Mammalian Systems biology: FANTOM5 promoters, enhancers and cell type specific regulation

Alistair Forrest [1][2]

[1] Harry Perkins Institute of Medical Research, University of Western Australia, Australia
[2] RIKEN Center for Life Science Technologies, Division of Genomic Technologies, Yokohama, Japan

We are complex multicellular organisms composed of hundreds of different cell types. The specialization of cell types and division of labour allows us to have coordinated complex functions such as responding to pathogens, movement and maintaining homeostasis. In the FANTOM5 project we have been interested in identifying the complete set of transcribed objects in the human genome and then predicting how they work together in the context of transcriptional regulatory networks (TRN). Each primary cell type runs a different version of the TRN based on the set of gene products it expresses. Not only this, but the FANTOM5 CAGE data reveal a wealth of cell-type-specific enhancers that are expressed in a very specific manner. Understanding the cell-type-specificity of these elements and promoters is key to building cell type specific TRNs. Lastly we go beyond the TRNs and examine cell-cell signaling within a multicellular organism. By identifying the sets of protein ligands and receptors expressed in any given human cell type we have made the first draft cell-cell communication network (CCCN) map.
Network Rewiring – analysing the variability, connectivity, and function of molecular interaction networks

Melissa Davis

Network analysis of molecular interactions (between proteins, transcription factors and their targets, or RNA molecules) usually exploits a canonical interactome assembled from a multitude of experiments and manually curated for maximum quality and coverage. However, advances in high throughput measurement technologies and increasing experimental resolution now enable us to capture biological context in these analyses. This work describes research into the mechanisms contributing to the rewiring of biological networks and illustrates the impact of these mechanisms on molecular interaction networks and their function.
Effectively using de novo assembled transcriptomes in non-model organisms and cancer

Alicia Oshlack

Next generation sequencing technologies have revolutionized gene expression studies by giving a comprehensive view of the RNA landscape without the need for prior understanding of genetic architecture. De novo transcriptome assembly is a powerful technique that enables full length gene sequences to be reconstructed from short read RNA sequencing (RNA-Seq). Many studies of the transcriptome, such as gene discovery, phylogenetics, differential gene expression and splicing analysis, may be performed using the assembled transcriptome as a reference. Because transcriptome assembly removes the need for a reference genome, it is routinely used for non-model organisms where an unreliable, or even no, reference genome is available.

Assembling transcripts is even more challenging than assembling genomes due to several factors including the enormous range in coverage created by differing expression levels and genes having multiple isoforms generated from the same genetics locus. Despite this, there are now several techniques that can effectively perform de novo assembly to produce a comprehensive transcriptional profile of a sample. However using these assemblies in downstream analysis is not always straightforward.

One common application of RNA-seq studies is to preform differential expression (DE) analysis between treatment groups to understand the functional consequences of a given perturbation. There are many methods dedicated to looking at this problem using a reference transcriptome but very few techniques for enabling DE using a de novo transcriptome assembly. We have developed a method called CORSET to allow for testing of differential expression in non-model organisms using data from de novo transcriptome assemblies. We show CORSET has excellent performance.

Another less explored opportunity of de novo transcriptome assembly is in tumours where the human reference genome is often a poor reflection of the true tumour genome. Here we have developed highly sensitive methods for using de novo transcriptomes to detect genomic rearrangements resulting in fusion genes. We show further work that demonstrates that analysis of assembled transcriptomes can identify important tumour specific variants.
Taking the meta out of metagenomics

Aaron Darling

Microorganisms play crucial roles in nearly every ecosystem, from sea to the soil. They inhabit the human body from the time of birth and appear to be essential for proper development of the immune system. Metagenomics and related high throughput DNA sequencing techniques have given us a first glimpse of our invisible and unculturable friends, leading to the recent discovery of dozens of new phyla. But metagenomic data has given rise to a number of unsolved computational challenges. One promising approach is to reduce metagenome analysis to another problem which is much better understood: isolate genome analysis. Doing so requires that we deconvolute metagenome sequences into individual genome sequences. To date however, it remains difficult to resolve the genomes of individual species or strains in a metagenomic sample. I will discuss emerging technologies that involve both computational and wet lab advances, that together will help take the meta out of metagenomics.
Deciphering root causes, driver mutations and therapeutic opportunities from Cancer Genomes.

Sean Grimmond

It is now well established that a cancer’s mutational burden drives tumour formation, influences disease progression and can dictate sensitivity to chemotherapy. Over the last 7 years, Professor Grimmond has pioneered cancer genome analysis of more than 1000 cancer patients, primarily as part of the International Cancer Genome Consortium (ICGC). These studies have sought to address 4 key questions: i) What are the root causes of somatic mutation that underlie each tumour and how do these evolve over time?, ii) What are the core pathways, that when mutated, lead to tumour formation & disease progression?, iii) How heterogeneous are pathologically uniform cohorts at a molecular level?, and iv) What insights can genome analysis provide that might improve patient outcome? A major focus of this effort has centred on Pancreatic Cancer. PC is currently the 5th most common cause of death in the USA and recent modelling suggests it will become the 2nd most common cause of death by 2030. There is an urgent need to better understand the aetiology and molecular pathology of this disease and gain a clearer understanding of therapeutic segments within this disease. A comprehensive integrated genomic analysis of >600 PC using a combination of Whole-Genome, Deep-Exome, whole transcriptome sequencing, and array based gene copy number, methylome and expression analysis has been performed. These data have been used to determine the mutational mechanisms and mutations causing for pancreatic carcinogenesis, identify distinct molecular subtypes that are clinically relevant, define the transcriptional networks and mutations driving theses networks and identify potentially druggable mutations in PC. These studies have also provided the foundations for large-scale cancer genome discovery (the ICGC’s Pan-cancer Analysis of Whole Genomes Project) and the pipelines built to capture patient relevant data provide an exemplar for scalable, genome-directed clinical trials in other recalcitrant cancers.
Cross-Species Systems Analysis of a Metabolic Disease

Jean Yang

Insulin resistance (IR) is perhaps the best predictor of future development of type 2 diabetes (T2D). Identification of differential profiles that can identify individuals with higher risk of developing T2D could enable early stage intervention. To reduce inherent variability that commonly associated with human expression data, we examine an approach that integrates human data with more stable data from a model organism. We will describe a three-level cross-species (human-mouse) analysis framework that examines information at single gene level, gene set level and network level. This cross species GE analysis identified an optimized GE motif (GEM) comprising 92 genes. This GEM was enriched in insulin signaling genes validating the approach. The GEM also contained genes related to beta Catenin and Jak1 function and a functional role for these pathways in muscle insulin action was validated using in-vitro studies. Together, these studies provide proof of principle evidence that integrated approaches can be used to acquire molecular information that is both highly diagnostic of human IR and that contains novel mechanistic information that contributes to our understanding of human IR.
Interrogating the genome-wide relationship between Replication Timing and DNA Methylation in Cancer

Qian Du [1], Saul Bert [2], Nicola Armstrong [3], Aaron Statham [2], Phuc Loi Luu [2], Liz Caldon [4], Clare Stirzaker [2], Susan Clark [2]

[1] 1 Epigenetics Laboratory, Genomics and Epigenetics Division, Garvan Institute of Medical Research, Darlinghurst, NSW, Australia [2] Epigenetics Laboratory, Genomics and Epigenetics Division, Garvan Institute of Medical Research, Darlinghurst, NSW, Australia [3] Mathematics and Statistics, School of Engineering and Information Technology, Murdoch University, Murdoch, Western Australia, Australia [4] Replication and Genome Stability, Cancer Division, Garvan Institute of Medical Research, Darlinghurst, NSW, Australia

DNA replication follows a highly organized ‘replication-timing’ program where genomic domains are replicated in a specific temporal order from early to late during S-phase. The transmission of epigenetic information is intimately associated with the replication fork. We hypothesise that changes to the replication-timing program during carcinogenesis is critically involved in disrupting the transmission of epigenetic information contributing to deregulation of the cancer epigenome. Here we investigate the relationship between cancer-related changes to both DNA methylation and the replication-timing program.

To determine the ‘replication-timing’ program, we performed Repli-Seq that utilizes whole-genome sequencing to map the order of newly-replicated DNA across the genome. By integrating Repli-Seq data with whole-genome bisulfite sequencing (WGBS) methylation data, we discovered that large domains of extensive hypomethylation in cancer are significantly localized to late-replicating regions of the genome.

To further interrogate this relationship, we modified Repli-Seq to create a novel technique RepliBis-seq. RepliBis-seq involves bisulfite treatment and whole-genome sequencing of nascent DNA to directly inform replication-coupled DNA methylation throughout replication-timing. Our preliminary data shows that maintenance methylation is less efficient at late-replicating DNA, thereby contributing to extensive hypomethylation. Moreover, to determine if hypomethylation drives a change in replication-timing, we performed Repli-Seq on HCT116 and DKO1 cells, a colorectal cancer cell model with knockouts of DNA methyltransferase proteins DNMT1 and DNMT3B. Importantly, we show that the changes in replication timing are commonly localized to regions of methylation change. Overall, our results show that replication-timing is an important mechanism associated with DNA methylation deregulation of the cancer genome.
Elucidating the mechanism of aberrant gene expression in MLL leukaemia

Enid Lam [1], Omer Gilan [1], Ester Cannizzaro [2], Meike Wiese [2], Chun Yew Fong [1], Sarah Ftouni [1], Dean Tyler [1], Kym Stanley [1], Margherita Ghisi [1], Marnie Blewitt [1], Andrew Bannister [2], Tony Kouzarides [2], Brian Huntly [2], Ricky Johnstone [1], Sarah-Jane Dawson [1], Mark Dawson [1]


MLL fusion protein (MLL-FP) leukaemia is a subset of acute leukaemia with a poor prognosis, caused by chromosomal translocations of the MLL gene. Central to the molecular pathogenesis of this disease is the abnormal co-optation of members of transcriptional complexes to drive abnormal gene expression. A number of targeted therapies against members of transcriptional complexes translocated to MLL have shown significant pre-clinical efficacy and are currently in clinical trials. Using a range of innovative proteomic methods we characterised several epigenetic proteins that are central to the pathogenesis of MLL leukaemia. These epigenetic proteins are part of separate transcriptional complexes and are amenable to therapeutic intervention. Treatment of MLL-FP leukaemia cell lines, primary human leukaemia cells and murine leukaemia models with small molecule inhibitors targeting these epigenetic proteins shows remarkable synergy in cell growth inhibition. To elucidate the molecular mechanisms by which these epigenetic regulators direct leukaemogenic transcription programs, we performed RNA-Seq and ChIP-Seq on a MLL-FP leukaemia cell line treated with small molecule inhibitors alone, or in combination. These data reveal a previously unrecognised functional collaboration between transcription complexes and provide novel insights into the regulation of malignant transcription programs. They also establish the molecular framework and rationale for combination epigenetic therapy in this poor risk disease.
Reconstructing signalling networks from dynamic phosphoproteomics data

Pengyi Yang

National Institutes of Health

Cell signalling controls various aspects of basic cellular processes including homeostasis, proliferation, and cell fate decisions, with defects underlying these processes associated with a wide range of diseases. Protein phosphorylation, characterized by the addition of a phosphate group by a protein kinase to a serine, threonine, or tyrosine residue on a substrate protein, has emerged as a key mechanism for signal transduction and integration. Recent advances in mass spectrometry-based technology makes it possible to profile proteome-wide phosphorylation events for investigating signal cascades. However, development of algorithms to analyse and identify signalling events from high-throughput phosphoproteomics data is still in its infancy. The first step toward reconstruction of signalling networks is to identify key kinases involved in signalling cascades. To this end, we have developed a knowledge-based CLUster Evaluation (CLUE) approach that utilizes known kinase-substrate annotations to identify kinases that are perturbed in the signalling cascades. We demonstrate the utility of the approach on two time-series phosphoproteomics datasets and identify kinases associated with embryonic stem cell differentiation and insulin stimulation. Another critical step for reconstructing signalling networks involves de novo substrate prediction for key kinases of interest. We have developed an ensemble approach based on positive unlabelled learning that integrates dynamic phosphoproteomics data with known kinase recognition motifs to predict novel substrates for a given kinase. We find that the integrated approach is more effective compared to using sequencing information alone. Together, these approaches will serve as a valuable resource for dissecting signalling cascades and making biological inferences from phosphoproteomics data.
brain-coX - a shining web-application to interrogate human brain co-expression networks

Saskia Freytag [1], Karen Oliver [1], Johann Gagnon-Bartsch [2], Terence Speed [3], Melanie Bahlo [4]


Gene prioritization tools can identify the most promising candidate genes generated from whole-exome or whole-genome sequencing. Having a list of credible candidates can speed up the process of finding the cause of a patient's disease. Unfortunately, results from these tools are not usually readily available to clinicians, as their generation and interpretation often requires trained Bioinformaticians. With the help of RStudio and R we have developed an interactive shiny tool - brain-coX - that allows clinicians to perform prioritization, and more, themselves. brain-coX is targeted towards neurogenetic diseases, overcoming the non-specificity of other prioritization tools. brain-coX combines these diverse datasets for downstream analysis. Gene names are automatically standardized and each dataset is cleaned with removal of unwanted variation taking the users' research interest into account. The application of this adaptive cleaning procedure has been shown to lead to greater consistency across different studies. In addition to gene prioritization brain-coX also offers network visualization, developmental analysis and correlation plots.
Unbiased detection of circular RNA in high throughput sequencing data

David Humphreys

Victor Chang Cardiac Research Institute

High throughput sequencing (HTS) has recently identified a new class of abundant transcripts called circular RNA (circRNA). Thousands have been identified from the junction of out of order mRNA exons (referred to as backsplice junctions). Interestingly many circRNAs have been identified to be evolutionary conserved and/or have tissue specific expression, suggesting these molecules may have biological function. In support of this there is a growing list of suggestive associations of circRNAs associated with disease or biological function.

To date the bioinformatics pipelines used in the literature to detect circRNA typically require the input of known exon junctions to identify candidates. Here I describe a bioinformatics pipeline that does not require the prior knowledge of exon junctions for the detection of circRNA transcripts. This pipeline uses the chimeric junction output of the STAR aligner and applying stringent filtering to identify candidate circRNAs. Subsequent annotation thereby identifies the origin of circRNAs. I have used these steps to analyse public RNA-Seq data from the brain, heart and liver. This unbiased pipeline identifies many circRNAs produced from canonical exon junctions and interestingly identifies some candidates that may not use canonical exon junctions.
Genome assembly in presence of its intimate associates - Leptospermum as a case study

Amali Thrimawithana, Ross Crowhurst, Elena Hilario, Bruce Smallfield, Cecilia Deng, Charles David, David Chagne, David Lewis, Ed Morgan, Helge Dzierzon, John Van Klink, Kevin Davies, Murray Boase, Nigel Perry, Ross Bicknell, Kathy Schwinn

The New Zealand Institute for Plant & Food Research Limited

Despite the significant advances in the field of genome assembly, the assembly of multi-genomic samples of higher organism systems poses a great challenge. The presence of parasitic and symbiotic counterparts in wild plants is common, and eradication of these additional taxa from plant tissue is not always viable. We have developed an approach to assemble the genome of Leptospermum scoparium, a native plant of eastern Australia and New Zealand. L. scoparium is commonly associated with one or more endophytic fungi and black sooty mould, because of the presence of scale insects. Our bioinformatic approach included the use of multiple classification workflows and mapping techniques to partition the data into different taxonomic levels, both at pre-assembly (read based partitioning) and post-assembly (scaffold based partitioning) stages; we compared these two methods. The tools used for read and scaffold classification included our in-house annotation system and the taxonomic classifiers Kraken and Blobology. In this presentation, we highlight how these tools were used within our pipeline to obtain a draft genome sequence for L. scoparium, and their strengths and weaknesses.
Disentangling Genetics and Methylation in the Major Histocompatibility Complex

Jovana Maksimovic, Damjan Vukcevic, Justine Ellis, Stephen Leslie, Alicia Oshlack

Murdoch Childrens Research Institute

Many GWAS of autoimmune diseases have identified SNP associations in the major histocompatibility complex (MHC) region, particularly within and proximal to the human leukocyte antigen (HLA) genes. Certain HLA alleles have also been shown to be associated with various autoimmune diseases, such as multiple sclerosis (MS). However, many individuals who harbour known risk alleles do not develop autoimmune disease, suggesting that the genetic risk is modulated by other factors.

