ISCB REGIONAL Student GROUP Australia

SYMPOSIUM 2017

COMBINE

Adelaide, 13 November

COMBINE Symposium 2017 Booklet

South Australian Health and Medical Research Institute (SAHMRI)
Adelaide, Australia, 13th Nov, 2017

LongReach QFAB eRSA Advancing Research Innovation ROYAL SOCIETY OF SOUTH AUSTRALIA for the advancement of science

Foreword

The COMBINE Symposium Committee would like to welcome you to the COMBINE Symposium 2017. COMBINE is a student-run organisation of Australian researchers in bioinformatics, computational biology and related fields. We run various workshops, seminars and social events each year all over Australia with the goal of bringing students and early career researchers together for networking and professional development. We are the student subcommittee of The Australian Bioinformatics And Computational Biology Society (ABACBS) as well as the official International Society for Computational Biology (ISCB) Regional Student Group (RSG) for Australia.

It is our great pleasure to welcome you to this ever-growing event with this year marking the sixth annual COMBINE symposium. For the first time, the symposium is held in Adelaide, South Australia. Today we are hosted by the South Australian Health and Medical Research Institute (SAHMRI) in Adelaide's new medical research hub.

We would like to thank today's judges and panelists for offering their time. We would also like to thank our very generous sponsors QFAB, eResearchSA, RSSA and LongReach who made the event possible. Thank you also to ABACBS, and to the 2017 ABACBS sponsors.

And finally, we would like to thank you, the delegates, for joining us in Adelaide. We hope you find the presentations engaging, the panel session informative and the social night enjoyable.

Klay Saunders

Klay Saunders

Chair of the 2017 COMBINE Symposium Committee

Symposium Committee Members

Klay Saunders (Symposium Chair)	University of South Australia
Shani Amarasinghe	The University of Adelaide
Luis Arriola	The University of Adelaide
Jimmy Breen	The University of Adelaide
Nikeisha Caruana	La Trobe University
Helen Dockrell	Flinders University
Josie Hyde	The University of Adelaide
Chen Li	Monash University
Ning Liu	The University of Adelaide
Yan Ren	The University of Adelaide
Ayla van Loenen	The University of Adelaide
Luke Zappia	The University of Melbourne Murdoch Children's Research Institute
Yiwen Zhou	The University of Adelaide

Sponsors

We greatly appreciate our sponsors for their generous support to our symposium.

Silver Sponsors

LongReach

QFAB

Bronze Sponsors

eRSA
Advancing Research Innovation

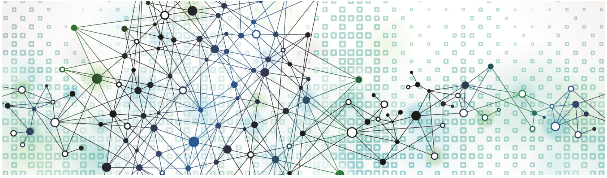
RSSA 1853
ROYAL SOCIETY OF
SOUTH AUSTRALIA
for the advancement of science

Website: https://www.longreachpb.com.au/

LongReach Plant Breeders is a commercial wheat breeding business established in 2002 to provide superior wheat varieties for Australian wheat growers and with the long term vision to be the leading wheat breeder in Australia. LongReach's national wheat breeding program is designed to deliver grain growers robust and high yielding varieties with attractive traits for all the key production zones of the Australian wheat belt. The aim is for good quality classification, high levels of disease resistance and high, stable yields which all leads to wide adaptation of our varieties.

QFAB Bioinformatics supports the production of high quality bioinformatics outcomes which deliver high impact publications and patents faster, through the provision of bioinformatics and biostatistics services for life science researchers to analyse and manage large-scale genomics, proteomics and clinical datasets. We have a strong track record of delivering results which are grounded in the team’s commitment to continuous learning and innovation, ensuring that our clients can trust in the knowledge and integrity of our support. We are the research data specialists: responsive, professional, secure and quality focused.

http://www.qfab.org/



Acknowledgment

We greatly appreciate the support from the following organisations, universities and companies.

AUSTRALIAN BIOINFORMATICS AND COMPUTATIONAL BIOLOGY SOCIETY

iSCB INTERNATIONAL SOCIETY FOR COMPUTATIONAL BIOLOGY

SAHMRI

Flinders UNIVERSITY

University of South Australia

illumina®

Centre for Cancer Biology

THE UNIVERSITY of ADELAIDE

EMBL Australia

Australian Genomics Health Alliance

PACIFIC BIOSCIENCES®

agrf

SHOTTESBROOKE

millennium science

scientifix

Stay Connected

Symposium Committee

Email: symposium@combine.org.au

Facebook:

