Programme Booklet

ISCB REGIONAL Student GROUP SYMPOSIUM 2019

COMBINE

Australia

Sydney, 9 December &

Asia & Australia's conference for students in bioinformatics and computational biology

Welcome

Welcome to the 2019 COMBINE/AYRCOB Symposium. This annual event is an opportunity for students and early career researchers to present their work to peers in a relaxed and supportive environment.

Regards, The COMBINE Team

Symposium Committee Members

Katarina Stuart (co-chair)	Tingting Gong	Anushi Shah	Noorul Amin
Yingxin Lin (co-chair)	Jiayuan Huang	Qing Wang	Jinxin Zhao
Jieun Hani Kim	Arindam Halder		

Sponsors

Australian BioCommons

INTERSECT RESEARCH FASTER

THE UNIVERSITY OF SYDNEY

BioTuring

About COMBINE

COMBINE is a student-run Australian organisation for researchers in computational biology, bioinformatics, and related fields. COMBINE is the official International Society for Computational Biology Regional Student Group for Australia and a subcommittee of The Australian Bioinformatics and Computational Biology Society (ABACBS). We aim to bring together students and early-career researchers from the computational and life sciences for networking, collaboration, and professional development.

Australia has many research institutes, each with their own cohorts of students. Aside from conferences, there are few opportunities that bring these students together, allowing them to discover the different kinds of research going on at other institutes. COMBINE aims to bridge this institutional divide by organising seminars, workshops and social events. Together, these events allow students to connect with each other and build a network in a casual environment.

Symposium Programme

8:00	**Registration**	
8:50	**Symposium welcoming address**	
Session 1 (Chair: Jinxin Zhao)		
9:00	**Roni Froumine** The University of Melbourne	Predicting CRISPR-Cas activity and exploring its relationship with antimicrobial resistance in Klebsiella pneumoniae.
9:15	**Gerry Tonkin-Hill** Wellcome Sanger Institute, UK	Preventing Pangenome Pitfalls with Panaro.
9:30	**Maria Satti** National Institute of Genetics, Japan	Compar ative genomic analysis of Bifidobacterium species isolated from Egyptian fruit bat Rousettus egyptiacus.
9:45	**Jieun Hani Kim** The University of Sydney	CiteFuse: a comprehensive toolkit for the analysis of CITE-seq data.
10:00	**Pablo Acera Mateos** Australian National University	Comprehensive identification of nucleotide biochemical modifications from nanopore signal data
10:15	**Morning tea**	
Session 2 (Chair: Anushi Shah)		
10:45	**Rachel Bowen-James** Children's Cancer Institute	Exploring somatic BAM compression.
11:00	**Richard Lupat** Peter MacCallum Cancer Centre	Application of Deep Learning Techniques in Extracting Features from Breast Cancer Genomic Data.
11:15	**Tingting Gong** Graven Institute of Medical Research	Refining somatic structural variant detection and annotation for precision oncology.
11:30	**Alessandra Whaite** GeneCology Research Centre, University of the Sunshine Coast	The role of proteomics in understanding the structure of natural protein fibres produced by molluscs and spiders.
11:45	**Nhi Hin** The University of Adelaide	RNA-seq analysis in a zebrafish model of Alzheimer's disease highlights the importance of iron homeostasis.
12:00	**Fast forward 16 talks (Chair: Yingxin Lin & Katarina Stuart)**	
12:20	**Lunch & Poster session**	
Session 3 (Chair: Tingting Gong)		

13:30	**Marina Reixachs** Australian National University	Ribosome profiling at isoform level reveals an evolutionary conserved impact of differential splicing on the proteome.
13:45	**Ahmad Zeeshan Siddiqui** The University of New South Wales	Tissue-specific expression of circular RNAs in healthy human adults.
14:00	**Qiuyi Li** The University of Melbourne	HIDTL, a new model of gene family evolution.
14:15	**Ilariya Tarasova** Walter and Eliza Hall Institute of Medical Research	Exploration of time and division dependent gene expression during B cell division.
14:30	**Xiangnan Xu** The University of Sydney	LC-N2G: A Local Consistency Approach for Nutrigenomics Data Analysis.
14:45	**Afternoon tea**	
15:15	**Welcome to Country/Conference**	
15:40	**Joint keynote of COMBINE, AYRCOB, ABACBS and GIW (Prof. Rafael Irizarry)**	
16:30	**COMBINE Awards, ABACBS Awards + ABACBS Award Q&A**	
17:10	**Welcome reception drinks**	
17:30	**Career Panel** at CPC Level 6 Seminar Room: Prof. Rafael Irizarry, Prof. Marcel Dinger, Dr. Denis Bauer, Dr. Emily Wong	
19:00	**Social Night** at CPC Level 6 Seminar Room	

Fast forward

	Title	Presenter
1	Gene signature-based predictive models suitable for clinical translation	Dharmesh D. Bhuva
2	ampir: an R package to predict antimicrobial peptides in genomes	Legana Fingerhut
3	The unique methylation profile of the placenta can be used for quality control.	Qianhui Wan
4	Investigation of de novo mutations in human genomes using whole genome sequencing datasets	Anushi Shah
5	CPOP: Cross-Platform Omics Prediction procedure enables precision medicine	Kevin Wang
6	Opportunities and challenges of analysing multi-regional tumour biopsies to characterise heterogeneity in cancer	Sebastian Hollizeck
7	Human transposons are an abundant supply of transcription factor binding sites and promoter activities in breast cancer cell lines.	Jiayue-Clara Jiang
8	A 9-gene score for predicting B-ALL risk of relapse and survival	Feng Yan
9	Orthogonal evidence for Olfactory Receptors can be used for agonist prediction	Amara Jabeen
10	sPLSDA-batch: Batch effect correction in microbiome data	Yiwen Wang
11	Comparison of algorithms for mtDNA variant discovery from whole-genome sequencing data	Eddie K.K. Ip
12	Human Brain eQTL: Analysis and Insights	Letitia Sng
13	Strain level genotyping of microbial communities using long and short read phasing	Rhys Newell
14	The dynamic genome behind the emergence of octopod novelties	Brooke Whitelaw
15	Bioinformatic analysis of genome-wide SNPs elucidates cryptic species boundaries and potential speciation of Australian toadlets (Myobatrachidae: Uperoleia)	Frederick Jaya
16	Predicting the functional effect of genetic changes using variant effect prediction algorithms	Emma Darling