Epigenetic marks such as methylation might play a role in modifying genetic risk; although, unlike DNA, methylation varies substantially between cell types. This is problematic if the affected tissue is not accessible or samples are from a mixed cell population, such as blood. To understand the effect of methylation on disease in a complex region such as the MHC, the relationship between the genetics, cell-type specific methylation and methylation differences between individuals need to be explored.

Using blood, cell-sorted 450k methylation data we show that within the MHC there is a subset of CpGs that are cell-type specific, whilst another subset varies amongst individuals but is relatively uniform between cell types, presumably due to different genetic backgrounds. A recently published set of MS-associated MHC CpGs falls almost exclusively in the subset of CpGs that vary between individuals in the cell-sorted data. We have obtained a cohort of individuals with both methylation array data and HLA alleles to elucidate whether methylation differences between individuals are completely associated with certain HLA haplotypes or whether there is within-haplotype variability that may moderate disease risk.
Differential Gene Expression Exploration using Degust

David Powell

Monash Bioinformatics Platform, Monash University, Victoria 3800, Australia

Degust is a web-based tool for exploring differential gene expression data. It was designed to be easily used by someone with little bioinformatics experience but has also proved very useful to experienced bioinformaticians. Degust consists of a backend that runs analyses using limma/voom or edgeR to perform the differential expression analysis; and of a front-end that is interactive, powerful and easily shared.

Degust includes an interactive MA-plot, and a parallel coordinates plot that is useful for showing gene expression changes over multiple conditions. An important advantage of Degust is that calculation is done within the browser to give fast feedback to the user based on their filtering. This includes an in-browser heatmap, and a novel interactive MDS plot that makes it possible to quickly explore the variation within an RNA-Seq experiment. This MDS plot is immediately recalculated when different subsets of genes are selected which rapidly allows a user to explore the genes that are driving the clustering seen in the plot. An interactive table of genes can be searched for a particular gene of interest, or downloaded for subsequent analysis.

Recent changes in Degust have added the ability to use limma with sample weights in the backend analysis, and to visualize particular gene sets and even kegg pathways.

Public server: http://www.vicbioinformatics.com/degust/
Degust homepage: http://victorian-bioinformatics-consortium.github.io/degust/
Statistical methods and software for functionally characterizing every single mutation in Your Favorite Gene

Alan Rubin [1], Stanley Fields [2], Terence P. Speed [3], Douglas M. Fowler [4]


Although high-throughput DNA sequencing has rapidly expanded catalogues of normal and disease-associated variation, the functional consequences of most mutations are unknown. In deep mutational scanning, selection for protein function applied to a library of protein variants is combined with high-throughput DNA sequencing, allowing direct measurement of the activity of hundreds of thousands of variants of the protein easily and cheaply. This approach helps to bridge the gap between variant identification and interpretation, enabling researchers to elucidate sequence-function relationships at high resolution. The resulting data can be used in a variety of contexts, from aiding the assessment of clinical variants to guiding protein engineering. Despite the growing popularity of deep mutational scanning, there are few formal statistical methods available to help analyze these complex datasets. Here we present a novel method for assigning functional scores and statistical significance to all variants in a deep mutational scanning dataset based on weighted regression. These methods are implemented as part of Enrich 2, a user-friendly software package that makes the initial data analysis accessible to experimental biologists while providing an extensible framework for bioinformaticians manipulating these large datasets. For illustration, we show the results of applying this method to deep mutational scans of diverse targets, including BRCA1 and the WW protein-binding domain of YAP65.
Efficient project management for better reproducible bioinformatics analysis

Cecilia Deng, Amali Thrimawithana, Helge Dzierzon, Eric Burgueno, Elena Hilario, Rebekah Mitter, Roy Storey

The New Zealand Institute for Plant & Food Research Limited

The Bioinformatics team at the New Zealand Institute for Plant & Food Research Limited (PFR) contributes to a complex assortment of projects for a wide range of species, including bacteria, fungi and insects, to plants. The next generation sequencing (NGS) data we process are exploding in volume, diverging in variety, and the rate of increase is accelerating. 'Omics studies are evolving from hypothesis-driven to data-driven approaches, and effective data analysis and interpretation are critical in NGS projects with big data. The complexity of the data types and the special analysis requests make project management and progress sharing difficult. Ad hoc practice without careful analysis planning, proper version tracking, and sufficient documentation, is not uncommon in bioinformatics. For best practice and better reproducible research, we developed a de facto standard protocol for bioinformatics analysis at PFR, including data provenance capture, resource management, workflow control, version control for reproducible analyses, to result transference to researchers, using in-house developed tools as well as open source software, such as GitHub and Waffle.io. This implementation enables effective project tracking and efficient reproducible research (RR).
XGSA: An unbiased statistical method to perform cross-species gene set analysis

Djordje Djordjevic [1], Kenro Kusumi [2], Joshua Ho [1]


Gene set analysis is a powerful tool for gaining biological insights from genome-wide gene expression data by comparing differentially regulated gene sets with gene set knowledge-bases, such as the gene ontology. Current gene set analysis methods are designed to compare gene sets that are from the same species, and do not facilitate comparing gene sets that are derived from different organisms. As a result, much of the knowledge that is gained in model organism studies is not easily transferred across the tree of life.

In this work, we show that failing to account for the complex homology structure when mapping between genes in two species will introduce a bias in subsequent gene set enrichment results. To overcome this bias, we develop a statistical approach that takes into consideration the homology mapping when testing the significance of enrichment between gene sets from different species. We show that our approach is significantly better than other current alternative approaches in terms of detection sensitivity and specificity.

Our method is implemented as an R package (XGSA) with a focus on simplicity for the end-user. By directly accessing homology information from the ENSEMBL database, our tool can easily integrate with gene expression analysis workflows in many model organisms. To demonstrate its real life applications, we use our tool to investigate conserved pathways that are involved in organ regeneration across diverse organisms such as zebrafish, axolotl and frogs.
Transcription and differential DNA repair underlies promoter mutation hotspots in cancer genomes

Rebecca Poulos

*University of New South Wales, NSW*

Promoters are DNA sequences that play an essential role in controlling gene expression. While recent whole cancer genome analyses have identified numerous hotspots of somatic mutations within promoters, most do not appear to be functional as they do not perturb gene expression. As such, positive selection does not adequately explain the frequency of promoter mutations in cancer genomes. We have found that increased mutation density at gene promoters is in fact linked to transcriptional activity and differential DNA repair. By analysing 1,163 cancer genomes, we find evidence for increased local density of somatic point mutations within the DNase I hypersensitive centre of gene promoters across 14 cancer types. Mutated promoters are strongly associated with transcriptional activity, with mutation density highest within transcription factor binding sites. By analysing genome-wide maps of nucleotide excision repair (NER), we find that NER is impaired within the DNase I hypersensitive centre of active gene promoters, inversely mirroring the increase in somatic mutation density. Thus, our analysis has uncovered the presence of a previously unknown mechanism linking transcription initiation and DNA repair, thereby implicating localised differential DNA repair as the underlying cause for the somatic mutation hotspots observed at gene promoters of cancer genomes.
**Identify functional patterns in high throughput binding assays**

Alexandra Essebier, Mikael Boden

*The University of Queensland*

Mapping transcription factor (TF) binding sites using high throughput assays such as chromatin immunoprecipitation (ChIP-seq) is essential to understanding the role TFs play in regulatory processes. ChIP-seq uses an antibody against the target TF to isolate bound regions of DNA and reports binding sites as read pile-ups or peaks. The interaction between TF and DNA can alter the distribution of reads or 'shape' of a peak. We hypothesise that ChIP-seq peaks contain identifiable patterns in read pile ups linked to different binding modes and functional patterns of TFs.

We developed a modelling technique to identify clusters based on the shapes of peaks across a ChIP-seq dataset. We investigated the significance of peak shape on TF binding location, epigenetic environment and sequence. We applied the clustering technique to ten different TF datasets performing statistical analyses to explore differences in biological features between clusters within each TF. We demonstrate that binding locations and epigenetic annotations differ significantly between clusters. We observe that clusters are enriched with different sequence motifs. Finally, we find that epigenetic environments for the binding locations of a TF are significantly altered between different cell types. Our model clusters ChIP-seq peaks based on read density or 'shape' then successfully links each cluster to unique biological features. The patterns we describe in ChIP-seq datasets have not previously been identified by available processing methods and can be linked to varied interactions between a TF and DNA.
FAST-FORWARD ABSTRACTS
Combining spatial and chemical information for clustering pharmacophores

Lingxiao Zhou, Renate Griffith, Bruno Gaeta

The University of New South Wales

Background
A pharmacophore model consists of a group of chemical features arranged in three-dimensional space that can be used to represent the biological activities of the described molecules. Clustering of molecular interactions of ligands on the basis of their pharmacophore similarity provides an approach for investigating how diverse ligands can bind to a specific receptor site or different receptor sites with similar or dissimilar binding affinities. However, efficient clustering of pharmacophore models in three-dimensional space is currently a challenge.

Results
We have developed a pharmacophore-assisted Iterative Closest Point (ICP) method that is able to group pharmacophores in a manner relevant to their biochemical properties, such as binding specificity etc. The implementation of the method takes pharmacophore files as input and produces distance matrices. The method integrates both alignment-dependent and alignment-independent concepts.

Conclusions
We apply our three-dimensional pharmacophore clustering method to two sets of experimental data, including 31 globulin-binding steroids and 4 groups of selected antibody-antigen complexes. Results are translated from distance matrices to Newick format and visualised using dendrograms. For the steroid dataset, the resulting classification of ligands shows good correspondence with existing classifications. For the antigen-antibody datasets, the classification of antigens reflects both antigen type and binding antibody. Overall the method runs quickly and accurately for classifying the data based on their binding affinities or antigens.
Circulating tumour DNA reflects clonal evolution in chronic lymphocytic leukemia


Cell free circulating tumour DNA (ctDNA) is an emerging biomarker in cancer management providing noninvasive opportunities for serial monitoring of tumour specific genomic alterations during disease progression and treatment. We evaluated the role of ctDNA to monitor clonal evolution in patients with chronic lymphocytic leukemia (CLL). Studies have reported the development of Richter's syndrome (RS) as a common method of relapse in patients receiving novel therapies for CLL. RS is a disease transformation into an aggressive diffuse large B cell lymphoma (DLBCL), which is often refractory to treatment and carries a poor prognosis. Prompt recognition of RS is clinically important, but currently there are no strategies that allow for early diagnosis and treatment. We studied 2 individuals who developed RS following treatment for relapsed/refractory CLL. To understand molecular events underpinning disease transformation, we performed whole exome sequencing (WES) on paired tumour tissue and plasma, at baseline and at the time of RS. Analysis of WES data including SNV and copy number analysis, revealed biologically distinct modes of transformation in each of the 2 cases; (i) CLL transformation to a clonally related DLBCL and (ii) the development DLBCL unrelated to the dominant CLL clone. In each case, quantification of allele fractions of selected mutations in plasma, through targeted deep sequencing, identified the early emergence of mutations associated with disease transformation, including key mutations in TP53. These findings reveal the potential of ctDNA analysis to track clonal evolution during disease progression providing a unique opportunity for noninvasive molecular disease monitoring in CLL.
**Fast-Forward #3**

**The SMRT Way to Sequence a Yeast Genome**


We have performed PacBio single molecule real time (SMRT) sequencing of three yeast whole genomes. A haploid reference yeast strain (S288C) and two novel diploid strains were sequenced as part of a larger functional genomics project. For each strain, 20kb SMRT Bell library preps were performed and sequenced on two SMRT Cells using the P6-C4 chemistry. Between 1.74 Gb and 2.55 Gb of usable sequence data was generated for each strain, with read lengths of up to 53.3 kb. Pure PacBio whole genome de novo assemblies were generated using the HGAP3 pipeline. We are using the S288C data to explore performance in comparison to the published genome as a reference. An initial assembly of S288C yielded over 99.9% genome coverage at 99.997% accuracy on only 29 unitigs, versus 17 reference chromosomes (16 nuclear chromosomes plus mitochondrion). Of these, 15 chromosomes were essentially returned as a single, complete unitig. We are now using the S288C data to optimise the assembly process and derive assembly settings for two novel strains. To this end, we have developed a new pipeline for the comparative assessment of high quality whole genomes against a reference. Genome assembly using the PacBio SMRT Portal is a two-step process, with HGAP generating a pre-assembly of error-corrected seed reads that are subsequently assembled using the Celera assembler. We explore the trade-off between accuracy and sequencing depth of this pre-assembly for different seed read length cutoffs and how this affects the final assembly.
Fast-Forward #4

The Microbial Genomics Virtual Laboratory

Simon Gladman, Nuwan Goonesekera, Clare Sloggett, Dieter Bulach, Torsten Seemann, Andrew Lonie

The uptake of genomics in public health and clinical microbiology laboratories is being slowed by the perceived requirement that each laboratory needs to, counterproductively, establish and evaluate their own tools and infrastructure which will result in a lack of standardisation of methods.

An easily instantiated computer image based around Galaxy with a defined set of microbial-specific tools and reference data is an ideal solution for enabling standardisation between laboratories. We have established the Genomics Virtual Laboratory [GVL: http://genome.edu.au] to empower laboratories to establish their own private operating environment to securely analyse their own data using software and analysis methods that are widely used for microbial genomics in a reproducible manner suited to government accreditation

The GVL consists of a set of machine images for performing genomics analyses in a scalable, reproducible manner, plus web tools for instantiating and managing the images on multiple cloud architectures. The images incorporate a number of pre-configured genomic analyses platforms including Galaxy, the Linux command line, RStudio and IPython Notebook.

The GVL images are constructed from Ansible scripts which make it straightforward to customise. Here we present a flavour of the GVL fully tailored to microbial genomics (MGVL) by incorporating various microbial analysis pipelines and tools for both the Galaxy environment and the command line.

The Genomics Virtual Laboratory project is funded by the federal NeCTAR and ANDS programs (http://nectar.org.au; http://ands.org.au)
Maize is an important agricultural crop - based on metric tons, maize is the #1 production grain crop in the world (http://faostat.fao.org). Largely as a result of plummeting sequencing costs, the growing availability of whole genome sequences has led to an increased demand for high-throughput omics studies. Interpretation of the results of these studies often requires the analysis of functional annotations of genes. The Gene Ontology (GO) is a widely used database that consists of terms that describe gene function. The majority (~99%) of the GO annotations in maize are inferred from electronic annotations by high-throughput pipelines such as Ensembl. On the contrary, only about half of the GO annotations in Arabidopsis are inferred from electronic annotations. Clearly there is a need to evaluate the confidence of existing GO annotations for maize and improve the overall quality of functional predictions in this well-studied species. Here we present a pipeline that annotates GO terms to maize gene models using multiple functional annotation methods along with an evaluation of confidence for these annotations. Our pipeline uses three approaches to assign GO terms: BLAST-based methods, domain-based methods, and advanced methods. Using a test dataset that contains high-quality manual annotations from MaizeGDB and reviewed annotations from UniProt, we are currently evaluating the performance of our approaches, comparing the performance to existing annotations, and creating a designated set of high-confidence functional annotations for maize genes. We are also designing a crowd-sourcing platform for the review of these annotations by the maize community.
Comparing algorithms to genotype short tandem repeats in next-generation sequencing data

Harriet Dashnow, Alicia Oshlack

Murdoch Childrens Research Institute

Short tandem repeats (STRs) are short (2-6bp) DNA sequences repeated in tandem, which make up approximately 3% of the human genome. These loci are prone to frequent mutations and high polymorphism. Dozens of neurological and developmental disorders have been attributed to STR expansions. STRs have also been implicated in a range of functions such as DNA replication and repair, chromatin organisation and regulation of gene expression.

Traditionally, STR variation has been measured using capillary gel electrophoresis. This process is time-consuming and expensive, and so has tended to limit STR analysis to a handful of loci.

Next-generation sequencing has the potential to address these problems. However, determining STR lengths using next-generation sequencing data is difficult. For example, many callers are limited by sequencing read lengths and polymerase slippage during PCR amplification introduces stutter noise.

Recently, a small number of software tools have been developed genotype STRs in next-generation sequencing data. We have performed a general comparison of the tools published to date, identifying their application domains, assumptions and limitations.

We have assessed the performance of some of the most popular STR genotyping tools on human next-generation sequencing data. When comparing STR callers we have observed drastic differences in which STR loci are identified as variant. Surprisingly, even for variant loci reported in common between tools, there is markedly low concordance between the specific genotype calls.