https://www.facebook.com/combine.australia/

https://www.facebook.com/events/1458961404161146

Website: https://combine.org.au

Twitter: follow @combine_au and #COMBINE17

Wi-Fi

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

Laptop Charging

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

Symposium Programme

8:00 AM	Registration Open (SAHMRI Entrance Atrium - Ground Level)	
8:50 AM	Welcoming Address (SAHMRI Auditorium - Ground Level)	
Session 1 (SAHMRI Auditorium - Ground Level) **Chair: Ning Liu and Nikeisha Caruana**		
9:00 AM	Andrew H. Buultjens	Comparative genomics suggest *Mycobacterium ulcerans* migration and expansion is aligned with rise of *Buruli ulcer* in south-east Australia
9:15 AM	Jiayuan Huang	Comparative analysis of phosphoethanolamine transferases involved in polymyxin resistance across ten clinically relevant Gram-negative bacteria
9:30 AM	Dharmesh Bhuva	A dynamical systems simulator to evaluate methods for inferring co-expression networks
9:45 AM	Kirsti Paulsen	Optimising intrinsic protein disorder prediction for short linear motif discovery
10:00 AM	Andrew Pattison[1]	Predicting the outcome of breast cancer using novel RNA-Seq analysis **(Poster ID: 10)**
10:15 AM	Harriet Dashnow[1]	STRetch: detecting and discovering pathogenic short tandem repeat expansions **(Poster ID: 8)**
10:30 AM	Morning Tea (SAHMRI Entrance Atrium - Ground Level)	
Session 2 (SAHMRI Auditorium - Ground Level) **Chair: Yan Ren and Luis Arriola**		
11:00 AM	Gustave Severin	Multi-omic Characterisation of a Novel Xylose Metabolising Strain of *Saccharomyces cerevisiae*
11:15 AM	Eddie Ip[1]	VPOT: a customisable tool for the prioritisation of annotated variants **(Poster ID: 9)**
11:30 AM	Nhi Hin[1]	Transcriptomic and proteomic characterisation of a zebrafish model of familial Alzheimer's disease **(Poster ID: 12)**
11:45 AM	John Salamon	Visualisation and analysis of spatially-resolved transcript data using InsituNet

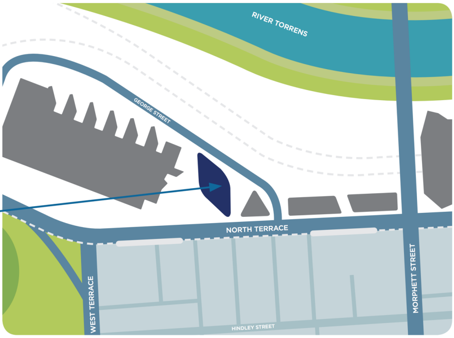
12:00 PM	Gabriel Foley[1]	SeqScrub: A web tool for automatic cleaning of FASTA file headers **(Poster ID: 11)**
12:15 PM	Fast Forward Talks[2] **Chair: Ayla van Loenen**	
12:35 PM	Lunch (SAHMRI Entrance Atrium - Ground Level)	
1:15 PM	Poster Session	
1:45 PM	Group Photo - Meet out front of SAHMRI	
Session 3 (SAHMRI Auditorium - Ground Level) **Chair: Luke Zappia and Helen Dockrell**		
2:00 PM	Pei Qin Ng[1]	Using genome-wide variants to determine the historical migration of chickens through South East Asia to the Pacific Islands **(Poster ID: 13)**
2:15 PM	Andrian Yang	Cloud-based single-cell transcript reconstruction using Falco
2:30 PM	Yuen Ting Wong[1]	Genome-wide study of 10,539 cancer samples reveals 27 novel associations between mutational processes and somatic driver mutations
2:45 PM	Wei Lu	Genome-wide SNPs modelling improved genetic risk prediction for psoriasis
3:00 PM	Greg Bass	Spatial statistics analysis of super-resolution protein co-localization data
3:15 PM	Career Panel and Afternoon Tea (SAHMRI Auditorium - Ground Level) **Chair: Jimmy Breen**	
5:00 PM	Awards and Closing Address (SAHMRI Auditorium - Ground Level)	
5:00 PM	ABACBS Registration (SAHMRI Entrance Atrium - Ground Level) and Dinner Break	
6:00 PM	ABACBS Keynote Lecture (SAHMRI Auditorium - Ground Level)	
7:15 PM	**COMBINE & ABACBS Social Events** COMBINE - The Edinburgh Castle ABACBS ECR - Duke of York ABACBS Professional - Cumberland Arms Hotel	

[1] These authors will also report the same studies in the poster session.

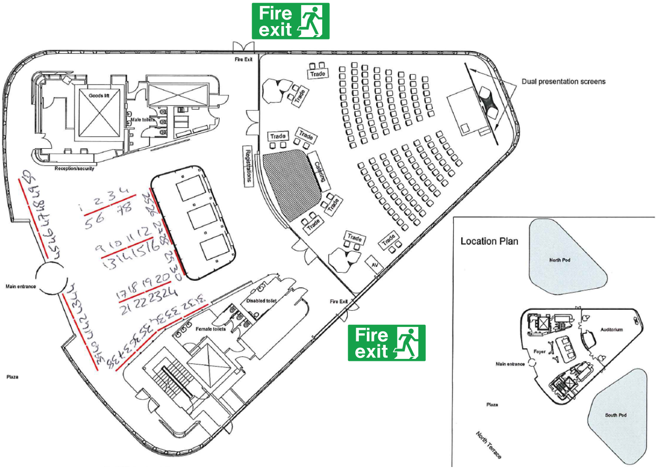
[2] Please refer to Page 28 for the topics in this session.

Symposium Venue Map

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



Venue layout



Career Panel Profiles

Dr Desmond Higgins
Desmond Higgins was educated at Trinity College, Dublin where he was awarded a PhD in 1988 for research on numerical taxonomy of Pterygote insects. He has an international reputation in bioinformatics as an innovator, a leader, and a practical provider of working solutions to key problems and is currently Professor of Bioinformatics at University College Dublin. Research in the Higgins laboratory focusses on developing new bioinformatics and statistical tools for evolutionary biology. The main focus is on the development and maintenance of the Clustal package for multiple sequence alignment. Originally written by Desmond in 1988, Clustal has gone one to be one of Nature's 10 most cited papers. Higgins also works on transcriptomics and proteomics data analysis and he is a PI in Systems Biology Ireland.

