zealand followed by postdoctoral studies at the Ruhr Universität Bochum (Germany) and in the Allan Wilson Centre for Molecular Ecology and Evolution (New Zealand). Prior to joining the University of Tasmania she worked as a Mathematics lecturer and researcher at Massey University. Since beginning her PhD she has enjoyed the challenge of working with biologists in trying to translate the problems they face into the language of mathematics. Biology is awash with data since the advent of DNA sequencing technology and this has opened up a range of very interesting research questions that require a combination of skills from mathematics, biology and computer science.

ABACBS Accepted Abstract List

Abstract #6

Jarny Choi (The University of Melbourne)
Creating interactive visualisation of gene expression data online using latest web technologies

Stemformatics.org is an established online gene expression data portal containing hundreds of carefully curated and annotated datasets. The website has recently been completely re-worked using latest software, and contains a suite of visualisation tools which enable easy exploration of the data, as well as an API server which can be used to access the data directly. A list of key features will be discussed, including the integrated data atlas where the user can project their own data and classify their cell types against a reference. Also discussed will be the lessons learnt from building an online data portal, including what resources may be required and which latest technologies may be most useful.

Abstract #12

Hasindu Gamaarachchi (Kinghorn Centre for Clinical Genomics, Garvan Institute of Medical Research, Sydney, NSW, Australia)
SLOW5: a new file format enables massive acceleration of nanopore sequencing data analysis

Nanopore sequencing is an emerging genomic technology with great potential. However, the storage and analysis of nanopore sequencing data have become major bottlenecks preventing more widespread adoption in research and clinical genomics. Here, we elucidate an inherent limitation in the file format used to store raw nanopore data, known as FAST5, that prevents efficient analysis on high-performance computing (HPC) systems. To overcome this, we have developed SLOW5, an alternative file format that permits efficient parallelisation and, thereby, acceleration of nanopore data analysis. For example, we show that using SLOW5 format, instead of FAST5, reduces the time and cost of genome-wide DNA methylation profiling by an order of magnitude on common HPC systems (up to >30X times), and delivers consistent improvements on a wide range of different architectures. With a simple, accessible file structure and significant reductions in size compared to FAST5 (~25% reduction for a typical human genome), SLOW5 format will deliver substantial benefits to all areas of the nanopore community. Example result: With the maximum resource allocation available on Australia's National Computing Infrastructure, genome-wide DNA methylation profiling on a single ~30X human genome sequencing dataset runs for >14 days at a cost of >$500 when using FAST5 files. We were able to complete whole-genome methylation profiling on a single 30X human dataset in just ~10.5 hours when using SLOW5 format as the input. SLOW5 format and all associated software are free and open source.
SLOW5 format specification documents: https://hasindu2008.github.io/slow5specs;
Slow5lib: https://hasindu2008.github.io/slow5lib;
Slow5tools:https://hasindu2008.github.io/slow5tools;
Pre-print: https://www.biorxiv.org/content/10.1101/2021.06.29.450255v1.

Abstract #13

Paola Cornejo-Paramo (Victor Chang Cardiac Research Institute)
TF binding motif composition of cis-regulatory regions in a sea sponge species suggest the existence of some constraints in metazoan regulatory grammar

Chromatin accessibility plays a major role in regulating animal gene expression and therefore it is key to decipher animal embryogenesis. Here we profile chromatin accessibility of six embryonic stages, along with a larval and adult stage in the sea sponge Amphimedon queenslandica, a member of the early branching phylum Porifera, in order to study the regulatory landscape of metazoan embryogenesis. On the grounds that most transcription factor (TF) families are conserved in eukaryotes, we examine the TF binding motif composition of Amphimedon cis-regulatory regions. We leveraged the high performance of tree-based machine learning models and tree ensemble interpretation algorithms to extract complex patterns and identify some TF binding motifs that are commonly enriched in Amphimedon cis-regulatory regions. We further tested the machine learning models trained on Amphimedon on other species data and were able to predict cis-regulatory regions in other animal species but not in the close relative of animals, *Capsaspora owczarzaki*. Overall, we find that animal cis-regulatory regions share some resemblance and constraints in motif composition besides the extreme divergent life strategies adopted by different phyla.

Abstract #17

Aditya Sarkar (School of Computing and Electrical Engineering, Indian Institute of Technology Mandi, North Campus, Kamand, Mandi, HP 175005, India)
Comprehensive analysis of microblogging landscape in science and medicine

Twitter is one of the most popular microblogging and social networking services, where users can post, retweet, comment, and engage in collaborative discussions. However, improper usage of Twitter can be detrimental to science and even have a negative impact on mental health. Thus, analysing tweets and Twitter data of various researchers will help us to deduce appropriate ways of using Twitter to advance in our research careers. Existing literature has analysed the activity of scientists on Twitter, such as studying the relationship between Twitter mentions and article citations, determining the benefits of Twitter in the development and distribution of scientific knowledge, relevant metrics for prediction of highly cited articles, and type of content that researchers tweet etc. For example, Eysenbach et al performed an analysis of tweets and citations and how one can predict citations using tweets. Most of the existing literature analysed a limited number of researchers, compromising the generalizability of derived results. In our study, we have taken a comprehensive and systematic approach to analyse 4,257,739 scientists who published research papers on PubMed using data-driven methods. We observed various parameters like number of followers, number of friends, citation count and K-index. There was a low correlation between number of followers and number of friends and on an average scientists have 2111.45 followers and 554.6 friends. Citation count and number of followers are also having low correlation and researchers have an average citation count of 292.37. Average K-index of researchers is 15.56 and there is no significant relationship between number of followers and K-index.

Abstract #18

Angelita Liang (UNSW Sydney)
A universal naming system for human alternative splicing

Our understanding of alternative RNA isoforms has progressed rapidly over the past decade. Advances in bioinformatic tools to analyse isoform usage from short-read RNA-seq and the advent of long-read RNA-seq have led to the rapid expansion of the RNA splicing field, with an ever-increasing number of annotated and novel alternative splicing events characterized both structurally and functionally. Yet, there is a lack of standard nomenclature to describe these splicing events. This issue is compounded by the fact that there is no stable system to refer to transcript isoforms by their HGNC gene symbol and existing naming conventions are ineffective as text representations of alternative exon structures. We propose a simple and compact shorthand notation to represent splicing events of any complexity without the aid of a figure, hereby termed the "Extended Delta Notation". An amalgamation of existing naming conventions already familiar to bioinformatics and molecular biologists, our proposed nomenclature uses HGNC gene symbols to improve readability whilst preserving the reproducibility of Ensembl, LRG and RefSeq reference transcripts. In addition, the notation conveys all structural information whilst remaining concise. To assist in the understanding, interpretation and generation of the Notation, we have an R Shiny web app currently in development, which will serve as a hub for the RNA splicing community. In turn, the adoption of consistent naming conventions in literature will form the basis of more uniform standards of reporting in the future, alleviating the reproducibility crisis.

Abstract #20

Patricia Sullivan (Children’s Cancer Institute, University of New South Wales, Randwick, Sydney, NSW, Australia)
Breaking the black box: Dissecting deep neural network models to reveal splicing motifs

SpliceAI is a state-of-the-art deep neural network that outperforms previous in silico approaches to model splicing. However, deep neural networks are notoriously black-box, making it challenging to elucidate biological relevance from the learned models. Explainable artificial intelligence (XAI) methods make it possible to interrogate a deep neural network's behaviour and can potentially uncover the biological patterns learnt by SpliceAI. We use an XAI method called activation maximisation (AM) by gradient ascent to explain individual nodes within the SpliceAI neural network. AM works by identifying which aspects of the input contribute most to the activity of that node. Using this method, we produced 160 position weight matrices (PWMs) representing the 103nt input sequences that contribute most to the activity of each node in SpliceAI's final CNN layer, and thus help characterise the hidden latent space of the model. Approximately one-third of the PWMs were classified as a known splicing motif, most of which being variations of the canonical acceptor and donor sites. Many unclassified motifs appeared to demarcate exon/intron boundaries, prompting us to investigate the features that influence SpliceAI's predictions. Using in silico perturbation experiments, we found that SpliceAI has implicitly learnt the codon triplet code. Given a notable absence of non-canonical motifs, we used validated splice-altering mutations to confirm that SpliceAI under-performs on non-canonical variants. Despite SpliceAI's state-of-the-art performance, these limitations could considerably impact the results of critical scientific applications. Thus, we highlight the importance of rigorously dissecting black-box neural networks as part of the model development and operationalisation workflow.

Abstract #21

Andy Tran (University of Sydney)
Using multimodal single cell data to predict regulatory gene relationships and to build a computational direct cell reprogramming model

Cell reprogramming offers a potential treatment to many diseases, by regenerating specialized somatic cells. Despite decades of research, discovering the transcription factors that promote direct cell reprogramming has largely been accomplished through trial and error, a time-consuming and costly method. A computational model for direct cell reprogramming, however, could guide the hypothesis formulation and experimental validation, to efficiently utilize time and resources. Current methods often cannot account for the heterogeneity observed in reprogramming, or they only make short-term predictions, without modelling the entire reprogramming process. Here, we present scREMOTE, a novel computational model for direct cell reprogramming that leverages single cell multiomics data, enabling a more holistic view of the regulatory mechanisms at cellular resolution. This is achieved by first identifying the regulatory potential of each transcription factor and gene to uncover regulatory relationships, then a regression model is built to estimate the effect of transcription factor perturbations. We show that scREMOTE successfully predicts the long-term effect of overexpressing two key transcription factors in hair follicle development by capturing higher-order gene regulations. Together, this demonstrates that integrating the multimodal processes governing gene regulation creates a more accurate model for direct cell reprogramming with significant potential to accelerate research in regenerative medicine.

Abstract #23

Rebecca C Poulos (ProCan, Children’s Medical Research Institute, Faculty of Medicine and Health, The University of Sydney, Westmead, NSW, Australia)
A pan-cancer proteomic map of 949 human cell lines

Proteomic data can reveal novel associations between genotype and phenotype, beyond what is apparent from genomics or transcriptomics alone. However, a lack of large proteomic datasets across a range of cancer types has limited our understanding of proteome network organisation and regulation. We produced a pan-cancer proteomic map derived from 949 human cancer cell lines. The map encompasses more than 40 cancer types derived from over 28 distinct human tissues. The samples were processed with a clinically-relevant workflow involving rapid and minimally complex sample preparation, quantifying 8,500 proteins. The raw proteomic data were acquired by data independent acquisition mass spectrometry (DIA-MS) at ProCan in Australia. The processed data were analysed with a bespoke deep learning-based pipeline (DeeProM) that integrates multi-omics, CRISPR-Cas9 gene essentiality and drug sensitivity information produced at the Wellcome Sanger Institute. First, our findings reveal pervasive post-transcriptional modification and thousands of putative protein biomarkers of cancer vulnerabilities. Second, DeeProM statistics show that a fraction of the proteome can confer similar predictive power to the entire transcriptome. This has key implications for the clinical application of proteomics in drug response prediction. Third, we demonstrate that a random proportion of the identified proteins can provide robust predictions of cancer cell phenotypes, underpinning the concept of pervasive co-regulation of protein networks. This pan-cancer cell line proteomic map is a comprehensive resource that expands our understanding of cancer proteomes. These data reveal principles of cancer cell phenotypes, including genetic vulnerabilities and drug sensitivities, that are important for developing novel targeted anticancer therapies.

Abstract #26

James M. Ferguson (Kinghorn Centre for Clinical Genomics, Garvan Institute of Medical Research, Sydney, NSW, Australia)

InterARTIC: an interactive web application for whole-genome nanopore sequencing analysis of SARS-CoV-2 and other viruses

InterARTIC is an interactive web application for the analysis of viral whole-genome sequencing (WGS) data generated on Oxford Nanopore Technologies (ONT) devices. A graphical interface enables users with no bioinformatics expertise to analyse WGS experiments and reconstruct consensus genome sequences from individual isolates of viruses, such as SARS-CoV-2. InterARTIC is intended to facilitate widespread adoption and standardisation of ONT sequencing for viral surveillance and molecular epidemiology. We demonstrate the use of InterARTIC for the analysis of ONT viral WGS data from SARS-CoV-2 and Ebola virus, using a laptop computer or the internal computer on an ONT GridION sequencing device. We showcase the intuitive graphical interface, workflow customisation capabilities and job-scheduling system that facilitate execution of small- and large-scale WGS projects on any common virus.

Abstract #27

Isobel J Beasley (Melbourne Integrative Genomics, The University of Melbourne, Royal Parade, 3010, Parkville, Victoria, Australia)

Predicting the cross-population portability of human expression quantitative trait loci (eQTLs)

Individuals of European ancestry disproportionately dominate participation in human genetic studies, to the detriment of scientific inquiry and the equitable translation of genomic research. One genomic study type, expression quantitative trait locus (eQTL) mapping, has been a valuable tool for understanding the regulatory consequences of disease associated genetic variants. However, not all associations between genotype at given loci and variation in gene expression (eQTLs) are shared across populations. Understanding the features of these eQTLs could enable a priori prediction of eQTLs in understudied populations. Here we use summary statistics from three previous multi-population eQTL studies to classify eQTLs as population-specific or shared between at least two populations (African American, European American, Indonesian etc.). We train machine learning models to predict whether an eQTL is specific to its discovery population using publicly available information on the evolutionary, functional, and expression features of these eQTLs. As per similar applications, we find trained random forest classifiers perform better than other algorithms on a held out test set of eQTLs (auROC > 0.8). Features of gene conservation (e.g. LOEUF, phyloP), Gene Ontology, allele frequency and eQTL effect size have the highest importance scores in our best performing models. This result suggests these features could be used to understand gene regulatory trait differences across populations. Since current Eurocentric biases in genomic resources are likely to persist for some time, our approach could be an important step toward a more equitable understanding of gene regulation, and hence more equitable personalised medicine.