Finally, we draw together our findings to comment on the considerations when choosing and running an STR genotyping tool, with emphasis on applications to human disease.
Assessing clonality in malaria parasites from massively parallel sequencing data.

Stuart Lee [1], Gabrielle Harrison [2], Natacha Tessier [2], Livingstone Tavul [3], Olivo Miotto [4], Peter Siba [3], Dominic Kwiatkowski [5], Ivo Mueller [6], Alyssa E. Barry [2], Melanie Bahlo [2]


Parasite resistance to drug treatments for malaria have begun to emerge in south-east Asian and Pacific populations, highlighting the importance of methods for detecting which genes in the parasites are under selection pressure. In highly endemic areas of malaria, hosts may harbour multiple parasite (clonal) infections at the same time. This multiplicity of infection (MOI) complicates detection of genes under selection because current methods implicitly assume that hosts have single clone infections. Samples with MOI > 1 may be removed from downstream analysis but this results in decreased power and systematic bias.

Here we describe a new approach for estimating MOI from massively parallel sequencing (MPS) data using probabilistic clustering. This technique takes advantage of the expected symmetry in the distribution of read-counts supporting single nucleotide variants (SNVs) under MOI. We evaluate our method by simulation of MPS data with known MOI under varying coverage, sequencing error and mixtures of clones. We also apply our method to a set of MPS data from the malaria parasite Plasmodium falciparum sampled from Papua New Guinea. To assess our method's performance we compare our estimates to several other available software packages and wet-lab data. Our simulation and estimation methods are available in an R package called moimix. This MOI detection method has potential applications as the basis for a selection detection method that does not require removal of samples with MOI > 1.
Reverse engineering viral evolution

Jan Buchmann, Edward Holmes

University of Sydney

We propose a new hypothesis for why some viruses are enveloped while others lack an outer lipid bilayer, a major question in viral evolution. Such questions have received relatively little attention, despite a growing amount of available biological sequence information. We reviewed the entry, transmission and exit pathways of all (101) viral families on the 2013 International Committee on Taxonomy of Viruses (ICTV) list and propose a new hypothesis for the existence of enveloped and non-enveloped viruses, in which the latter represent an adaptation to cells surrounded by a cell wall while the former are an adaptation to animal cells where cell walls are absent. In particular, cell walls impair viral entry and exit, as well as viral transport within an organism, all of which are critical way points for successful infection and spread. Further, we propose a bioinformatic approach to investigate viral evolution by applying "reverse engineering" to viral protein structures.
Fast-Forward #9

Real-time Identification of Pathogenic Antibiotics Resistance Profiles with Oxford Nanopore MinION Sequencing

Minh Duc Cao, Devika Ganesamoorthy, Alysha Elliott, Matthew Cooper, Lachlan Coin

Institute for Molecular Bioscience, The University of Queensland

Clinical pathogen sequencing has significant potential to drive informed treatment of patients with unknown bacterial infection. However, the lack of rapid sequencing technologies with concomitant analysis has impeded clinical adoption in infection diagnosis. In early 2014, Oxford Nanopore Technologies released MinION, a USB-stick sized sequencing device that provides the potential of sequencing in real-time, but with considerably higher error rates than existing technologies. It is currently unclear the extent to which clinically actionable data can be obtained within a few hours using this device. We have developed streamline algorithms to analyse MinION sequencing data on-the-fly and bioinformatics techniques to overcome the high error rates of the data. Here we demonstrate the potential of analysing clinical microbial samples in real-time of this miniature sequencing device. Using our pipeline, we have been able to accurately identify the bacterial species and strains in the sample within half an hour of loading the sample into the sequencer and the antibiotic resistance profile has been characterized within few hours of sequencing time. We anticipate nanopore sequencing and associated analysis methods will become useful clinical tools to guide appropriate therapy in time-critical clinical presentations such as bacteraemia and sepsis.
Whole-proteome prediction of protein localisation in malaria parasites

Benjamin Woodcroft [1], Robert Radloff [2], Kristie-Lee Scanlon [2], Terry Speed [3], Stuart Ralph [1]


The phylum Apicomplexa is the most important group of eukaryotic pathogens, including causative agents of malaria and cryptosporidiosis, and many parasites in this phylum have well annotated genomes. Understanding sub-cellular localisation of proteins encoded by these genomes is critical to understanding parasite proteomes, but genetic tools in apicomplexans are limited. In the absence of empirical data on localisation, inferences have largely been drawn from protein homologues in well-characterised animals or fungi. Our analysis shows that inferences over such extreme evolutionary distances are unsafe. We assembled a database of known localisations for over 1000 experimentally verified, published protein localisations from 28 apicomplexan species. The database, called ApiLoc, is available at http://apiloc.biochem.unimelb.edu.au. Mapping localisations between homologues showed that the localisation status of a distant homologue was poorly predictive of localisation for corresponding apicomplexan proteins. Alternative tools are therefore required to predict localisation. Existing bioinformatic tools are trained mainly on animals, yeast and occasionally plants, and fail to deal with the many sub-cellular compartments that are not shared between protist parasites and animals. We therefore used data for ApiLoc to train a prediction tool, called Plasmarithm, which attempts to assign localisation among seven different subcellular destinations for Plasmodium parasites, causative agents of malaria. As inputs, we used a combination of sequence and non-sequence data, including proteomic data, timing-of-expression data, and phyletic profiles of proteins. Prediction accuracy was markedly better than existing generic tools and demonstrates that many novel data-types not previously used for localisation prediction can be highly informative as inputs.
**Fast-Forward #11**

**Functional transcriptome annotation by clustering structural RNA motifs**

Martin Smith [1], Stefan Seemann [2], Xiucheng Quek [3], John Mattick [1]


Long noncoding RNAs compose the most abundant and diverse class of transcriptional products in mammalian cells. However, their functional annotation is confounded by the poorly characterized molecular mechanisms in which they partake. A unifying feature of all ncRNAs is their propensity to fold into compact secondary and tertiary structures, which ultimately govern their function. We developed an ultra-fast, memory efficient algorithm for the pairwise comparison of RNA secondary structures that considers both the sequence composition and the sub-optimal base pairing probabilities of any two input sequences. We present the algorithm's genesis, parameter optimization, and performance benchmarking on well-curated RNA structure alignments. Finally, we present preliminary results from transcriptome-wide clustering of homologous RNA structural motifs, opening the door to systematic, high-throughput annotation of putative functional elements in lncRNAs. Amongst other applications, this work will facilitate the annotation of disease-associated SNPs from genome wide association studies (80% of which fall in non-coding regions).
Proteomic Validation of Transcript Isoforms, Including Those Assembled from RNA-Seq Data

Aidan Tay [1], Chi Nam Ignatius Pang [1], Natalie Twine [1], Gene Hart-Smith [1], Linda Harkness [2], Moustapha Kassem [2], Marc Wilkins [1]

[1] School of Biotechnology and Biomolecular Sciences, The University of New South Wales [2] Endocrine Research Laboratory (KMEB), Department of Endocrinology and Metabolism, Odense University Hospital & University of Southern Denmark

Human proteome analysis now requires an understanding of protein isoforms. We recently published the PG Nexus pipeline, which facilitates high confidence validation of exons and splice junctions by integrating genomics and proteomics data. Here we comprehensively explore how RNA-seq transcriptomics data, and proteomic analysis of the same sample, can identify protein isoforms. RNA-seq data from human mesenchymal (hMSC) stem cells were analyzed with our new TranscriptCoder tool to generate a database of protein isoform sequences. MS/MS data from matching hMSC samples were then matched against the TranscriptCoder-derived database, along with Ensembl and the neXtProt database. Querying the TranscriptCoder-derived or Ensembl database could unambiguously identify ~450 protein isoforms, with isoform-specific proteotypic peptides, including candidate hMSC-specific isoforms for the genes DPYSL2 and FXR1. Where isoform-specific peptides did not exist, groups of nonisoform-specific proteotypic peptides could specifically identify many isoforms. In both the above cases, isoforms will be detectable with targeted MS/MS assays. Our analysis also revealed that some isoforms will be difficult to identify unambiguously as they do not have peptides that are sufficiently distinguishing. We covisualize mRNA isoforms and peptides in a genome browser to illustrate the above situations. A graphical user interface (GUI) of the PG Nexus pipeline is available for download via the BitBucket repository: https://bitbucket.org/aidantay/pg-nexus-gui/src.
**Fast-Forward #13**

**Bioinformatics Software Testing**

Amir Kamali, Eleni Giannoulatou, Joshua Ho

*Victor Chang Cardiac Research Institute*

With recent advances in computational technologies the result of many medical and bioinformatics research programs now relies on programs that are often performing computation on a massive amount of data. Therefore correctness of bioinformatics programs is a must and false output in such programs can lead to serious biological incorrect decisions and could misguide downstream experimentation. Common software testing routines consist of execution and verification of a set of test cases. However, many bioinformatics programs suffer from the oracle problem - it is often not possible to evaluate the correctness of the results produced by these large and complex programs. Therefore, verification of such programs is considered very difficult or sometimes impossible. There is a need to adopt state-of-the-art software testing strategies to systematically verify and validate these bioinformatics programs.

We evaluate the application of Metamorphic testing (MT) on two widely used DNA sequencer aligners: BWA and Bowtie. To test the effectiveness of MT, we artificially introduced 'errors' (i.e., mutations), into the BWA source code. These mutant programs are then compiled and subjected to MT. We found that MT is effective at identifying mutants, yet some mutants are harder to be identified, suggesting the need to construct effective test cases.
Evidence for viral causes of cancer in prostate cancer transcriptomes and genomes

Timothy G. Amos, James S. Lawson, Wendy K. Glenn, Noel J. Whitaker, Richard J. Edwards

The University of New South Wales Australia

Viruses are known to cause 10-15% of human cancers. Some ongoing viral infections such as HPV in cervical cancer can be easily detected in the cancer transcriptome. However, in other situations viruses contribute to cancer causation without sustained high expression levels. PCR has detected HPV in prostate cancers but RNA-Seq has not shown strong evidence of continuing infections. We analysed prostate cancer transcriptomes and genomes from The Cancer Genome Atlas to further investigate HPV's role in prostate cancer causation. Previous studies have filtered out all reads that might not be viral and then set thresholds for real infections by comparing to positive controls. In contrast, we found all possible viral reads and then evaluated multiple streams of evidence that they were genuine. Sequences were evaluated on the uniqueness of alignments to human, viral and vector sequences. They were also evaluated for sequence complexity, sequence quality, alignment quality, relative alignment locations of paired-end reads and the presence of chimeric reads that indicate viral integration sites. By screening cancer transcriptomes and genomes against all viruses in the NCBI database (c. N = 5766), including non-human viruses, we are also able to compare the strength of evidence against known false positives. We discuss the results of 22 viral candidates for oncogenesis in 558 prostate cancer paired RNA-Seq and WGS datasets.
Pros and cons of sequencing with Agilent's Haloplex target enrichment system

David Goode [1], Mandy Ballinger [1], David M Thomas [2]

[1] Peter MacCallum Cancer Centre  [2] Kinghorn Cancer Centre, Garvan Institute of Medical Research

The Haloplex target enrichment system from Agilent combines aspects of capture and amplicon sequencing. It combines restriction digestion of DNA with fragment capture by labelled probes. It can provide a cost-effective balance between sample size and genome coverage for sequencing studies with well-defined analysis goals but limited budgets.

Haloplex's unique enrichment strategy requires special data analysis considerations. Because restriction enzymes cut at defined sites, all reads from a given target region start at the same base, creating irregular coverage profiles and complicating removal of PCR duplicates. Haloplex also has a distinct error profile, with adaptor read-through errors being particularly prevalent in shorter amplicons. Many of these problems can in part compensated for by rigorous filtering and leveraging the fact most sites in target regions are covered by multiple amplicons.

We used Haloplex to identify germline risk variants in 1153 sarcoma patients. Our filtering yielded >90% specificity and sensitivity for SNVs. Haloplex is also useful for detection of short indels, but finding longer indels, complex indels and CNVs remains challenging.

Another challenge for genetic studies is the lack of publicly available Haloplex data sets for comparison. We have devised a scheme that uses read depth and variation at linked neutral sites to correct for ascertainment biases between data sets obtained with different capture technologies. Application of this method has revealed several novel sarcoma risk genes.

Our experiences underscore the importance of spending a significant amount of time examining sources of errors and biases when employing a new sequence capture technology.
POSTER ABSTRACTS
Host microRNAs are promising biomarkers of infectious diseases

Christopher Cowled, Cameron Stewart, Chwan-Hong Foo, Andrew Bean

CSIRO

Molecular diagnostic tests for infectious diseases may be either direct or indirect. Direct tests target pathogen-derived molecules (such as protein antigens or nucleic acids), while indirect tests target host-derived molecules, such as antibodies. Direct tests can face problems with fast-evolving pathogens such as viruses, or highly localized infections, or agents that replicate at very low levels. In contrast, antibodies are generally undetectable for the first week of infection, frequently cross react against related pathogens, and cannot easily distinguish current from past infections. There is therefore a constant need for new and better diagnostic tests. Quantitative 'omics' including transcriptomics, proteomics and metabolomics are well suited for identification of host-derived biomarkers. The chemical uniformity of RNA makes it a particularly attractive target, and RNA-Seq enables detailed measurement of transcripts over a large dynamic range. One promising category of biomarkers are microRNAs, which exhibit several attractive characteristics: Easy extraction from blood or serum, reduced transcriptional complexity (no splicing and moderate diversity), and each sequencing read corresponds to exactly one count (no need to adjust for transcript length). Despite these advantages, difficulties still remain. In particular, microRNA quantification is somewhat platform-dependent. Reasons for this include difficulty with normalization due to skewed sample composition (the single most abundant miRNA can make up > 95% of some samples), and the natural occurrence of isomiRs can influence abundance estimates. Nevertheless, host microRNAs are a promising class of infectious disease biomarkers.
**Poster #2**

**kWIP: The k-mer Weighted Inner Product, a de novo alignment free genetic clustering metric**

Kevin Murray [1], Cheng Soon Ong [2], Christfried Webers [2], Norman Warthmann [1]


Modern techniques in population genomics generate unprecedented quantities of data within which complex genetic histories reside. The scale and complexity of these data require the development of new approaches to the analysis of genetic data. We present the k-mer Weighted Inner Product, a de novo, alignment free measure of genetic similarity between samples in a population. kWIP, is an efficient tool implementing this metric that can determine the genetic relatedness between samples without alignment or assembly. We show kWIP can reconstruct the true relatedness between samples directly from sequencing reads generated with various modern sequencing platforms, as well as from simulated data.

kWIP works by decomposing sequencing reads to short k-mers, hashing these k-mers using a constant-memory data structure, and performing pairwise distance calculation between these sample k-mer hashes. The power of kWIP comes from the weighting applied across different hash values, which decreases the effect of erroneous, rare or over-abundant k-mers while focusing on k-mers which give the most insight into the similarity of samples. kWIP is free, open source software implemented in C++ and released under the GNU LGPL v3.
**Poster #3**

**The effect of ADAR3 deficiency on RNA editing in mouse hippocampus**

Dessislava Mladenova, Lotta Avesson, Guy Barry, Mark Pinese, Bryce Vissel, John Mattick

*Garvan Institute of Medical Research*

RNA editing refers to the deamination of adenosines or cytosines to alter the sequence of RNA, and is most active in the brain. RNA editing may have played an important role in cognitive evolution, as it has expanded during mammalian evolution, especially in the primate lineage. Adenosine to inosine (A-to-I) editing is catalyzed by three members of adenosine deaminase acting on RNA (ADAR) protein family.

ADAR3 catalytic deaminase activity has not been detected in vitro, and ADAR3 is thought to act by blocking the action of ADAR1 and/or ADAR2. Here we asked how ADAR3 deficiency affects RNA editing. We generated Adar3 knockout mice and performed RNA sequencing from mouse hippocampus.

On average 49.8 million reads per mouse were obtained. 800 A-to-I(G) sites met the criteria of sufficient coverage (>20 reads) and an average editing level of at least 5% per sample. Additional criteria were applied to obtain 282 high confidence candidate editing sites. There was no global change in the editing level of ADAR3-deficient compared to control mice; however there were pronounced differences in the editing frequencies of individual sites.