Dr Emily Hackett-Jones
Emily Hackett-Jones is an applied mathematician who has recently transferred into bioinformatics, working at the Centre for Cancer Biology, University of South Australia. Since completing her PhD in mathematical physics at the University of Durham, U.K., she has worked in diverse fields including string theory, ecology, mathematical biology and data science for digital marketing. She particularly enjoys working with experimentalists on problems that can be translated to mathematics. Currently she is working with the Goodall lab on micro RNAs, circular RNAs and the epithelial to mesenchymal transition, the process that drives cancer metastasis. Hackett-Jones beings a wealth of experience in academia, both here and overseas, and in industry.

Mr Jeremy Hack
Jeremy Hack is a Systems Administrator at eRSA working predominantly with Linux systems, High Performance Computing and the Nectar Research Cloud. His background in analytical chemistry, mass spectrometry and signal analysis lead him to Linux system administration and signal processing using open source software. Having experienced the challenges of "big-data" first hand, he's committed to helping users access the tools and resources they need in their computing environment. Jeremy is able to provide insight into the differences between academia and working in an academic support role.

Dr Maely Gauthier
Maely Gauthier is a genome bioinformatician working in diagnostics at SA Pathology. She completed a PhD and post-docs in marine and clinical genomics at The University of Queensland and the University of Zurich, contributing to multiple scientific publications. Maely has also worked briefly as a computational scientist for QFAB providing bioinformatics support to research groups, before taking her current position with SA Pathology. Her everyday job entails developing and maintaining diagnostic pipelines, and providing bioinformatics support to medical scientists, researchers and geneticists.

ABSTRACTS

Oral Presentations

Session 1-1

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

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

The University of Melbourne; Institute of Tropical Medicine; Victorian Infectious Diseases Reference Laboratory; Austin Health, Microbiology Diagnostic Unit

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

Session 1-2

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

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

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

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

Dharmesh Bhuva and Melissa Davis

The University of Melbourne; Walter and Eliza Hall Institute of Medical Research

Inferred gene regulatory networks can provide useful insight around genetic co-regulation during disease progression and such methods have identified novel pathological genes through ‘guilt by association’. Numerous methods are available to infer such networks, with most recent approaches attempting to infer context specific or differential co-expression networks. Due to the sparsity of known regulatory interactions, however, there is a need for simulated data to properly assess different methods, especially for the latest inference methods.

We have repurposed a simulator that uses models based on Boolean logic and systems of differential equations to simulate expression data. Activation signals are modelled by a single regulator using normalised Hill functions where the dissociation coefficients have been replaced with a more intuitive parameter that reflects the concentration required to achieve half maximal activation (EC50). The simulator provides five alternative classes of activation functions to choose from: linear, linear-like, sigmoidal, exponential and mixed types. Regulation of a target by multiple inputs is modelled using logic equations which specify the regulation mechanism using AND, OR and NOT functions. One major improvement over previous simulators is the simplicity with which a user can specify a model. Older simulators required users to specify a number of parameters which cannot be easily determined, such as dissociation rates for each interaction.

We believe that this simulator will enable more thorough evaluation of inference methods under varying conditions, and we are in the process of developing a BioConductor package using S4 classes for easy implementation by end users. The improved inference of regulatory networks in disease may ultimately have implications in drug regimen stratifications and improve our understanding of complex diseases.

Session 1-4

Optimising intrinsic protein disorder prediction for short linear motif discovery

Kirsti Paulsen, Sobia Idrees, Åsa Pérez-Bercoff and Richard Edwards
University of New South Wales

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

Session 1-5

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

Andrew Pattison, Paul Harrison and Traude Beilharz
Monash University

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

(Poster ID: 10)

Session 1-6

STRetch: detecting and discovering pathogenic short tandem repeat expansions

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

Murdoch Children's Research Institute; Broad Institute of MIT and Harvard; PathWest Laboratory Medicine, QEII Medical Centre; Royal Perth Hospital; University of Western Australia; Massachusetts General Hospital

Short tandem repeat (STR) expansions have been identified as the causal DNA mutation in dozens of Mendelian human diseases. Traditionally, pathogenic STR expansions could only be detected by single locus techniques, such as PCR and electrophoresis. These methods are expensive and most diagnostic tests only genotype the most common known events. In addition these methods do not scale to the whole genome and so cannot be used to identify new pathogenic STR loci.

The ability to genotype STRs directly from next-generation sequencing data has the potential to discover new causal STR loci and to reduce both the time and cost to reaching a diagnosis. Most existing tools for detecting STR variation are limited to repeat lengths that fit within the read length, and so are unable to detect the majority of pathogenic expansions.

We present STRetch, a new genome-wide method for detecting pathogenic STR expansions and estimate their approximate size directly from short read sequencing. STRetch takes the approach of adding STR decoy sequences to the reference genome prior to mapping reads. Reads mapping to the decoys are assigned back to their genomic position using read-pair information. Each locus is assessed for expansion using a statistical test based on coverage of the decoy chromosome.