Abstract #36

Venta Terauds (University of Tasmania)

Genome algebras in action

We demonstrate some features of our recently introduced genome algebra framework by applying it to echinoderm mitochondrial genome data. In particular, the framework facilitates a fine-grained consideration of rearrangement models. We show how varying the underlying rearrangement model and the choice of genomic distance measure can affect the phylogeny that is produced.

Abstract #39

Kaitao Lai (Ancestry & Health Genomics Laboratory, Charles Perkins Centre, School of Medical Sciences, University of Sydney, New South Wales, Australia)

Shotgun microbial profiling reveals geographic disparities in aggressive prostate cancer

The significant geographic disparity has been identified in prostate cancer (PCa). Specifically, men from Sub-Saharan Africa have the highest mortality rates, over double the global average. Additionally, rates for infection-associated cancers within Sub-Saharan Africa are also over double global averages. While PCa risk is associated with inflammation, no pathological agent has been identified. Our hypothesis, a microbial pathogen is contributing, at least in part, to adverse PCa outcomes for men in Africa.We have investigated the prostate tissue microbiome from 176 men from Sub-Saharan Africa (African) and Australia (European ancestry), with or without PCa, and a bias towards aggressive disease. In the differential analysis, the relative abundances of different bacterial taxa have been found to be associated with the risk level of PCa and population with sampled countries. The relative abundances of Klebsiella pneumoniae, Comamonas testosterone and Moraxella osloensis, significantly increased with high-risk PCa, were associated with the risk of high-grade PCa; while Akkermansia muciniphila and Anaerostipes hadrus increased with men with benign prostate, have been considered potential probiotic. Escherichia coli, identified from all patient groups, are crucial pathogens and may cause significant morbidity and mortality of PCa5. The highest mortality rates in Sub-Saharan Africa may be associated with the Human mastadenovirus C, found in all Africa sampled groups (except the Australia group). Geographic disparity is significant in bacteria diversity from prostate tissue for the PCa patients. The PCa tissue microbiota profile could be a novel method for identifying pathogens, the prediction of high-risk PCa and for developing the treatment of PCa.

Abstract #40

Ulf Schmitz (College of Public Health, Medical & Vet Sciences, James Cook University, Sydney, Australia)

Multi-omics data analysis identifies epigenetic regulators of alternative splicing

The phenomenon of widespread and dynamic intron retention (IR) programs in cells of vertebrate species has recently gained increasing attention. It has been shown that IR is involved in a multitude of cell-physiological processes, while aberrant IR profiles have been associated with numerous human diseases including several cancers. Despite consistent reports about intrinsic sequence features that predispose introns to being retained, conflicting findings about cell type or condition-specific IR regulation by trans-regulatory and epigenetic mechanisms demand an unbiased and systematic analysis of IR in a controlled experimental setting. We integrated matched mRNA sequencing (RNA-seq), whole genome bisulfite sequencing (WGBS), nucleosome occupancy methylome sequencing (NOMe-Seq), and chromatin immunoprecipitation sequencing (ChIP-seq) data from primary human myeloid and lymphoid cells. Using these multi-omics data and machine learning we trained two complementary models to determine the role of epigenetic factors in the regulation of IR in cells of the innate immune system. Our results suggest that intrinsic characteristics are key for introns to evade splicing and that epigenetic marks can modulate IR levels. However, cell type-specific IR profiles are largely caused by changes in chromatin accessibility, whereby predisposed introns in nucleosome free regions are more likely to be retained. This study is the first to demonstrate the important role of nucleosome occupancy in IR regulation. Our results have profound implications for the analysis of other forms of alternative splicing as well. Since an increasing number of studies describe pathogenic alterations in splicing regulation and therapeutic approaches targeting aberrant splicing, our findings will inform novel epigenetic therapy development.

Abstract #41

Jeffrey Pullin (University of Melbourne)

Benchmarking methods for selecting marker genes in single cell RNA sequencing data

Over the past ten years, the advent of single cell RNA sequencing (scRNA-seq) technology has allowed scientists to study the transcriptomes of individual cells in unprecedented detail. A crucial step in all scRNA-seq analysis is to identify

the cell types present in the sample. Currently, this step is often performed using so-called marker genes. First, cells are partitioned into clusters using a clustering algorithm. Then, computational methods are used to select a small subset of genes which distinguish each cluster of cells: the marker genes. The data for these genes can be interpreted by a biologist to identify the cell type of each cluster. Today, there are over 40 methods for identifying marker genes, ranging from simple statistical tests to recent methods developed explicitly for the task. However, no independent studies comparing and scrutinizing the competing methods currently exist. In my talk I will describe the space of existing methods for selecting marker genes, and then present the results of a comprehensive benchmarking of the methods. We compare the performance of the methods on over 50 datasets simulated using the Splatter package with parameters estimated from 8 publicly available datasets. Methods are also compared on the predictive performance of the marker genes they select and their ability to recapitulate known marker genes across a range of real datasets. In addition, I will highlight unexpected implementation details of the widely used methods implemented in the Scanpy and Seurat packages. I will conclude with recommendations for best practice selection of marker genes.

Abstract #42

Jiru Han (Population Health and Immunity Division, The Walter and Eliza Hall Institute of Medical Research, Melbourne, VIC, Australia)

Population-level genome-wide STR typing in Plasmodium species reveals higher resolution population structure and genetic diversity relative to SNP typing

Short tandem repeats (STRs) are highly informative genetic markers that have been used extensively in population genetics to study the population structure and genetic diversity in Plasmodium species. In this study, we performed the first large-scale in-silico STR typing study using HipSTR (Willems et al. 2017) in more than 3,000 P. falciparum (MalariaGEN Plasmodium falciparum Community Project) and 174 P. vivax (Plasmodium vivax Genome Variation project and P. vivax data from Hupalo et al. 2016) WGS samples sourced from global malaria hot-spots. We developed a multivariable logistic regression model for the measurement and prediction of the quality of STRs based on gold standard SNP genotyping data, which provides a high fidelity set of STRs from across the whole genome (6,768 from P. falciparum and 3,496 from P. vivax). Additionally, we used this set of genome-wide high-quality STRs to study parasite population genetics and compare them to genome-wide SNP genotyping data, revealing both high consistency with SNP based signals, as well as identifying some signals unique to the STR marker data. We also explored the biological importance of STR variation in different populations and identified STR loci that may be linked to antimalarial drug resistance. These results demonstrate that the identification of highly informative STR markers from large population screening is a powerful approach to study the genetic diversity, population structures and genomic signatures of selection in P. falciparum and P. vivax.

Abstract #45

Zhaoxiang Cai (Zhaoxiang Cai - ProCan, Children’s Medical Research Institute, Faculty of Medicine and Health, The University of Sydney, Westmead, NSW, Australia)

Integrating multi-omics data with biological knowledge by Transformer-based deep learning

Omics data analysis, powered by machine learning, has significantly improved cancer diagnosis and prognosis. However, most machine learning methods consider each gene as an independent feature, failing to integrate experimentally-acquired gene regulation and pathway information. The benefit of utilising this information increases in the era of multi-omics, because gene regulation is the key mechanism that links different omic layers together. Here, we present an interpretable deep learning model, DeepPathNet, which uses cancer-specific pathway information for both single and multi-omics data analysis. DeePathNet leverages the cutting-edge deep learning technique, Transformer, which is derived from the field of natural language processing, to model complex interactions between pathways from omics data. The computation of self-attention in the Transformer module allows DeePathNet to learn the encoding of pathways to achieve superior predictive performance and interpretability. Techniques such as drop out layers are also integrated into DeePathNet to maximise its generalisability for unseen data. Moreover, DeePathNet supports any number of omics layers and can handle missing values. Using multiple evaluation metrics, we demonstrate that DeePathNet robustly outperforms traditional methods for predicting drug response and cancer type on four publicly available datasets, namely COSMIC Cell Lines, Genomics of Drug Sensitivity in Cancer (GDSC), Cancer Cell Line Encyclopedia (CCLE) and Cancer Therapeutics Response Portal (CTRP). DeePathNet also provides reliable model interpretation, potentially enabling biomarker discoveries at the pathway level. Using the Transformer, DeePathNet is the first method that supports multi-omics data analysis, integrates cancer pathway knowledge into modelling, and provides pathway-level model explanation.

Abstract #47

Andreas Zankl (Children's Hospital Westmead, University of Sydney, Garvan Institute)
Ontoclick: a web browser extension to facilitate biomedical knowledge curation

The world's biomedical literature has achieved a scale that the human brain cannot possibly process. Computers could assist with searching through this vast sea of knowledge, but this requires knowledge curation, converting unstructured text into well-defined units of information, such as a set of Human Phenotype Ontology terms, that a computer can process. Knowledge curation can be a repetitive and laborious process but without knowledge curation, vast amounts of knowledge in the biomedical literature will remain untapped. The paucity of user-friendly tools is one of the reasons for the lack of widespread adoption of good biomedical knowledge curation practices. Here we present Ontoclick, a web browser extension that streamlines the process of annotating a text span with a relevant ontology term, a key component of knowledge curation. We hope this tool will make biocuration more accessible to a wider audience of biomedical researchers.

Abstract #53
Jiayue-Clara Jiang, Chenwen Hu, Sonia Shah (The University of Queensland)
Using transcriptomic and statistical genomic approaches to investigate anti-depressive effects of statins

Statins are a class of lipid-lowering medications commonly prescribed for treating cardiovascular diseases. Statins exert their pharmacological effects via inhibition of the HMG-CoA reductase (HMGCR). Observational studies have reported both therapeutic and adverse unintended effects of statins, including conflicting evidence on their association with depression risk. The effect of statins on depression is an important consideration when treating the increasing prevalence of comorbid individuals with both cardiovascular disease and depression.||In this study, we leveraged the power of transcriptomics and genome-wide association studies (GWAS) to interrogate the anti-depressive effects of statins. Using the LINCS L1000 data, we profiled transcriptomic perturbations induced by commonly prescribed medications, including statins. The statin-induced transcriptomic signatures were found to exhibit similarity to the transcriptomic impacts of antidepressants. Functional enrichment analysis of genes commonly perturbed by statins and antidepressants indicated functional involvement in immune processes, which have been shown to play an important role in depression. ||To determine if the observed perturbation of immune pathways was an on-target (mediated through HMGCR inhibition) or off-target (mediated through unintended targets, specifically ITGAL and HDAC2) effect, we queried published large-scale GWAS summary statistics to determine whether eQTLs of these genes were associated with immune-related traits, and conducted Mendelian randomisation analysis to determine the causal effect of each target gene on depression risk. ||Overall, we showed that while statin use was not associated with the risk of developing depression, transcriptomic and genomic analyses revealed an association of statins with changes in immune response pathways, which were also perturbed by antidepressants. We also observed an association between HMGCR and ITGAL eQTLs with haematological parameters, suggesting that both on-target and off-target effects could potentially contribute to modulating disease symptoms.

Abstract #62

Sebastian Burgstaller-Muehlbacher (Center for Integrative Bioinformatics Vienna, Max Perutz Labs, University of Vienna and Medical University of Vienna, Vienna, Austriaustria)
Model selection on empirical data using deep learning

Selecting the correct model of sequence evolution (mSE) for a multiple sequence alignment (MSA) constitutes the first step of tree reconstruction. State of the art approaches for inferring nucleotide models mostly apply maximum likelihood (ML) methods. Here, we demonstrate that neural networks can infer the correct mSE including the shape parameter α of the Γ-distribution. A Residual Neural Network (Resnet) was trained with the six most frequently used mSE (JC, K2P, F81, HKY, TN93 and GTR), whereas a Long Short-Term Memory (LSTM) network with attention was trained to determine **α**. Our results show that the CNN correctly identifies the mSE in a range of 51.94% to 100%, depending on the true mSE. Thus, it is comparable to results of IQ-Tree. Similar accuracies were obtained for α. However, the trained networks need substantially less computing time (up to a 60x speedup), depending on the size of the MSA. We demonstrate, for the first time, that neural networks can be used to identify the correct mSE as well as rate heterogeneity of an MSA. Furthermore, we intend to generalize our approach so all mSE of relevance in phylogenetics can be inferred using neural networks, conferring a substantial reduction in computational requirements to model selection.