16 of the candidate editing sites showed statistically significant differences in editing level between the Adar3 wild-type and knockout mice. ADAR3 deficiency resulted in a reduction of the editing level of 13 sites, indicating that ADAR3 can enhance the editing frequency of specific sites in vivo.

In conclusion this is the first report indicating that ADAR3 can modulate the RNA editing levels of specific genes in vivo.
Better Detection of Natural Selection in Whole Exome Data

Benjamin Kaehler [1], Von Bing Yap [2], Gavin Huttley [1]


Estimation of selective pressure on protein-coding sequences is a key comparative genomics approach for de novo prediction of lineage specific adaptations. Selective pressure is measured on a per gene basis by comparing the rate of non-synonymous substitutions to the rate of neutral evolution, typically assumed to be the rate of synonymous substitutions. We previously demonstrated that under time-reversible models of nucleotide evolution, the number of substitutions is systematically biased toward overestimation when sequence composition changes.
We extend these findings to the case of time-reversible codon substitution models and demonstrate overestimation of selective pressure (ie exaggerating the magnitude of purifying natural selection) over whole exome-scale data drawn from mammals, insects, frog, and fish. All published codon substitution models have been time-reversible and thus assume that sequence composition does not change over time. Our basis for comparison is a non-stationary codon substitution model that allows sequence composition to change.
We demonstrate by application of model selection techniques that our new codon model tends to fit the data better. We further show through direct measurement of non-stationarity that bias in selective pressure and rate of neutral evolution increase with the extent to which non-stationarity is violated.
With an increased rate of genomic sequencing across the tree of life, the de novo estimation of natural selection assumes even greater importance for identifying functionally important genes and processes. We have shown that the inference drawn under time-reversible models is affected by compositional divergence. Our new codon model provides a foundation for more robust inferences.
Poster #5

Dynamic expression of long noncoding RNAs and repeat elements in synaptic plasticity


Long-term potentiation (LTP) of synaptic transmission is recognized as a cellular mechanism for learning and memory storage. Although de novo gene transcription is known to be required in the formation of stable LTP, the molecular mechanisms underlying synaptic plasticity remain elusive. Non-coding RNAs have emerged as major regulatory molecules that are abundantly and specifically expressed in the mammalian brain. By combining RNA-seq analysis with LTP induction in the dentate gyrus of live rats, we provide the first global transcriptomic analysis of synaptic plasticity in the adult brain. Expression profiles of mRNAs and long noncoding RNAs (lncRNAs) were obtained at 30 minutes, 2 hours and 5 hours after high-frequency stimulation of the perforant pathway. The temporal analysis revealed dynamic expression profiles of lncRNAs with many positively, and highly, correlated to protein-coding genes with known roles in synaptic plasticity, suggesting their possible involvement in LTP. In light of observations suggesting a role for retrotransposons in brain function, we examined the expression of various classes of repeat elements. Our analysis identifies dynamic regulation of LINE1 and SINE retrotransposons, and extensive regulation of tRNA. In sum, these experiments reveal a hitherto unknown complexity of gene expression in long-term synaptic plasticity involving the dynamic regulation of lncRNAs and repeat elements. These findings provide a broader foundation for elucidating the transcriptional and epigenetic regulation of synaptic plasticity in both the healthy brain and in neurodegenerative and neuropsychiatric disorders.
**Poster #6**

**Narrowing Down On CIPK16 Orthologues: A Gene Involved In Enhanced Salt Tolerance**

Shanika Amarasinghe [1], Nathan Watson-Haigh [1], Matthew Gilliham [2], Stuart Roy [1], Ute Baumann [1]


Calcineurin B-Like Protein Interacting Protein Kinases (CIPKs) are key regulators of pre-transcriptional and post-translational responses to abiotic stresses. Many plant species have numerous forms of CIPKs, which appear to have very specific roles during abiotic stress. From a forward genetics screen, Arabidopsis thaliana CIPK16 (AtCIPK16) has been previously identified as an important gene responsible for reduced shoot salt accumulation and improved salinity tolerance in Arabidopsis and transgenic barley. However, the existence of AtCIPK16 homologues in cereals and the mode of action of AtCIPK16 in them are still unknown. Thus, this study aimed to understand the prevalence of CIPK16, and in return find an AtCIPK16 orthologue in cereals. A phylogenetic analysis has been performed on CIPK genes obtained from several species related to A. thaliana and cereal crops. The resulting phylogenetic tree revealed insights into the evolution of CIPK16 which has important consequences for breeding salt tolerance in cereal crops. One particular observation was that the clade including AtCIPK16 had a sister clade which contained two segmentally duplicated genes AtCIPK5 and AtCIPK25. Furthermore, we find AtCIPK16 contains an indel which is conserved in Brassicaceae.
**Poster #7**

**In Silico analysis of Single Nucleotide Polymorphisms (SNPs) in human FANCA gene**

Abubaker Mohamed [1], Ozaz Mohammed [2], Hadeel Yousif [3]


Single-nucleotide polymorphisms (SNPs) play a major role in the understanding of the genetic basis of many complex human diseases. Also, the genetics of human phenotype variation could be understood by knowing the functions of these SNPs owing to the importance of FANCA gene in a post replication repair or a cell cycle checkpoint function. In this work, we have analyzed the genetic variation that can alter the expression and the function of the FANCA gene using computational methods. Genomic analysis of FANCA was initiated Polyphen and SIFT server used to retrieve 16 harmful mutations, among of these 16 nsSNPs damaged SNPs five non-synonymous SNPs showed very damaging by higher PSIC score of the Polyphen server with a SIFT tolerance index of 0.00-0.01 (R318M, I493T, A610T, P739L, R1117G). Protein structural analysis with these amino acid variants was performed by using I-Mutant and Modeling amino acid substitution with chimera software to check their stability and the effect of the native and mutant residues protein and structure for all 16 nsSNPs damaged. Screening for these SNPs variants in coding region may be useful for Fanconi anemia disease molecular diagnosis. Of the total 229 SNPs in 3'UTR region of FANCA gene, 24 SNPs were found in the 3' UTR contain alleles can be disrupts a conserved miRNA site, therefore might change the protein expression levels.
Poster #8

Tracking clonal evolution in cancer

Christoffer Flensburg

wehi

Cancer cells evolve over time due to the accumulation of new somatic mutations. These new mutations can have a marked impact, driving disease progression, stifling drug response, or stimulating metastasis. With access to genomic data from multiple time-points, it is possible to use somatic mutations to identify and track the rise and fall of various cancer subpopulations, but the analysis is challenging.

We have developed an analysis platform to identify subpopulations within cancer samples using exome sequencing data. As well as considering point mutations and small indels, our platform uses changes in coverage and SNP allele frequencies to generate genotype aware copy number calls. Uncertainties are estimated for SNVs and CNVs, and are propagated throughout the analysis, taking biological and technical uncertainties into account. The final phylogenetic tree is assessed for self-consistency.

Applying this approach on real data identified instances where the same mutation occurs within an individual patient on multiple separate occasions, providing evidence of convergent evolution.
Poster #9

Batch effect detection, correction and characterisation in Illumina HumanMethylation450 Beadchip array data.

Jason Ross, Susan van Dijk, Peter Molloy, Yalchin Oytam

CSIRO

As with other genomic data, Illumina HumanMethylation450 (450K) arrays can be subject to significant batch effects. Batch effects are known to reduce experimental power and to potentially create false positive results. To study batch effects on 450K arrays, we took advantage of a dataset on 372 peripheral bloods from EpiSCOPE, a large multi-institutional epigenetics project. The arrays were structured in a balanced-block design, such that biological variation in experimental factors was not confounded with technical variation.

The 450K arrays encompassed 31 glass slides - with 12 arrays per slide. Analysis showed seven of the 31 slides had a noticeable batch effect. To correct this, we input normalised M-values into our batch correction software "Harman". Post-analysis demonstrated Harman was highly successful in removing the batch effect across samples.

Typically every fourth slide exhibited inordinate batch effect relative to other slides, suggesting a structured difference in processing. At the individual CpG probe level, we find most probes are very robust, however for 0.3% of probes, the batch effect was very large, with 10% to 65% corrections made in methylation (Beta) values. Overwhelmingly we find sensitivity to a batch effect is a function of probe melting temperature (Tm). While most probes have an in silico determined Tm ~75°C, the probes with a Tm of less than 71°C show high between-sample and between-batch variation.

Our findings suggest that approximately 20,000 probes on the 450K array with a low Tm can be highly sensitive to hybridisation conditions and these probes should be treated with caution.
Dissecting the evolution of melanoma through exome sequencing

Ismael A Vergara, Stephen Q Wong, Shahneen Sandhu, Christopher Mintoff, Gisela Mir Arnau, Sarah-Jane Dawson, Mark Shackleton, Anthony T Papenfuss

Peter MacCallum Cancer Centre

Each year, 12,500 people are diagnosed in Australia with melanoma, an aggressive type of cancer characterized by high metastatic potential to various organs as well as low survival rates after metastasis is initiated. Previous studies have shown that the mutational burden underlying melanoma is among the highest of all cancers, generating a complex scenario for the elucidation of the genetic basis of its metastatic potential and resistance to therapy. The Cancer Tissue Collection After Death (CASCADE) project at Peter MacCallum Cancer Centre aims to address these questions by multiple sampling of primary and metastatic sites from patients at the time of autopsy. These samples correspond to different organs at the sites of metastasis, as well as multiple cores from primary tumours. Here, I will present results from whole exome sequencing of 37 primary, regional and distant metastasis samples from four BRAF-mutant melanoma patients. Analysis of SNVs, InDels and allelic imbalance patterns provides insight into the extent of intratumour heterogeneity within primary sites of disease, differences within and across organs with metastasis, the evolutionary relationships between primary melanomas and metastases and the impact of treatment on the mutational profile observed across samples.
Poster #11

Combining multiple tools outperforms individual methods in gene set enrichment analyses

Monther Alhamdoosh [1], Milica Ng [1], Nicholas J. Wilson [1], Michael J. Wilson [1], Matthew E. Ritchie [2]


Gene set enrichment analysis (GSEA) allows researchers to efficiently extract biological insight from long lists of differentially expressed genes, by interrogating them at a systems level. In recent years there has been a proliferation of GSEA methods, with each method having unique advantages and disadvantages dependent on the particular characteristics of the data set of interest. It is therefore becoming increasingly difficult for researchers to select an optimal GSEA tool based on their particular data set. Moreover, the majority of GSEA methods do not allow researchers to simultaneously compare gene set level results between multiple experimental conditions.

A novel method, which utilizes ten prominent GSEA tools in order to generate more robust gene set rankings, is proposed. The ensemble of GSEA (EGSEA) combines the enrichment analysis results of multiple methods and calculates collective gene set scores to improve the biological relevance of the highest ranked gene sets. The EGSEA tool enables researchers to interrogate more than 20,000 gene sets collected from public databases. It has multiple visualization options, including the ability to superimpose gene expression fold changes on pathway maps, and generate heat maps of gene signatures across multiple conditions. The significance of a gene set can also be viewed in the gene set space via summary plots. Importantly, the performance of the individual GSEA methods can be compared for each contrast. EGSEA has been tested on a number of human and mouse data sets and, based on biologists' feedback, it consistently outperforms the individual tools that have been combined.
Transcription factors (TFs) are one of the main players to regulate gene expression in different eukaryotic cell types. Differential binding of TFs to DNA regulatory regions plays a major role in determining cell-type specific gene expression in normal cells. Conversely, aberrant TF binding is potentially associated with disruption of gene expression in cancer development. Identification of functional TFBS across the genome will provide a better understanding of the complex gene expression patterns in different cell types. ChIPseq is the main experimental approach to profile TFBS, but limited to a single TF for each experiment. The recent integration of genome-wide sequencing and computational approaches makes it possible to uncover potential regulatory regions via the identification of TF footprints in DNase I hypersensitivity sequencing data. There are several algorithms developed for this data type and/or combining with other epi/genomic data. These algorithms significantly improve cell-specific predictions of whole-genome TFBS and reduce cost as well. In parallel, we have developed a novel method to predict direct genome-wide TFBS as TF footprints using NOMEseq* data. The algorithm is applied to a normal cell and its counterpart cancer cell to map all the direct TFBS. From the identified binding sites, the corresponding TFs are also determined for further downstream analysis. Moreover, the analysis of differentially binding sites between cancer and normal is performed to reveal the TF profile of cancer cell. The method can be applied to new cell types with NOMEseq data.
Poster #13

Biological networks reveal the genetic response to antifungal drug synergy

Chi Nam Ignatius Pang [1], Yu-Wen Lai [2], Leona Campbell [2], Sharon Chen [3], Dee Carter [2], Marc R. Wilkins [1]

[1] Systems Biology Initiative, The University of New South Wales, Australia
[2] School of Molecular Bioscience, University of Sydney, Australia
[3] Institute for Clinical Pathology and Medical Research, Westmead Hospital, Westmead, Australia

Invasive fungal infections (IFIs) are difficult to treat. Few effective antifungal drugs are available and many have problems with toxicity, efficacy and drug-resistance. To overcome these challenges, existing therapies may be enhanced using more than one agent acting in synergy. Iron chelators combined with antifungals can improve the clearance of some IFIs, but the mechanism of synergy is complex and poorly understood. Using checkerboard assays, we found amphotericin B (AMB) and the iron chelator, lactoferrin (LF), were synergistic against Saccharomyces cerevisiae and Cryptococcus neoformans. We extracted mRNA from S. cerevisiae treated with i) AMB only, ii) LF only, iii) a combination of AMB + LF or iv) corresponding matching controls. RNA-Seq was performed using Illumina HiSeq 2000 and/or NextSeq with biological triplicates. Using a generalized linear model (EdgeR and RUVSeq), we tested the effects of AMB alone and AMB + LF synergy on transcript expression. Cytoscape was used to construct biological networks to co-visualize with the differential gene expression data and identify the cellular response to drug synergy. AMB treatment alone caused the up-regulation of iron responsive transcription factor Aft1 and its targets, while unexpectedly, AMB + LF treatment decreased their expression and also down-regulated expression of the zinc responsive transcription factor Zap1 and its target genes. S. cerevisiae Aft1 and Zap1 knock-out mutants had increased susceptibility to AMB, which has also been seen when the Zap1 homologue was disrupted in C. neoformans. Zap1 has druggable sites and may be a potential target for synergistic therapy with AMB.
Development of Trawler, a web based motif discovery tool

Louis Dang [1], Henry Chiu [2], Jerico Revote [3], Mirana Ramialison [1]


Identification of enriched motifs in downstream analysis of DNA binding assays is one of many biological computational problems. As a result, a variety of motif discovery tools and algorithms have been developed to accurately identify the correct binding motif present in the sequenced regions. We have previously developed Trawler, a fast motif-discovery tool that searches for de novo binding motifs against a biological background. To make Trawler user friendly, we aim to develop a web version with the option of using bed files as input and automation of the background data.

Firstly, the user inputs a bed file used to retrieve sequences. A "bedtofasta" script has been designed to extract FASTA sequences from locally stored chromosome data. Second, generation of the random background takes calculations of each input region to the nearest transcription start site. A gene list and chromosome lengths are required to retrieve the corresponding sequences for the background data. In addition, Trawler displays conservation scores for every instance of the motif identified in the sample sequences, a feature unique to Trawler.

Overall, the process of generating a random background has been streamlined and automated into the current version of software while allowing the use of bed files as input, which is not available with other motif discovery tool. It is hoped that this new implementation of Trawler will help researchers to not only detect the primary binding motif, but also accurately identify secondary motifs.
Anscombe's 1948 variance stabilizing transformation for the negative binomial distribution is well suited to RNA-Seq expression data.