We apply STRetch to the analysis of 97 whole genomes to reveal variation at known STR loci. We further demonstrate the application of STRetch to solve cases of patients with undiagnosed disease. A key advantage of STRetch over other STR detection tools is that it assesses expansions at all STR loci in the genome and so can be used to detect novel disease-causing STR loci.

STRetch is open source software, available from github.com/Oshlack/STRetch. The preprint can be found at http://biorxiv.org/content/early/2017/07/04/159228.abstract. **(Poster ID: 8)**

Multi-omic characterisation of a novel xylose metabolising strain of *Saccharomyces cerevisiae*

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

University of New South Wales; Microbiogen Pty Ltd

With growing demand for improved biofuel production the need for efficient conversion of xylose to ethanol is vital. Wild *Saccharomyces cerevisiae* (Baker's yeast) are commonly used in the production of biofuels, however they are unable to efficiently grow on xylose as a sole carbon source. Microbiogen Pty Ltd has evolved a novel xylose metabolising *S. cerevisiae* using a 15 year process of breeding and selection on xylose. To identify the genes that allow this strain to grow efficiently on xylose we have used a combination of PacBio whole genome sequencing, Illumina population resequencing, and RNA-Seq transcriptomics.

Two loci were determined to be under significant positive selection when grown on xylose minimal media in competition with the s288c (reference yeast) variants of these genes. Both loci contain unique variants at the genomic and protein coding level, compared to known yeast genomes. One gene was identified as a master regulator of transcription. The second gene was identified as a dehydrogenase. RNA-Seq analysis of our xylose-metabolising strain was used to identify genes with significantly increased expression on xylose versus glucose minimal media. This highlighted two further candidate genes, previously shown to substitute for the key xylose metabolic proteins xylose reductase (XYL1) and D-xylulokinase (XYL3). Combined with the previously mentioned dehydrogenase, these may explain a complete and novel xylose metabolic pathway. Future work will focus on the confirmation on the role of these genes and their requirement for efficient growth on xylose.

Session 2-2

VPOT: a customisable tool for the prioritisation of annotated variants

Eddie Ip, Sally Dunwoodie and Eleni Giannoulatou
Victor Chang Cardiac Research Institute

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

(Poster ID: 9)

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

Nhi Hin, Morgan Newman, Michael Lardelli and Stephen Pederson
The University of Adelaide

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

(Poster ID: 12)

Session 2-4

Visualisation and analysis of spatially-resolved transcript data using InsituNet

John Salamon, Xiaoyan Qian, Mats Nilsson and David Lynn
South Australian Health and Medical Research Institute; SciLifeLab

Gene expression studies typically homogenise samples before sequencing, discarding spatial information on where transcripts are expressed. In contrast to this, in situ sequencing is a novel technique for generating spatially-resolved, in situ RNA localization and expression data that preserves the spatial context of transcripts. Gene-specific barcodes allow data for up to 40 different transcripts/genes at an almost single-cell resolution to be generated in situ, resulting in images that display the location and intensity of a million or more individual transcripts in a tissue section. Despite the obvious potential of in situ sequencing, few methods currently exist to analyse and visualize the complex relationships that exist between these transcripts or identify how these transcriptional profiles change in different regions of the tissue or across different tissue sections. Here, I present InsituNet, an innovative new application that converts in situ sequencing data into interactive network-based visualisations, where each transcript is a node in the network and edges represent the spatial co-expression relationships between transcripts. InsituNet identifies co-expressions that occur between transcripts both significantly more, and less, than statistically expected given their relative frequencies within the tissue to allow intelligent filtering of the resulting visualisations. An automated sliding window function allows the generation of networks representing each individual section of the tissue and these networks enable users to quickly and easily identify regions where the transcriptional profiles are altered (e.g. regions associated with pathology). Alternatively, the user can also select irregularly-shaped regions of interest in the section for comparison to other regions. When multiple networks are constructed, their layouts may be spatially synchronised to facilitate comparison. Synchronisation allows one to easily observe how transcriptional relationships change across different tissue sections and conditions. InsituNet has been developed for the popular Cytoscape visualisation platform, and is available for download from within the Cytoscape app store.

Session 2-5

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

Gabriel Foley and Mikael Bodén

School of Chemistry and Molecular Biosciences, University of Queensland

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

(Poster ID: 11)

Session 3-1

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

Pei Qin Ng

The University of Adelaide

Chickens were domesticated from wild jungle fowls and were among the commensals transported during human migration eastwards across the Pacific Ocean. However, there is no detailed documentation of the exact origin of domesticated chickens, although several domestication centres have been suggested, based on archaeological and biomolecular evidence.

A previous study by Thomson *et al.* (2014) on chicken mitochondrial DNA suggested the possible origin and ancestry of modern Pacific chickens to be from the South East Asia jungle fowl. As mtDNA is maternally inherited as a single locus, this study could not consider complex evolutionary events such as introgression. To confirm previous findings and identify possible gene flow, three wild and five domestic samples of four different Gallus spp. were sequenced using whole genome sequencing and analysed using a bioinformatics workflow.

We confirm the hypothesis that the modern Pacific chickens originated from the Philippines *Gallus gallus* (red jungle fowl). Our results confirm the phylogeny of the wild species, with *Gallus varius* (green jungle fowl) basal to both *G. lafayettii* (Ceylon jungle fowl) and *G. sonneratii* (grey jungle fowl), consistent with findings of previous studies. Gene flow observed within the domestic chicken samples suggests a pattern of dispersal eastwards across the Pacific. Our analysis also suggests putative introgression of grey jungle fowl into domestic chickens. This could be explained through a recent common ancestor before the Pacific Island radiation that is not identified in this study.

