COMBINE 2022 Symposium

Symposium Abstracts

Ten-minute presentations

Modelling Systematic Inter-Replicate Bias in the Integration of ChIP-seq experiments

Oliver Hughes

University of Queensland

Large databases such as ENCODE enable the integration of multiple datasets from functional genomics assays such as ChIP-seq. Integration of these datasets aims to mitigate noise and systematic factors influencing individual experiments. Unlike other technologies such as RNA-seq, there are no established strategies for the optimal integration of independent ChIP-seq datasets. Externally validated sites cannot be relied upon to be discovered across every replicate, and recurring false positive signals can be present across replicates. These issues lead to ambiguities in the ideal permissiveness of integrative methods. To address these issues, we use statistical evidence from a range of experiments to demonstrate systematic patterns of disagreement between ChIP-seq replicates (n=16 comparisons). We demonstrate that these patterns occur both within and between labs, but are exacerbated by inter-lab variability. Further, by constructing a linear sequence of binding sites or “peaks”, we show that this variability forms regions of disagreement that can be identified using Hidden Markov Models. Motif analysis of regions of disagreement reveal that they can be either regions of increased false positives, or reproducible regions worth salvaging. More data is needed for robust interpretation of these patterns, however their potential utility is clear. Overall, our findings indicate that systematic inter-replicate bias exists in a regional form, is readily modellable, and can be used to increase precision in the integration of multiple ChIP-seq replicates.

Distinguishing transcriptional and genetic heterogeneity in metastatic prostate cancer

Sirui Weng

Peter MacCallum Cancer Centre

Prostate cancer that develops treatment resistance and metastasizes is known as metastatic castration-resistant prostate cancer (mCRPC), with patients surviving only 1-2 years. Disease heterogeneity is critical in explaining the ineffectiveness of treatments for these patients. While genetic heterogeneity caused by the accumulation of somatic mutations drives tumour evolution, transcriptional plasticity also likely plays a role in the differential response to treatment. However, how much genetic mutation or transcriptional plasticity contributes to mCRPC heterogeneity remains unknown. We performed whole-genome sequencing and single-nucleus RNA sequencing (snRNAseq) on seven mCRPC lesions taken from two patients from the CASCADE rapid autopsy program. We calculated the effect of copy number alterations (CNA), point mutations, and gene fusions on

differential gene expression (DE) across lesions and tumour subclones. We found that somatic mutation profiles were similar across metastases of an individual patient, with significant differences between patients. We identified 2-4 subclones in 6/7 samples using the CNA profiles derived from our snRNAseq data. While most clones showed an enrichment of DE genes in CNA regions, the proportion of DE genes in CNA regions ranged considerably, from 4.1% to 89.8% . Furthermore, they lacked enrichment of biological pathways or regulons, suggesting most CNAs unlikely result from a process of selection. Interestingly, major phenotypes observed, such as neuroendocrine differentiation, could be assigned to specific tumour subclones but not be explained by genetic alterations, suggesting epigenetic alterations attached to particular clones, or microenvironmental differences. In conclusion, we found mCRPC inter- and intra-tumoral heterogeneity is mostly transcriptionally driven, which the accumulation of somatic mutations likely providing background noise. The link between epigenetic states and genetic clones suggests a previously unappreciated association between genetic and epigenetic regulation. The results enhance our understanding of mCRPC heterogeneity and pave the way for futures studies focusing on the epigenetic regulation of mCRPC.

Hybrid Sequencing of Staphylococcus aureus from Chronic Rhinosinusitis Patients Reveals Widespread Beta-Lactamase Carrying Plasmids Associated with Compensatory Chromosomal Mutations in Virulence Genes

George Bouras

University of Adelaide

Background: Chronic rhinosinusitis (CRS) is a common inflammation affecting the nose and paranasal sinus mucosa. Staphylococcus aureus colonisation has been implicated as a disease driver an aggravating factor in not only CRS but also other chronic membranous inflammations such as asthma and atopic dermatitis. A better understanding of the genetic variation of S. aureus in nasal colonisation is required in elucidating CRS pathogenesis. Methods: S. aureus isolates colonising the sinuses of CRS and healthy control patients were cultured and sequenced using hybrid long-read and short-read technologies. Chromosome assemblies were created using long-read only assembly and short-read polishing tools. Extra-chromosomal plasmids were assembled using Plassembler (https://github.com/gbouras13/plassembler), a tool developed to accurately assemble plasmids and estimate plasmid copy numbers. Putative plasmids were manually curated using the PLSDB database and annotated to create a ‘pan-plasmidome’. Plasmids were hierarchically clustered using pairwise Mash distances. Chromosomes of isolates were tested for chromosomal differences between clusters using statistical genomics framework. Results: From 175 isolates analysed, 118 plasmids were present in 98 isolates. Of these 118 plasmids, 85 were clustered into a “beta-lactamase like” cluster present in 82 isolates and often harboured beta-lactamase, cadmium resistance, lactococcin and bacteriophage repressor genes. Isolates carrying such plasmids were significantly less likely to contain chromosomal encoded beta-lactamase genes compared to isolates without plasmids and were likely to have chromosomal mutations in virulence genes relating to fibronectin and collagen binding and capsular polysaccharides. Conclusion: Staphylococcus aureus isolates colonising the sinuses of patients with CRS commonly carry extra-chromosomal plasmids harbouring beta-lactamase genes. Those plasmids associate with chromosomal mutations in virulence genes that may be involved in mediating chronic inflammation.