Abstract #63

Yunwei Zhang (School of Mathematics and Statistics, The University of Sydney, Sydney, Australia)
HISP: cohort heterogeneity identification and risk factors detection for survival prediction

Cohort heterogeneity exists in many clinical and omics datasets. Discovering sub-cohort specific risk factors is crucial for personalised medicine and providing accurate survival prediction. While numerous unsupervised clustering methods have been applied to identify potential sub-cohorts in the population data, the detection of the associated risk factors and the corresponding survival modelling are conducted in separate steps. Therefore, cohort specific prediction is not guaranteed to be superior than fitting the full model. To address this research gap, we design Heterogeneity Identification and Survival Prediction (HISP), which is a hybrid workflow that uses survival model predictability to guide sub-cohort identification using both unsupervised and supervised learning strategies. Two key innovative ideas in HISP involve (1) using model predictability to guide ensemble methods selection; (2) using the first layer of the data (recipient features) for sub-cohort identification while using all three layers of data (recipients, donors and compatibilities variables) in the prediction. Here we show that by applying HISP to the Australian and New Zealand dialysis and transplant registry data (ANZDATA, 2008-2017), different recipient subgroups are identified with different sets of risk factors driving their survival outcomes to provide accurate post-transplant graft survival prediction. HISP has the capacity and flexibility to be applied to various data modalities such as proteomics, genomics and metabolomics to identify sub-cohorts in a risk population to offer precise survival predictions.

Abstract #67

Aidan P. Tay (Australian e-Health Research Centre, Commonwealth Scientific and Industrial Research Organisation, New South Wales, Sydney, Australia)
INSIDER: alignment-free detection of foreign DNA sequences

External DNA sequences can be inserted into an organism's genome either through natural processes such as gene transfer, or through targeted genome engineering strategies. Being able to robustly identify such foreign DNA is a crucial capability for health and biosecurity applications, such as anti-microbial resistance (AMR) detection or monitoring gene drives. This capability does not exist for poorly characterised host genomes or with limited information about the integrated sequence. To address this, we developed the INserted Sequence Information DEtectoR (INSIDER). INSIDER analyses whole genome sequencing data and identifies segments of potentially foreign origin by their significant shift in k-mer signatures. We demonstrate the power of INSIDER to separate integrated DNA sequences from normal genomic sequences on a synthetic dataset simulating the insertion of a CRISPR-Cas gene drive into wild-type yeast. As a proof-of-concept, we use INSIDER to detect the exact AMR plasmid in whole genome sequencing data from a Citrobacter freundii patient isolate. INSIDER streamlines the process of identifying integrated DNA in poorly characterised wild species or when the insert is of unknown origin, thus enhancing the monitoring of emerging biosecurity threats.

Abstract #68

Chelsea Mayoh (Children's Cancer Institute, Lowy Cancer Centre, UNSW Sydney, Kensington, New South Wales, Australia)
Utilising the transcriptome to functionally resolve molecular aberrations in precision medicine

The ZERO childhood cancer program integrates whole genome sequencing (WGS) and transcriptome sequencing (RNA-seq) to provide an integrated multi-omics, precision medicine pipeline to improve treatment outcomes in high-risk paediatric cancer. We have developed a comprehensive RNA-seq pipeline that integrates with WGS to accurately identify molecular aberrations and infer their predicted function and pathogenicity. WGS identifies complex structural rearrangements however can have difficulty resolving the expressed fusion product, its translation and associated retained protein domains which can change diagnosis or treatment. RNA-seq was able to resolve these complex structural rearrangements and identified pathogenic fusions that WGS was unable to identify. For example, a CIC-DUX4L fusion unidentified from WGS and an EML4-ALK fusion where WGS identified a multi-hop complex structural rearrangement. Mutation calling from RNA-seq identified 90% of the aberrations reported from WGS. Of the 10% not identified, the gene is not expressed in 53% and 21% were either nonsense mediated decay or allele specific expression. RNA influenced the clinical actionability of 55 (11%) tumour suppressor gene mutations identified as heterozygous in DNA but homozygous in RNA, having a direct clinical impact. Furthermore, RNA identified 34 (7%) splice site mutations where 88% were aberrantly spliced in the RNA either increasing or confirming the pathogenicity of the variant. Here we will present the added clinical utility that a comprehensive RNA-seq pipeline provides to a precision medicine program on 477 patients enrolled on ZERO. In addition, we explore the question can we use RNA-seq as a stand-alone platform in precision medicine?

Abstract #77

Feargal Ryan (Precision Medicine Theme, South Australian Health and Medical Research Institute, Adelaide, SA 5001, Australia)
Long-term perturbation of the peripheral immune system months after SARS-CoV-2 infection

Background Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) is a highly infectious respiratory virus which is the causative agent behind the Coronavirus Disease 2019 (COVID-19) pandemic. It is increasingly clear that recovered individuals, even those who had mild COVID-19, can suffer from persistent symptoms for many months after infection, which is popularly referred to as, "long COVID". However despite the plethora of research on COVID-19 relatively little is known about the molecular underpinnings of these long-term effects. Methods We have undertaken an integrated analysis of immune responses in blood at a transcriptional, cellular, and serological level at 12-, 16-, and 24-weeks post-infection (wpi) in 69 patients recovering from mild, moderate, severe, or critical COVID-19 in comparison to uninfected controls. Results Antibody responses were largely stable up to 24wpi and correlated with disease severity. Deep immunophenotyping revealed significant differences in multiple innate (NK cells, LD neutrophils, CXCR3+ monocytes) and adaptive immune populations (T helper, T follicular helper and regulatory T cells) in convalescent patients compared to controls, which were most strongly evident at 12 and 16wpi. RNASeq revealed perturbations to gene expression in COVID-19 convalescents until 6 months post-infection, a subset of which was associated with ongoing "long COVID" related symptoms. Conclusion Variation in the rate of recovery from infection at a cellular and transcriptional level may explain the persistence of symptoms associated with long COVID in some individuals.

Abstract #81

Alejandro Casar (The University of Melbourne, School of Mathematics and Statistics, Victoria, Australia)
Investigating how transcriptional plasticity drives drug resistance in cancer by combining computational modelling with single cell genomics

Transcriptional plasticity is a phenomenon where cells reversibly change molecular states by altering their gene expression. Recent evidence suggests transcriptional plasticity is a key driver of the emergence of drug resistance.
We designed a computational model that simulates Acute Myeloid Leukemia (AML) as a stochastic process at a discrete time resolution. It is based on the rates of DNA mutation, cell division, cell death and switching rates between sensitive and resistant states. We model tumors with initial growth rates between 0.001% and 0.1%; with driver mutation rates

between 3.4*10^(-8) and 3.4*10^(-5); testing switching rates between 4.0*10^(-7) and 9.0*10^(-6) from sensitive to resistant, and 0.5 to 50000 times those values for the resistant to sensitive switching rate. Results so far suggest that low rates of transcriptional plasticity have little effect on the growth and development of the cancer pre-treatment. However, as plasticity increases, we can see the tumor develop differently. Significant correlation was observed between tumor fitness, representing the weighted average net growth rate of all the different subpopulations of cells within in the tumor, and the number of resistant cells, which will be explored further. We are following up to understand the interactions between each independent simulation parameter, transcriptional plasticity, and clinical outcomes. To understand how transcriptional plasticity drives drug resistance in cancer, different drug treatment regimens for AML will be simulated and model predictions validated in mouse models of AML using SPLINTR, a synthetic expressed barcoding strategy that allows for cell lineage tracking.

Abstract #83

Tyrone Chen (School of Biological Sciences, Monash University, Clayton, VIC, 3800, Australia)
Identifying the complete functional and regulatory signatures driving antimicrobial resistance in sepsis.

Sepsis, a bloodstream infection, has a high mortality rate and is a common comorbidity in cancer. Rising antimicrobial resistance (AMR) globally amplifies this medical condition. Recently, the rapid growth in genotyping and multi-omics datasets [1] has provided an opportunity to investigate AMR at a higher resolution. We therefore investigated available heterogeneous datasets to study organism-specific resistance mechanisms of AMR. A major obstacle to a unified analysis of such diverse datasets are their distinct formats. To overcome this, we are refining a universal, annotation-less framework based on our previous work which showed that it is viable to use genomic data as raw fastq sequences [2], k-mers or abundance matrices [3,4] to recover signals of interest such as unannotated biomarkers [1]. We performed an integrative analysis on matched transcriptome, proteome and metabolome profiles of bacterial strains known to cause sepsis. Next, we re-coded the genome using natural language processing techniques, allowing us to overlay regulatory information on the functional omics layers. Using this technique, we recovered informative markers of pathogenicity. Scaling this method will allow us to identify complete functional and regulatory signatures driving AMR in sepsis. Furthermore, our approach is not restricted to a single biological system, and can be scaled out to study a different disease-pathogen interaction. [References:] [1] Chen T and Tyagi S. GigaScience 9.6 (2020). [2] Chen T, et al. [version 1; not peer reviewed] F1000Research (2019) (poster). [3] Chen T et al. Briefings in Bioinformatics (2021). [4] Chen T et al. [version 1; awaiting peer review] F1000Research (2021).

Abstract #95

Sabrina Yan (Computational Biology, Children's Cancer Institute, Sydney, Australia)
Helium: Automatic variant prioritization for surfacing cancer drivers

Zero Childhood Cancer is a precision medicine program at the Children’s Cancer Institute that aims to improve clinical outcomes for cancer patients through identifying optimal, personalised treatments based on their tumours unique genetic characteristics. This is done using bioinformatic tools that identify millions of genetic changes in the tumour. The major challenge is in identifying the medically-relevant mutations that could lead to better drugs being recommended for treatment. To address this challenge, the Computational Biology group has developed Helium, a cancer variant prioritisation tool. Helium `floats` the most relevant variants to the top for cancer experts to manually curate and tag those responsible for the patient’s disease. Helium scores variants based on combinations of expertly tagged annotations. Both the combinations of annotations used and their relative weighted contribution to the score are completely customisable using a single configuration file. This flexibility allows Helium to be a versatile tool with many potential use cases. It is written as an npm cli tool and package, allowing anyone with command line experience to run this with any csv, tsv or json input file. The tool has also been dockerized and wrapped in CWL for integrating into larger workflows. Helium can score 500 variants per second and is fast enough to incorporate as a live scoring tool directly into our variant curation web portal, ZeroDash, or be used to pre-score variants in our database for easy retrieval. We hope that the configurability and flexibility of Helium makes it a useful variant ranking tool for many in the precision medicine world.

Abstract #98

Andreas Halman (Peter MacCallum Cancer Centre, Melbourne, VIC, Australia)
STRipy: a graphical application for detecting short tandem repeat expansions in all known pathogenic loci

There are a handful of tools available for genotyping short tandem repeats (STRs) from high-throughput sequencing data. However, there are several limitations for application in a clinical setting. Firstly, many of these tools are command-line based. Secondly, without configuration, they can genotype only half of the known disease-causing loci and finally genotyping lengths are often limited to either read or fragment length of the data which is less than pathogenic cut-off for some diseases. Moreover, repeat lengths in a general population are not well defined for many STR loci which complicates the interpretation of results. To address these issues, we have created a new software called STRipy that has an intuitive graphical interface which significantly simplifies STRs genotyping from human short read sequencing data, making it accessible for everyone. STRipy is configured to target all known disease-causing STRs loci, including the rarest and most recently discovered. We have incorporated an established tool, ExpansionHunter, as a part of the software to perform accurate genotyping and created additional functionality to enable detection of long alleles.
Additionally, we have analysed over two and half thousand samples from various populations and acquired genotypes for each pathogenic STR locus which will provide further insights, particularly for newly discovered loci, into the normal range of STR lengths. This, along with information sourced from the literature has been also used to curate a comprehensive database of pathogenic STRs which is made freely accessible online at stripy.org.

Abstract #102

Daniel Cameron (Walter and Eliza Hall Institute of Medical Research)
Single breakends: a new paradigm for structural variant calling

The reliable detection of structural variants is critical to understanding the role genome architecture plays in health and disease. A fundamental obstacle to high quality structural variant calling is the inability to call breakpoints involving low mappability or non-reference sequence. Here, I show single breakend variant calling improves the accessibility of such regions.Using the Hartwig Medical Foundation cohort of 3,782 deeply sequenced WGS tumour/normal samples, I demonstrate the power of single breakend variant calling has in the detection and resolution of somatic genomic rearrangements. I show that GRIDSS2, the only SV caller to report single breakends, achieves a somatic FNR lower than possible for any purely breakpoint-based caller. I demonstrate the ability to detect viral integration into centromeric and telomeric sequences with VIRUSBreakend, a single-breakend based viral integration detection tool. I use LINX, a single-breakend-aware genomic rearrangement classification and interpretation tool, to resolve complex centromere-overlapping events. Finally, I posit a novel somatic centromeric rearrangement signature. Single breakends fundamental change the nature of breakpoint calling and provide insights into the “dark” regions of the genome currently considered inaccessible to short read sequencing.