Our findings elucidate a possible migration pathway of the chickens, which can be used to infer the potential route of human dispersal to the Pacific Islands and its impact on the genetic diversity of chickens as a commensal. Understanding of the phylogenetic relation and further investigation on the underlying genetic variations between the wild and domesticated samples can provide information on improving the commercial chicken breeds.

(Poster ID: 13)

Session 3-2

Cloud-based single-cell transcript reconstruction using Falco

Andrian Yang, Abhinav Kishore and Joshua Ho
Victor Chang Cardiac Research Institute; The University of New South Wales

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

Here we present a single-cell transcript reconstruction extension for the cloud-based Falco framework. The Falco framework enables highly parallellised processing of scRNA-seq data using big data technologies of Apache Hadoop and Apache Spark. The new transcript reconstruction pipeline allows for sharing of information across samples and are highly scalable to allow for efficient and timely analysis of large number of cells in scRNA-seq data.

Session 3-3

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

Yuen Ting Wong, Rebecca C. Poulos, Regina Ryan and Jason W. H. Wong
Prince of Wales Clinical School, Faculty of Medicine, University of New South Wales

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

Session 3-4

Genome-wide SNPs modelling improved genetic risk prediction for psoriasis

Wei Lu and Gad Abraham
The University of Melbourne

Psoriasis, a chronic skin disease affecting ~2% of Western populations, is caused by multiple genetic factors. Genome-wide association studies (GWAS) have identified multiple psoriasis susceptibility loci (single nucleotide polymorphisms, SNPs). A simple model that counts the number of significantly associated risk alleles in an individual has been used to predict disease genetic risk. However, this model suffers from poor prediction accuracy since it doesn't include all potential SNPs that are associated with the disease.

In this study, we applied penalised linear models to genome-wide SNPs for psoriasis genetic risk prediction. These models employ genome-wide SNPs and estimates their genetic effects simultaneously. To select the optimal model we compared different regression methods (ordinary least squares, least absolute shrinkage and selection operator (LASSO), Ridge) to estimate SNP effects. We also compared the prediction accuracy between models using genotyped SNPs and imputed SNPs. The prediction accuracy is measured by Pearson's correlation coefficient between observed disease risk and predicted genetic risk for case-control study.

We tested the models using three large psoriasis datasets (~2500 cases and controls each) for model selection and validation. We found that models including genome-wide SNPs led to increases of ~10% in prediction accuracy compared with models using SNPs from chromosome 6 only. We also found including imputed SNPs into model increased prediction accuracy by a small amount. It suggests that genetic risk prediction should include modelling genome-wide SNPs for complex diseases. Prediction accuracy could also be improved by using alternative SNP effect estimation methods. We also expect that the statistical models can help researchers and clinicians get a better understanding of genetic causes of Psoriasis.

Session 3-5

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

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

Systems Biology Laboratory, Department of Biomedical Engineering, University of Melbourne; Imaging Facility, Babraham Institute; Laboratory of Experimental Cardiology, Department of Cardiovascular Sciences, KULeuven; Cell Structure and Mechanobiology Group, Department of Biomedical Engineering, University of Melbourne; Systems Biology Laboratory, ARC Centre in Bio Nano Science and Technology, School of Mathematics and Statistics

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

ABSTRACTS

Fast Forward Talks

1

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

David Chisanga, Sushma Anand, Shivakumar Keerthikumar, Suresh Mathivanan and Naveen Chilamkurti
La Trobe University; Peter MacCallum Cancer Centre

Exosomes are small (30-150nm in diameter) membranous vesicles of endocytic origin. They have been implicated in a range of biological functions such as intercellular communication through the transmission of macromolecules such as proteins, nucleic acids and lipids, as well as in the pathogenesis and progression of diseases such as cancer. As such, there has been growing interest in understanding the biogenesis, functions, and applications of exosomes in both physiological and pathological conditions.

The biogenesis of exosomes has long been associated with the endosomal sorting complex required for transport (ESCRT) machinery together with other accessory proteins. However, the mechanisms behind exosome biogenesis are still poorly understood and the proteins involved in the process of exosome biogenesis have not all been characterised. Here we therefore, attempt to identify novel proteins that regulate the process of exosome biogenesis through the ESCRT pathway and improve our understanding of exosome biogenesis and exosomes in general. To achieve this, network analysis methods are applied to a Protein-Protein Interaction (PPI) network of the ESCRT machinery. To counter the bias that exists in PPIs due to false positives stemming from experimental errors in techniques used to identify them, we extended the network analysis method by using physical coherence, a technique that quantifies the connectedness of a PPI network due to topological changes. Using this technique, STAMBP and NEDD4 were identified as potential novel regulators of exosome biogenesis. It was found that STAMBP increased the physical coherence of the ESCRT machinery network while NEDD4 reduced the physical coherence of the ESCRT machinery network. To validate our findings, SDCBP, a protein that has been previously shown to regulate exosome biogenesis was also found to change the physical coherence of the ESCRT machinery. Further analysis using CRISPR-Cas9 based knockout cells of NEDD4 and STAMBP confirmed their active role in exosome biogenesis.