Automating discovery of tandem duplications in cancer

Briana Robson

University of Melbourne

Structural variants (SVs) are a type of genetic mutation that rearranges approximately 50 or more base pairs of the genome. SVs include deletions, duplications, insertions, inversions, and translocations, and are frequently found in tumours. Internal tandem duplications (ITDs) and partial tandem duplications (PTDs) are an under-explored subcategory of SV, despite known variants having great clinical significance. For example, ITDs of the FLT3 gene is found in nearly 20% of acute myeloid leukemia (AML) cases and is associated with poor patient prognosis, making it an important marker and therapeutic target. More recently a tandem duplication of the UBTF gene was discovered, which could be used to define a sub-type of AML. While numerous computational methods have been developed to detect fusion genes from RNA-sequencing data, SVs such as ITDs and PTDs are difficult to identify, and only a few computational methods exist for their detection. MINTIE is an example of a recently developed pipeline that can perform unbiased detection of transcribed SVs, including ITDs and PTDs. However, a challenge with MINTIE is the high rate of reported events, which must be examined manually to identify ITDs or PTDs that could be candidate cancer driving mutations. This research will focus on the development of an automated candidate driver ITD and PTD classifier, based on SVs found with MINTIE, and other SV callers for RNA-Sequencing. This will be achieved by identifying defining features of candidate disease driving ITDs and PTDs, using publicly available RNA sequencing data from AML patients, cell lines and simulation. An algorithm to automatically sort candidate disease driving ITDs and PTDs from likely non-driver tandem duplications and other large insertions will be developed. One approach being considered is the training of a machine learning classifier. The significance of this work will be that it should produce a fast and accurate variant finding method, which can then be applied to large cancer cohorts to reveal new knowledge on the frequency and characteristics of ITDs and PTDs within various cancers.

Building an RNA-seq subtype classifier for T-cell acute lymphoblastic leukaemia

Allen Gu

Peter MacCallum Cancer Centre

T-cell acute lymphoblastic leukaemia (T-ALL) is an aggressive malignancy affecting both children and adults, characterised by uncontrolled proliferation of T-cell lymphoblasts. Official classifications of T-ALL subtypes do not currently reflect its heterogeneity, with the World Health Organisation recognising only one provisional subtype. However, recent studies have proposed potential subtypes based on transcriptomic analyses. We aim to develop a machine learning classifier that utilises bulk RNA-sequencing gene expression data to assign T-ALL samples to subtypes suggested by Liu et al. (2017) and Dai et al. (2022). This builds on our previous work designing the ALLSorts classifier for B-cell ALL subtypes (Schmidt et al. (2022)). The algorithm currently features a logistic-regression model trained on 202 paediatric samples from two sources. Various pre-processing steps and regularisation are included within the model. Samples are classified into seven subtypes corresponding to overexpression of NKX2, TAL1, TLX1 or TLX3, fusions involving KMT2A or MLLT10, and a diverse category including samples with dysregulation of LMO1/2. Testing the model on holdout samples demonstrated excellent recall and precision, as did testing performed on samples

originating from other studies not included in the training data, indicating robustness against batch effects. As there is no publicly available software for T-ALL subtype classification, we hope to create a tool that may aid with molecular classification and risk stratification, which may be of benefit in the clinical setting. Ongoing work aims to improve model performance, incorporate more samples for training/testing, and prepare the software for release. In addition, subtypes can be further investigated for deeper understanding of the driver mechanisms.

Cytocipher detects significantly different populations of cells in single cell RNA-seq data

Brad Balderson

University of Queensland

Identification of novel and known cell types with single cell RNA-seq (scRNA-seq) is revolutionising the study of multicellular organisms. Typical scRNA-seq analysis involves many preprocessing steps and represents an abstraction of the original measurements, often resulting in clusters of single cells that may not display distinct gene expression. To mitigate this, cell clusters are typically validated as cell types by re-examination of the original expression measurements, often resulting in post-hoc manual alteration of clusters to ensure distinct gene expression. However, distinct cell populations may exist that are not clearly demarcated by a single marker gene, but instead co-express a unique combination of genes; a phenomenon which is difficult to detect by manual examination. Furthermore, manual examination of genes and post-hoc cluster editing is time-consuming, error-prone, and irreproducible. Here, we present Cytocipher, an scverse compatible bioinformatics method and software that scores cells for unique combinatorial gene co-expression and statistically tests whether clusters are significantly different. Application to both simulated and real data demonstrates that the combinatorial gene expression scoring outperforms existing per-cell gene enrichment methods, such as Giotto Parametric Analysis of Gene Set Enrichment and Scanpy-score. Furthermore, Cytocipher cluster-merging identified distinct CD8+ T cell subtypes in human peripheral blood mononuclear cells that were not identified in the original annotations. Cytocipher cluster-merging was also able to identify distinct intermediate states corresponding to cell lineage decisions and branch points in mouse pancreas development, not previously identified in the original annotations. Identification of significantly different clusters of cells is an important new methodological improvement to the existing analysis pipeline of scRNA-seq data. Utilisation of Cytocipher will thus ensure that single cell atlas mapping efforts provide distinctly different and programmatically reproducible cell clusters. Cytocipher is available at https://github.com/BradBalderson/Cytocipher.