Paul Harrison

Monash Bioinformatics Platform

The negative binomial distribution has been proposed as a good fit to RNA-seq read count data, taking into account noise due to both the count-based nature of the data and biological variation. It is possible to use this distribution directly for GLM-based differential expression (as in EdgeR or DESeq2), but for many other statistical techniques and visualizations a simpler error model with a constant level of variance is assumed. When working with RNA-Seq data it is also useful to be able to speak in terms of fold-changes, so a log transformation is often used, however this overly inflates variation for genes with small counts. What is needed is a moderated log transformation, behaving like log for large counts but more moderately for small counts. One commonly used moderated log transformation is log(x+c) where c is thought of as a "prior count". Anscombe in 1948 gave another moderated log transformation which provides uniform variance over a wider range of expression levels. He also gives a reasonable choice of value for the prior count c in the log(x+c) form based on the dispersion (which may be estimated using software such as EdgeR).
**Poster #16**

**Gene Ontology enrichment analysis for gene subsets of distinctive function**

Firoz Anwar, William Wilson, Jason Ross, Peter Molly, Denis Bauer

CSIRO

Functional enrichment of gene sets, such as Gene Ontology (GO) enrichment, can provide insight into the underlying biology. Typically, a list of interesting genes, e.g. differentially expressed (DE) genes, is used to determine whether a GO term is over/under-represented by comparing its frequency against a baseline (background). The choice and size of the background can influence the identified terms and their significance, potentially resulting in misleading biological interpretation. This problem is specifically pronounced for subsets of the genome with specific function, e.g. mitochondrial proteins. We propose a tailored approach for the analysis of gene subsets with distinctive functions. We demonstrate that performing GO enrichment analysis using goseq [1] on a set of 93 DE mitochondrial genes between visceral and subcutaneous adipocyte cells in human (3 matched biological replicates) using the traditional approach of including all genes as background yields 161 enriched terms, which mostly cover broad mitochondrial function. Whereas, adjusting the background to only include genes related to mitochondrial functions, defined by MitoCarta [2], yields no enrichment. This necessitates the need for constructing a purpose-built background. By including only GO terms associated with MitoCarta genes as a tailored background and a pruned GO tree we found 226 enriched terms, identifying biology processes around beta-oxidation of fatty acids and acetyl-CoA metabolism as differential. This finding suggests that by constructing a purpose-built background set of genes and a pruned GO tree we could identify some fundamental difference in the catabolism of fats between visceral and subcutaneous adipocytes.
Mutations detection between a mutant population and its wild-type parent is challenging when no reference genome is available. In wheat, analysis of mutant sequences is even harder due to the polyploid nature of its genome. One way of reducing complexity of the analysis is to capture the exonic regions of the genome only and sequence these. NimbleGen array technology was applied to the genome of tetraploid durum wheat that captures 110 million bp of the 13 Gbp of whole genomic space. In this study we have developed statistical tools and methods using Java and R to analyse and estimate capture target efficiency, average and cumulative coverage, mutation probability and mutation type. We also devised a statistical test to detect the zygosity of a mutation. We applied our methods to exome capture data of a durum wheat cultivar (as control) and 5 EMS mutants from an M2 generation. We compared our results to a similar study on an EMS-induced population in rice where a full reference genome was available. We found very similar rate of mutation and zygosity. Finally, we selected a diverse subset of predicted mutations in terms of mutation type and zygosity and for verification in the lab by PCR amplification and subsequent Sanger sequencing. Even though we evaluated our methods for tetraploid wheat, they can be applied to any other polyploidy species.
Detecting sample swaps in cancer cohorts

Jan Schroeder [1], Vincent Corbin [1], Anthony Papenfuss [2]


Research in cancer genomics often deals with large cohorts of matched tumor and normal samples, and sometimes with replicate tumor samples. The process from data collection, to sequencing or ChIP genotyping, to analysis involves a large degree of manual handling of data. It is not uncommon that small or large mistakes happen in this handling of data: sample mix-ups, contamination of DNA, mislabelling of samples, etc.

We have developed a simple, but effective method to automatically compare tumour and germ line variant calls across the cohort. The method is based on comparing homozygous SNPs, which are not affected by changes in copy number and very unlikely to change between two related samples (such as a tumour and its matched germ line). A mixture model of the concordance of homozygous variants between any two samples in the cohort can automatically distinguish related samples from unrelated.

This allows the user to quickly identify (i) wrong pairings (low correlation for the paired germ line), (ii) alternative pairings (high correlation for an unexpected germ line or sample), and (iii) issues of contamination (correlation values that do not fit the established model well). We propose this method as a simple sanity check for larger cohorts to avoid laborious back tracking through the analysis process when things do not work out as planned.
Poster #19

Mapping human variation across the exome to identify regions intolerant to change and assist predictions of deleteriousness.

Michael Silk

Walter and Eliza Hall Institute of Medical Research

Personal genomics is rapidly advancing in its efficacy in identifying causative variants of Mendelian disorders. To isolate these variants from the background of benign variation, in silico scores of deleteriousness are used to prioritise those most likely to affect protein function. While progress has been made in developing these scores, no score or combination of scores has a level of accuracy truly fit for clinical use.

Many in silico scores currently in use measure how conserved a variant's position is at the amino acid and genetic levels. This approach is confounded by the necessity to compare across species, where functionally relevant mutations in humans may not be conserved in other species. Instead, we propose a novel scoring approach using a newly released variant database of human exomes to identify regions intolerant to variation. We explore different approaches to summarising variability within predefined regions, within and across genes and exons, and discuss the complexities of this strategy. We also consider ways to assess our results using well-studied genes where the consequences of variation are well known. We expect that this information can be used to improve in silico scoring metrics to better patient diagnoses from personal genomic pipelines.
Poster #20

SLiMScape 3.x: a Cytoscape 3 app for discovery of Short Linear Motifs in protein interaction networks

Emily Olorin [1], Kevin O'Brien [2], Nicolas Palopoli [3], sa Perez-Bercoff [1], Denis Shields [2], Richard Edwards [1]


Short linear motifs (SLiMs) are small protein sequence patterns that mediate a large number of critical protein-protein interactions, involved in processes such as complex formation, signal transduction, localisation and stabilisation. SLiMs show rapid evolutionary dynamics and are frequently the targets of molecular mimicry by pathogens. Identifying enriched sequence patterns due to convergent evolution in non-homologous proteins has proven to be a successful strategy for computational SLiM prediction. Tools of the SLiMSuite package use this strategy, using a statistical model to identify SLiM enrichment based on the evolutionary relationships, amino acid composition and predicted disorder of the input proteins. The quality of input data is critical for successful SLiM prediction. Cytoscape provides a user-friendly, interactive environment to explore interaction networks and select proteins based on common features, such as shared interaction partners. SLiMScape embeds tools of the SLiMSuite package for de novo SLiM discovery (SLiMFinder and QSLiMFinder) and identifying occurrences/enrichment of known SLiMs (SLiMProb) within this interactive framework. SLiMScape makes it easier to (1) generate high quality hypothesis-driven datasets for these tools, and (2) visualise predicted SLiM occurrences within the context of the network. To generate new predictions, users can select nodes from a protein network or provide a set of Uniprot identifiers. SLiMProb also requires additional query motif input. Jobs are then run remotely on the SLiMSuite server (http://rest.slimsuite.unsw.edu.au) for subsequent retrieval and visualisation. SLiMScape can also be used to retrieve and visualise results from jobs run directly on the server.
Cancer cells have lost many of the main features of multicellularity, such as an effective cell-to-cell adhesion, communication and differentiation. One explanation is that cancer cells undergo a process of atavism, where there is a coordinated reactivation of ancient transcriptional programs that evolved during the emergence of unicellular organisms. During this process, cells would lose part of their identity as multicellular cells.

First, we assessed the 'ancientness' of the transcriptomes of tumour and normal samples with the Transcriptome Age Index (TAI), which uses the point of emergence of the oldest ortholog of each gene as weights of their expression levels. We calculated the TAI of normal and tumour samples of 8 tissue types available in TCGA, adjusting by loess correction for contamination of surrounding tissue in the samples. We found that all 8 tumour types had a consistently older transcriptome than their normal counterparts due to the upregulation of pre-metazoan genes, and the downregulation of metazoan genes. Additionally, this effect was more pronounced in higher-grade, less-differentiated tumour samples.

Next, we identified differentially expressed pathways in the tumours, and found a core of consistently upregulated primitive pathways that date back to unicellularity, whereas pathways related to multicellularity showed a differential inactivation.

Overall our results suggest that cancer cells are undergoing a process of atavism where the cells differentially break free from the constraints of multicellularity, which converges to a core of reactivated processes of unicellular cells.
**Poster #22**

**Deleterious Passenger Mutations as a Marker for Progression towards Liver Cancer**

Magdalena Budzinska [1], Thomas Tu [1], Fabio Luciani [2], Nicholas Shackel [1]


Hepatocellular carcinoma (HCC) is associated with hundreds of passenger mutations, which have generally been ignored but can alter cell survival and thus may act as marker for cancer progression. A previous study has shown that deleterious passenger mutations (DPMs) occur frequently and accumulate in tumours. Therefore, we aim to detect DPMs in liver disease progression and hypothesise that DPMs accumulate in hepatocytes in precancerous conditions leading up to HCC. We have performed whole exome sequencing of pre-cancerous liver tissues: 12 patients with limited level of liver injury and 6 HCV-positive patients with liver cirrhosis. We have also analysed 3 publically available datasets of paired HCC and surrounding non-tumour liver (total of 148 samples). Further, we determined whether these DPMs were likely to alter genes expressed in the hepatocytes by filtering out those genes not detected in liver tissue by microarray analysis.

Increasing numbers of DPMs were observed in patients with progressively worse liver disease leading up to HCC. The pattern of observed DPMs in HCC is consistent over multiple algorithms for scoring deleterious effect, in multiple aetiologies of HCC, and in multiple datasets. This strongly suggests that DPM accumulation is a general mechanism in tumour evolution. Moreover, precancerous alterations were found in non-tumour tissue, despite being used in prior studies as normal paired controls.

In summary, we have shown that DPM accumulation could act as a potential biomarker for risk towards HCC development without having first to identify rare and unknown HCC driver mutations.
Poster #23

3D reconstruction of genome-wide gene expression and regulation in mouse hearts

Nathalia Tan [1], Mirana Ramialison [1], Jose Maria Polo [2]


An understanding of spatial and temporal gene expression and regulation is key to uncovering developmental and physiological processes, and accordingly, disease processes. Numerous techniques exist to gain gene expression and regulation information, but very few utilise intuitive true-to-life imaging methods to analyse and present their results. Mapping gene expression and regulation information onto a three-dimensional model will greatly aid in the visualisation of gene expression patterns and allow researchers to see which areas in an organ or organism express which genes in the genome, and vice versa. Furthermore, analysing this data using cluster analysis may uncover novel genes relating to development and disease. Hearts make a good model for this study as they are structurally and developmentally complex with distinct sub-compartments. Furthermore, mice as an animal model are ideal as they are genetically similar to humans and may have clinical translation. We hypothesise that finding the three-dimensional position and regulation of every gene at a given time during development will reveal the specific subset of genes that play an essential role in specific sub-compartments.
Identification of chromatin accessibility domains and the DNA binding motifs they contain, in breast cancer cells.

Kristine Hardy [1], Sherry Tu [1], Tara Boulding [1], Angelo Theodoratos [2], Sudha Rao [1]


The epithelial-mesenchymal transition (EMT), the process whereby epithelial cells gain migratory and invasive properties characteristic of mesenchymal cells has been linked to the formation of cancer stem cells (CSCs). Activation of Protein Kinase C (PKC) by phorbol esters can induce the luminal MCF7 cell line to undergo EMT, with a proportion of the population becoming CSC. To determine how PKC-induced alterations in the epigenome influence EMT and CSC formation in MCF-7 cells, we employed a combination of expression profiling and Formaldehyde Assisted Regulatory Elements (FAIRE)-sequencing in order to reveal novel links between gene expression and DNA accessibility changes after PKC activation. Various approaches to normalisation of the different FAIRE-seq samples were explored, with a loess approach ultimately chosen. We found that during the EMT increases in accessibility principally occurred in regions distant from transcription start sites that were low in CpG, enriched with chromatin marks of enhancer elements and motifs for FOX, AP1, TEAD and AP2. Increases in FOX and AP-1 motif accessibility were associated with genes that exhibited increased expression in CSC, while increased AP-2 accessibility was associated with genes that had higher expression in non-CSCs. This study revealed novel regions of DNA accessibility induced by PKC that contribute to the understanding of how epigenomic plasticity of cells undergoing EMT leads to the activation of genes that drive the CSC transcriptional program.
Circulating tumor cells (CTCs) are cancer cells that can detach from the primary tumor and enter the blood circulation. Most cancer cells that leave the tumor are destroyed by immune system, however some survive and can migrate to other tissues. They can then become phenotypically different and cause micro or macro metastasis. Thus, research into CTCs is crucial for developing more effective therapies against metastasized cancer. In our current research we have a combination of breast cancer tumor and circulating tumor cell samples. These samples are obtained from human tumor xenografts in immunodeficient mice. Xenograft mouse models are known to be challenging as they suffer from the scarcity of human cells. The major difficulty of working with these xenografts is the mixed human and mouse information. Most of the CTC samples mapped to the mouse genome with higher mapping proportions than they mapped to human genome.

Here we outline a bioinformatics strategy called XenoSplit adopted for this data. A major challenge is to distinguish human from mouse reads. To classify ambiguous reads, XenoSplit compares human mapping scores with mouse genome mapping scores. First, reads were mapped to human and mouse genomes. Then reads were allocated to a species based on the number of correctly aligned bases. This identifies a sufficient number of human reads to proceed with downstream analyses. For further comparative analysis of CTC and tumor samples we used R packages limma and edgeR and the results revealed numerous differentially expressed genes.
The control of gut motility remains poorly defined even in healthy individuals and this limited understanding therefore makes it difficult to treat disorders in patient populations. This is particularly true in the human colon, where we still have very little information on normal or abnormal motor patterns, and even less information on the mechanisms that control them. A preliminary computational model of the human colonic mucosa has been developed to characterise the three dimensional spread of serotonin which is influenced by pressure changes, tissue structure and serotonin release rates. The model integrates data of the kinetics of serotonin secretion from individual human enterochromaffin cells with optical fiber readings of intracolonic pressure patterns and assumes that the mucosal mesentery acts as a compressible porous medium. Preliminary findings indicate that serotonin concentrates in the mucus and moves quickly to the lumen, rather than moving deeper into the colonic tissue. The concentration of serotonin in the mucus may explain the contrast between serotonin concentrations measured in whole colon and those measured in single-cell analyses. Future development of this model will involve alternative mathematical modeling of the mucosal mesentery, further statistical robustness of source data and ideally, refinement until a link between serotonin release kinetics, muscle movement pattern and patient presentation is found. Alternatively, separate characterisation of serotonin release kinetics and muscle movement patterns will be elucidated, for individual application to patient presentation.
Genetic Characterisation of the Evolution of a Novel Metabolic Function in Yeast

Asa Perez-Bercoff [1], Tonia L. Russell [2], Philip J.L. Bell [3], Paul V. Attfield [3], Richard J. Edwards [1]


We have a unique opportunity to study how new biological pathways have evolved to produce a yeast with a novel metabolic activity. Thirty Saccharomyces cerevisiae strains were grown as a mixed population, and adapted on a specific media for 1463 days, undergoing sexual mating every two months. As a pilot study, three of the ancestral strains from the starting population, including two diploids, were selected for new PacBio RSII long-read single molecule real time (SMRT) sequencing at the Ramaciotti Centre for Genomics. High quality complete genomes were assembled de novo using the hierarchical genome-assembly process (HGAP3) using only PacBio non-hybrid long-read SMRT sequencing data, and corrected using Quiver. In addition, we have shotgun metagenomic Illumina data from a population exhibiting early adaptation to growth on the selective media. We are developing methods to map these short-read data onto several high quality ancestral genomes in order to estimate the relative contribution of each ancestor's genetic variation to the evolved population, and identify possible sites of recombination. In addition, metagenomic data is being mapped against the official S. cerevisiae reference strain S288c to conduct variant calling to identify single nucleotide polymorphisms (SNPs) that are not present in any of our ancestral or reference genomes. These will be compared to the publicly available "100 yeast genomes", and partitioned into natural variation and candidates for novel mutations. By doing these comparisons we hope to elucidate how the population has evolved to acquire its novel characteristics.
Poster #28

Identification of novel cardiac NKX2-5 co-factors using machine learning approaches

Ashley J. Waardenberg [1], Bernou Homan [2], Stephanie Mohamed [2], Richard P. Harvey [3], Romaric Bouveret [3]