(Poster ID: 14)

2

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

Kristy Horan, Joseph Cursons, Momeneh Foroutan and Melissa Davis

The University of Melbourne; Walter and Eliza Hall Institute of Medical Research

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

(Poster ID: 15)

3

The causative role of serine and glycine on Macular Telangiectasia - a mendelian randomization approach

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

Walter and Eliza Hall Institute of Medical Research; MRC Epidemiology Unit, University of Cambridge, Cambridge, UK; Department of Research and Development, Moorfields Eye Hospital, London, United Kingdom; The Emmes Corporation, Rockville, Maryland, United States; The MacTel Consortium, The Lowy Medical Research Institute, Medical Retina Department, Moorfields Eye Hospital NHS Foundation Trust, London, UK; UCL Institute of Ophthalmology, University College London, London, UK

Macular telangiectasia type 2 (MacTel), is a rare and often underdiagnosed degenerative eye disease that may result in blindness. In 2017, we published the first five genetic loci involved in this disease discovered from genome-wide association study. Four out of the five loci were previously being connected with the glycine/serine metabolism pathway. In the same study, we identified glycine, serine and threonine to be the mostly differently abundant metabolites in MacTel. Here, we present the causative analyses on the role of these metabolites in MacTel disease.

Genetically Predicted Metabolites" (GPMs) constructed from SNP data can be used to test causality between metabolites and diseases. By analysing SNP data on 476 MacTel cases and 1733 controls we constructed GPMs for 140 metabolites. From our analysis, serine and glycine were the only two metabolites to have a causative role on MacTel. Genetically predicted serine showed a higher association with the disease (p=1.5E-31) when compared with glycine (p=3.6E-20). Although highly significant in the differential abundant analysis, threonine did not appear to be causally associated with the disease (p=0.902). Further, to assess their effect on disease progression, we performed a case-only analysis in 455 MacTel patients testing the association between GPMs and specific disease phenotypes. Genetically predicted serine was associated with a higher risk for the progression of various macular abnormalities, while glycine only had a small effect on one disease characteristic.

Our results confirm that serine and glycine have a causative role in the development and progression of MacTel disease. These results will help elucidate the disease mechanism in future work leading to better prognosis and future treatments for MacTel patients.

(Poster ID: 16)

4

Reference-free methods for genomic prediction and selection

Kevin Murray and Justin Borevitz
Australian National University

Genomic prediction uses knowledge of the population and family genetic relatedness to explain and predict variation underlying complex quantitative traits. This process has accelerated the breeding of crops and livestock as selection can occur on generally more accurate predicted phenotypes and can be extended to predictions on unobserved individuals.

Genomic prediction using gBLUP currently relies on relatedness data as determined from SNP genotypes mapped to a reference genome. This can induce bias, and precludes its use in non-model species, where reference genomes are either missing or poor. Additionally, incorporating the predictive power of the microbiomes associated with crops or livestock is arduous at best, requiring assembly of complex metagenomes.

We propose to capture this missing predictive power using a new method, which incorporates the weighted covariance between *k*-mer counts into traditional gBLUP-based genomic prediction approaches.

(Poster ID: 17)

5

Epigenetic differential DNA methylation analysis in monozygotic twins discordant for depression

Yan Ren, Jimmy Breen, Stephen Pederson and Sarah Cohen-Woods
The University of Adelaide Bioinformatics Hub; Robinson Research Institute; Flinders University

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

(Poster ID: 18)

Investigating computational analysis pipelines and genomic proximity interactions in T lymphocytes

Ning Liu, Timothy Sadlon, Stephen Pederson, Simon Barry and Jimmy Breen
The University of Adelaide

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

(Poster ID: 19)

7

Utilising mixture models for unveiling patterns in scRNA-Seq data

Yingxin Lin, Shila Ghazanfar, Pengyi Yang and Jean Yang
The University of Sydney

Single cell RNA-Sequencing (scRNA-Seq) has enabled unprecedented insight into the behaviour of individual cells on the scale of the entire transcriptome. Such precision offers an opportunity to explore cell-specific heterogeneity, however two distinct features arise from such data: (1) hyperinflation of identically zero counts for the majority of genes for any given cell, and (2) an apparent bimodal distribution of non-zero counts. Both features are unique to scRNA-Seq, and warrant further development of statistical tools in order to answer biological questions of interest.

We propose a mixture modelling framework to classify cells into three transcriptional states for each gene: (1) no, (2) low, and (3) high gene expression. This approach has the potential to reveal the cell-specific dynamics of RNA transcription (bursting) and degradation, as well as acting as a cross-dataset standardisation. We conducted a comparison of four particular models using either gamma-gamma or gamma-normal mixture models, and either performed independently across genes or constrained to ensure the first gamma component (lowly expressed) parameters are common across all genes. Comparison was conducted using metrics such as the Bayesian Information Criterion (BIC) to identify the most parsimonious mixture model type across all profiled genes. As a result, in addition to a standardised dataset, specific gene features can be obtained via the estimated parameters of each mixture model fit and used for further characterisation of genes, e.g. to identify especially highly or lowly variable genes.

We utilised a number of publicly available scRNA-Seq datasets, stemming from mouse neuronal cell populations, to perform the mixture model comparison, assess highly and lowly variable genes, and to estimate cell networks via a uniqueness thresholding.