Abstract #103

Emily Swan (Genome Informatics Laboratory, Department of Immunology, The John Curtin School of Medical Research, The Australian National University, Canberra, Australia)
Most disease phenotypes are actually very subtle: Deep learning with a massive, mouse, mutant cellular dataset to detect sub-clinical phenotypic variation

The effect size range of single mutations has been predominantly studied in very simple model systems that include viruses and single-cell organisms. However, with the modern imperative to interpret personal genome sequences for clinical application, this doesn’t provide the kind of information needed to stratify variation to predict complex and whole-organism phenotypes. Here, we present the analysis of a historical dataset of fluorescence-activated cell sorting (FACS) experiments collected from 40,000+ individual, inbred mice that harbour 20-60 randomly-induced missense mutations per individual. With this dataset, we can observed the phenotypic variation caused by a population of missense mutations, assayed at the cellular level with a consistent and sensitive analytical procedure. We have trained both classical machine learning (random forest) models and convolutional neural network models to predict whether the

immune cell population present in a given, single mouse belies the presence of a phenotypic mutant. We identify that different mutant genotypes give rise to a continuous scale of disruption of mouse immune cell subsets. We use our deep learning model to sensitively identify subtle mutant phenotypes that are consistently overlooked by non-automated, human observations of multi-dimensional FACS data. Most mutant genotypes produce little to no phenotypic effect, as expected, yet the proportion of slightly phenotypic and sub-clinical variants is substantially larger than previously appreciated. The magnitude of sub-clinical genetic variation we have identified has implications for our understanding of the combinatorial nature of phenotypic alleles that contribute to complex disease in humans.

Abstract #105

Ariane Mora (School of Chemistry and Molecular Biosciences University of Queensland, St Lucia QLD 4072, Australia)
Integrated gene landscapes uncover multi-layered roles of repressive histone marks during mouse CNS development

A prominent aspect of most, if not all, central nervous systems (CNSs) is that anterior regions (brain) are larger than posterior ones (spinal cord). Studies in Drosophila and mouse have revealed that the Polycomb Repressor Complex 2 (PRC2), a protein complex responsible for applying key repressive histone modifications, acts by several mechanisms to promote anterior CNS expansion. However, it is unclear what the full spectrum of PRC2 action is during embryonic CNS development and how PRC2 integrates with the epigenetic landscape. We removed PRC2 function from the developing mouse CNS, by mutating the key gene Eed, and generated spatio-temporal transcriptomic data. To decode the role of PRC2, we developed a method that incorporates standard statistical analyses with probabilistic deep learning to integrate the transcriptomic response to PRC2 inactivation with epigenetic information from ENCODE. This multi-variate analysis corroborates the central involvement of PRC2 in anterior CNS expansion, and reveals layered regulation via PRC2. These findings uncover a differential logic for the role of PRC2 upon functionally distinct gene categories that drive CNS anterior expansion. To support the analysis of emerging multi-modal datasets, we provide a novel bioinformatics package that integrates transcriptomic and epigenetic datasets to identify regulatory underpinnings of heterogeneous biological processes.

Abstract #108

Bennet McComish (Menzies Institute for Medical Research, University of Tasmania)
Using signatures of natural selection to focus the search for causal genetic variants in multiple sclerosis.

Multiple sclerosis prevalence shows a heterogeneous geographical pattern, with higher prevalence in populations of European ancestry, as well as increasing with distance from the equator within those populations. This pattern has likely been shaped by both natural selection and neutral genetic drift. Identifying genes that have undergone selection at MS risk loci will improve our understanding of the causative mechanisms behind the disease. Population genomics can be used to identify functional variation that has been subject to natural selection at loci associated with MS risk. We carried out genome-wide scans for natural selection using cross-population extended haplotype homozygosity in population genomic data. MS-related selection was localised by targeting genes prioritised in the large genome-wide association study carried out by the International Multiple Sclerosis Genetics Consortium. Strong signatures of natural selection in European and Asian populations were identified in several MS risk genes. Further analysis of these genes is underway to identify likely causes of selection and mechanisms by which they may contribute to MS risk. This approach allows us to narrow down candidate genes and pinpoint a small number of top causal candidates and mechanisms. This may enable more informed targeting of the molecular mechanisms behind the disease.

Abstract #110

Letitia Sng (Health and Biosecurity, Commonwealth Scientific and Industrial Research Organisation, Australia)
Novel Coronary Artery Disease Variants and Epistatic Interactions Identified with Machine Learning Pipeline using VariantSpark and BitEpi

Cardiovascular disease (CVD) is the leading cause of mortality worldwide. Although behavioural risk factors are important, there is a strong genetic component in CVD aetiology too. Genome-wide association studies have identified hundreds of loci associated with CVD risk, but account for less than 50% of CVD heritability. Epistasis, the combinatorial effect of multiple genetic variants, may explain part of this missing heritability. Yet, the nature of epistasis analysis poses multiple challenges for parametric statistical methods, such as computational demand and the high multiple testing burden. VariantSpark is a cloud-based machine-learning platform that can identify complex interactions between millions of SNPs from thousands of samples efficiently. We have applied VariantSpark to the UK Biobank dataset and have identified 47 significant SNPs associated with coronary artery disease (CAD). These SNPs map to known CAD genes including LPA, CDKN2B, and CELSR2 as well as novel genes including ARAP3 and FBN2. For comparison, we ran a logistic regression model on the same dataset, resulting in significant SNPs that have all been associated to CAD. We then used our novel epistasis platform, BitEpi, to search for interactions between these significant SNPs. We found that all the novel SNPs identified by VariantSpark were involved in higher order epistatic interactions with known CAD SNPs. Further exploration of these interactions found that interactions were occurring between key CAD pathways. The incorporation of these epistatic interactions into risk prediction models can improve the predictive ability of existing genetic risk scores which are based on variants with additive effects only.

Abstract #120

Mikhail Gudkov (Victor Chang Cardiac Research Institute, Australia and St Vincent's Clinical School, UNSW Sydney, Australia)
Quantifying negative selection on synonymous variants

Most disease sequencing studies tend to focus primarily on missense and potential loss-of-function 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 mutations is potentially more important than was previously thought. For instance, synonymous SNVs (sSNVs) may create nonoptimal codons, thus affecting the stability of the produced mRNA and the overall translational efficiency. It has also been shown that optimality-reducing sSNVs undergo purifying selection, the extent of which, nonetheless, remains unknown. Here we quantify the intensity of the negative selection acting on all possible sSNVs using the Mutability-Adjusted Proportion of Singletons (MAPS) metric. We found that optimality-reducing sSNVs are subject to stronger selection than optimality-increasing ones, which was confirmed using conservation analysis. Furthermore, we found that purifying selection affects sSNVs in an amino acid-dependent manner, with glutamine being particularly intolerant to such mutations. Apart from the effect on mRNA stability, other key mediators of this selection include the reduction in tRNA availability of the mutated codon and presence of amino acid homorepeats. Incidentally, we also propose an improved version of MAPS for sSNVs. We believe that the results of this work will help to further elucidate the role of synonymous mutations in health and disease and, in particular, will aid in variant prioritisation for finding the true causes of genetic disorders.

Abstract #122

Hieu T. Nim (Australian Regenerative Medicine Institute and Systems Biology Institute Australia, Monash University, Clayton, VIC, Australia)
Finding needles in a haystack: identifying cardiac disease-causing genes from cis-regulatory elements

BACKGROUND
Congenital heart diseases (CHD) are the major cause of death in newborns, but the genetic aetiology of this developmental disorder is not fully known. The conventional approach to identify the disease-causing genes focuses on screening genes that display heart-specific expression during development. However, this approach would have discounted genes that are expressed widely in other tissues but may play critical roles in heart development.
RESULTS
We report an efficient pipeline of genome-wide gene discovery based on the identification of a cardiac-specific cis-regulatory element signature that points to candidate genes involved in heart development. With this pipeline, we retrieved 76% of the known cardiac developmental genes, and predicted 35 novel genes that previously have no

connectivity to heart development. Functional validation of these novel cardiac genes by RNAi-mediated knockdown of the conserved orthologs in Drosophila cardiac tissue revealed that disrupting the activity of 71% of these genes led to adult mortality, and among these genes, RpL14, RpS24 and Rpn8 were associated with heart phenotypes. CONCLUSIONS
Our pipeline has enabled the discovery of novel genes with roles in heart development. This workflow, which relies on screening for non-coding cis-regulatory signatures, is amenable for identifying developmental genes for an organ without constraining to genes that are expressed exclusively in the organ of interest.

Abstract #123

P Acera-Mateos (The John Curtin School of Medical Research, Australian National University, Canberra, Australia)
RNA methylation detection at single-molecule resolution uncovers isoform-specific modifications in neurodevelopmental disorders

The expanding field of epitranscriptomics has been setting to rival the epigenome in the diversity of biological process involvement, prominently linking biochemical modifications of the RNA to development and disease onset. However, current methods based on short-read high-throughput sequencing to detect these modifications are often complicated and present significant accuracy limitations. Here we describe CHEUI, a new computational approach to identify N6-methyladenosine (m6A) and 5-methylcytidine (m5C) using signals from Nanopore direct RNA sequencing reads at single-nucleotide and single-molecule resolution. CHEUI uses a two-stage neural network to accurately predict methylation in individual reads and transcriptomic sites in a single condition, as well as differential methylation between any two conditions. Using extensive benchmarking with Nanopore data derived from in vitro modified and non-modified transcripts as well as cells with or without methylation enzymes, CHEUI showed higher accuracy than other existing methods in the prediction of m6A and m5C sites and their stoichiometry levels, while maintaining a lower number of false positives. We applied CHEUI to Nanopore RNA data derived from mouse embryonic pre-frontal cortex and uncovered isoform-specific modifications that are established and change stoichiometry during neuronal development. In a mouse model of autism spectrum disorder (ASD), we identified specific alterations in modifications that are associated with neuronal development, highlighting the biological potential of the isoform specificity of RNA modifications in neurodevelopmental disorders. CHEUI's ability to detect RNA modifications with high accuracy and resolution can be expanded to other modifications to unveil the full span of the epitranscriptome in normal and disease conditions.

Abstract #127

Natalie Charitakis (Murdoch's Children's Research Institute, Parkville, VIC & Department of Paediatrics, University of Melbourne, Parkville, VIC)
Benchmarking methods for the identification of spatially variable genes in spatial transcriptomics datasets.

Spatially transcriptomics (ST) is a novel, disruptive technology set to push the boundaries in exploring gene regulatory networks while providing both spatial and temporal resolution. It is expected to be exponentially adopted by the transcriptomics community after being named Nature's Method of the Year in 2020. Despite this, analysis packages for ST datasets are still in their infancy, with a clear forerunner yet to emerge. A comprehensive review of the performance of commonly used packages on the same datasets is lacking and there is a need to determine their performance in correctly labelling spatially variable genes (SVGs) in a tissue section. Various studies have established the ability of ST to uncover novel genetic determinants in different biological tissues and conditions; therefore, ensuring the most accurate analysis is performed is critical. To establish which of the current packages is most effective in identifying spatially variable genes (SVGs) within data generated using the same experimental protocol, I am creating a benchmarking process by testing a combination of publicly available, simulated, and novel 10X Visium datasets generated from cardiac tissues. Preliminary results demonstrate less than half of SVGs identified by different packages overlap. Upon incorporation of the most appropriately matched single cell RNA-Seq data for accurate cell-type deconvolution, identification of the cell types expression SVGs will be possible allowing for future experimental validation. This will offer a clear workflow to be implemented in future ST experiments and better understanding of mechanisms underlying disease and tissue development.

Abstract #128

Dillon Hammill (Division of Genome Sciences and Cancer, John Curtin School of Medical Research, Australian National University, Canberra, ACT)
CytoExploreR: Next-Generation Open-Source Software for Cytometry Data Analysis

Cytometry continues to be the preferred technique employed in clinical and research settings to classify, characterise, and quantify heterogeneous suspensions of cells. Recent technological advancements in cytometry have resulted in an unprecedented increase in the size and dimensionality of cytometry data sets. Unfortunately, this rapid increase data complexity has not been adequately met with innovations in commercially available software platforms to efficiently analyse these immense data sets. Consequently, many cytometry users have resorted to developing their own computational tools, in the form of R packages, to provide the additional tools required to analyse their data. Despite the quality of these packages, adoption of these tools within the broader cytometry community is limited, due to a lack of prerequisite coding knowledge, interactivity, and coherence between packages required for end-to-end analysis. Accordingly, there is exists an urgent need to develop a robust unified framework for cytometry data analysis, that is intuitive, interactive, efficient, extensible, and freely accessible. At ABACBS 2021, I am excited to announce the official release of CytoExploreR, the next generation of open-source software for cytometry data analysis. During this condensed presentation, we will endeavour to explore the plethora of computational tools implemented within the new CytoExploreR framework for end-to-end cytometry data analysis.