Poster

	Poster Presenter	Title
1	Marco Montes de Oca	Comparative analysis of alternative promoters in immune and non-immune tissues by RNA-seq and H3K4me3 ChIP-seq data
2	Urwah Nawaz	Transcriptome profiling of individuals with compromised nonsense-mediated mRNA decay implicates immune system and neuronal cell dysfunction in neurodevelopmental disorders
3	Victor Wei Tse Hsu	Multi-breed comparison of canine lymphoma susceptibility
4	Shweta S. Joshi	Benchmarking short and long read Capture sequencing techniques to identify novel transcripts in neuropsychiatric disorder risk genes
5	Frederick Jaya	Assessing the performance of recombination detection methods using simulated viral sequences
6	Mikhail Gudkov	ConanVarvar: a versatile tool for the detection of large syndromic copy number variation from whole genome sequencing data
7	Ning Liu	Identifying candidate cis-regulatory variants in regulatory T cells for Type 1 Diabetes research
8	Melanie Smith	A comprehensive miRSeq profile of miRNA in the human placenta across early gestation.
9	John Salamon	Investigating Tumour Heterogeneity using Computational Modelling of Patient-Specific Network Rewiring
10	Yujie Cao	Development of a one-stop thalassaemia screening method by next generation sequencing
11	Ahmad Mollazadeh Taghipour	Functional studies of plastid-targeted protein 1 (PTP-1) gene associated with desiccation tolerance in the resurrection plant Craterostigma plantagineum
12	Stephen Cristiano	Genome-wide cell-free DNA fragmentation in patients with cancer
13	Dean Southwood	Benchmarking state-of-the-art genome assembly methods for eukaryotic genomes
14	Patricia Sullivan	Introme: Identifying atypical splice-altering mutations as drivers of high-risk paediatric cancer
15	Thomas Geddes	Autoencoder-based cluster ensembles for single-cell RNA-seq data analysis
16	Claire Sun	Integrative Approaches in Functional Genomics to Identify Genetic Dependencies in Paediatric Cancer
17	Ruining Dong	SVEnsemble: an algorithm for ensemble structural variant calling using re-evaluated quality scores via probabilistic random forest
18	Aedan Roberts	A Bayesian hierarchical model for detecting differential gene expression distributions for RNA-seq data
19	Akanksha Srivastava	ReorientExpress: reference-free orientation of nanopore cDNA reads with deep learning
20	Yue Cao	scDC: Single cell differential composition analysis
21	Xining Li	Understanding the in vivo function of A-to-I RNA editing by ADARs
22	Emma Darling	Predicting the functional effect of genetic changes using variant effect prediction algorithms
23	Xiunan Fang	FlowGrid: A python package for fast clustering for millions of single cell transcriptomic profiles
24	Pei Qin Ng	Using tRNA-seq, RNA-seq, and proteomics data analysis to investigate the importance of tRNA modifications ‚Äö√Ñ√¨ A case study of tRNA Guanine and Inosine-N1-methyltransferase TRM5 in Arabidopsis thaliana.
25	Megan Soon	Single-cell transcriptomic and epigenomics map effector to memory CD4+ T cell transition in vivo.
26	Rui Chen	openMTB: A System for Evidence-Driven Personalized Cancer Treatments in Molecular Tumor Boards

27	Zainab Noor	A Mix-And-Match Library Approach for Enriching Plasma Proteome Discovery
28	YUPEI YOU	Accurate identification of mRNA splice sites using Oxford Nanopore sequencing
29	Angelita Liang	The role of kinases in the differentiation of bone marrow stromal cells into osteoblasts: a systematic analysis by knockdown
30	Nicolas Canete	Spatial Analysis of Highly Multiplexed Microscopy Data
31	Katarina Stuart	Using genomics to reveal drivers of invasion success
32	Yingxin Lin	Multiscale hierarchical classification of single cells into cell-types and sample size learning
33	Yuan Gao	The relationship between maternal gut microbiota during pregnancy and the offspring,Äôs immune phenotype: a birth cohort study.
34	Gordon Qian	Discovery of perturbation gene targets via free text metadata mining in Gene Expression Omnibus
35	Samuel Old	A Case Study of Cutting-Edge Immunology: Interrogating Results from a Large-Scale RNA-Sequencing Experiment
36	Agus Hartoyo	Inference of the parsimonious mechanism underlying the emergence of alpha-blocking in human electroencephalography
37	Javier Ortega	Identifying the genes that yield the benefits of dietary restriction without the costs.
38	Jinxin Zhao	Integration of transcriptomics data in a genome-scale metabolic model to decipher the mechanisms of polymyxin resistance in Acinetobacter baumannii
39	Dharmesh D. Bhuva	Gene signature-based predictive models suitable for clinical translation
40	Frederick Jaya	Bioinformatic analysis of genome-wide SNPs elucidates cryptic species boundaries and potential speciation of Australian toadlets (Myobatrachidae: Uperoleia)
41	Legana Fingerhut	ampir: an R package to predict antimicrobial peptides in genomes
42	Qianhui Wan	The unique methylation profile of the placenta can be used for quality control.
43	Anushi Shah	Investigation of de novo mutations in human genomes using whole genome sequencing datasets
44	Kevin Wang	CPOP: Cross-Platform Omics Prediction procedure enables precision medicine
45	Sebastian Hollizeck	Opportunities and challenges of analysing multi-regional tumour biopsies to characterise heterogeneity in cancer
46	Jiayue-Clara Jiang	Human transposons are an abundant supply of transcription factor binding sites and promoter activities in breast cancer cell lines.
47	Feng Yan	A 9-gene score for predicting B-ALL risk of relapse and survival
48	Amara Jabeen	Orthogonal evidence for Olfactory Receptors can be used for agonist prediction
49	Yiwen Wang	sPLSDA-batch: Batch effect correction in microbiome data
50	Eddie K.K. Ip	Comparison of algorithms for mtDNA variant discovery from whole-genome sequencing data
51	Letitia Sng	Human Brain eQTL: Analysis and Insights
52	Rhys Newell	Strain level genotyping of microbial communities using long and short read phasing
53	Brooke Whitelaw	The dynamic genome behind the emergence of octopod novelties

Abstracts: Oral Presentation

Predicting CRISPR-Cas activity and exploring its relationship with antimicrobial resistance in Klebsiella pneumoniae.

Roni Froumine
The University of Melbourne

Klebsiella pneumoniae (Kp) is a major cause of bacterial healthcare-associated infections worldwide. The Kp population comprises hundreds of lineages (clones), a subset of which are particularly concerning because they have accumulated high frequency and diversity of antimicrobial resistance (AMR) genes through horizontal gene transfer (HGT) and cause infections that are extremely difficult to treat. CRISPR-Cas are adaptive immune systems present in ~33% of Kp. They can block HGT and limit acquisition of AMR genes. Therefore, we hypothesise that Kp clones enriched for AMR are less likely to possess an active CRISPR.

While several tools exist for identifying the presence of CRISPR-Cas among genome assemblies, none distinguish active and degraded systems. Here we describe a novel approach to predict CRISPR activity. Cas proteins detect and degrade incoming DNA, and insert copies of short recognition motifs, 'spacers', into the CRISPR array in the host chromosome. Spacer sequences are preferentially incorporated at the array's leader end thus their ordering provides important temporal information. We exploit this by comparing CRISPR array spacers among Kp in the same clone. Their spacer sequences are extracted, aligned and mapped to the clone's phylogenetic tree. Ancestral state reconstruction is used to quantify activity based on the predicted amount and location of spacer acquisition or loss in the CRISPR arrays.

We use this approach to identify putatively active CRISPR-Cas in 31 Kp clones and test for association with the presence of AMR genes. Our findings have the potential to inform novel control strategies and CRISPR-based therapeutics for Kp infections.

Preventing Pangenome Pitfalls with Panaro.

Gerry Tonkin-Hill
Wellcome Sanger Institute, UK

Improvements in genome sequencing have led to larger population samples, but counterintuitively this has led to an increase in the number of gene annotation errors. The automated annotation of draft genomes remains difficult and resulting errors tend to propagate across databases. This has profound consequences for defining the pangenome (the set of all genes found in a species), as most algorithms for clustering genes do not account for such errors. This can lead to an artificial inflation in estimates of pangenome size and can complicate downstream analyses. We have developed Panroo, a pangenome graph based clustering tool that is able to correct for many of the sources of error introduced during annotation by fragmented assemblies, contamination, diverse gene families and mis-assemblies. We verified our approach through extensive simulation of de novo assemblies using the infinitely many genes model and by analysing a diverse set of large bacterial genome datasets. Using a highly clonal dataset containing 1,419 Pseudomonas aeruginosa genomes from a single patient we demonstrate that failing to account for annotation errors leads to a 40% increase in the pangenome size estimated with Roary, a popular pangenome method. Accounting for such errors with Panaroo reduces this error rate to 1.8%. The improved accuracy of Panaroo enabled us to quantify gene gain and loss rates across 13,454 isolates from the major global clades of Streptococcus pneumoniae. Furthermore, our underlying graph based representation allowed for an association study between genes, structural rearrangements and antibiotic resistance in S. pneumoniae.