NKX2-5 is a key transcription factor required for normal heart development and has been implicated in a range of congenital and adult onset heart diseases. Identifying genomic target regions for NKX2-5 is important for understanding how NKX2-5 controls regulatory networks involved in heart development and their susceptibility to dysregulation in congenital heart disease. We previously identified NKX2-5 genomic target regions in cultured HL-1 cardiomyocytes using the DamID method. Here, we apply the least absolute shrinkage and selection operator (lasso) algorithm to generate classification models of NKX2-5 target regions for predicting NKX2-5 co-factors based on known and de novo generated position weight matrices. We assess model performance using leave-one-out cross validation as well as a completely independent experiment. Our models, built on sequence composition alone, confirmed the importance of previously described co-factors, including GATA, HAND and TBX families, as well as a number of previously unrecognised co-factors, for predicting NKX2-5 binding in vivo. Predicted novel NKX2-5 co-factors were validated using the yeast-2-hybrid assay, providing new insights into NKX2-5 complexes involved in gene regulation. These findings demonstrate the utility of machine learning algorithms for detecting co-factors from DNA sequence data alone and the importance of protein-protein interactions for directing transcription factors to cis-regulatory elements.
Automated mining of gene perturbation data sets in Gene Expression Omnibus

Shu Kwan, Djordje Djordjevic, Joshua Ho

Victor Chang Institute

Over 60,000 gene expression data series are currently publicly available in NCBI's Gene Expression Omnibus (GEO). These data sets could be generated by a perturbation, case-control and/or time-series experiment that aim to investigate various biological questions. Our previous research showed that gene perturbation experimental data are particularly informative in terms of reconstructing causal gene regulatory relationships. This opened up the idea that large-scale mining of deferentially regulated genes in perturbation experiments will enable construction of causal gene regulatory networks (GRNs). The first step is to identify as many perturbation experimental data sets as possible. Harnessing the large amount of data and meta-data in GEO, we aim to build an automated pipeline to identify perturbation data-sets in GEO. Here we present a machine learning pipeline that automates this process. We extracted textual features from the meta-data of each GEO data series (such as title, summary, experimental design, sample names), and performed feature selection to identify the most informative features for identifying perturbation data sets. We then employed a support vector machine (SVM) with a radial basis kernel to perform supervised classification. Based on cross-validation of 141 manually curated data series, we show that our pipeline can accurately identify perturbation data series from others (AUROC of ~0.9). As an application, we use this automated pipeline to discover perturbation data sets that could be used to generate a GRN for mammalian heart development.
Ruminants are very successful herbivorous mammals, in part due to specialized forestomachs, the rumen complex, which facilitates the conversion of feed to soluble nutrients by micro-organisms. Sixteen gene expression clusters were identified from 11 tissues covering the sheep gastrointestinal tract (GIT), two stratified epithelial tissues and controls. The clustering of the rumen, skin and tonsil was driven by genes from the epidermal differentiation complex, and genes encoding stratified epithelium keratins and innate immunity proteins. Consistent with its high turnover rate the whole GIT showed a marked enrichment of cell cycle process genes (P=1.4E−46), relative to liver and muscle, with highest expression in the cecum followed by colon and rumen. The expression patterns of several membrane transporters (Chloride, Zinc, nucleosides, amino acids, fatty acids, cholesterol, bile acids and lactate) along the GIT was very similar in sheep, pig and humans. In contrast, short chain fatty acid uptake and metabolism appeared to be different between the species and different between the rumen and colon in sheep. The importance of nitrogen and iodine recycling in sheep was highlighted by the highly preferential expression of SLC14A1-urea (rumen), RHBG-ammonia (intestines) and SLC5A5-iodine (abomasum). The gene encoding a novel, poorly characterized member of the maltase-glucoamylase family (MGAM-like), predicted to play a role in the degradation of starch or glycogen, was highly expressed in the small and large intestines. The rumen is probably not a modified stomach or colon with a cornified epithelium, but may be a modified oesophagus with some liver-like and other specialized metabolic functions.
**Poster #31**

**Modeling Tumour Evolution to Inform Therapeutic Failure**

Luis Eduardo Lara-Gonzalez [1], David Goode [2]


Solid tumours are evolving ecosystems that display extensive intratumor heterogeneity, which is a marker of poor prognosis and drug resistance. In silico modelling coupled with sequencing is providing a framework to explore the evolutionary dynamics of tumour growth. Our aim is to extend reported models in the context of tumour response and relapse caused by chemotherapy.

We modelled tumour evolution as a discrete time branching process, starting with a single clone which spawns clonal diversity as driver mutations are acquired. New clones gain a slight proliferation advantage relative to their parent, and randomised fluctuations of their mutation rates. During tumorigenesis, proliferation is subject to a size-dependent penalties. Once the tumour attains a diagnosable size (1 to 3 billion cells) a new penalty is added to model the effects of chemotherapy, implemented as a mitotic phase-specific penalty impacting clonal proliferation rates.

Preliminary results show diverse tumour development rates, in which genomic instability promotes clonal diversification, thereby leading to oncogenic traits. This parallels with the morphological and physiological properties of inter and intratumour heterogeneity faced by clinicians. Despite these, few available tools study chemotherapeutic schemes.

The current aim, is to compare existing and hypothetical treatment regimes to determine which are best at suppressing the emergence of highly resistant clones—which would take several years to perform in the clinic. Exhaustive simulation, and data analysis are required to finalise our model. The ultimate goal is to better understand dynamics of tumour evolution in response to drugs to design more effective therapies.
Poster #32

Cloud-Based Testing-as-a-Service for Genomic Sequencing Pipeline Software

Michel Troup, Andrian Yang, Joshua Ho

Victor Chang Cardiac Research Institute

In the field of bioinformatics, genomic sequencing and subsequent comparison to a reference human genome can help identify points of variation between individuals. Genetic sequencing has important applications in the field of targeted genetic medicine, where the importance of correctly functioning analysis software cannot be underestimated.

The software that analyses the sequencing data can be a collection of proprietary, open-source, or custom-coded components that when combined together form an analysis "pipeline". Testing of such pipelines has lacked a readily accessible systematic approach to determining the correctness of the resulting output. In addition to this, smaller bioinformatics departments may also be challenged by the scale and complexity of implanting a suitable testing framework. The Victor Chang Cardiac Research Institute has done previous work in software testing that has been applied to certain parts of a genetic sequencing pipeline, which involved employing innovative methods such as Metamorphic Testing (MT). MT compares the output resulting from multiple related input test data, to determine if known relationships hold.

This research describes the implementation of a cloud-based testing framework that uses MT to test a supplied sequencing pipeline. The solution requires minimal user expertise in the field of either cloud computing or MT. The provision of an automated cloud-based testing framework enables a large number of test cases to be run in a scalable computational environment that can be tuned to the users speed and price requirements. The testing framework has been applied to a genetic sequencing pipeline comprised of industry-standard open-source components.
**Poster #33**

**Combining multiple methods to improve differential expression testing**

Luke Zappia [1], Fanny Grillet [2], Christina Molck [3], Kym Pham [3], Julie Pannequin [2], Graham Taylor [3], Frederic Hollande [3], Arthur L. Hsu [3]


RNA-seq has rapidly become the experiment of choice for many biological investigations. Perhaps the most common analysis conducted on RNA-seq data is differential expression testing in order to identify changes in gene product regulation between conditions. This type of analysis typically consists of several stages: alignment to a reference genome, summarisation by the feature of interest, normalisation to minimise batch effects, testing for differential expression and functional analysis to examine the biological implications. At each stage the analyst must grapple with the choice of which of many existing bioinformatics tools to use.

A range of factors can affect this decision. Is the tool easy to use and readily available? Has it been used in previously published work? Has it been designed for this data? Has it been based on good theory? Ideally choices should be made based on proven real-world performance, however this is hard to establish given the broad range of experiments and the lack of a ground truth to test against. Often analysts, particularly those new to the field, are left to pick a tool and trust that it is being used correctly and produces the required results.

Focusing on the testing stage I will demonstrate some of the ways in which results can be affected by tool choice through analysis of a complex colorectal cancer dataset. I will present my experience in selecting which tool to use and suggest that where time permits it may be preferable to run multiple methods in parallel and combine the results.
**Poster #34**

**Culture-independent genome sequencing of clinical samples reveals an unexpected heterogeneity of infections by Chlamydia pecorum**

Nathan Bachmann [1], Mitchell Sullivan [1], Martina Jelocnik [1], Garry Myers [2], Peter Timms [1], Adam Polkinghorne [1]


Chlamydia pecorum is an important global pathogen of livestock and it is also a significant threat to the long-term survival of Australia’s koala populations. This study employed a culture-independent DNA capture approach to sequence C. pecorum genomes directly from clinical swabs samples collected from koalas with chlamydial disease as well as from sheep with arthritis and conjunctivitis. Investigations into single nucleotide polymorphisms within each of the swab samples revealed that a portion of the reads in each sample belonged to separate C. pecorum strains, suggesting that all of the clinical samples analyzed contained mixed populations of genetically distinct C. pecorum isolates. Using the genomes of strains identified in each of these samples, whole genome phylogenetic analysis revealed that a clade containing a bovine and a koala isolate is distinct from other clades comprised of livestock or koala C. pecorum strains. Providing additional evidence to support exposure of koalas to Australian livestock strains, two ‘minor’ strains assembled from the koala swab samples clustered with livestock strains rather than koala strains. Culture-independent probe-based genome capture and sequencing of clinical samples provides the strongest evidence yet to suggest that naturally occurring chlamydial infections are comprised of multiple genetically distinct strains.
Poster #35

A cloud-based pipeline for processing single-cell transcriptomic data

Andrian Yang, Joshua Ho

Victor Chang Cardiac Research Institute

Our current understanding of genes and their regulation is based on studies on a population of cells, typically in the order of thousands or millions of cells. While the resulting analysis is no doubt informative, we are disregarding the transcriptional heterogeneity of individual cells within the population. For example, this confounds the analysis of cell differentiation, i.e. due to variable proportion of cells, and overlooks the spatial patterning within tissues. Thus, it is important that we examine transcriptomic variation at a single cell level for a clearer understanding of biological processes.

With the advances in single-cell isolation technology, we are now able to combine the power of single-cell and NGS technologies to generate high resolution genome-wide transcriptomic profiles of thousands of single cells simultaneously in one experiment. To illustrate the scalability issue, consider the problem of aligning the RNA sequencing data of 1,000 single-cell simultaneously. If alignment were to be performed sequentially (one sample at a time using one CPU), it will take 10.5 days to finish aligning all samples - a task that will only take 15 minutes if we can adaptively use 1,000 CPUs simultaneously.

We have implemented a cloud-based pipeline for processing single-cell transcriptomic data to solve this embarrassingly parallel problem. Our solution takes advantage of a new Big Data framework, Apache Spark, in order to process large quantity of transcriptomic data on a distributed computing environment which can be scaled to meet user's requirement.
Classification of true and false positive variants within a single sample of next generation sequencing.

Eleni Giannoulatou [1], Steven Phan [2], Joshua Ho [1]


The emergence of high-throughput Next Generation Sequencing technologies has resulted in multiple novel disease gene discoveries and has therefore revolutionised the clinical diagnosis of human genetic diseases. Sequencing the whole genome or exome of an individual can yield millions of variants ranging from single nucleotide variants (SNVs) to complex variants such as short insertions/deletions (indels) or copy number variation (CNV).

The discovery of disease-causing variants requires the application of methodologies with high accuracy and precision. However, many sources of false positives have been identified; false positive variants can be the result of sequencing errors, alignment errors as well as variant calling errors. Currently the identification of these false positive variants in a sequencing experiment of a single sample is very difficult. Genotyping algorithms can provide quality scores and other metrics for each variant called but each one of these metrics might not be sufficient to determine whether a variant is real. Many sequencing studies use large cohorts of control samples (or patient samples from different diseases) to help prioritizing variants found in a single disease sample.

We have developed a hierarchical logistic regression model that can help classifying variants found in a single sample into false positives (FPs) and true positives (TPs). By applying our model to an exome sequencing run of NA12878 and using the gold standard variant callset of Genomes in a Bottle, we were able to identify multiple features that can distinguish TPs from FPs. Our model achieved higher accuracy and precision than current state-of-the-art methods (AUC=86%).
**Poster #37**

**Characterising mutation-expression network relationships in multiple cancers**

Shila Ghazanfar, Jean Yee Hwa Yang

*The University of Sydney*

Data made available through large cancer consortia like The Cancer Genome Atlas make for a rich source of information to be studied across and between cancers. In recent years, network approaches have been applied to such data in attempts to uncover the complex interrelationships between mutational and expression profiles, but these methods lack direct testing for expression changes via mutation. In this pan-cancer study we analyse mutation and gene expression information by considering the networks generated through testing directly for differences in expression in association with specific mutated genes. On evaluation of these mutation-expression networks, we found that our generated networks were significantly enriched for known cancer-related genes, such as skin cutaneous melanoma (P<0.01 using Network of Cancer Genes 4.0). Our framework identified that while different cancers contained commonly mutated genes, there was little concordance among the associated gene expression changes between cancers. Our pan-cancer application of this approach suggests that while mutations are frequently common among cancer types, the impact they have on the surrounding networks via gene expression varies. Despite this finding, there are some cancers for which mutation-associated network behaviour appears to be similar, suggesting a potential framework for uncovering related cancers for which similar therapeutic strategies may be applicable. This work is integrated into an interactive R Shiny application, PAn Cancer Mutation Expression Networks (PACMEN), containing dynamic and static network visualization of the mutation-expression networks, and also featuring additional tools for further examination of network topology characteristics among cancers.
PTMOracle: a Cytoscape app for visualizing post-translational modifications and protein interaction networks

Aidan Tay, Chi Nam Ignatius Pang, Marc Wilkins

School of Biotechnology and Biomolecular Sciences, The University of New South Wales

The majority of proteins are modified by at least one post-translational modification (PTM). PTMs act as key regulators of protein activity, such as protein-protein interactions (PPIs). Many proteins are capable of carrying multiple modification sites, either by the same or different types of modifications. However, establishing detailed relationships between PTM combinations and PPIs remains a challenge due to combinatorial complexity. To address this, we have developed a new Cytoscape app, PTMOracle. With the PTMOracle app, users can co-visualize PTM and structural information in the context of PPI networks. We have developed several tools to query the PPI network for proteins of interest. This includes searching for proteins with pairs of modifications that are close to each other, count their frequency and quickly highlight these proteins in the Cytoscape network. We also map PTMs to the protein primary structure and co-locate them in domains, motifs and disordered regions, as well as allow comparative analysis of PTM patterns between proteins. We illustrate how the PTMOracle can be used for studying the inter-relationship between multiple PTMs in Saccharomyces cerevisiae. The PTMOracle is open-source and available on the Cytoscape app store: http://apps.cytoscape.org/apps/ptmoracle
Poster #39

Machine learning approaches for the prediction of leaderless secretory proteins in Arabidopsis thaliana

Andrew Lonsdale

The University of Melbourne

Proteomic studies of secreted proteins in plants often find proteins without a signal sequence. The occurrence of these proteins could be due to contamination from other parts of the cell during sample preparation, or as the result of unconventional (non-classical) protein secretion (UPS). Proteins without a signal peptide that undergo UPS are called leaderless secretory proteins (LSPs). Bioinformatics can be used to classify proteins found in such studies in order to distinguish between contaminants and LSPs. The accuracy of existing approaches is limited, as the bioinformatics tools commonly applied to plant secretomes in the search for LSPs were trained on mammalian data. This research uses various machine learning techniques, trained on plant protein data, to create a plant specific prediction tool for LSPs. Work in progress with Arabidopsis thaliana proteins as a case study will be presented.
PhosphoPICK: predicting kinase-specific phosphorylation sites through protein context and sequence

Ralph Patrick [1], Kim-Anh Le Cao [2], Bostjan Kobe [1], Mikael Boden [1]


The determinants of kinase-substrate binding can be found both in the substrate sequence, and the surrounding cellular context. Activating/mediating proteins, sub-cellular localization and cell-cycle specific expression of kinases and their substrates all contribute towards maintaining kinase-substrate specificity. While much work has focused on methods to predict kinase-specific phosphorylation sites from sequence, the presence of a valid binding pattern is no guarantee that a kinase will come into contact with the protein. Few studies have attempted to leverage the cellular context that kinases operate in, and incomplete coverage and variable certainty amongst available context data means incorporating context features into a model is non-trivial.