Abstract #130

Feng Yan (Australian Centre for Blood Diseases, Central Clinical School, Monash University, Melbourne, VIC, Australia)
DNA Methylation in pre-leukemic stem cells in a T-cell acute lymphoblastic leukemia mouse model

The role of DNA methylation in the initiation and evolution of cancer remains poorly understood due to lack of studies of the purified early pre-malignant state. We have analysed three stages of leukemogenesis using a Lmo2 transgenic mouse model of T-cell acute lymphoblastic leukemia (T-ALL). Purified pre-leukemic stem cells (pre-LSCs), LSCs, T-ALL and wild-type controls were profiled with RNA-seq and enhanced reduced representation bisulfite sequencing for DNA methylation. Hierarchical clustering of DNA methylation showed the greatest change between pre-LSCs and LSCs. Hypermethylation predominated in pre-LSCs and LSCs, with hypomethylation predominantly in T-ALL. In pre-LSCs, differentially methylated regions (DMRs) occurred mostly in CpG open seas marked by H3K27Ac/H3K4me2. In contrast, DMRs in LSCs were predominantly in CpG islands marked by H3K27me3/H3K4me3, similar to those reported in cancers and ageing. Finally, in T-ALL new hypomethylated regions emerged in CpG open seas. To focus on the clonal evolution in T-ALL, we calculated an epi-polymorphism score indicating sample heterogeneity. The epi-polymorphism increased in pre-LSCs and LSCs but decreased in T-ALL indicating a positive selection after disease onset. Differentially heterogenous epialleles (DHEs) did not overlap with DMRs in pre-LSCs, but half of DHEs in pre-LSCs overlapped with DMRs in LSCs suggesting methylation pre-seeding in pre-LSCs. Interestingly, DHEs showed stronger correlation with gene expression compared to DMRs. In conclusion, we have used mouse model of T-ALL to describe the DNA methylation at clonal level and associated gene expression during leukemogenesis. This will provide new insights into the mechanism and role of DNA methylation in cancer development.

Abstract #133

Yuzhou Feng (Peter MacCallum Cancer Centre)
SPIAT: Novel Computational and Statistical Tools for the Simulation and Analysis of Spatial Data

Platforms for spatial profiling are becoming more popular and are soon to become commonplace. Unfortunately, there is a significant lag in the methods for spatial analysis, making it challenging to get the most out of the data. In the cancer space, this limits our ability to understand the tumour microenvironment and develop quantitative biomarkers based on spatial patterns. To address this, we have developed SPIAT, an R package with a suite of spatial analysis algorithms based on spatial statistics, GIS and ecological principles, integrated with a novel simulator of spatial data. SPIAT can quantify diversity, spatial distribution, spatial heterogeneity and cell-cell interactions of different cell types, allowing a deep, orthogonal profiling of cell colocalization, attraction and repulsion. For the analysis of tumour samples, SPIAT allows the automatic detection of tumour borders and the quantification of immune populations relative to structures, allowing us to quickly calculate spatial properties of the tumour microenvironment, independent of user input or manually set thresholds. Our analyses in prostate cancer and melanoma have found a strong association between the novel spatial metrics in SPIAT and key clinical features, including prognosis and disease progression. SPIAT is currently compatible with single-cell spatial data generated from platforms such as OPAL, CODEX and MIBI. SPIAT provides a rich resource to profile spatial profiles, and derive novel patterns associated with clinical outcomes. This work will also allow the discovery of new biomarkers and provide novel insights into therapeutic opportunities.

Abstract #139

Alice R. Whitehead (A.R.W.: Florey Institute of Neuroscience and Mental Health, The University of Melbourne, Victoria 3052, Australia)
Characterising expression of RXFP1 splice variants across human cell types and tissues

Introduction: Relaxin Family Peptide Receptor 1 (RXFP1) is a hotly pursued therapeutic target for various cardiovascular and fibrotic conditions, due to its vasodilatory and anti-fibrotic effects when activated by the peptide hormone H2 relaxin. Therapeutic targeting of this G-protein coupled receptor has been hampered by a lack of knowledge of its complex mechanisms of action, including the roles of its alternatively spliced isoforms. In the present study we assess the expression profile of RXFP1 transcript isoforms across tissues and cell types using Oxford Nanopore sequencing. Methods: Oxford Nanopore sequencing was conducted on cDNA amplicons from a range of human cell types and tissues prepared using RXFP1-targeting oligonucleotides. Sequencing reads were aligned to the human genome (Hg38) using minimap2, and a high confidence set of transcript isoforms was determined using FLAIR. Results: The expression and relative abundance of RXFP1 splice variants varied across tissues and cell types. All samples were found to express multiple transcript isoforms of RXFP1. The canonical RXFP1 isoform was the highest expressed isoform in most, but not all, cell types and tissues. Of the top 10 most highly expressed variants across all samples, 7 were novel transcripts. Discussion: Characterising expression of RXFP1 splice variants marks a critical first step in leveraging the distinct RXFP1 isoforms in therapeutic targeting efforts. The novel variants will require functional characterisation to reveal the contribution of splice variants to RXFP1 physiology in health and disease.

Abstract #144

Nicolas P. Canete (The Westmead Institute for Medical Research, University of Sydney, Westmead, NSW, Australia; Sydney Medical School, University of Sydney, Sydney, NSW, Australia)
spicyR: Spatial analysis of in situ cytometry data in R

Recently, there have been advances in high parameter histological techniques such as imaging mass cytometry and multiplexed ion beam imaging by time of flight, and spatial-based transcriptomic techniques such as High-Definition Spatial Transcriptomics and sequential fluorescence in situ hybridisation. These imaging methodologies have allowed for the identification of a variety of distinct cell types within an image, providing a comprehensive overview of the tissue environment. This allows the complex cellular architecture and environment of diseased tissue to be explored. While spatial analysis techniques have revealed how cell-cell interactions are important within the disease pathology, there remains a gap in exploring changes in these interactions within the disease process. Specifically, there are currently few established methods for per

Abstract #146

Yingxin Lin (The University of Sydney)
Large scale single-cell multi-sample multi-condition data integration using scMerge2

Technological advances such as large-scale single-cell profiling such as single-cell RNA-seq (scRNA-seq), Cytometry by Time-Of-Flight (CyTOF) and imaging mass cytometry have exploded in recent years and enabled unprecedented insight into the identity and function of individual cells. The recent emergence of multi-condition and multi-sample single-cell large cohort studies further allows researchers to investigate cells from the same subpopulation in multiple cell states. Effective integration of multiple collections of multi-condition large cohort studies promises biological insights of cells under different conditions that can not be uncovered with individual study. Here, we present scMerge2, a scalable algorithm that allows data integration of large-scale multi-sample multi-condition single-cell studies. We have generalised scMerge2 to enable the merging of millions of cells from single-cell studies generated by various single-cell technologies, including scRNA-seq, CyTOF, and imaging mass cytometry. Leveraging pseudo-bulk to perform factor analysis of stably expressed genes and pseudoreplicates, scMerge2 is able to integrate 200,000 cells and 10,000 genes within an hour using a single core. Using a large COVID-19 data collection with 2,740,510 cells from 779 samples of 6 studies, we demonstrate that scMerge2 enables multi-sample multi-condition scRNA-seq data integration from multiple cohorts and reveals distinct cell states of COVID-19 patients with varying degrees of severity. We further illustrate that the data integration of different cohorts enables characterisation of cell-cell interaction for COVID-19 patients, which can be used as potential signatures for discriminating between moderate and severe patients.

Abstract #148

David Humphreys (Victor Chang Cardiac Research Institute)
Overlapping transcripts within gene models can influence bioinformatic analysis.

Many bioinformatic pipelines are dependent on existing gene models for quantification and/or annotation. Transcript information of gene models are efficiently compiled within GTF formatted files which enables the reconstruction and interpretation of complex gene structures. In compiling expression patterns of “high confidence heart developmental” genes we came across one gene, GDF1, that had unexpected low signal in many compiled RNA-Seq databases. A closer examination of GDF1 revealed that this gene belongs to a rare class of bicistronic transcripts that are difficult to describe in GTF format as GDF1 exons are shared with the other cistron annotated as CERS1. As GDF1 and CERS1 have near identical exon structure they are interpreted as being overlapping by many bioinformatic pipelines and are therefore often ignored. We therefore re-analysed single cell datasets and have been able to correctly capture the expression patterns of GDF1, which involves a switch from single cistron to bicistronic transcript isoform. For the first time we also identify the key cell types responsible for this expression pattern. Furthermore we compiled a list of other overlapping transcripts that exist within gene models and explore the implications for various bioinformatic pipelines. From this we identified that 10x single cell RNA-Seq bioinformatic pipelines are susceptible for not quantifying reads that align to overlapping transcripts regions and highlight how this could have important implications for downstream analysis.

Abstract #151

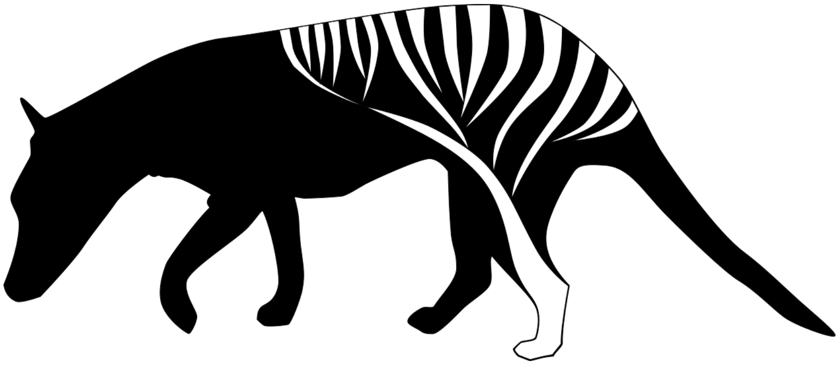
Aaron Chuah (Genome Informatics Laboratory, Department of Immunology and Infectious Disease, The John Curtin School of Medical Research, The Australian National University, Canberra, Australian Capital Territory)
Genome-wide estimation of the change in protein stability due to missense mutation

It is widely accepted that decreased protein stability due to missense genetic variation is strongly associated with inherited genetic disease. However, the interpretative power of instability effects of mutation on proteins has been limited by the need for structural data covering the observed variation. Consequently, analysis of the influence of genetic variation on protein stability is not routinely performed as part of clinical genomic workflows to detect pathogenic variation. The availability of AlphaFold predicted structures for whole genomes of proteins, including that of humans, now allows genome-wide assessment of the stability effects of genetic variation. Importantly, this now allows routine assessment of personal genetic variation for missense variants that potentially have clinical interpretive value. Here, we present both a web-tool and a standalone software package that allows prioritisation of genetic variation to identify just those variants that lead to strong destabilisation effects in essential proteins. Our methodology innately down-weights variation is non-essential and redundant proteins, as these genes inherently harbour more variation deleterious to protein function. Our predictive metrics are trained on the presumed non-disease-associated variation present in the GnomAD dataset. For each missense variant from an incident clinical genome, a test statistic is calculated that assesses the magnitude of the destabilisation effect with respect to all other non-disease associated variation identified in this protein in the GnomAD dataset. Missense variants in a personal genome are prioritised such that variation that produces unusually large stability effects in the encoded protein, given the population variation observed in that protein, are easily identified. The results of this methodology are assessed and contrasted between the HapMap and two clinical cohorts.

Abstract #154

Richard J Edwards (Evolution & Ecology Research Centre, School of Biotechnology and Biomolecular Sciences, UNSW Sydney, Sydney, New South Wales, Australia)
DepthSizer and DepthKopy: genome size and copy number prediction using single-copy long-read depth profiles

A fundamental part of any genome project is establishing the genome size of the organism being sequenced. The gold standard for genome size measurement is flow cytometry, but this is not available to all groups and can give surprisingly variable results. Popular bioinformatic approaches predict genome size using kmer frequency profiles from high-accuracy (e.g. illumina or hifi) sequencing reads, or the mean depth of coverage reads mapped to an assembly. Both of these approaches can be adversely affected by repetitive regions of the genome. Mean sequencing depth is also highly reliant on assembly completeness. Here, we present DepthSizer (https://github.com/slimsuite/depthsizer), which refines this approach by estimating sequencing depth based on single-copy complete BUSCO genes. DepthSizer works on the principle that genuine single-copy regions will tend towards the same, true, single-copy read depth. In contrast, assembly errors, collapsed repeats within those genes, or incorrect BUSCO predictions, will give inconsistent read depth deviations. The modal read depth across single-copy BUSCO genes, calculated from a depth density profile of these regions, should therefore provide a good estimate of the true depth of coverage. The method is benchmarked on model organism data and corrections for possible contamination, biases/inconsistencies in read mapping and/or raw read insertion/deletion error profiles are discussed. We also present DepthKopy (https://github.com/slimsuite/depthkopy), which uses the same read depth approach to estimate the copy number of assembly regions. This can be useful for identifying haplotigs, and collapsed repeat regions.



Phylomania Abstracts

(In presentation order)

Abstract # 49
Barbara Holland (University of Tasmania)
Some possibly bad ideas for detecting convergent selection

Given a phylogeny, a character matrix and a binary trait of interest that has been gained/lost multiple times across the phylogeny, how might we go about detecting characters that are associated with the trait of interest? In this talk I will survey three different ideas for approaching this problem.