Comparative genomic analysis of Bifidobacterium species isolated from Egyptian fruit bat Rousettus egyptiacus.

Maria Satti

National Institute of Genetics, Japan

Bifidobacterium is an important probiotic genus. The species in the genus were previously isolated from various hosts like cow, rabbit, pig, non-human primates, and human being; our group has recently isolated two novel Bifidobacterium species from Egyptian fruit bat. The host diet contributes to the development of intestinal microbial communities, and the bat dietary habits should affect the development of important probiotic bacterial species like bifidobacteria. The aim of this study was to investigate the genetic biodiversity of bifidobacteria from bat compared to bifidobacterial species from human and non-human primates by decoding genome sequences. The description of the genomic features in different niches is fundamental in clarifying repertoire of genes that have caused their evolutionary differentiation. Such genomic analyses support the hypothesis that bat strains have been subjected to genetic adaptations to their host environment such as a peculiar diet heavily based on sugars. The comparative analysis of bifidobacterial species revealed that bifidobacteria in bat possess the higher genomic similarity with non-human primates than human or other mammals. Bat strains report the presence of unique GHs i.e. GH 59 and GH 88 classes. Plant-dietary metabolising GHs such as GH 28, GH 53, GH 78, GH 105, GH 115 and GH 146 were found to be specific for bat and non-human primate species. The comparative analysis in this study has revealed the important features of bifidobacteria in bat such as their contribution in metabolizing the host dietary carbohydrates. Bat and non-human primate specific GHs corresponding to the metabolism of their dietary carbohydrates suggest the dietary association between these groups.

CiteFuse: a comprehensive toolkit for the analysis of CITE-seq data.

Jieun Hani Kim
The University of Sydney

Multi-modality profiling of single cells represents one of the latest technological advancements in molecular biology. Among various single-cell multi-modality strategies, cellular indexing of transcriptomes and epitopes by sequencing (CITE-seq) allows simultaneous quantification of two distinct species: RNA and surface marker proteins (ADT). Here, we introduce CiteFuse, a streamlined package consisting of a suite of tools for the pre-processing, modality integration, clustering, ADT evaluation, RNA-ADT network construction, differential expression analysis, and interactive web-based visualization of CITE-seq data. We show the integrative capacity of CiteFuse to fuse the two data types and its relative advantage against data generated from single modality profiling. Furthermore, we illustrate the pre-processing steps in CiteFuse and in particular a novel doublet detection method based on a combined index of cell hashing and transcriptome data. Collectively, we demonstrate the utility and effectiveness of CiteFuse for the integrative analysis of transcriptome and epitope profiles from CITE-seq data.

Comprehensive identification of nucleotide biochemical modifications from nanopore signal data

Pablo Acera Mateos
Australian National University

Nanopore sequencing is currently the only technology able to sequence ultra-long DNA and RNA molecules in their native forms: hence, potentially enabling the detection naturally occurring covalent modifications in nucleotides. This raises the opportunity of creating algorithms that can detect such modifications. These tools will help us fill the gaps in the understanding of the role of nucleotide modifications in the regulation of transcription, RNA splicing, RNA translation, RNA transport, RNA degradation...etc. Currently, available tools are able to detect a reduced number of modifications with limited accuracy. These algorithms are mainly of two types: supervised machine learning models that show acceptable accuracy but are limited to specific modifications the algorithms have been trained on, and methods based on statistical tests on the nanopore signal distribution that usually have a high false positive rate but are able to identify modifications in a more unbiased fashion. Here, we describe a new algorithm that combines both principles, making use of deep-learning and speech recognition strategies to accomplish high accuracy and unbiased detection of DNA/RNA modifications. As our model is not trained with explicit information from nucleotide modifications, it can potentially be used to detect a broad range of them.

Exploring somatic BAM compression.

Rachel Bowen-James
Children's Cancer Institute

ISCB REGIONAL Student GROUP SYMPOSIUM 2019
COMBINE
Australia
Sydney, 9 December &
Asia & Australia's conference for students in bioinformatics and computational biology

The application of next-generation sequencing technologies to the study of cancer reveals the specific mutations that give rise to an individual,ÄövÑv¥s cancer. Hence, the volume of data generated for somatic mutation analysis is growing rapidly. The cost-effective storage of this data is becoming increasingly dependent on compression methods that do not compromise the detection of low variant allele frequency (VAF) alterations. Here we investigate and optimise BAM compression methods for somatic variant detection. 'Truth' sets of variants were generated using Strelka2 on uncompressed BAM files, acting as a benchmark for the evaluation of the impact of compression. The BAM files were compressed using default and customized levels of Crumble ,ÄövÑv¨ a lossy compression tool developed for germline data. Strelka2 variants produced from the compressed files were then compared to the truth sets. Initial parameter optimisation was conducted using 2-chromosome BAM files. We analysed 9 standard Crumble levels and developed 18 optimised levels. The top performing optimised levels were used to compress two whole genome datasets. The best optimised level achieved a mean F-score of 0.95, and mean decrease in file-size of 36.68%. The most aggressive standard Crumble level achieved a marginally better mean decrease in file-size of 38.95% but gave less accurate results, with a mean F-score of 0.78. We have developed optimised Crumble parameters for the compression of somatic data with fewer negative impacts on variant calling. We minimised the loss of variants due to compression and achieved a false call rate similar to the disparity between different variant callers.

Application of Deep Learning Techniques in Extracting Features from Breast Cancer Genomic Data.

Richard Lupat
Peter MacCallum Cancer Centre

Rapid advancement in genomic technologies has produced a vast amount of clinical genomic data across different levels of omics variables. Some of these data are accessible through public repositories such as The Cancer Genome Atlas (TCGA). However, the enormous volume of data requires the application of specialised techniques for data mining, integration and interpretation to provide valuable insights. There have been various machine learning algorithms, supervised and unsupervised, successfully applied to these data and led to clinically relevant conclusions. However, these algorithms often rely on prior biological studies or limited to a selected number of most significant features in the data. In this project, we applied an unsupervised deep learning based method, known as Autoencoders, to extract complex patterns from breast cancer genomic data independent of prior known biology. We designed an autoencoder using features derived from gene expression and copy number data from the TCGA cohort. We used 746 samples to train and validate this model, which extracted 128 features from the combination of all input variables. To evaluate the performance of this method, these extracted features are used as part of our dimensionality reduction step for our downstream-supervised classifiers of ER status and PAM50 intrinsic subtype. The classifiers were applied to the same training datasets and accuracy of each was assessed on the validation set (70:30 datasets split). This combination of autoencoder and feed-forward neural network classifier distinguished ER status (92% accuracy), Basal-like vs. non-Basal-like (94% accuracy) and able to predict samples,ÄövÑv¥ breast cancer PAM50 intrinsic subtype (88% accuracy). This initial result provides a good foundation for our further study in developing a deep-learning based prognosis model.

Refining somatic structural variant detection and annotation for precision oncology.