We present Bayesian network models that integrate these sequence and context factors in order to firstly determine the probability of a kinase targeting a query protein, and secondly, identify the most likely binding sites within the protein's sequence. The models integrate known kinase-substrate relationships, protein-protein interactions, and protein abundance across the cell cycle. Kinase binding patterns are represented using position-specific amino acid and k-mer frequencies. Our method, PhosphoPICK, shows reliable prediction accuracy, obtaining an average AUC of 0.81 for kinase substrate prediction. When comparing its ability to predict kinase-specific phosphorylation sites with alternative methods, we find PhosphoPICK obtains superior performance for most comparisons made particularly with methods operating on sequence alone. The results demonstrate that PhosphoPICK can ably integrate context and sequence, and provide a robust description of the regulation of kinase-protein phosphorylation.
Poster #41

One in a million: choice of immune cell receptor dictates T-cell fate

Zoe Welham [1], Stuart Archer [2], Stephen Daley [1]


Every T-cell displays on its surface a unique T-cell receptor (TCR), produced by recombination of randomly selected TCR components at the DNA level, to form a mature TCR-coding gene. Hundreds of latent TCR components are present in the TCR locus in immature T-cells prior to recombination, thus many millions of combinations are possible for each mature T-cell clone, which underlies the immune system's adaptability to target new pathogens. Here, diversity of mature TCRs, and their effect on T-cell fate, was examined by targeted deep sequencing. This type of TCR profiling data represents a statistical challenge, as it is extremely skewed and sparse. Even in mice where TCR diversity was genetically constrained, the majority of TCRs were still unique to each animal. Nonetheless, by studying similarities between receptors, consistent effects of individual mature TCRs and their constituent components were inferred. TCR profiling from sub-populations of T-cells demonstrated that use of certain TCR components definitively directs the T-cell towards particular cell fates. Importantly, some TCRs strongly and consistently bias the T-cell towards reacting to self-antigens, and these clones are removed by a process of T-cell vetting and programmed cell death in healthy individuals. This process is likely to be perturbed in individuals with autoimmune disease, unleashing T-cells with TCRs that react against endogenous molecules in the body.
Poster #42

Benchmarking Membrane Protein Helix Predictions

Emma Rath [1], Bret Church [1], Dominique Tessier [2], Alexander Campbell [3], Hong Ching Lee [4], Tim Werner [5], Noeris Salam [6], Lawrence Lee [7]


Helical membrane proteins are critical proteins, yet the methods for their three-dimensional structure determinations have not been as yielding as for soluble proteins. Predicting the location of membrane helices in protein amino acid has been considered over several decades. Comparisons have also been limited to global assessments, as specialised benchmarks for predicting specific classes of membrane proteins were inadvisable and not previously performed.

We present a comprehensive benchmark server that implements 53 approaches, albeit some are alternative implementations of the one method. It uses recent high resolution protein structural data to provide a comprehensive assessment of the accuracy of membrane helix prediction methods. Further, predictions can be uploaded permitting the comparison of novel methods. Benchmark metrics include sensitivity and specificity of predictions for membrane helix location and orientation, and many others. Interest may range from focus on determination of a particular location or specific membrane boundary in a protein, to ascertaining data for deriving an overall three-dimensional model. Customised evaluations, such as assessing prediction method performances for specific helical membrane protein subtypes is also possible.

The best performing method depends on the measure being benchmarked. Nevertheless, the OCTOPUS membrane helix prediction method is consistently one of the highest performing methods across all measures in the benchmarks. This benchmark server can be employed to determine the prediction methods most suitable for the specific purpose. This work represents the only current and comprehensive benchmark tool for evaluating transmembrane helix prediction methods, for global or specialised assessments.

Exploring the evolution of Mycobacterium ulcerans and emergence of Buruli ulcer in Australia with whole genome sequencing

Andrew Buultjens [1], Torsten Seemann [2], Janet A. M. Fyfe [3], Caroline Lavender [3], Koen Vandelaanoot [4], Paul D. R. Johnson [5], John Hayman [6], Kirstie M. Mangas [1], Jessica L. Porter [1], Timothy P. Stinear [1]


A significant obstacle for controlling the spread of Buruli ulcer is that the mode of transmission and infectious reservoir of Mycobacterium ulcerans - the causative agent - remains poorly understood. To help address these key research questions we retrospectively investigated the molecular epidemiology of this pathogen in the Australasian region, paying special attention to south-eastern Australia, the focus where Buruli ulcer was first described. Associations between phylogeny and geography were examined by comparing geospatial variables against a bacterial population phylogenomic tree. The trees were inferred from alignment of core genome single nucleotide polymorphisms revealed by whole genome sequencing of 110 Australasian M. ulcerans isolates. Short read alignment against the Agy99 reference genome, revealed 5,752 nucleotides of variation (0.10% of reference) among the Australasian isolates and 241 (0.004% of reference) among Victorian isolates. The analyses showed there has been a significant westward radiation of M. ulcerans across south-east Australia, a phenomenon that fits with the well-documented spread of Buruli ulcer in this region.

This study has provided high-resolution insights into the population structure and evolution of M. ulcerans in south-east Australia, and show that pathogen spread is a prerequisite for the expansion of Buruli ulcer endemicity, rather than emergence of a quiescent local clone. Under this scenario, public health interventions aimed at halting pathogen spread would provide a means of disease prevention. This research also provides the basis for testing specific hypotheses aimed at elucidating the natural ecology of M. ulcerans, which in turn may allow for more effective disease control.
Applying cheminformatics strategies to design better immunogenic T cell epitopes

Ari Hardianto, Shoba Ranganathan

Macquarie University

Adaptive immune responses are governed by major histocompatibility complexes (MHC) binding to specific short antigenic peptides and then this peptide bound major histocompatibility complex (pMHC) being recognized by the T cell receptor (TR) which activates the T cells. The use of critical sequence-structure-function information to understand the principles underlying MHC specific peptide binding is well established and the focus is now on understanding TR recognition of pMHC complexes. Three-dimensional X-ray structures of pMHC complexes bound to the TR that are today characterized in good numbers facilitate structural analysis further. It is thus possible to predict better T cell epitopes for vaccine design by utilizing information derived from available experimental structures. We have used a cheminformatics approach to predict peptide epitopes which bind stronger than the native epitope. We have systematically carried out single and double amino acid substitution of residues in the epitopes from experimental MHC-peptide-TR crystallographic structures. For each modified peptide, we carried out docking simulations of the mutated peptides to the original MHC and then to TR. The results show that it is possible to enhance both peptide-MHC binding as well as pMHC-TR binding. The results have implications for rational epitope-based vaccine design.
Poster #45

Functional annotation of the truffle proteome

Mohammad Islam, Abidali Mohamedali, Mark Baker, Shoba Ranganathan

Macquarie University

The black Perigord truffle (Tuber melanosporum Vittad.) is a highly prized food today, with its unique scent (i.e., perfume) and texture. Despite these attributes, it remains relatively poorly studied, lacking "omics" information to characterize its biology and biochemistry, especially changes associated with freshness and the proteins/metabolites responsible for its organoleptic properties. In this study, we have functionally annotated the truffle proteome from the 2010 T. melanosporum genome comprising 12,771 putative nonredundant proteins. Using sequential BLAST search strategies, we identified homologues for 2587 proteins with 2486 (96.0%) fungal homologues (available from http://biolinfo.org/protannotator/blacktruffle.php). A combined 1D PAGE and high-accuracy LC-MS/MS proteomic study was employed to validate the results of the functional annotation and identified 836 (6.5%) proteins, of which 47.5% (i.e., 397) were present in our bioinformatics studies. Our study, functionally annotating 6487 black Perigord truffle proteins and confirming 836 by proteomic experiments, is by far the most comprehensive study to date contributing significantly to the scientific community. This study has resulted in the functional characterization of novel proteins to increase our biological understanding of this organism and to uncover potential biomarkers of authenticity, freshness, and perfume maturation.
Lipid rafts are dynamic, cholesterol-rich membrane micro-domains that regulate complex molecular interactions. Many tumor suppressors and oncogenes localize to lipid rafts, including the tumor suppressor, opioid binding protein cell adhesion molecule (OPCML), which is frequently inactivated in epithelial ovarian cancer. Since OPCML is an extracellular GPI-anchored protein with no cytoplasmic domains, we hypothesized that it acts by modulating lipid raft composition. To dissect raft-mediated mechanism of tumor suppressor effect of OPCML in ovarian cancer, and compare with other cancer raft proteomics studies, subcellular quantitative proteomics coupled with network analysis was performed on ovarian cancer SKOV3 cells expressing OPCML wild type or partially active P95R mutant. RaftProt database was used for comparative analysis with other cancer rafts. Enrichment and interaction network analysis revealed altered lipid raft-cytoskeleton interaction upon expression of OPCML. Interestingly, wild type but not P95R OPCML significantly decreases vimentin and desmin expression in lipid rafts. Independent meta-analysis of three lipid raft proteomics datasets modeling progression in breast cancer, renal cell carcinoma and melanoma showed that over half of the commonly-altered lipid raft proteins were cytoskeleton associated. Taken together, these results suggest enhanced cytoskeletal-membrane raft interaction as a common feature of aggressive progression in tumors.
De novo Assembly of the Koala Genome Using Short Reads

Zhiliang Chen [1], Ryan Mathew Salinas [1], Matthew Hobbs [2], Ana Pavasovic [3], Andrew G King [2], Peter Prentis [4], Mark Eldridge [5], Don Colgan [5], Adam Polkinghorne [6], Cheyne Flanagan [7], Amber Gillett [8], Jon Hanger [9], Kathy Belov [10], Peter Timms [6], Rebecca Johnson [11], Marc Wilkins [12]


Background
The koala, Phascolarctos cinereus, is a biologically unique and evolutionarily distinct Australian marsupial. The koala transcriptome has been recently sequenced and assembled in 2014 by the Koala Genome Consortium (Hobbs et al. BMC Genomics 2014, 15:786). The goal of this study was to sequence and generate a draft genome sequence of the koala, as a complement to the existing transcriptome, to provide information for future genetic and conservation studies of the koala.

Results
The genome sequence was obtained from a female koala 'Pacific Chocolate', from Port Macquarie (New South Wales, Australia). A total of 3.35 billion paired-end reads and 732 million mate-pair reads, from 4 libraries of different lengths, were generated using Illumina HiSeq2000. The paired-end reads were first de novo assembled by ABySS, the mate-pair reads were then used for scaffolding by SSPACE. A 3.4 gigabase draft genome sequence of koala, comprising 24,932 scaffolds longer than 2000 bp, was successfully generated.

By use of CEGMA (Core Eukaryotic Genes Mapping Approach), the assembled scaffolds are estimated to cover approximately 95.97% of the koala genome, including 84.68% of the core eukaryotic genes. By aligning to the koala transcriptome to the genome and by comparison to the complete human proteome, 6,225 complete genes and their gene structure were confidently predicted.

Conclusions
In addition to the published transcriptome, this genomic dataset will provide a unique resource for future research into Koala and marsupial biology.
RNA-Seq technology is capable of detecting all forms of RNA transcribed from the genome (including mRNA coding for proteins, microRNA, snoRNA, lincRNA, and mtRNA). Until recently RNA-Seq experiments were generally conducted using a single stranded protocol. Using this protocol it was not possible to detect whether transcription had occurred from the forward or reverse strand of DNA. However, recent developments in sequencing chemistry have made it feasible to routinely identify the DNA strand from which any RNA transcript has originated (strand-specific protocol). Here we compare the outcomes of read mapping and feature counting using both a non-stranded and a strand-specific protocol. Our data is derived from 2 studies using human primary hematopoietic cells. Sequencing was performed using the Illumina NextSeq platform and the RNA libraries were prepared using Illumina's TruSeq Stranded Total RNA kit with Ribozero pulldown. Over 40 million 75 bp paired-end reads were produced per sample. Reads were mapped to the human reference (Ensembl GRCh37) using Tophat2 and reads mapping to features were counted using HTSeq. Here we present a breakdown of the types of RNA, coding and non-coding RNAs detected using the strand-specific and non-strand specific protocols and compare the outcomes of differential expression analysis using these protocols.
Poster #49

Novel genetic markers define a subgroup of pathogenic Escherichia coli strains belonging to the B2 phylogroup

Nandan Deshpande [1], Marc Wilkins [1], Hazel Mitchell [2], Nadeem Kaakoush [2]

[1] Systems Biology Initiative, School of Biotechnology and Biomolecular Sciences, The University of New South Wales
[2] School of Biotechnology and Biomolecular Sciences, The University of New South Wales

The B2 phylogroup of Escherichia coli contains important pathogens such as extra-intestinal pathogenic, adherent-invasive, and uropathogenic strains. In this study, we used comparative genomics and statistical methods to identify genetic variations, which define a subset of pathogenic strains belonging to the B2 phylogroup. An initial proof of concept analysis indicated that five of the 62 E. coli strains available in the KEGG database showed close association with B2 adherent-invasive E. coli, forming a subgroup within the B2 phylogroup. The tool, kSNP which uses a k-mer approach, and the statistical phenotype prediction tool PPFS2 were then employed to identify 29 high-resolution SNPs, which reaffirmed this grouping. PPFS2 analysis also provided indications that the clustering of this subgroup was highly consistent, and thus, could have a strong phenotypic basis rather than being only evolutionary. Protein homology analyses identified three proteins to be conserved across this subgrouping, two CRISPR-Cas proteins and a hypothetical protein. Functional analyses of these genetic and protein variations may provide insights into the phenotype of these strains.
Identifying subtle gene-expression signals in the residual memory of induced pluripotent stem cells

Fernando J Rossello [1], Sue Mei Lim [1], Paul F. Harrison [2], Jose M. Polo [1], David R. Powell [2]

[1] Department of Anatomy and Developmental Biology, Monash University, Victoria 3800, Australia

Induced pluripotent stem cells (iPSCs) share many features with embryonic stem cells. Despite this, recent studies have purported that low passage iPSCs maintained a residual "memory" originating from the donor somatic cells. We generated 9 different iPSC lines from the three germ layers and concurrently performed RNA-Seq and in vitro differentiation assays back into derivatives of the mesoderm, ectoderm and endoderm lineage. Whole transcriptome sequencing analysis of iPSCs aimed to identify unique gene expression patterns within each germ layer that could be guiding the differentiation of iPSCs towards derivatives of each germ layer.

The dataset proved challenging to analyse due to its inherent experimental structure of 3 donor mice, 9 cell types across 3 germ layers. Initial analysis showed a strong effect from the individual mouse but no consistent patterns attributable to the germ layer of origin was evident in clustering nor statistical testing after removing the identified mouse batch-effect. RUV normalisation was used to remove both the batch effect of the mouse donor and the cell-type effect while leaving the more subtle effect from the germ layer. Clustering of this normalised data showed clear grouping of the 3 germ layers. Subsequent differential gene expression testing successfully produced several differentially expressed transcripts with significant difference among the germ layers compared.

Differentiation assays agreed with the analysis above and demonstrated that low passage iPSCs retained a “cellular memory” of the somatic cell from which they were derived from and that germ layer of origin does appear to bias the differentiation outcome.
Poster #51

Detecting Differential Exon Usage in RNA-Seq Experiments

Yunshun Chen, Gordon Smyth
WEHI

Alternative splicing is a process where exons are differentially combined or skipped, resulting in multiple protein isoforms encoded by a single gene. It generates diverse transcripts and provides an opportunity for gene regulation. The proteins translated from alternatively spliced mRNAs will contain different or even opposite biological functions. It has been discovered that more than 95% of human multi-exon genes express multiple splice isoforms. Alternative splicing can affect various functions in cellular processes, tissue specificity, developmental states and disease conditions.

High throughput technologies, such as RNA-Seq, have been proven to be very powerful tools in studying alternative splicing. Meanwhile, it is also very tempting and challenging for statisticians and bioinformaticians to develop statistical methods for detecting alternatively splicing events. Here we present a method for testing for differential exon usage between different experimental conditions or different genetic backgrounds. In particular, we compare the change of the expression level of each exon to the change of the expression level of the gene containing that exon under a certain comparison. Similarly to the gene level analysis of RNA-Seq data, we use negative binomial distribution to model exon counts. An empirical Bayes information sharing strategy is applied to estimate the variation between biological replicates. Statistical tests are performed at both the exon level and the gene level. The method has been implemented in the edgeR package.
**Poster #52**

**De novo identification of differentially methylated regions in the human genome**

Tim Peters [1], Mike Buckley [2], Aaron Statham [3], Susan Clark [1], Peter Molloy [4]


The identification and characterisation of differentially methylated regions (DMRs) between phenotypes in the human genome is of prime interest in epigenetics. We present a novel method, DMRcate, that fits replicated methylation measurements from the Illumina HM450K BeadChip (or 450K array) spatially across the genome using a Gaussian kernel. DMRcate identifies and ranks the most differentially methylated regions across the genome based on tunable kernel smoothing of the differential methylation (DM) signal. The method is agnostic to both genomic annotation and local change in the direction of the DM signal, removes the bias incurred from irregularly spaced methylation sites, and assigns significance to each DMR called via comparison to a null model.