(Poster ID: 20)

8

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

Andrew Lonsdale, Melissa Davis, Monika Doblin and Tony Bacic
ARC Centre of Excellence in Plant Cell Walls, School of BioSciences, The University of Melbourne;
Walter and Eliza Hall Institute of Medical Research

Leaderless secretory proteins (LSPs) are proteins that are secreted into the extracellular space yet lack the canonical N-terminal signal peptide sequence. The routes these proteins take are not fully understood, and it is an active area of research. Using proteomics to study LSPs in plant cells is further complicated by the open compartment nature of the cell surface, cell wall and apoplastic space. Typically a combination of destructive and non-destructive lab methods are used in the preparation stages in order to maximise the coverage of proteins. Both of these methods are thought to lead to the high number of potential LSP proteins found in plant cell wall proteomes.

Finding sequences like these in a proteomic study poses a challenge. They could be genuine LSPS. They could be intracellular contamination. Knowing which is which requires either (a) follow up experiments using biochemical approaches such as immuno-localisation, or (b) comparison to known LSPS. Since (a) takes time and money, there are very few (b) to compare to. Bioinformatics techniques to predict likely LSPs from candidates would be a good solution, but how do we build a prediction tool without positive sample data?

This work proposes using data from experimental observation, protein features relevant to the secretory environment, gene ontology terms and, protein interaction networks to distinguish likely candidate LSPs from contamination. Features associated with secretory and non-secretory proteins respectively are used classify potential LSPs and a create a reference set for further work in predicting leaderless secretory proteins in plants.

(Poster ID: 21)

9

SWATH-MS spectral reference library species conversion with the R package “dialects”

Madeleine J Otway, Peter G Hains and Phillip J Robinson
Children's Medical Research Institute

Mass spectrometry (MS) based proteomics is a methodology used to measure the relative abundance of proteins in biological samples. Proteins are extracted from tissue, enzymatically cleaved into peptides and sequenced by the fragmentation of selected peptides. “Shotgun” mass spectrometry (data-dependent acquisition; DDA) approaches produce biased results, where only the most abundant peptides are measured. SWATH-MS is a relatively new approach that measures all theoretical peptide fragments. The resulting file is extremely complex and identification of peptides relies on a separately generated spectral reference library (SRL). This is a table of peptide fragments, defined though a series of DDA-MS runs, which are used to search the SWATH-MS data.

Creation of an SRL is a lengthy process with large computational requirements. Re-searching of MS files with a species other than that of the original tissue is often avoided. This precludes the use of a large well characterised SRL designed in one species from being applied to SWATH-MS data generated in another species. To overcome this, we developed an R package named “dialects” (Data Independent Acquisition Library Editing to Convert The Species) for the conversion between species in an SRL.

This package has five core functions:
1. Import a UniProt formatted protein sequence database (fasta file)
2. Import a PeakView/OneOmics or OpenSWATH formatted SRL
3. Perform an in silico trypsin digestion on the proteins from the UniProt database
4. Swap the species of the SRL to that of the digested protein sequence database, only for peptides with full sequence homology occurs
5. Export the newly created SRL in either PeakView/OneOmics or OpenSWATH format

This package aids the creation of comprehensive SRLs, without the need for repeated MS runs and/or reprocessing of MS searches. This reduces the time to convert between species of a SRL and expands utility of large well characterised SRLs.
(Poster ID: 22)

ABSTRACTS

Poster Presentations

Poster ID: 1

Understanding the mechanism of action of in-feed antibiotics for chicken

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

Bioinformatics Institute A*STAR; BIOMIN Research Center; Biomin Holding GmbH; Singapore Bioinformatics Institute

Feed additives are products used in animal nutrition to improve the quality of feed and to improve the animal's performance and health. Antibiotics have been used since the mid 1940s, but the spread of antibiotic resistance in zoonotic bacteria poses a threat to health and hence have been banned in several countries. This has led to the need of developing viable alternatives to improve performance and protect animal health. To develop effective alternatives it is crucial to understand the mechanism of action of in-feed antibiotics.

In this project we used next generation sequencing and omics technologies to decipher the mechanism of action and to understand what signaling pathways are enriched on the usage of in-feed antibiotics.

For this purpose RNA from the mid-ileum tissue of chickens fed with no in-feed additives (control set) and with avilamycin (antibiotics treatment set) for 35 days, was extracted and used for RNA sequencing. Five biological replicates and two technical replicates from each set were used for RNAseq (Illumina HiSeq 4000, PE data, 28-50M reads, 150 bp). Comparison was carried out between the treatment set and control set to obtain the genes changing due to the antibiotic treatment. The pathways enriched with these differentially expressed genes were determined to understand the mechanism of action of antibiotics.

The number of differentially expressed genes was around 237. The most enriched pathways were related to inflammatory and immune responses, such as: IL-6, IL-10, IL-22 signaling, LXR/RXR, FXR/RXR activation, Communication between Innate and Adaptive Immune Cells, Production of Nitric Oxide and Reactive Oxygen Species in Macrophages, Th1 Pathway, Graft-versus-Host Disease Signaling.

Such kind of studies, will promote the development of safer alternatives to antibiotics. It would be interesting to study the differences in the mechanism of actions of alternative treatments to antibiotic treatment.

Poster ID: 2

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

Sarah Romanes and John Ormerod
The University of Sydney

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

Poster ID: 3

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

Kevin Wang, Jean Yang, Samuel Mueller and Garth Tarr

The University of Sydney

Gene annotation and pathway databases such as Gene Ontology Kyoto Encyclopedia of Genes and Genomes are important tools in Gene Set Test (GST) that describe gene biological functions and associated pathways. GST aims to establish an association relationship between a gene set of interest and an annotation. Importantly, GST tests for over-representation of genes in an annotation term. One implicit assumption of GST is that the gene expression platform captures the complete or a very large proportion of the genome. However, this assumption is neither satisfied for the increasingly popular boutique array nor the custom designed gene expression profiling platform. Specifically, conventional GST is no longer appropriate in this new context due to the gene set selection bias induced during the construction of these platforms.