Tingting Gong
Graven Institute of Medical Research

Somatic structural variants (SVs) play a significant role in cancer development and evolution. While next generation sequencing (NGS) has facilitated the detection of somatic variants in cancer genomes, accurate detection of somatic SVs is still limited by short-read NGS data, as well as low tumour purity and high tumour genome heterogeneity commonly observed in clinical samples. We aim to evaluate variables impacting our ability to accurately detect somatic SVs and facilitate further SV breakpoint feature and pattern recognition for a better understanding of their functional impact and formation mechanisms. In this study, we evaluated single and combinatoric effects of SV caller, SV types and sizes, variant allele frequency (tumour purity), sequencing depth of coverage, and variant breakpoint resolution. Using a generalized additive model allowed predictions of sensitivity and precision to be made for any combination of predictors. The prediction model was implemented in a web-based application, called Shiny-SoSV, which is freely available at https://hcpcg.shinyapps.io/shiny-sosv. Shiny-SoSV provides an interactive and visual platform for users to easily explore the impacting variables on somatic SV detection,

thereby enabling users to rapidly make informed sequencing and bioinformatics decisions early on in their study design. Using clinically-derived whole genome NGS data for a hyper-duplicated prostate cancer genome, we evaluated the performance of three best-performing SV callers (Lumpy, Manta and GRIDSS). While 263/669 (40%) of DUPs were identified by all three callers, after visual inspection 421 (63%) were validated at single nucleotide resolution. Interestingly, downstream annotation showed that 74% of the DUP breakpoints interrupted a total of 416 genes. Assembly of the DUP breakpoints revealed 88% of DUPs are composed of 706 transposable elements. This single patient example, highlights the significance of somatic SVs in carcinogenesis and the need to further refining somatic SV detection.

The role of proteomics in understanding the structure of natural protein fibres produced by molluscs and spiders.

Alessandra Whaite
GeneCology Research Centre, University of the Sunshine Coast

Natural protein fibres are secreted by many species including insects, molluscs, fish and arachnids. Silk is best known as the material procured from the mulberry silkworm, Bombyx mori (Linnaeus, 1758), yet silk is also made by bees, wasps, pseudoscorpions and spiders. Mussels, clams and oysters are all bivalve molluscs that produce a marine version of these proteinaceous fibres called byssal threads, or collectively, a byssus or ‚ÄövÑvƒbeard‚ÄövÑvπ. Proteinaceous threads made by invertebrates have become a research focus due, not only to the strength of such materials, but also the potential for use in biomedical fields due to low immunogenicity. Applying Bioinformatics to elucidate the molecular structure of natural protein fibres is challenging because of the repetitiveness of silk sequence motifs and the dearth of genome databases available for many marine species that produce proteinaceous fibres. This study has utilised RNA-seq and proteomics approaches to facilitate the assembly of secretory tissue transcriptomes and thread proteomes in spiders and bivalve molluscs. Understanding the molecular structure of such materials underpins our ability to mimic these diverse materials for use in biomedicine.

RNA-seq analysis in a zebrafish model of Alzheimer's disease highlights the importance of iron homeostasis.

Nhi Hin
The University of Adelaide

Analysing gene expression data from diseased and normal brain tissue has been valuable for exploring molecular mechanisms contributing to Alzheimer's disease and how these diverge from normal aging. Our laboratory has used gene editing technologies to introduce familial Alzheimer's-like mutations into zebrafish, followed by RNA-sequencing of wild-type and mutant brains at young and old age under both normal and low-oxygen conditions in a 2x2x2 full-factorial design. This design has offered us a unique opportunity to study the molecular basis of the disease in its early stages in the young brains, and augment our knowledge with how aging and brain oxygen levels relate to disease progression. We describe exploratory analyses from RNA-seq data from the brains of these zebrafish and how they allow us to explore broad-scale disruptions in molecular processes, in addition to more detailed analyses testing whether processes hypothesised to be important in Alzheimer's disease (iron homeostasis in particular) showed disruption at the regulatory level. We also emphasise the importance of data visualisation in facilitating hypothesis generation and communicating findings in an accessible way.

Ribosome profiling at isoform level reveals an evolutionary conserved impact of differential splicing on the proteome.

Marina Reixachs
Australian National University

The differential production of transcript isoforms from gene loci is a key cellular mechanism. Yet, its impact in protein production remains an open question. Here, we describe ORQAS (ORF quantification pipeline for alternative splicing) a new pipeline for the translation quantification of individual transcript isoforms using ribosome-protected mRNA fragments (Ribosome profiling). We found evidence of translation for 40-50% of the expressed transcript isoforms in human and mouse, with 53% of the expressed genes having more than one translated isoform in human, 33% in mouse. Differential analysis revealed that about 40% of the splicing changes at RNA level were concordant with changes in translation, with 21.7% of

changes at RNA level and 17.8% at translational level conserved between human and mouse. Furthermore, orthologous cassette exons preserving the directionality of the change were found enriched in microexons in a comparison between glia and glioma, and were conserved between human and mouse. ORQAS leverages ribosome profiling to uncover a widespread and evolutionary conserved impact of differential splicing on the translation of isoforms and in particular, of microexon-containing ones. ORQAS is available at https://github.com/comprna/orqas

Tissue-specific expression of circular RNAs in healthy human adults.

Ahmad Zeeshan Siddiqui
The University of New South Wales

Circular RNAs (circRNAs) are a naturally occurring class of RNA molecules formed by the backsplicing mechanism -where a pre-mRNA 3 splice site of a downstream exon is covalently linked with the 5,ÄövÑv¥ splice site of an upstream exon, to form a covalently closed circular transcript. CircRNAs are dysregulated in certain cancers and have the potential to serve as candidates for functional biomarkers and therapeutic targets, due to their tissue-specific expression. However, there is limited understanding of circRNA expression in healthy human tissues and their association with diseases. Further in-depth analysis of molecular pathology requires a collection of reference circRNAs specific for individual healthy tissues. The identification and characterisation of circRNAs was the goal of our study. We systematically investigated circRNA expression in five healthy human adult tissues; kidney, liver, lung, colon and stomach. We discovered there is similar abundance of circRNAs across these tissues except for in stomach (much lower abundance). Altogether, we present a reference list of 13 tissue-specific circRNAs that have been chosen according to our filtering criteria; present in all datasets analysed and expression level of > 0.1 CPM. Overall, we identified highly expressed and tissue-specific circRNAs that will facilitate future investigation into circRNAs and their relationship with disease development in specific human tissues.

HIDTL, a new model of gene family evolution.

Qiuyi Li
The University of Melbourne

The evolution of gene families is an important aspect of molecular evolution and also crucial when inferring the relationships among species. Gene families evolve through a complex process involving evolutionary events such as speciation, gene duplication, horizontal gene transfer, and gene loss. Furthermore, when a population of individuals undergoes several speciations in a relatively short time, there can exist polymorphisms maintained throughout the time which eventually fix in different descendant lineages. This phenomenon is called incomplete lineage sorting (ILS). Due to these evolutionary processes, there are often topological differences between a gene tree and its corresponding species tree. Reconciliation methods are developed to explain these differences. Accurate gene and species reconciliation is fundamental to infer the evolutionary history of a gene family. Any reconciliation method is built on a model of gene family evolution. A few gene family evolution models have been proposed over the last decade, for example the duplication-loss model, the locus tree model and the haplotype tree model. However, little attention has been paid to the presence of hemiplasy, which occrs when a newly created locus does not fix in all descendant species. In this talk, we review the existing models of gene family evoluation, and then introduce a new probabilistic gene family evolution model, HIDTL, which combines all the advantages of the existing models and additionally allows hemiplasy. We compare HIDTL with the existing models to show that HIDTL can model more complex scenarios, and so should be used for testing the accuracy of reconciliation methods.

Exploration of time and division dependent gene expression during B cell division.