Idea 1: Detect characters (or suites of characters) that are more compatible with each other (and the trait of interest) than expected conditional on their level of homoplasy on the phylogeny. (A recycling of an old idea from Holland, Spencer, Worthy, & Kennedy (2010).)
Idea 2: Seed a phylogenetic mixture model with components whose edge weights vary systematically depending on the ancestral character estimate of the binary trait. (With Rob Lanfear and Stephen Crotty.)
Idea 3: Model convergent selection explicitly using convergence-divergence networks. (Joint with Jonathan Mitchell)

Abstract # 45

Zhaoxiang Cai, Rebecca C Poulos, Qing Zhong (The University of Melbourne)

Integrating multi-omics data with biological knowledge by Transformer-based deep learning

Omics data analysis, powered by machine learning, has significantly improved cancer diagnosis and prognosis. However, most machine learning methods consider each gene as an independent feature, failing to integrate experimentally-acquired gene regulation and pathway information. The benefit of utilising this information increases in the era of multi-omics, because gene regulation is the key mechanism that links different omic layers together. Here, we present an interpretable deep learning model, DeepPathNet, which uses cancer-specific pathway information for both single and multi-omics data analysis. DeePathNet leverages the cutting-edge deep learning technique, Transformer, which is derived from the field of natural language processing, to model complex interactions between pathways from omics data. The computation of self-attention in the Transformer module allows DeePathNet to learn the encoding of pathways to achieve superior predictive performance and interpretability. Techniques such as drop out layers are also integrated into DeePathNet to maximise its generalisability for unseen data. Moreover, DeePathNet supports any number of omics layers and can handle missing values. Using multiple evaluation metrics, we demonstrate that DeePathNet robustly outperforms traditional methods for predicting drug response and cancer type on four publicly available datasets, namely COSMIC Cell Lines, Genomics of Drug Sensitivity in Cancer (GDSC), Cancer Cell Line Encyclopedia (CCLE) and Cancer Therapeutics Response Portal (CTRP). DeePathNet also provides reliable model interpretation, potentially enabling biomarker discoveries at the pathway level. Using the Transformer, DeePathNet is the first method that supports multi-omics data analysis, integrates cancer pathway knowledge into modelling, and provides pathway-level model explanation.

Abstract # 13

Paola Cornejo-Páramo, Emily S Wong, Katherine Roper, Sandie M Degnan, Bernard M Degnan
(Victor Chang Cardiac Research Institute, School of Biological Sciences, University of Queensland)

Transcription Factor-binding motif composition of cis-regulatory regions in a sea sponge species suggest the existence of some constraints in metazoan regulatory grammar

Chromatin accessibility plays a major role in regulating animal gene expression and therefore it is key to decipher animal embryogenesis. Here we profile chromatin accessibility of six embryonic stages, along with a larval and adult stage in the sea sponge Amphimedon queenslandica, a member of the early branching phylum Porifera, in order to study the regulatory landscape of metazoan embryogenesis. On the grounds that most transcription factor (TF) families are conserved in eukaryotes, we examine the TF binding motif composition of Amphimedon cis-regulatory regions. We leveraged the high performance of tree-based machine learning models and tree ensemble interpretation algorithms to extract complex patterns and identify some TF binding motifs that are commonly enriched in Amphimedon cis-regulatory regions. We further tested the machine learning models trained on Amphimedon on other species data and were able to predict cis-regulatory regions in other animal species but not in the close relative of animals, Capsaspora owczarzaki. Overall, we find that animal cis-regulatory regions share some resemblance and constraints in motif composition besides the extreme divergent life strategies adopted by different phyla.

Nicholas Fountain-Jones *et al.* (University of Tasmania)

Hunting alters viral transmission and evolution in a large carnivore

Hunting can fundamentally alter wildlife population dynamics, but the consequences of hunting on pathogen transmission and evolution remain poorly understood. Here we present a study that leverages a unique landscape-scale quasiexperiment coupled with pathogen transmission tracing, network simulation, and phylodynamics to provide insights into how hunting shapes feline immunodeficiency virus (FIV) dynamics in puma (Puma concolor). We show that removing hunting pressure enhances the role of males in transmission, increases the viral population growth rate, and increases the role of evolutionary forces on the pathogen compared to when hunting was reinstated. Changes in transmission observed with the removal of hunting could be linked to short term social changes while the male puma population increased. These findings are supported through comparison with a region with stable hunting management over the same time period. This study shows that routine wildlife management can have impacts on pathogen transmission and evolution not previously considered.

Abstract # 43

Nehleh-Fatemeh Kargarfard, Aaron Darling, Mathieu Fourment (ithree Institute, University of Technology Sydney)

PhiloBacteria: A new tool to infer phylogenetic trees from recombinant bacterial genomes

One of the vital services to maintain public health is detecting and tracking outbreaks of infectious diseases. Phylogenetic tools are one of the most efficient tools to analyze the source, evolutionary history of epidemic outbreaks. According to the rapidly growing number of genomes, it's essential to develop new methods to apply to increasingly large datasets. On the other hand, as asexual organisms, Bacteria have a clonal reproduction; however, their genomes can evolve by different mechanisms such as point mutation and recombination. When recombination occurs in bacteria genomes, several nucleotides can change together, contributing to considerable evolutionary leaps, for example, helping pathogens develop resistance mechanisms against antibiotics. Practically, it's not only swapping genetic information among organisms, but also it's exchanging their evolutionary histories. Hence, the evolutionary analysis would be different and challenging in the presence of recombination, and ignoring the effect of recombination can result in misleading in phylogenetic. Furthermore, detecting the boundary of recombination events and reconstructing a global tree to illustrate the underlying evolutionary pattern of biological sequences has never been a straightforward problem in terms of accuracy and scalability. We introduce PhiloBacteria, a new tool that uses a hidden Markov model to detect recombination and infer the phylogeny of related organisms simultaneously. Specifically, the algorithm‚Äôs goal is to reconstruct the true clonal evolutionary history of the organisms while avoiding misleading evolutionary signals originating from recombination events. The method aims at quickly and accurately analyzing large datasets of closely related bacterial genomes. Using simulated data, we investigate the accuracy of our algorithm and compare our results to other established recombination detection methods.

Abstract # 54

Folagbade Abitogun (University College Hospital Ibadan, Nigeria)

Identification and Analysis of Novel Putative Drug Targets in Hypervirulent Klebsiella pneumoniae Proteomes

The last 5 years have seen increasing cases of infection with hypervirulent Klebsiella pneumoniae, a strain defined by the presence of many biomarkers, with increasing resistance to extended-spectrum beta-lactams and carbapenems. The rapid evolution of this pathogen has led to the continued failure of antibiotics to combat it, calling for alternative therapies that will combat the pathogen and reduce antibiotic resistance which this study aimed to achieve. A series of In silico analyses were employed to identify putative drug targets from the proteome of two strains of Klebsiella pneumoniae- HvKp TK421 and HvKp1. A combined total of 1528 Hypothetical proteins were extracted using a Perl script from the proteomes of both stains. Of these, 159 essential pathogen-specific proteins that are also not homologous to any known human protein were further extracted. Sixty eight (68) of the essential proteins were cytoplasmic and therefore chosen as potential drug targets, with the majority having catalytic activities. Sixty (60) of the potential drug targets showed no interactions with any known drug and were considered ‘novel’. Further, 11 of the novel drug targets showed considerable interactions with human host proteins with 9 of them showing favorable docking with compounds of established antibacterial activities. This study identified and analyzed putative drug targets from the proteomes of two emerging hypervirulent Klebsiella pneumoniae strains which also share highly related and evolving proteins, thereby providing data which will prove indispensable for monitoring pathogen evolution and drug development towards preventing antibiotic resistance.

Abstract # 74

Abhinay Thakur, Dikshita Bansode, Pragati Ghare, Shrutika Sakpal*

***In-silico* based approach of FDA Approved Drugs for targeting against nsp10 (6W75) of 2019-nCoV (novel coronavirus)**

The 2019-nCoV has triggered a global public health emergency due to its rapid spread, resulting in a pandemic situation. The nsp10 plays a crucial role in viral transcription by stimulating exoribonuclease of nsp14 (3'-5') and nsp16 (2'-O-methyltransferase) activities thereby playing an essential role in viral mRNAs cap methylation. The current study involves the molecular-docking of the nsp10 (6W75) using PyRx for FDA-approved drugs available for the treatment of SARS-1 and MERS, with the hypothesis that these drugs could be suggested for the treatment of 2019-nCoV or not. The Phylogenetic evolutionary relationship between SARS-1, MERS, and SARS-2 confirmed the validation. The docking process was carried out through results revealed that the top five drugs with the highest binding affinity rate are also used for Hepatitis-C virus treatment, and the Molecular Dynamic Simulation was carried out for the drug highest binding affinity rate Elbasvir, using GROMACS. The results indicated that Elbasvir could be used as a potential target against the 2019-nCoV nsp10 protein.

Abstract # 55

Eike Steinig, Izzard Aglua, Sébastien Duchêne *et al.* (University of Melbourne, Sir Joseph Nombri Memorial-Kundiawa General Hospital, Kundiawa, Simbu Province, Papua New Guinea, *and many more*)

Signatures of epidemic growth in the emergence of community-associated MRSA

Community-associated, methicillin-resistant Staphylococcus aureus (MRSA) lineages have emerged in geographically distinct regions around the world during the past thirty years. However, the genomic and epidemiological drivers behind their emergence have remained unclear. Here, we applied consistent phylodynamic methods across multiple community-associated MRSA lineages to describe and contrast their patterns of emergence and dissemination, including novel isolates from neglected populations and regional outbreaks in northern Australia, Papua New Guinea, and Pakistan. We observed surges in the effective reproduction number (Re > 1) at the divergence of antibiotic resistant clades from America, Asia, Australia, and Europe, coinciding with their establishment in urban host populations. Our data suggest that the emergence of community-associated S. aureus lineages in the late 20th century was driven by a combination of antibiotic resistance acquisition and changes in host epidemiology, leading to noticeable shifts in transmission dynamics and epidemic growth of emerging lineages.

Abstract # 114

Swapnil Tichkule, Simone M. Cacciò, Guy Robinson, Rachel M. Chalmers, Samantha J. Emery-Corbin, Kevin M. Tyler, Cock van Oosterhout, Aaron R. Jex (Population Health and Immunity, Walter and Eliza Hall Institute of Medical Research; Department of Infectious Disease, Istituto Superiore di Sanità, Rome, Italy; Cryptosporidium Reference Unit, Public Health Wales Microbiology, Singleton Hospital, Swansea, UK;; Biomedical Research Centre and School of Environmental Sciences, University of East Anglia, Norwich, UK)

Genomic landscape of diversification, selective sweeps, and demographic history of an anthroponotic parasite

Cryptosporidium is a significant public health problem and one of the primary causes of diarrhoea in humans, particularly in very young children living in low- and middle-income countries. The zoonotic Cryptosporidium parvum and anthroponotic C. hominis species account for most cases globally, but the latter is predominant in low- and middle-income countries. Here, we present a comprehensive whole genome study of C. hominis, comprising 114 isolates from 16 countries in five continents. We detect two highly diverged lineages with a distinct biology and demography that have diverged circa 500 years ago. We consider these lineages as two subspecies, and provisionally propose the names C. hominis hominis (clade 1) and C. hominis acquapotentis (clade 2). C. hominis hominis is mostly found in low-income countries in Africa and Asia, and it appears to have recently undergone population contraction and selective sweep. In marked contrast, we reveal a signature of population expansion in C. hominis acquapotentis found in high-income countries, mainly in Europe, North America, and Oceania. Moreover, we detect genomic regions of introgression representing gene flow after a secondary contact between the subspecies from low- and high-income countries. Furthermore, we show that this gene flow resulted into genomic island of high diversity and divergence, and we find that diversity at potential virulence genes is maintained by balancing selection, which suggests that they are involved in a coevolutionary arms race.

Abstract # 109

Amarinder S. Thind, Bruce Ashford, Dario Strbenac, Ruta Gupta, Jonathan R Clark, N. Gopalakrishna Iyer, Jenny Mitchellm, Jenny Lee, Simon A Mueller, Elahe Minaei, Jay Perry, Marie Ranson (School of Medicine, University of Wollongong, Wollongong, NSW; Illawarra Health and Medical Research Institute, Wollongong; Illawarra Shoalhaven Local Health District Wollongong; Sydney Medical School, The University of Sydney; Royal Prince Alfred Hospital, Sydney, *and many more*).

Potential non-coding regulations in cSCC Metastasis cancer

There is limited published data exploring whole-genome sequencing (WGS) in metastatic cutaneous squamous cell carcinoma (cSCC). We used WGS on matched tumour and blood DNA to detect somatic variants from 25 patients with regional metastases of head and neck cSCC. Computational analyses at coding and non-coding levels are performed utilising a combination of bioinformatic tools to interrogate their clinical impacts on metastasis across the cohort. miRNA binding locations in 3'UTR were significantly functionally altered in EVC (71%), PPP1R1A (71%) and LUM (24%). Recurrent variation was observed in the tumour suppressing lncRNA LINC01003 in 68% of specimens. Recurrent copy number loss in tumour suppressor genes KANSL1 and PTPRD and gain in CALR, CCND1 and FGF3 was observed. Single nucleotide variation with the greatest functional impacts was most frequently observed in TP53, CDKN2A, ZNF750, OR51S1 and TET2. Metastatic cSCC is characterized by a highly mutated genome, most pronounced in the non-coding region, with recurrent patterns of variation in key regulatory elements, and recurrent copy number and short variation in cancer-associated genes.

Abstract # 30

Joshua Stevenson, Venta Terauds, Jeremy Sumner (University of Tasmania)

Rearrangement Events on Circular Genomes

In the context of estimating genome rearrangement distances, genomes are often represented by signed permutations which form a group under composition. Further, the symmetry of genomes and rearrangement events can also be described algebraically. I will give an introduction to genome rearrangement modelling, and explore some of the questions that naturally arise when viewing this problem from an algebraic perspective, with the help of Python/SageMath for demonstrations.