We propose bcGST, a bias-corrected Gene Set Test method, by introducing bias correction terms in the contingency table needed for calculating the Fisher's Exact Test (FET). The adjustment method works by estimating the proportion of genes captured on the array with respect to the genome in order to assist filtration of annotation terms that would otherwise be falsely included or excluded. We illustrate the practicality of bcGST and its stability through multiple differential gene expression analyses in melanoma and TCGA cancer studies. The bcGST method is made available as a Shiny web application.

Poster ID: 4

MicroRNA regulatory networks in cancer progression

Holly Whitfield, Melissa Davis and Joseph Cursons
Walter and Eliza Hall Institute of Medical Research

MicroRNAs are small, endogenous, non-coding RNAs which participate in gene regulation through the repression of mRNAs. MicroRNAs (miRNAs) play a fundamental role in regulating both normal cellular development and in the progression of disease. They can exert control over these phenotypes by the coordinated effects of multiple miRNA which includes the additive effects of multiple miRNA co-targeting individual mRNA, as well as a single miRNA targeting multiple mRNAs. For example, the mutual repression between the miR-200 family and transcription factors ZEB1 and ZEB2 form regulatory feedback loops which are known to contribute to cancer progression. It is the complex interactions between cell constituents, such as miRNA and mRNA, which drive cellular function rather than any individual molecule. Hence, by capturing these topological features in networks allows for a systems-level exploration of complex biological systems, as well as a powerful visualization tool. These regulatory networks can be constructed either from observed experimental data, or through computational inference. Here, these approaches will be integrated to construct miRNA regulatory networks for the identification of novel regulatory interactions.

Networks will be constructed using information from both binding site prediction tools, and from TCGA breast cancer data. Using what is understood about miRNA:mRNA binding, databases such as TargetScan and DIANA-microT can predict putative relationships between miRNA and mRNA. The TCGA data will be used to establish associations between miRNA and mRNA, which when integrated with relationships from prediction databases will provide a framework for a tentative network. A revised network can then be permuted with a condition or drug of interest to explore the regulatory system. Network analysis methods will be used to identify novel putative relationships, and regulatory mechanisms which drive the progression of cancer.

Poster ID: 5

Precision medicine: a clinical perspective on genome data

Gulrez Chahal, Sonika Tyagi and Mirana Ramialison

Australian Regenerative Medicine Institute, Monash University Clayton VIC; SBI Australia, Monash Bioinformatics Platform, Monash University Clayton VIC

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

Poster ID: 6

Integrative analysis of lipid metabolic pathways in prostate cancer reveals DECR1 as a key cancer-related gene that promotes tumour cell survival

Chui Yan Mah, Max Moldovan, Zeyad Nassar, David Lynn and Lisa Butler
The University of Adelaide; South Australian Health and Medical Research Institute

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

Poster ID: 7

Computational prediction of serotonin distribution in the human and rodent colon

Helen Dockrell, Phil Dinning, Damien Keating and Lukasz Wiklendt

Flinders University

Approximately 95% of serotonin is produced in the intestines by enterochromaffin cells where it has been shown to modulate muscular contraction pattern and contraction force. Perturbation of intestinal muscle contraction and of serotonin concentrations underlie intestinal pathologies including Irritable Bowel Syndrome and Crohn’s Disease, where changes in total digestion time, gastrointestinal emptying and gastrointestinal motility are observed.

Serotonin concentration fluctuations are studied in relation to muscle contraction using ex vivo human and animal colonic samples. However, it has proved difficult to identify the complex relationship between these factors. This project collates the experimental data in a three-dimensional computational model of the intestines. This allows potential interactions between serotonin and muscle contraction to be explored and the physics constricting possible interactions to be applied consistently to inform hypothesis development.

Our model replicates a transverse segment of human colon ex vivo study containing a buffer solution in the lumenal space. All parameters applied to the model are experimentally determined, and are physiologically present. We have found that the addition of involutions in the mucosal epithelium, termed crypts, results in the formation of distinct serotonin concentration patterns in the mucosal pool with respect to the mucus and buffer pool concentrations ($P < 2.2e-16$, Man-Whitney-Wilcoxon test). The mucus sitting in the crypts acts as a well of serotonin, stabilising serotonin release from the mucus into the lumenal buffer solution. As a result, mucus and lumenal serotonin concentrations remain comparatively steady during muscle contraction, where serotonin concentrations in the mucosa change rapidly.

This correlates with ex vivo experimental findings as well as neural signalling theory, which together predict that serotonin concentrations remain high in the lumenal space, but fluctuate at lower concentrations within the mucosa in order to prevent neural desensitisation and communicate complex information to the nerves affecting muscle contraction.

NOTES

CONTACT US

Symposium Committee
Email: symposium@combine.org.au
Facebook: https://www.facebook.com/combine.australia/
https://www.facebook.com/events/1458961404161146
Website: https://combine.org.au
Twitter: follow @combine_au and #COMBINE17

ABACBS Student Sub-committee

SYMPOSIUM 2017 COMBINE
Adelaide, 13 November

ISCB REGIONAL Student GROUP Australia