We show that, for both simulated and real data, the predictive performance of DMRcate is superior to those of Bumphunter and Probe Lasso, and commensurate with that of comb-p. For the real data, we validate all array-derived DMRs from the candidate methods on a suite of DMRs derived from whole-genome bisulfite sequencing called from the same DNA samples, using two separate phenotype comparisons.
RNA-seq mixology: designing realistic control experiments to compare protocols and analysis methods

Aliaksei Holik, Ruijie Liu, Charity Law, Marie-Liesse Asselin-Labat, Matthew Ritchie

The Walter and Eliza Hall Institute of Medical Research

Recent publication of the SEQC project (Nature Biotechnology, August 2014) has highlighted the utility of control experiments to benchmark different platforms and analysis workflows. Current RNA-seq control data sets are however atypical of real experimental situations in that they are very clean (i.e. have no biological variation) and simulate large, widespread expression changes by comparing RNA from different tissues (e.g. brain versus universal reference). To obtain a more realistic control data set, we designed an experiment that uses two different cell-lines of the same cancer type. These were mixed in the usual way to induce predictable fold-changes across the series. We also included samples that were intentionally degraded to simulate poor quality samples, which are commonly observed in practice. In this presentation we demonstrate the usefulness of this data set for benchmarking different library preparation kits (total RNA versus polyA mRNA) and various RNA-seq analysis pipelines for detecting differential gene expression, differential splicing and gene fusion events.
Promoter mutation hotspots in cancer genomes are associated with increased transcriptional activity and reduced access by DNA repair machinery

Rebecca C. Poulos [1], Dilmi Perera [1], Anushi Shah [1], Dominik Beck [1], John E. Pimanda [2], Jason W. H. Wong [3]


With increasing numbers of whole cancer genomes available, there is significant interest in the identification of causal mutations that affect gene regulation. Recent pan-cancer analyses of mutations from hundreds of whole cancer genomes have identified numerous hotspots of somatic mutations in promoter elements. Nevertheless, few promoter mutations appear to be functional as they show no significant association with altered gene expression. While recurrence of mutations is often associated with positive selection, we have sought to identify an alternative explanation for the lack of correlation between promoter mutations and gene expression. By analysing somatic mutations from over 1,000 cancer genomes across 14 cancer types, we have found that mutations in many cancers, particularly melanoma, lung and ovarian cancers, show increased density of somatic mutations at core promoters. In melanoma, the increase in mutation density is correlated with overall sample mutation rate and the presence of a UV signature. To determine the cause of increased mutation density, we compared the regulation and expression of genes with and without promoter mutations, finding that mutated core promoters are associated with stronger transcriptional activity and transcription factor binding sites. This suggests that transcriptional regulatory mechanisms at the core promoter may play a role in abrogating DNA repair. By analysing genome-wide maps of nucleotide excision repair (NER), we have found that NER activity is decreased within the DNase I hypersensitive centre of active gene promoters, a phenomenon which inversely mirrors the increase in somatic mutation density. Taken together, our analyses have uncovered the presence of a previously unknown mechanism which links transcription initiation and DNA repair. We thereby implicate localised differential DNA repair as the underlying cause of the somatic mutation hotspots that we have observed at gene promoters of cancer genomes.
Poster #55

Probabilistic Evolutionary Kernel Function for Analysis of Protein Sequence-function Relationships

Julian Zaugg, Mikael Boden, Yosephine Gumulya

Historically protein engineering methods, such as directed evolution, have faced issues in how to efficiently explore sequence space - multiple simultaneous mutations exponentially increase the number of possible sequence variants to be screened, or evolutionary trajectories that can be taken, in each round of an engineering experiment. A number of statistical methods have been developed to capture and analyse the rich information inherently available in the sequences of existing proteins, allowing protein engineers to make more informed decisions regarding the optimal experimental strategy and protein design. However these methods often assume mutations have simple additive effects on protein fitness, treating them independently and ignoring epistatic influences. Such assumptions are inappropriate when trying to model complex protein properties, e.g., enantioselectivity or activity.

Here we present an ongoing investigation into how to model such higher-order relationships between mutation positions, ultimately with the goal of predicting functional properties of uncharacterised mutants and to provide insight into possible preferences for individual evolutionary trajectories through sequence space. Using a probabilistic evolutionary kernel function, we have trained a Support Vector Machine (SVM) to predict the enantioselectivity of mutants from a directed evolution experiment of an Aspergillus niger epoxide hydrolase, achieving a Spearman rank correlation score of 0.82 ± 0.02 from ten-fold cross-validation. Such a result is comparable to that obtained with alternative statistical methods, i.e., pls-regression. Performing a kernel principle component analysis on the resulting feature space illustrated the sensitivity of the kernel function to variations in both amino acid content and ordering of mutations. Such sensitivity could assist in the identification of regions of this evolutionary space that may be displaying higher epistatic influence or, when property values are mapped to the first few principle components, where desirable properties may be found.
Poster #56

Modelling Tcf7l2 DNA Binding

Tomasz Szczesnik, Rich Sherwood, Joshua Ho

The binding location of transcription factors throughout the genome, along with their effect on surrounding gene expression, varies significantly according to cell type. Specifically, the transcription factor Tcf7l2 binds a largely disparate set of sites between mouse embryonic stem cells and intestinal endoderm cells. We hypothesise this is due to interactions with surrounding transcription factors, and hence dependent on the combination of transcription factor binding sites present, along with the relative orientation and spacing between them. To probe these differences in an unbiased way, we use a large library of 100bp sequences that vary in the presence and position of various cofactors, or show differential binding between cell types. These are integrated into specific loci within mouse embryonic stem cells using CRISPR/Cas9, and assayed for Tcf7l2 binding through DamID. For predicting and designing Tcf7l2 binding in new sequences, we would like to learn the dependencies present between these cofactors, while avoiding the combinatorial explosion and overfitting caused by fitting a parameter to every arrangement. To alleviate this, we abstract over possible combinations by using a kernel based on the "distance" between two sequences, and rely on a gaussian process to estimate the relevant binding affinities.
Making a genome sequence work

Alexie Papanicolaou [1], NESCent Manual Curation Working Group [2]


The molecular biology community has overcome a major bottleneck: access to near-complete genome sequence from multiple species. Such diversity of sequence information underpins a new type of genomic capability and greatly benefits the comparative genomics field. How do we, however, can make best use of this information? Here, outline the bioinformatic tools and best practices we developed and used: First, best-practices for designing a sequencing strategy for insects. Second, the Just_Annotate_My_Genome, a platform that allows nascent bioinformaticians to structurally annotate genomes at the same standard provided by NCBI. Third, I will showcase DEW and the Just_Annotate_My_Proteins software, a visualisation platform for functional annotation via gene expression and sequence comparisons that biologist without computational training can actually use. Fourth, the progress of curating major genome sequencing efforts, namely the Mediterranean fruit fly, Helicoverpa armigera, H. zea and Heliconius erato genome projects. The software is open-source and continually under development and therefore feedback on improvements and future work will be greatly appreciated.
Completing microbial draft genomes with Oxford Nanopore sequencing data.

Son H Nguyen, Minh D Cao, Devika Ganesamoorthy, Lachlan Coin

IMB, UQ

The abundance of repetitive elements create a challenge in assembling genomes using high-throughput sequencing data. Recent advanced sequencing technology allows the generation of ultra-long reads giving potential to constructing genome scaffolds that are closer to completion. However, data of these technologies are highly erroneous. In this work, we present a hybrid assembly approach that takes advantage of long spanning reads from Oxford Nanopore MinION for improving draft assemblies of bacteria. Throughout experiments with different microbial datasets, we show that contigs from assemblies on MiSeq Illumina data can be further connected toward completion by using only raw Nanopore long reads as backbone. Furthermore, our algorithm is able to work efficiently with only 30-fold or less coverage of this noisy data on a quick processing time. Our tool provides a resources-saving but reliable method to help reduce the complication of microbial genome assembly problem.
Efforts to move high throughput sequencing into the clinic must confront many challenges including meeting clinical standards for cost, reproducibility, quality, ethical and privacy considerations. In this work we present Cpipe, an open source bioinformatics analysis pipeline created by the Melbourne Genomics Health Alliance as a clinical grade pipeline for analysis of exome and targeted sequencing data. The pipeline is based on the Bpipe toolkit and includes support for simultaneously analysing multiple diagnostic target regions, prioritised gene sets, regions blacklisted against incidental findings, automatic exclusion of sequencing artefacts and population variants, PDF provenance and quality reports. It allows genes to be prioritised at both a disease cohort and individual level and combines this with output from annotation tools to produce a clinically interpretable report that is available as an Excel spreadsheet or optionally can be imported into any LOVD (Leiden Open Variant Database) instance for curation. Being based on Bpipe, it offers powerful features for running many samples in parallel either on a cluster or dedicated computing resources as well as full traceability through auditable provenance of every file.
Poster #60

Transcriptome Analysis Of Human Rett Syndrome Brain Identifies Reduced Expression Of Complement Complex C1Q Genes

Peijie Lin, Laura Nicholls, Zhiming Fang, Timothy Amos, Richard Edwards, Irina Voineagu

MECP2, the gene mutated in the majority of Rett syndrome cases, is a transcriptional regulator that can activate or repress transcription. Although the transcription regulatory function of MECP2 has been known for over a decade, it remains unclear how transcriptional dysregulation in the brain of patients carrying MECP2 mutations leads to the neurodevelopmental disorder. Notably, little convergence has been observed between the genes abnormally expressed in the brain of Rett syndrome mouse models and those identified in human studies. Here we carried out a comprehensive transcriptome analysis of human brain tissue from Rett syndrome cases using two independent methods: RNA-seq and microarrays. By effectively controlling covariates, including the cellular composition of brain tissue samples, we identified over two hundred differentially expressed genes, and identified the complement complex C1Q genes (C1QA, C1QB and C1QC) as a point of convergence between gene expression changes in human and mouse Rett syndrome brain.
**Poster #61**

**Evaluating microbial community deconvolution when leveraging high-throughput chromosome conformation capture (3C) techniques (HiC, meta3C).**

Matthew Z DeMaere, Aaron E. Darling

The sampling of information by way of chromosome conformation capture (3C) has been used effectively in chromosome spatial organisation studies and the inferential modelling of chromatin 3D structure in eukaryotes and more recently prokaryotes. In a metagenomic setting, capturing state prior to cell lysis provides crucial information on DNA cellular co-locality which is lost in traditional metagenome shotgun DNA sequencing. The deconvolution of microbial metagenomes with 3C data has already been explored, while further applications are readily envisioned (e.g. fine-scale SNP linkage analysis). Much of the field's advance will fall into the computational domain, demanding a firm understanding of the applicability of existing algorithms and the development of new ones. Here, the accuracy and precision of two previously applied hard clustering algorithms (MCL, Louvain) and two novel soft clustering algorithms (SR-MCL, OclustR) were assessed by way of an end-to-end 3C/WGS microbial community simulation pipeline. We sought to understand whether individual genomes could be resolved from mixtures of closely related strains using existing techniques, exploring the experimental parameter space for two simple phylogenetic topologies (ladder, star). Under control were: community parameters (evolutionary divergence, relative abundance), important experimental parameters (WGS and 3C read depths) and algorithm related. Quantification of accuracy and precision of the resulting cluster solutions was measured by external indices, using both hard-clusters aware (F-measure, V-measure) and soft-clusters aware (FBcubed). The interplay between evolutionary divergence and abundance was found to be the primary factor affecting solution quality and was not counteracted significantly by increasing depth of coverage.
**Poster #62**

Identification of dysregulated microRNAs in soft-tissue sarcomas through integrative expression analysis

Regina Ryan, David Goldstein, Philip J. Crowe, Jia-Lin Yang, Jason W. H. Wong

Soft-tissue sarcomas (STSs) are malignant tumours that occur as a result of aberrant differentiation of mesenchymal stem cells (MSCs). Current theories suggest the loss of post-transcriptional control in gene expression leads to the development of sarcomas. However, the genetic mechanisms initiating STSs are poorly understood. Post-transcriptional regulators, specifically microRNAs, are involved in differentiation and aberrantly expressed in STS. The aim of this investigation was to identify microRNAs related to development and progression of STSs as well as patient outcomes. High-throughput sequencing was used to characterise microRNA expression profiles of MSCs undergoing adipogenesis (GEO20699). Differential expression (fold change >2, p-value <0.05) was used to determine microRNAs of interest. Log-rank test was performed on 109 candidates in 251 STS patients from The Cancer Genome Atlas. Upregulation of miR-148b, miR-186, miR-192, miR-335, miR-424, and miR-503 correlated with high patient fatality. Eleven differentially expressed and predicted gene targets were mapped to 145 pathways, several of which are associated with tumorigenesis. The analysis identified key microRNAs in sarcoma based on poor patient prognosis. Further validation using in vitro techniques may provide evidence to suggest microRNA involvement in STS pathogenesis. Thus, the microRNAs identified have the potential as biomarkers, prognostic and therapeutic tools for sarcoma patients.
Characterization of skin and wound microbial communities in diabetic patients using a Random Forest Classifier

Sunaina Bansal [1], Catherine Burke [1], Melissa Gardiner [2], Mauro Vicaretti [3], Jill Sparks [4], Elizabeth Harry1 [1], Aaron Darling [5]


Diabetes is one of the fastest growing chronic diseases, causing a range of medical problems including the development of chronic non-healing wounds. These wounds are costly to treat and impact severely on a patient's quality of life. The purpose of this study was to assess whether chronic condition such as diabetes is likely to affect the structure and function of the skin microbial community, with possible downstream effects on infection and wound healing. A group of 8 diabetic subjects with chronic wounds and 8 healthy subjects was selected. We carried out skin & wound microbiome profiling using 16S rDNA amplicon sequencing and taxonomic analysis for samples obtained over a period of 12 weeks. Differential analysis and a random forest classifier revealed the skin microbial communities that differed significantly between healthy and diabetic subjects. Certain bacterial relative abundances were found to be characteristic of the taxonomical structure of diabetic and healthy skin.
Poster #64

Scrutiny of parametric models: a crystal ball for anticipating inconclusive data

Jason M. Whyte

MASCOS and ACEMS at The Department of Mathematics and Statistics, The University of Melbourne

A parametric mathematical model provides a means of predicting the behaviour of unobservable aspects of a biological system. First we must determine parameter values for which predictions best reproduce observations. This exercise may return multiple estimates for one or more parameters, and distinguishing between these may be impossible. This is problematic when alternative estimates lead to very different predictions, as we cannot use the model with confidence. This renders the experimental study an unproductive use of time and materials. However, inspection of the mathematical model prior to data collection may anticipate non-uniqueness of parameter estimates. This result is likely to occur if the model lacks the property of global a priori identifiability.

We consider biomolecular interactions studied via optical biosensor experiments and a state-space model in which parameters represent rate constants. Inspection of this model shows that it is not globally a priori identifiable. However, the result suggests a reformulation that yields a globally a priori identifiable model. While it is not guaranteed that parameter estimation will return a unique estimate for this new model, it is at least possible. This result demonstrates the benefit of including a scrutiny of model properties in the planning for experiments.
The muddle of metagenome assembly/binning from an enriched nitrifying community.

Andrew Elohim Laloo, Qilin Wang, Inka Vanwonterghem, Philip Hugenholtz, Zhiguo Yuan, and Philip L Bond.

Advanced Water Management Centre, The University of Queensland, St. Lucia, Brisbane, QLD 4072, Australia

An enriched nitrifying community grown in a sequencing batch reactor was fed with increasing ammonia concentrations over time. The bacterial family Nitrosomonadaceae constitutes approximately 60% of the microbial community and is the main ammonia oxidising bacteria (AOB) converting the ammonia in the feed autotrophically to nitrite. Metagenome assemblies from this enriched and simplified nitrifying reactor reveal the difficulty in getting good bins from this abundant Nitrosomonadaceae family using normal metagenomic pipelines. This is not uncommon especially in more enriched microbial communities or ones that have a large spread of coverage. Algorithmic limitations of assemblers could also prevent genomes having strain heterogeneity/variation from assembling properly. Herein, we describe the use of different strategies and tools to examine the effect it has on metagenome assembly and binning of species from this dominant family.