Ilariya Tarasova
Walter and Eliza Hall Institute of Medical Research

B cells, a type of lymphocyte, play a critical role in the adaptive immune system. They have a sophisticated decision-making mechanism of differentiation into different cell types. As the cells divide, levels in key regulatory molecules alter and guide the differentiation path of each individual cell. However, the mechanisms governing this process and how cells transit to new cell types are still poorly understood at the molecular level. In my talk I will discuss how we are using RNA-seq to undertake a comprehensive inventory of which gene changes, occurring during B cell response programs, are affected by time (and not

division), by division (and not time), and by a combination of time and division. We are developing an intuitive statistical evaluation method to quantify the effect of our two variables. From preliminary analysis, we found that many gene changes are controlled by time and unaffected by division. Combining our method with pathway analysis has the potential to unravel which type of genes (time or division dependent) regulates specific biological processes how this all together affects B cells fate. For example, most cell cycle genes are time-dependent, as expected, and immunoglobulin gene family are not specific in any category.

LC-N2G: A Local Consistency Approach for Nutrigenomics Data Analysis.

Xiangnan Xu
The University of Sydney

Nutrigenomics aims at understanding the interaction between nutrition and gene information. Due to the complex mechanism of nutrients and genes, their relationship exhibits non-linearity. One of the most effective and efficient methods to explore their relationship is the nutritional geometry framework which fits a response surface for the gene expression over two prespecified nutrition variables. However, when the number of nutrients involved is large, it is challenging to find combinations of informative nutrients with respect to a certain gene and to test whether the relationship is stronger than chance. Methods for identifying informative combinations are essential to understanding the relationship between nutrients and genes. To address this challenge, we introduce Local Consistency Nutrition to Graphics (LC-N2G), a novel approach for ranking and identifying combinations of nutrients with gene expression. In LC-N2G, we first propose a model-free quantity called Local Consistency statistic to measure whether there is non-random relationship between combinations of nutrients and gene expression measurements based on the similarity between samples in the nutrient space and their difference in gene expression. Then combinations with small LC are selected and a permutation test is performed to evaluate their significance. Finally, the response surfaces are generated for the subset of significant relationships. Evaluation on simulated data and real data shows the LC-N2G can accurately find combinations that are correlated with gene expression. The LC-N2G is practically powerful for identifying the informative nutrition variables correlated with gene expression. Therefore, LC-N2G is important in the area of nutrigenomics for understanding the relationship between nutrition and gene expression information.

Abstracts: Fast Forward

Gene signature-based predictive models suitable for clinical translation

Dharmesh D. Bhuva
Walter and Eliza Hall Institute of medical research

Transcriptomic signatures are useful in understanding the molecular phenotypes of cells, tissues, and patient samples, for example, they have been successfully applied to stratify breast cancer patients into molecular subtypes. In most cases, gene expression signatures are developed using whole-transcriptome scale measurements. This means that methods for matching signatures to samples typically require samples to be measured on the same scale. The need for relatively large amounts of starting material, and sequencing cost for whole-transcriptome measurement limits clinical applications, and accordingly thousands of existing gene signatures are unexplored in a clinical context. Genes within a signature carry most of the information about the molecular phenotype they represent, so an efficient assay and scoring method would quantify the abundance of these genes with few additional measurements needed. We have modified our method, singscore, to quantify and summarise relative expression levels of signature genes from individual samples through the inclusion of additional "stably-expressed genes". We identified genes

with stable expression across different abundances and with a preserved relative order across large numbers (thousands) of samples to allow signature scoring, as well as to support general data normalisation. We show that signature scores computed from whole-transcriptome data are comparable to those calculated using only values for signature genes and our panel of stable genes. This opens up the potential to develop panel-type tests for gene expression signatures that can support clinical translation of the powerful insights contributed by transcriptomic studies in cancer.

ampir: an R package to predict antimicrobial peptides in genomes

Legana Fingerhut
Tropical Bioinformatics and Molecular Biology, JCU, Townsville

Antimicrobial peptides (AMPs) are key components of the innate immune response. AMPs regulate pathogens and the microbiome and are essential for the hosts health. Rather than just relying on bioactivity screening, the discovery of novel AMPs can be enhanced by in silico prediction. However, a number of limitations apply to the software currently available for this purpose. First, in general, these predictive tools are not easily applicable to large datasets, such as whole genome sequences. Second, selective exclusion of more challenging sequences from published training datasets means that many existing AMP predictors perform sub-optimally when applied to ‚ÄövÑvƒreal‚ÄövÑvπ biological data. To explore AMP evolution and to facilitate genome-wide studies of AMPs in host organisms, a fast, high-throughput R package, called ampir, was developed to predict AMPs. Predictions in ampir are based on a Support Vector Machine (SVM) model trained using 26 features calculated from ~8K protein sequences of known AMPs and random genome-wide proteins. The SVM model performed with 91% sensitivity, 95% specificity, and a 97% Area Under the Receiver Operating Characteristic Curve (AUC-ROC) measurement when tested on an independent validation dataset (~2K sequences). This talk will reveal the design process for ampir including specific steps to improve prediction accuracy, user interface design and performance to suit whole genome scans. It will also compare ampir to other existing AMP predictors and demonstrate the effects of selective exclusion on model performance. Finally, it will discuss downstream applications enabled by ampir including the evolution of AMPs and AMP repertoires.

The unique methylation profile of the placenta can be used for quality control.

Qianhui Wan
The University of Adelaide

The purity of tissue samples can affect accuracy and utility of DNA methylation array analyses. This is particularly important for the placenta since placental villous tissue from early pregnancy terminations can be difficult to separate from non-villous tissue, resulting in potentially inaccurate results. The placenta is globally hypomethylated and also contains large partially methylated domains (PMDs) that make it distinct from other surrounding tissues. To identify potential tissue impurities in placenta samples we developed a clustering method by applying principal component analysis (FactoMineR R package) and multivariate unsupervised clustering with mixtures of Gaussian distributions (mclust R package) to 408 public and our own Illumina 450K methylation array data from different regions of the placenta and surrounding tissues. We identified 11 potential mixed placenta samples from 379 placenta samples. PMDs in the mixed placenta tissue samples were hypermethylated instead of partially methylated as in pure placenta tissue samples. Also, the DNA methylation of placenta-specific imprinted genes was decreased in mixed placenta samples compared to pure placenta tissue samples. Furthermore, compared with other sample quality control

methods found in 'ewastools', our method more accurately classified pure and mixed placenta tissue samples. Mixed samples, often from early gestational ages, can lead to inaccurate DNA methylation profiles in the placenta and we demonstrate the value in implementing sample quality control methods to every DNA methylation study using Illumina arrays.

Investigation of de novo mutations in human genomes using whole genome sequencing datasets

Anushi Shah
UNSW

De novo mutations (DNMs) are genetic alterations occurring for the first time in a family member, which could be germline or somatic. DNMs have been shown to be a major cause of severe developmental genetic disorders. With the advent of next generation sequencing technologies, accurately detecting DNMs is crucial. A number of de novo variant callers to call DNMs from whole genome sequencing(WGS) data have been developed that differ in algorithms, filtering strategies and output. However, there is no study which has systematically evaluated these tools. We evaluated four DNM calling tools TrioDenovo, PhaseByTransition, DenovoGear and VarScan with regards to their concordance and accuracy in calling DNMs. Validation gold standard dataset consists of Illumina Hiseq WGS data of one CEU trio from 1000 Genomes Project. We also performed evaluation using simulated trio WGS datasets spiked-in with known DNMs for which we independently developed ‚ÄövÑvʃTrioSim‚ÄövÑvπ, an automated framework to generate simulated genomic datasets for trios. Our analysis on CEU 1000G trio dataset shows 3.5% DNM concordance amongst 4 DNM callers, while 8.8% to 33.5% of DNMs were called as unique to each caller. Of these, only between 3.7% to 9.9% of calls were real when compared to 1001 known DNMs confirmed in CEU trio 1000G dataset. Our analysis on simulated trio dataset spiked-in with 100 DNMs show 1.9% concordance while 0.6% to 66.9% DNM calls were unique to each caller. This shows large false positives detected by these tools and stringent post-filtering is required to obtain high confidence DNMs.