Abstract #36
Venta Terauds (University of Tasmania)
Genome algebras in action
We demonstrate some features of our recently introduced genome algebra framework by applying it to echinoderm mitochondrial genome data. In particular, the framework facilitates a fine-grained consideration of rearrangement models. We show how varying the underlying rearrangement model and the choice of genomic distance measure can affect the phylogeny that is produced.

Abstract # 62
Sebastian Burgstaller-Muehlbacher, Stephen Crotty, Tamara Drucks, Heiko Schmidt, Arndt von Haeseler (Center for Integrative Bioinformatics Vienna, Max Perutz Labs, University of Vienna and Medical University of Vienna; University of Adelaide)
Model selection on empirical data using deep learning
Selecting the correct model of sequence evolution (mSE) for a multiple sequence alignment (MSA) constitutes the first step of tree reconstruction. State-of-the-art approaches for inferring nucleotide models mostly apply maximum likelihood (ML) methods.
Here, we demonstrate that neural networks can infer the correct mSE including the shape parameter α of the Γ-distribution. A Residual Neural Network (Resnet) was trained with the six most frequently used mSE (JC, K2P, F81, HKY, TN93 and GTR), whereas a Long Short-Term Memory (LSTM) network with attention was trained to determine γ.
Our results show that the CNN correctly identifies the mSE in a range of 51.94% to 100%, depending on the true mSE. Thus, it is comparable to results of IQ-Tree. Similar accuracies were obtained for γ. However, the trained networks need substantially less computing time (up to a 60x speedup), depending on the size of the MSA.
We demonstrate, for the first time, that neural networks can be used to identify the correct mSE as well as rate heterogeneity of an MSA. Furthermore, we intend to generalize our approach so all mSE of relevance in phylogenetics can be inferred using neural networks, conferring a substantial reduction in computational requirements to model selection.

Abstract # 126
Jiahao Diao, Barbara Holland, Malgorzata O'Reilly (University of Tasmania)
A Stochastic model of evolution to improve the prediction of independent or dependent gain and loss of genes
We consider a model inferring functional gene links. The model can detect the instances of independent or correlated gain and loss of pairs of genes from species genomes. We start at a simple case that the phylogenetic trees generated by the birth-death process, which has a pair of genes in the evolutionary process. The presence ('1') and absence ('0') of these two genes are affected by two different rate classes. The possible states of these two genes in a phylogeny are applying R package corHMM from (Boyko & Beaulieu, 2020). We aim to obtain the best fit of rates in these two different rate classes and how the transition between two rate classes would be more realistic to understand the gene gain and loss in the evolutionary process.

Jonathan Mitchell, Elizabeth Allman, John Rhodes (University of Tasmania, University of Alaska)
Consistent inference of species networks from gene trees under the multispecies coalescent using a generalized AIC

We developed an algorithm that builds on the NANUQ method for statistically consistent inference of species networks under the multispecies coalescent model of incomplete lineage sorting. NANUQ uses hypothesis tests to determine whether quartets better fit a tree or a network, then combines all sets of quartets to infer an n-taxon species network. We increased the scope of NANUQ, allowing quartets to also be analyzed using a generalized AIC. This generalized AIC outperforms the AIC in models with parameter space boundaries or singularities, particularly at or near those points. Such models are common in many real world applications, including phylogenetics. Simple examples of models with boundaries or singularities are those with parameters that are constrained to be positive or non-negative, such as evolutionary times or rates. Using NANUQ we analyzed a dataset of xiphophorus fish with expected extensive reticulate evolution.

Abstract #33
Martin Smith (Durham University)
Improving consensus trees by detecting rogue taxa
“Rogue” taxa of uncertain phylogenetic position can confound attempts to summarise phylogenetic results. By reducing resolution and support values in consensus trees, rogues potentially obscure strong evidence for relationships between other taxa.

My new, information-theoretic measure of the congruence between a set of trees and their consensus allows rogue taxa to be identified more effectively, thus producing reduced consensus trees that are better resolved, more accurate, and more informative than those generated by existing methods.

Abstract # 76
Caitlin Cherryh, Minh Bui, Robert Lanfear (Research School of Biology and Research School of Computer Science, Australian National University)
Does Filtering Recombinant Loci Improve the Estimation of Species Trees?
Genome-scale data with thousands of loci are now routinely used to estimate species trees. Both concatenation and coalescent methods can be used to estimate species trees, and both methods assume that there is no recombination within loci. This assumption is largely ignored within empirical analyses, but recombination is present within loci from many empirical datasets. In this work, we examine how removing loci that show evidence of recombination affects species tree estimation for both coalescent and concatenation methods for four empirical datasets. We find in some datasets (like those from mammals) with limited recombination, filtering loci has no effect on species tree topology. However, for other datasets (e.g., from closely related plants), filtering loci can have dramatic effects on the species tree.

Abstract #111
Ana Serra Silva, Mark Wilkinson (Natural History Museum, London, UK University of Bristol, UK)
From islands of trees to clumps: can we generate informative clusters of partially overlapping trees?
Post-processing of trees often focuses on (multi)sets of phylogenetic trees on the same leaf set, which can be effectively summarised using a variety of consensus and clustering methods. One such approach is identifying islands of trees, or the disconnected components of a graph where vertices correspond to trees and edges connect sufficiently similar trees. However, this approach cannot be easily applied to (multi)sets of trees with partially overlapping leaf sets, such as those generated for phylogenomic studies, due to the confounding effects introduced by small trees (with small distances to potentially disparate larger trees) and by pairs of trees with non-overlapping leaf sets. In an attempt to identify informative clusters of trees with partially overlapping leaf sets, we define a new subsetting approach, "clumps of trees", based on the distance between any tree in a set and the set's supertree. While clumps were developed with the goal of minimising the negative effects introduced by small trees and avoiding comparisons of trees with non-overlapping leaf sets, they can be applied to (multi)sets of trees with identical taxonomic sampling. Time permitting we will introduce the concept of clumps, an analytical pipeline and an empirical example of clumping.

Abstract # 116

Sarah Alver, James Degnan (University of Mexico)

Improvement to GLASS/Maximum Tree Method of Species Tree Inference from Estimated Gene Trees Using Measurement Error Modified Single Linkage Clustering

The Global LAteSt Split (GLASS)/Maximum Tree (MaxTree) method of species tree inference performs well and is statistically consistent when inferring the species tree from known gene trees. When using estimated gene trees, the inferred tree is the maximum likelihood tree and is still a consistent estimator of the species tree under certain conditions. The method is implemented in software such as STEM (Species Tree Estimation using Maximum likelihood). Unfortunately, it has been shown to perform relatively poorly when the input is gene trees estimated from DNA sequences, and sufficient conditions for statistical consistency in this case can be unrealistic. We propose a modification to the STEM tree; the modification is an application of the method of clustering in general measurement error models described by Su, Reedy and Carroll in 2018. The proposed method replaces the estimated pairwise coalescence times used by STEM with randomly generated realizations from the estimated distribution of the true pairwise coalescence times. This distribution is estimated through measurement error modeling. As with STEM, the minimum of these realizations is taken over all loci for each pairwise distance. These minimums then form a distance matrix, and single linkage clustering is performed to infer the species tree. Our simulation studies find that the new method outperforms STEM in terms of Robinson-Foulds distance and branch score distance from the true species tree.

Abstract # 52

Lena Collienne, Alex Gavryushkin, David Bryant (University of Otago; University of Canterbury)

Distances between phylogenetic time trees

Phylogenetic time trees are evolutionary histories where evolutionary events, i.e. internal nodes, are timed. To measure the similarity of phylogenetic trees, including time trees, tree rearrangement operations are often used. These operations apply local changes to a tree, that are motivated by biological processes. The distance between two trees is then defined to be the minimum number of tree rearrangement operations needed to transform one tree into the other.

Most popular are the tree rearrangements Nearest Neighbour Interchange (NNI), Subtree Prune and Regraft (SPR), and Tree Bisection and Reconnection (TBR). The downside is however that the distance between trees under any of these tree rearrangement operations (NNI, SPR, or TBR) is NP-hard to compute and therefore not suitable to be used in practice. Other distance measures like the Robinson-Foulds distance can be computed efficiently, but lack biological interpretability. Moreover, none of the above mentioned distance measures takes times of internal nodes of phylogenetic trees into account.

In this talk, we introduce tree rearrangement operations for ranked phylogenetic trees. They are called ranked nearest neighbour interchange (RNNI) operations, as they are based on the classical Nearest Neighbour Interchange (NNI) operations. We show that RNNI distances between ranked trees can be computed efficiently, and provide a polynomial time algorithm to do so. Afterwards, we generalise our results to trees with integer-valued branch lengths.

Gleb Zhelezov (University of New Mexico)

Trying out a million genes to find the perfect pair with MTrip

Consensus methods can be used for reconstructing a species tree from several gene trees which exhibit incompatible topologies due to incomplete lineage sorting. Motivated by the fact that there are no anomalous rooted gene trees with three taxa and no anomalous unrooted gene trees with four taxa in the multispecies coalescent model, several contemporary methods form the gene tree consensus by finding the median tree with respect to the triplet or quartet distance — i.e., estimate the species tree as the tree which minimizes the sum of triplet or quartet distances to the input gene trees. These methods reformulate the solution to the consensus problem as the solution to a recursively-solved dynamic programming problem. We present an iterative, easily-parallelizable approach to finding the exact median triplet tree. By taking advantage of the fact that gene trees associated with the same species tree often have distinct bipartitions of identical taxa subsets, this implementation finds the exact median tree of many gene trees faster than comparable methods, has better scaling properties with respect to the number of gene trees, and has a smaller memory footprint.

Abstract # 112

Benjamin D Kaehler, Gavin A Huttley (University of New South Wales;Australian National University)

Phylogenetics if you stop ignoring non-stationary evolution

Probabilistic phylogenetic models almost ubiquitously ignore that molecular evolution is usually a non-stationary process, to their demonstrable detriment. Largely the problem of fitting non-stationary models has been relegated to something that happens after a tree has been built. If, however, non-stationary models are used to build trees, it is possible to fit rooted trees, which drastically changes the algorithmic landscape, and could resolve some phylogenetic controversies that look suspiciously like model misspecification problems. A long-standing barrier to adoption is that fitting models with lots of parameters is challenging. We overcome that problem with the scalpel of model selection and the hammer of GPU acceleration, which we show to yield orders of magnitude improvement in model fitting throughput. We then dust off some historically relegated rooted tree building algorithms, see how they perform in the harsh light of noisy data, and update them with the benefit of over a decade of further development in graphical algorithms.

Abstract # 107

Cassius Manuel Perez, Arndt von Haeseler

A test for phylogenetic saturation

In phylogenetic inference, an alignment is called saturated if the observed distances between sequences substantially underestimate the evolutionary distance (branch length). Saturation is typically discussed in the context of pairwise sequence comparison. However, even for this simple instance, a quantitative theory of saturation is to the best of our knowledge missing.

Here, we introduce a statistical test for saturation that assumes a reversible model of evolution and a phylogenetic tree. Our test is straightforward for two sequences. Moreover, we show how to generalize the test for arbitrary branches in a phylogenetic tree. Such tests provide an additional quality control of phylogenetic branching patterns.

As an illustrative example, we show that artifacts like long branch attraction or long branch repulsion can be detected using our test for saturation.

Abstract # 121

Qin Liu, Barbara Holland, Michael Charleston, Shane Richards

Is AIC An Appropriate Metric For Model Selection In Phylogenetics?

Partition models and mixture models are two common approaches to accommodate heterogeneity in genomic sequencing data. Both models have been shown to provide a better fit to the data than the conventional homogeneous models. Assessing the adequacy of these models is important since incorrect inferences can be made if the mechanism modelled is incorrect. Akaike Information Criterion (AIC) is an estimator of Kullback-Leibler divergence (KLD) and is a popular tool to select conventional homogeneous models in phylogenetics. A couple of recent papers have questioned the effectiveness of AIC in phylogenetics, but there has not been any investigation of mixture or partition models. We are interested in whether AIC is an appropriate tool to compare models, in particular, to compare a partition model and a mixture model in phylogenetics. In my talk, I am going to show that, under non-standard conditions, AIC underestimates the expected KLD and always chooses partition models over mixture models. Non-standard conditions occur when the branch lengths and/or the sequence lengths are small. I am also going to show some of the preliminary results from the research that we have done on evaluating AIC between a partition and a mixture model.