CPOP: Cross-Platform Omics Prediction procedure enables precision medicine

Kevin Wang
The University of Sydney

Risk prediction models separately constructed on two independent omics data for the same biological question are typically not "transferable" as the two independent omics data rarely share the same scale. That is, due to variations in statistical distribution between omics data, models tend to have poor prediction power across datasets and platforms. There are two classical approaches to ensure transferability across different omics data: data normalisation and refitting model parameters for additional data. However, both of these are not practical in clinical implementations when both the data and model are placed in lock-down. Such a situation is particularly common in a prospective testing experiment where there is an additional challenge that prediction must be made on individual samples in isolation from a large study cohort. To this end, we propose a new procedure, Cross-Platform Omics Prediction (CPOP) for building clinically implementable models using omics data. CPOP first resolves between-data scaling difference through a feature engineering step that conforms to abovementioned clinical restrictions. Then, CPOP preferentially selects the most statistically stable features across different omics datasets with an additional step that also stabilises model estimates. Application of our methods on four melanoma datasets demonstrates that CPOP can select features that are stable across multiple data without model re-training. Furthermore, we will also illustrate the practicality of CPOP in prospective experiment sample testing

without re-normalisation of additional data. Together, we show that CPOP can construct reproducible models that ultimately strengthens precision medicine research.

Opportunities and challenges of analysing multi-regional tumour biopsies to characterise heterogeneity in cancer

Sebastian Hollizeck
Peter MacCallum Cancer Center

Even though tumour heterogeneity is a widely accepted fact, the possibilities to study this phenomenon and its impact on the treatment of patients are limited. The CASCADE program enables rapid autopsies after death of the patient and therefore allows unique insights into the clonality and emergent resistance mechanisms of the different metastasis. This however creates a set of new bioinformatics challenges to manage the amount of data available for each of the patients and the combined analysis of this data spanned by the different samples. In our current work we explore variant calling capabilities of different methods in a multi-tumour-matched-normal sample scenario, to allow the reconstruction of evolutionary trajectories of all the tumour sites in the metastatic process. In this analysis we have utilised multi-regional tumour samples from 5 patients with advanced non-small cell lung cancer (n=4 EGFR mutant and 1 EGFR non-mutant) who underwent rapid autopsy through the CASCADE program. An average of 7 samples were analysed per patient by either whole exome or whole genome sequencing. To ensure high confidence variants we have used a consensus method of three variant callers. First an adapted version of the somatic variant calling with Freebayes from the BCBioinformatics pipeline, second our own developed 2-pass variant calling workflow with Strelka2 and lastly the newly developed joint calling capabilities of Mutect2 from GATK. This work aims to develop improved approaches for sensitively characterising the diverse mutational processes governing treatment resistance in non-small cell lung cancer.

Human transposons are an abundant supply of transcription factor binding sites and promoter activities in breast cancer cell lines.

Jiayue-Clara Jiang
The University of Queensland

Transposons, a type of repetitive DNA, occupy approximately 45% of the human genome and were predominantly viewed as junk DNA since their discovery by Barbara McClintock. However, recent progress has revealed the extensive co-option of transposons for the transcriptional regulation of human genes. In particular, transposons can become epigenetically deregulated in epithelial cancers, and may act as promoters that activate oncogenes and contribute to tumorigenesis. By bioinformatic analyses of various publicly available sequencing data, we investigated transposon-derived oncogenic transcription factor (TF) binding sites (TFBS) in breast cancer (BC) cell lines, predicted transposon-derived promoters, and validated these predictions using wet-lab approaches. Our results demonstrated that transposons were an abundant source of TF binding sites in BC, where ~38% of all TFBSs of MYC, E2F1 and C/EBP≈í‚â§ were harboured by transposons. We identified 268 transposon subfamilies as significantly enriched in the oncogenic TFBSs, suggesting a widespread role of these subfamilies in cancer transcriptional networks. TF-bound transposons were also associated with active histone modifications, further supporting their regulatory activity in BC cell lines. Luciferase assays in triple negative BC cell lines revealed that the promoter activities of SYT1, UCA1, AK4 and PSAT1 oncogenes were significantly reduced, and in some cases, almost completely abolished

following transposon deletion. Transposons with strong promoter activity were also found to be epigenetically deregulated in BC, characterised by hypomethylation and/or increased DNAse sensitivity. Our results provide an insight into the contribution of transposons to BC transcriptional regulation, and may facilitate the development of novel diagnostic biomarkers for BC.

A 9-gene score for predicting B-ALL risk of relapse and survival

Feng Yan
Monash University

Background: B cell acute lymphoblastic leukemia (B-ALL) is the most common malignancy in children. Although overall survival (OS) for pediatric B-ALL is around 85% at 5 years, ultimately more than 20% of affected children succumb to relapse. Leukemia stem cells (LSCs) which cause relapse and chemo-resistance are believed to relate to patient prognosis. We aim to use LSC gene signature to develop an easy way to predict patient risk at diagnosis.
Methods: Publicly available LSC RNA-seq for B-ALL was obtained from GEO and analysed using RNAsik and edgeR. Training data was obtained directly from TARGET website. Test datasets were microarray from GEO and TARGET website. LASSO regression was done on training data with 10 folds cross validation (CV). CV was performed 100 times randomly to select top 3 models with most occurrence. All models generated were then tested in all three test datasets for validation based on hazard ratio and p value. Survival analysis was based on Cox Proportional hazard model. Result: Differentially expressed genes from LSCs were enriched in pathways related to immune response and cell cycle arrest. Genes upregulated in LSCs with significant adverse survival impact were selected for LASSO regression. The final model is 0.065*S100A10+0.051*ZMAT3+0.017*PSAT1+0.108*RIMS3+0.01*LRRC25+0.015*H1FX+0.04*TSPO+0.029*NID2+0.014*CCDC69. It was validated in training data and all three test data including 2 paediatric and 1 adult B-ALL from different platforms. Moreover, it not only worked in full dataset, but also in a subset of patients with uninformative cytogenetics.
Conclusion: We are able to develop a 9-gene score to fast calculate the risk of patient. The score is agnostic to platform, patient age and uninformative cytogenetics.

Orthogonal evidence for Olfactory Receptors can be used for agonist prediction

Amara Jabeen
Macquarie university

Proteins are biological macromolecules critical for structure, function, and regulation of human cells and tissues. The human genome draft is available since 2003 but until now not all the coding genes have known protein products. Human Proteome Project (HPP), launched in 2010 with the aim of mapping entire human proteome. The HPP community has identified 88.62% of the coding genes as protein products. The remaining 11.38% are missing. Since most of the missing proteins are membrane proteins which might have clinical implications, therefore it is important to identify these proteins for utilizing their therapeutical potential. Several technical challenges make the missing protein (MP) identification through mass spectrometry (MS) difficult. Olfactory receptors (ORs), the superfamily of G-protein coupled receptors (GPCRs) are the largest MPs family. There is no convincing MS evidence available for even single OR. Four of the ORs are given the protein status based on orthogonal evidence. Therefore, we collated the available orthogonal evidence for ORs from published literature. Particularly, available ligand evidence can be used for novel agonist prediction. We have applied different classical ML and

deep learning methods to an ectopic OR, with a broad ligand spectrum. OR1G1 (UniProt ID: P47890) is ectopically expressed in gut enterochromaffin cells (normal and tumors) where it is known to be responsible for serotonin release. Based on classifier performance, we applied the naive Bayes classifier to a large test dataset, resulting in high probability predictions. Such an approach will assist in collecting experimental evidence for the missing olfactory proteome.

sPLSDA-batch: Batch effect correction in microbiome data

Yiwen Wang
University of Melbourne

Microbial communities have been increasingly studied in recent years to investigate their role in ecological habitats, as these little creatures can facilitate people to live an easy life. However, microbiome studies are difficult to reproduce or replicate as they may suffer from different sources of batch effects that are unavoidable in practice. These batch effects may confound the effects of interest. Batch effect correction is challenging in microbiome data, as these data have some inherent characteristics, such as sparsity, overdispersion, correlational dependency among variables and compositional nature. Traditional statistical methods used in microarray or RNA-seq data for batch effect removal are therefore not suitable for microbiome data. We propose a novel method called sPLSDA-batch. It is a promising alternative for batch effect correction in microbiome data compared with other methods such as ComBat, batch mean centering and removeBatchEffect (from LIMMA package), as this method is multivariate that can account for the correlation structure in the data and is non-parametric that can handle the skewed distribution caused by sparsity and overdispersion.