Samuel Davis, Gabriel Foley & Mikael Boden (The University of Queensland)

Inference of biochemical and biophysical parameters for ancestral proteins

Ancestral reconstruction of proteins is a powerful approach with potential benefits for several fields. A number of available tools reliably infer the sequences of ancestral proteins, with broad evidence indicating the robustness of inferred sequences in laboratory resurrections. This ability greatly increases the available sequence space in the selection of candidates for protein engineering. Despite this ability, the cost associated with synthesising and characterising ancestral variants is significant when performed for many potential candidates. Improving the means by which ancestors are selected for synthesis would ultimately reduce the overall cost of this process. Existing methods for inferring ancestral states are largely focused on species-level traits and rely on continuous-state stochastic processes to model evolution. Such models may be inappropriate for the evolution of biochemical and biophysical parameters, which are capable of dramatic change resulting from small numbers of mutations. We have designed an application for ancestral inference of continuous-valued parameters which address these constraints. Our approach utilises latent, discrete states at each node in the tree which evolve via Markov chains along branches. Real property values are generated from Gaussian features corresponding to each of these latent states, the parameters of which can be learned via expectation maximisation. Marginal inference at a given node takes the form of a weighted sum of Gaussian feature densities, with weights corresponding to the posterior probability distribution of the latent state. We tested the performance of the application on both real and simulated datasets, with it out-performing existing methods under certain circumstances.

Abstract #89

Conrad J.Burden, Robert C. Griffiths

Stationary Distributions of Neutral Multi-type Branching Diffusions

The stationary asymptotic properties of the diffusion limit of branching processes in which individuals within the population can mutate between different allele types is considered. For the critical and subcritical processes the interesting limits are those of quasi-stationary distributions conditioned on non-extinction. Limiting distributions for supercritical and critical processes are found to collapse onto rays aligned with stationary eigenvectors of the mutation rate matrix, in agreement with known results for discrete multi-type branching processes. For the sub-critical process the quasi-stationary distribution is obtained to first order in the overall mutation rate, which is assumed to be small. The sampling distribution over allele types for a sample of given finite size is found to agree to first order in mutation rates with the analogous sampling distribution for a Wright-Fisher diffusion with constant population size.

Abstract #

Qiuyi Li, Celine Scornavacca, Yao-ban Chan

The effect of copy number hemiplasy on gene family evolution

The evolution of gene families is complex, involving evolutionary events such as gene duplication, horizontal gene transfer, and gene loss (DTL), and other processes such as incomplete lineage sorting (ILS). Because of this, topological differences often exist between the trees of genes and the species which contain them. A number of models have been developed recently to explain these discrepancies, the most realistic of which attempt to consider both DTL and ILS. When unified in a single model, the interaction between ILS and DTL can cause polymorphism in gene copy number; we refer to this as copy number hemiplasy (CNH).

We study the effect of CNH on gene family evolution, by comparing two such models: MLMSC (MultiLocus MultiSpecies Coalescent), which models CNH, and DLCoal (Duplication, Loss, and Coalescence), which does not. We generate comparable gene trees under both models, showing significant differences in various summary statistics; most importantly, CNH reduces the number of gene copies greatly. If this is not taken into account, the traditional method of estimating duplication rates (by counting the number of gene copies) becomes inaccurate. We also use these trees for species tree inference with the summary methods ASTRAL and ASTRAL-Pro, demonstrating that their accuracy, based on simulations calibrated on real data, may have been overestimated.

Abstract # 82

Albert Ch. Soewongsono, Barbara R. Holland, Małgorzata M. O'Reilly (University of Tasmania)

The Shape of Phylogenies Under Phase-Type Distributed Times to Speciation and Extinction

Phylogenetic trees are widely used to understand the evolutionary history of organisms. Tree shapes provide information about macroevolutionary processes. However, existing macroevolutionary models are unreliable for inferring the true processes underlying empirical trees. Here, we propose a flexible and biologically plausible macroevolutionary model for phylogenetic trees where times to speciation or extinction events are drawn from a Coxian phase-type (PH) distribution. First, we show that different choices of parameters in our model lead to a range of tree balances as measured by Aldous' β statistic. In particular, we demonstrate that it is possible to find parameters that correspond well to empirical tree balance. Next, we provide a natural extension of the β statistic to sets of trees. This extension produces less biased estimates of β compared to using the median β values from individual trees. Furthermore, we derive a likelihood expression for the probability of observing any tree with branch lengths under a model with speciation but no extinction. Finally, we illustrate the application of our model by performing both absolute and relative goodness-of-fit tests for two large empirical phylogenies (squamates and angiosperms) that compare models with Coxian PH distributed times to speciation with models that assume exponential or Weibull distributed waiting times. In our numerical analysis, we found that, in most cases, models assuming a Coxian PH distribution provided the best fit. In addition, this model allows us to fit hazard rate for speciation and we found evidence that speciation rates had changed through time in some clades of the squamate phylogeny.

Abstract # 25

Matthew Macaulay, Mathieu Fourment, Aaron Darling (The University of Technology Sydney)

Hyperbolic Tree Embeddings for Bayesian Phylogenetic Inference

Hyperbolic embeddings have successfully provided low dimensional representations of complex tree-like data. For phylogenetics, tree embeddings could offer a way to overcome the super-exponential number of tree combinations. However, such approaches are missing from Bayesian phylogenetics. In this talk, I will discuss embedding trees in hyperbolic space to perform Bayesian analysis.

We begin with a Markov Chain Monte Carlo (MCMC) to compare the quality of embeddings to state-of-art methods. We find that hyperbolic space offers low dimensional embeddings for MCMC and that proposals between different trees - considering both continuous parameters and discrete topologies - can be given by straightforward continuous distributions with fast sample methods.

Then we investigate the possibility of embedding variational distributions to make use of speed-up from gradient-based optimisation. Variational inference in the embedding space is a challenging problem, primarily due to the non-differentiability of the phylogenetic likelihood and prior as tree topologies change. We examine the extent that variational approximations capture salient features of the posterior distribution and discuss possibilities to refine these approximations. The prospect of hyperbolic embeddings for Bayesian phylogenetic inference is promising. It naturally envelops both discrete and continuous tree parameters of the posterior in a unified manner.

Abstract # 34

Ruriko Yoshida, Shelby Cox

Tree Topologies along a Tropical Line Segment

Tropical geometry with the max-plus algebra has been applied to statistical learning models over tree spaces because geometry with the tropical metric over tree spaces has some nice properties such as convexity in terms of the tropical metric. One of the challenges in applications of tropical geometry to tree spaces is the difficulty interpreting outcomes of statistical models with the tropical metric. This talk focuses on combinatorics of tree topologies along a tropical line segment, an intrinsic geodesic with the tropical metric, between two phylogenetic trees over the tree space and we show some properties of a tropical line segment between two trees. Specifically we show that a probability of a tropical line segment of two randomly chosen trees going through the origin (the star tree) is zero if the number of leave is greater than four, and we also show that if two given trees differ only one nearest neighbor interchange (NNI) move, then the tree topology of a tree in the tropical line segment between them is the same tree topology of one of these given two trees with possible zero branch lengths.

Abstract # 136

Benjamin Wilson (Lateral)

Learning phylogenetic trees as hyperbolic point configurations

We propose a novel method for the inference of phylogenetic trees that utilises point configurations on hyperbolic space as its optimisation landscape. Each taxon corresponds to a point of the point configuration, while the evolutionary distance between taxa is represented by the geodesic distance between their corresponding points. The point configuration is iteratively modified to increase an objective function that additively combines pairwise log-likelihood terms. After convergence, the final tree is derived from the inter-point distances using a standard distance-based method. The objective function, which is shown to mimic the log-likelihood on tree space, is a differentiable function on a Riemannian manifold. Thus gradient-based optimisation techniques can be applied, avoiding the need for combinatorial rearrangements of tree topology.

Abstract # 12

Hasindu Gamaarachchi, Hiruna Samarakoon, Sasha P. Jenner, James M. Ferguson, Timothy G. Amos, Jillian M. Hammond, Hassaan Saadat, Martin A. Smith, Sri Parameswaran, Ira W. Deveson (Kinghorn Centre for Clinical Genomics, Garvan Institute of Medical Research, Sydney; University of New South Wales, University of Montreal)

SLOW5: a new file format enables massive acceleration of nanopore sequencing data analysis

Nanopore sequencing is an emerging genomic technology with great potential. However, the storage and analysis of nanopore sequencing data have become major bottlenecks preventing more widespread adoption in research and clinical genomics. Here, we elucidate an inherent limitation in the file format used to store raw nanopore data — known as FAST5 — that prevents efficient analysis on high-performance computing (HPC) systems. To overcome this, we have developed SLOW5, an alternative file format that permits efficient parallelisation and, thereby, acceleration of nanopore data analysis. For example, we show that using SLOW5 format, instead of FAST5, reduces the time and cost of genome-wide DNA methylation profiling by an order of magnitude on common HPC systems (up to >30X times), and delivers consistent improvements on a wide range of different architectures. With a simple, accessible file structure and significant reductions in size compared to FAST5 (~25% reduction for a typical human genome), SLOW5 format will deliver substantial benefits to all areas of the nanopore community.

Example result: With the maximum resource allocation available on Australia's National Computing Infrastructure, genome-wide DNA methylation profiling on a single ~30X human genome sequencing dataset runs for >14 days at a cost of >$500 when using FAST5 files. We were able to complete whole-genome methylation profiling on a single 30X human dataset in just ~10.5 hours when using SLOW5 format as the input.

SLOW5 format and all associated software are free and open source.
SLOW5 format specification documents: https://hasindu2008.github.io/slow5specs;
Slow5lib: https://hasindu2008.github.io/slow5lib;
Slow5tools: https://hasindu2008.github.io/slow5tools;
Pre-print: https://www.biorxiv.org/content/10.1101/2021.06.29.450255v1.

Abstract # 8
Michael Charleston (University of Tasmania)
Landscapes of split space part 2: the quest for utility.
The idea of attacking the problem of phylogenetic inference by divide- (or split-) and-conquer is appealing. Tree space grows super-exponentially with the number n of taxa, as O(2^n n!), so breaking a problem in half, solving that and then rejoining the solved pieces could present a massive time saving. Non-trivial bipartitions of taxa are often referred to as *splits* and there are a number of potential score functions by which we could judge whether a split is "good": e.g., parsimony-based, cluster-based, or based on (sub)flattenings. The number of splits is still large (around 2^n) so we need a heuristic search to hunt for good ones. We consider the combinatorial landscape of splits as vertices (nodes) in a graph, weighted by the score function, and connected by edges representing adjacencies between splits. Searching for a good split will be easier if the landscape is smooth, and useful if locally optimal splits are mostly correct.
This talk is an update from my talk at Phylomania 2020 and provides some progress in my ongoing quest for some use of this approach.

Abstract # 147
Luke Cooper, Barbara Holland, Michael Charleston (University of Tasmania)
Models of biomolecular network evolution
The various types of biomolecular networks such as gene regulatory networks and protein interaction networks, have the potential to provide a plethora of useful information about extant, and ancestral, organisms. Given a pair of networks from two or more extant species that share a common ancestor, it is useful to know which nodes are the 'same', or are homologous, between the networks. With this knowledge, we can gain an understanding of interactions (edges) in the networks that are conserved, and are different, across the two networks.
Algorithms which are guaranteed to obtain 'the best' alignment are NP-hard, so we work with efficient heuristics. To measure the quality of an alignment produced by such an algorithm, we have to use a biologically informed measure of correctness. We propose such a measure that compares the alignment of each node to an "ideal" alignment that minimises distances between two nodes in a "gene duplication forest" where the leaves represent the nodes in the extant networks. We simulate the gene duplication history of networks via biologically motivated random graph models, and use them to compare the ability of network algorithms to produce a biologically accurate alignment using topological information alone with our new measure.

Mareike Fischer, Andrew Francis, Kristine Wicke (Greifswald University, University of Western Sydney)
Phylogenetic Diversity Rankings in the Face of Extinctions: the Robustness of the Fair Proportion Index
Planning for the protection of species often involves difficult choices about which species to prioritize, given constrained resources. One way of prioritizing species is to consider their "evolutionary distinctiveness", i.e. their relative evolutionary isolation on a phylogenetic tree. Several evolutionary isolation metrics or phylogenetic diversity indices have been introduced in the literature, among them the so-called Fair Proportion

index (also known as the “evolutionary distinctiveness” score). This index apportions the total diversity of a tree among all leaves, thereby providing a simple prioritization criterion for conservation. Here, we focus on the prioritization order obtained from the Fair Proportion index and analyze the effects of species extinction on this ranking. More precisely, we analyze the extent to which the ranking order may change when some species go extinct and the Fair Proportion index is re-computed for the remaining taxa.
In my talk, I show that for each phylogenetic tree, there are edge lengths such that the extinction of one leaf per cherry completely reverses the ranking. Moreover, I show that even if only the lowest ranked species goes extinct, the ranking order may drastically change.

Abstract # 92
Andrew Francis, Peter Jarvis

The perfect picnic project: Sandwiches, eggs, and trees

Persi Diaconis and Susan Holmes showed in 1998 that binary phylogenetic trees on *n* leaves can be encoded by matchings on a set of (2*n*-2) elements, that is, a partition of the set into pairs. In this talk we show how this matching can be extended to a correspondence with unbalanced Brauer diagrams (and, more generally, partition diagrams) that preserve elements of the structure of the tree. Such diagrams have an associated natural action of the symmetric group, and through Green's relations of semigroup theory, we can construct “eggbox diagrams” displaying equivalence class structures. Another semigroup construction, the sandwich product, allows us to define a product on the set of trees, which is given relative to a chosen tree. With this, we can consider various related semigroup substructures, such as the regular subsemigroup, or idempotents, in the context of phylogenetic trees.