Comparison of algorithms for mtDNA variant discovery from whole-genome sequencing data

Eddie K.K. Ip
Victor Chang Cardiac Research Institute

Mitochondrial DNA (mtDNA) is maternally inherited and exists in 100-10000 copies in a cell. Two types of mtDNA variation exist, heteroplasmy, where the alternate allele is shared across all copies of mtDNA and homoplasmy, where the alternate allele is shared only across some copies of mtDNA. Mutations in mtDNA have contributed to human disease across a range of severity, from rare, highly penetrant mutations causal for monogenic disorders to mutations with milder contributions to phenotypes. Whole-genome sequencing (WGS) has allowed the investigation of mtDNA variation. However, a comprehensive evaluation of current mtDNA algorithms for WGS data has not been performed. We compared three state-of-the-art mtDNA algorithms for WGS data: MitoSeek, MitoCaller and mtDNA-server. Both heteroplastic and homoplastic variants were assessed in 97 families with congenital heart disease (CHD) that have been WGS. Accuracy was assessed by determining the maternal inheritance percentage of homoplastic mtDNA variants. The concordance of all mtDNA variants across the methods was also calculated. The homoplastic mtDNA variants identified by mtDNA-server had a maternal inheritance rate of 88%. This was 11% higher compared to MitoCaller. mtDNA-server also called 33% more homoplastic variants than the other algorithms. The concordance rate of heteroplasmies was low overall, ranging from 0.6% to 31% of variants by mtDNA-server. MitoSeek failed to call most heteroplastic variants using default parameters. In our assessment, mtDNA-server, has shown greater accuracy in calling mtDNA variants compared to the other two tools. Our study provides a guidance for best practices in evaluating mtDNA variation from WGS data.

Human Brain eQTL: Analysis and Insights

Letitia Sng
The University of Sydney

Expression quantitative trait loci (eQTL) analysis has been applied at the genome-wide level of the human brain to understand the underlying transcriptomic mechanisms relating to neurodegenerative diseases. In comparison to previous studies that have focussed on cis-acting eQTL at the transcript level, the current study investigates eQTL at the transcript/exon-level and cis vs trans-acting eQTL across ten regions of the human brain from the publicly available United Kingdom Brain Expression Consortium (UKBEC) dataset. We have found that when an eQTL is acting across multiple regions (MR-eQTL), it tends to be cis-acting and have very similar effects on gene expression in each of these regions. This indicates that MR-eQTLs tend to have an impact on genes which have important functions for the brain as a whole. Conversely, eQTLs specific to a single region (SR-eQTL) tend to be trans-acting with the cerebellum having six times more SR-eQTLs compared with other regions. This suggests that there are complex and unique systems of interaction between genes that regulate activity within specific brain regions. Furthermore, a large number of trans-acting eQTL were found, and overall their effect sizes were, on average, larger than those for cis-acting eQTLs at both transcript and exon-levels. These effect sizes differed between chromosomes and brain regions. Through our in-depth analysis, we have gained more insights into the patterns of genome-wide eQTLs in the human brain, especially in terms of trans-acting eQTLs and multi-region patterns.

Strain level genotyping of microbial communities using long and short read phasing

Rhys Newell
The University of Queensland

Detailed analysis of the microbial communities that drive the Earth,ÄövÑv¥s nutrient cycles and underpin human health can be carried out through the recovery of metagenome assembled genomes (MAGs). Hybrid genome assemblers combine the accuracy of short read sequences with long reads capable of resolving longer repeat regions. Methods for creating hybrid assembly MAGs are still under active development but can often be used to recover high quality species level genomes. These MAGs represent consensus from a population of related lineages. Even long-read sequences cannot bridge regions identical between highly similar microbial strains, since these areas of commonality may be arbitrarily long. Evidence to suggest that specific microbial strains can perform differentiating functions within a community come from a variety of sources including medical microbiology (e.g. strain-specific drug resistance phenotypes), culturing, 16s rRNA sequence composition, genomic variation analyses at high resolution, and single-cell genomics. Here we present a method to resolve genotypes based on finding single nucleotide polymorphisms (SNPs) and other structural variants within hybrid assembly MAGs. Variations are grouped together to form strain level genotypes using their co-variation in abundance across related microbial communities, and by observing their co-occurrence within individual short and long reads. We show the utility of this general method through its application to microbial communities derived from a climate-critical thawing permafrost environment.

The dynamic genome behind the emergence of octopod novelties

Brooke Whitelaw
James Cook University

Cephalopods (octopus, squids and cuttlefish) are characterized by many organismal novelties, such as prehensile limbs lined with chemosensory suckers along with the largest neuronal system among invertebrates. To reveal the genomic correlates of organismal novelties, we conducted a comparative study of three octopod genomes Hapalochlaena maculosa, Octopus bimaculoides and Callistoctopus minor. We present the first genome of a blue ringed octopus, the southern blue-ringed octopus (Hapalochlaena maculosa). This is the only known octopod genus to store large quantities of the potent neurotoxin tetrodotoxin (TTX) within their tissues and venom gland. We reveal highly dynamic genome evolution at both non-coding and coding organizational levels. We demonstrate expansions of zinc finger and cadherin gene families associated with neural functions/tissues in both H. maculosa and C. minor are congruent with the previously observations in O. bimaculoides, suggesting an octopod specific trait. Similarly, transposable element families LINE and SINE are consistent with an octopod specific expansion. Examination of tissue specific genes in the posterior salivary/venom gland (PSG) revealed the putative venom proteins, serine proteases dominate expression in O. bimaculoides and C. minor, while representing a minor component in H. maculosa. Voltage-gated sodium channels (Nav) in H. maculosa contain a resistance mutation previously documented in pufferfish and garter snakes to confer 15-fold resistance to TTX. No known resistance mutations were identified in either O. bimaculoides or C. minor. Analysis of the PSG microbiome revealed a diverse array of bacterial species present including genera of which some species can produce TTX.

Bioinformatic analysis of genome-wide SNPs elucidates cryptic species boundaries and potential speciation of Australian toadlets (Myobatrachidae: Uperoleia)

Frederick Jaya
UTS

The Austro-Papuan Uperoleia are a morphologically cryptic genus of frogs. The closely related U. borealis, U. crassa and U. inundata represent a species complex spanning the Top End and Kimberley in Northern Australia. Typically, Uperoleia species are delimited using genetic or acoustic data. However, relationships within the species complex were unable to be resolved using Sanger-sequencing data. Here we investigated the species boundaries, population structuring and interspecies hybridisation within the species complex. Bayesian inference and maximum likelihood estimation tools were used to analyse genome-wide, high-throughput Diversity Array Technology markers (silicoDArT and SNPs) and three mtDNA genes. Utilising population genetic tools (STRUCTURE, NewHybrids) and phylogenetic tools (IQ-TREE, BEAST), we delineate the species complex into two major groups, between U. borealis and U. crassa, with strong support to synonymise U. inundata into U. crassa. F1 hybrids between the two groups were identified with no evidence for backcrossing between hybrids and assigned parental species. We find variation in mtDNA population structuring and acoustic mating calls, which are congruent with the two-group division and interspecies hybridisation. Importantly, we find support for potential speciation, where acoustic mating calls are more divergent where species are sympatric.

Predicting the functional effect of genetic changes using variant effect prediction algorithms

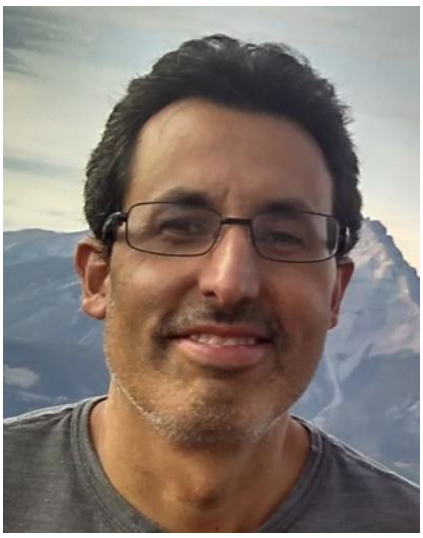
Emma Darling
The University of Melbourne

An enormous number of genetic variants have been reported and the collection continues to grow at a rapid rate. However, the functional consequences of all but a small fraction of observed variants are unknown. In large part due to the challenges associated with developing and conducting functional assays to generate supporting data. This absence of data has led to the development of algorithms for predicting the effect of variants on normal protein function. In turn, allowing inference of the variant, influence on disease development or progression. Predicting the functional effect of variants allows scientists and clinicians to focus on a small subset of variants that present as influential candidates for a given phenotype rather than expending time, effort and money investigating those unlikely to have a phenotypic consequence. The most widely used current prediction methods extract predictive gene-related features from curated genomic datasets. These features are used to train statistical models to predict variant functional effects. However, a major drawback with current methods is that their self-reported prediction accuracies are inflated, due to biases and lack of independence in input data, as discussed in multiple studies. A parallel track of research aims to predict protein function using protein-structure data. However, current algorithms are constrained in applications of variant prediction by their lack of consideration of population genomic data. This work aims to bridge this gap in the research by using orthogonal predictive features and new datasets, to build a new tool for variant effect prediction.

ISCB REGIONAL Student GROUP SYMPOSIUM 2019
COMBINE
Australia
Sydney, 9 December &
Asia & Australia's conference for students in bioinformatics and computational biology

Career Panel Profile

Professor Rafael Irizarry



Rafael Irizarry is a Professor of Applied Statistics at Harvard and Chair of the Department of Data Sciences at Dana-Farber Cancer Institute. His research focuses on Genomics and he teaches several Data Science courses.

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

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

Professor Irizarry has developed several online courses on data analysis that are offered by HarvardX and which have been completed by thousands of students. These courses are divided into two series: Data Analysis for the Life Sciences and Genomics Data Analysis. Much of the material is included in a book with an online version available for free.

ISCB REGIONAL Student GROUP SYMPOSIUM 2019
COMBINE
Australia
Sydney, 9 December &
Asia & Australia's conference for students in bioinformatics and computational biology

Dr Denis Bauer

Dr Denis Bauer is an internationally recognised expert in machine learning, specifically in processing big genomic data to help unlock the secrets in human DNA. Her achievements include developing an open-source, artificial intelligence-based cloud-service that accelerates disease research and contributing to national and international initiatives for genomic medicine funded with over $500M.

As CSIRO's transformational bioinformatics leader, Denis is frequently invited as a keynote at the international medical and IT conferences including Amazon Web Services Summit 2019, International conference on Frontotemporal Dementia '18, Alibaba Infinity Singapore '18 and Open Data Science Conference India '18. Her revolutionary achievements have been featured in international press such as GenomeWeb, ZDNet, Computer World, CIO Magazine, the AWS Jeff Barr blog, and was in ComputerWeekly's Top 10 IT stories of 2017.

Denis holds a BSc from Germany and PhD in Bioinformatics from the University of Queensland, and has completed postdoctoral research in both biological machine learning and high-throughput genetics. She has 38 peer-reviewed publications (18 as first or senior author), with over 1000 citations and an H-index 16.

Denis advocates for gender equality in IT, and is active on CSIRO's Inclusion and Diversity committee.

ISCB REGIONAL Student GROUP SYMPOSIUM 2019
COMBINE Australia
Sydney, 9 December &
Asia & Australia's conference for students in bioinformatics and computational biology

Professor Marcel Dinger

Marcel Dinger is Professor and Head of School for Biotechnology and Biomedical Sciences at UNSW Sydney.

Professor Dinger has more than 20 years experience in genomics as both an academic and entrepreneur. He was the founding CEO of Genome.One, one of the first companies in the world to provide clinical whole genome sequencing services. He is also founder of two successful IT companies and is the recent co-founder of a precision healthcare start-up, Pryzm Health. Marcel was the inaugural Head of the Kinghorn Centre for Clinical Genomics (KCCG) at the Garvan Institute of Medical Research from 2012-2018.

Professor Dinger's research laboratory seeks to establish new links between phenotype and genotype, particularly between rare and complex disease and underexplored regions of the genome, such as pseudogenes, repetitive elements, and those folding into non-canonical DNA structures or are transcribed into noncoding RNAs. Harnessing the potential of population scale genomic datasets, and sophisticated data science methods, the laboratory aims to bring an objective perspective to better understand how the genome stores information and how it is transacted in biology.

Professor Dinger graduated with a PhD in Biochemistry and Molecular Biology from the University of Waikato, New Zealand in 2003. Attracting more than 15,000 citations, he has (co)-authored more than 120 papers, many of which appear in the most high profile journals in the life sciences. In 2016, Marcel was admitted as a Fellow into the Faculty of Sciences of the Royal College of Pathologists of Australasia and is a Graduate of the Australian Institute of Company Directors.

ISCB REGIONAL Student GROUP SYMPOSIUM 2019
COMBINE
Australia
Sydney, 9 December &
Asia & Australia's conference for students in bioinformatics and computational biology

Dr Emily Wong

Dr Emily Wong is a newly joined group leader at the Victor Chang Cardiac Research Institute, Sydney. Her lab is interested in the regulatory basis of cell and organism phenotype.

Emily has a degree BSc (Hons) at the University of New South Wales, majoring in Biology. After a Masters degree in Bioinformatics at the University of Sydney, Emily continued to undertake her PhD in comparative genomics in the Faculty of Veterinary Sciences. There she studied the evolution of mammalian immune gene families under the supervision of Professor Kathy Belov and Professor Tony Papenfuss.

After a brief period at the Institute for Molecular Biosciences, Emily was recruited to the laboratory of Paul Flicek at the European Bioinformatics Institute (EMBL-EBI), Cambridge, UK, where she studied mammalian regulatory genomics using mouse models in close collaboration with Duncan Odom at Cancer Research.

In 2019, Emily returned to Australia to take up an ARC DECRA Fellowship at the School of Biological Science at the University of Queensland to work in the laboratory of Prof. Bernie Degnan. Broadly, her work aims to understand the impact of non-coding variation in regulatory evolution using varied approaches. Emily has been supported by two EMBO Fellowships and a University of Queensland Fellowship.