A computational framework to detect ceRNA cross-talks in pan-cancer studies

Yi-Wen Hsiao

University of Melbourne

Competitive endogenous RNA (ceRNA) is a novel mechanism of gene regulation involved in pathological processes in human diseases, including cancer. However, many potential interactions of miRNA-mRNA triplets make experimental validations difficult. The recent advances in computational methods have enabled the inference of ceRNA interactions at a genome-wide scale from expression

data. To understand the etiology of cancers, we developed a new tool, ceRNAR to depict the landscape of ceRNA patterns. We demonstrate that ceRNAR has high sensitivity and low false positive rate to detect ceRNA across 31 cancers from TCGA repository (9,464 samples). We first collected 7,535 putative miRNA-targets from nine published databases, including both experimental validation and computational prediction. Our algorithm then iteratively evaluates whether each mRNA pair is a potential ceRNA event through three main steps: (1) ceRNA pairs filtering based on a rank-based running sum correlation statistic, (2) sample clustering based on gene-gene correlation values, and (3) peak merging to support specific ceRNA event based on the most relevant sample patterns. We show that the most relevant ceRNA events are either common or unique in cancer types. We also characterized the underlying biological mechanism and the prognostic role of such ceRNA network-based signature for further pathological interpretation. ceRNA-mediated gene regulation involved in the occurrence and progression of cancers plays a key role in diagnostic, prognostic, or therapeutic aspects of clinical applications. Our computational framework implemented in the ceRNAR package can further support the community in studying such regulation in cancer biology.

Characterising gene co-expression changes facilitating the loss of features of multicellularity driving Prostate Cancer progression

Mikhail Dias

Peter MacCallum Cancer Centre

The transition to multicellularity involved evolution of gene regulatory networks (GRN) to coordinate and maintain cellular processes in order to promote organism-level fitness. Transcriptomic analysis of data from The Cancer Genome Atlas has revealed networks acquired during the transition to multicellularity are often broken down in cancer leading to tumorigenesis. We aim to uncover how these pathways are rewired in Prostate cancer (PC) to evade treatment. We have developed Evolutionary Network Analysis (ENA) a unique multi-omics approach combining evolutionary analysis, transcriptomics and network biology to investigate how GRNs acquired during the transition to multicellularity are rewired in cancer. Applying ENA to PC patient samples stratified by progression of benign to malignant and primary to metastatic tumours, created a comprehensive landscape of changes in gene co-expression during PC progression. Our analysis reveals as PC advances to higher Gleason grade groups, genes acquired during the transition to multicellularity become progressively more rewired. Further analysis using gene expression and functional enrichment revealed, new connections facilitate the activation of more ancient unicellular pathways associated with proliferation, angiogenesis, protein trafficking and metabolic processes. This study presents a new paradigm in cancer biology investigating how genes cooperate in complex networks to derive tumour progression and evade drug treatment. We have demonstrated how utilizing gene co-expression signatures can be used to gain a comprehensive molecular landscape of PC, which is immensely valuable for the development of more robust therapeutic strategies.

Deciphering heterogeneous patient-derived bulk RNA-seq data using single-sample pathway perturbation analysis and cell-type deconvolution

Wenjun Liu

University of Adelaide

Performing bulk RNA sequencing on patient-derived tumour samples is a cost-effective way to explore biological mechanisms driving disease progression and study treatment response in a clinically relevant context. However, cancers not only have a high degree of diversity between patients, but also exhibit a high level of intra-tumour heterogeneity, which bulk sequencing approaches fail to capture. To tackle these challenges, sSNAPPY, a Single-Sample directioNAl Pathway Perturbation analYsis method has been developed (available on Bioconductor) and incorporated with scRNA-seq directed cell-type deconvolution to unravel changes in biological activities that cannot be detected using existing approaches. sSNAPPY differs fundamentally from conventional pathway enrichment testing methods. Instead of detecting over-representation of pre-defined gene-sets in a group of samples, sSNAPPY utilises pathway topology and changes in gene expression between paired samples to compute a perturbation score for each pathway within each paired sample, where directionality of scores indicates activation or repression of the biological process. Additionally, sSNAPPY does not rely on detection of differentially expressed genes, which can be masked by confounding factors and hence not detected as significant in patient-derived bulk RNA-seq data. This talk will provide the audience with an overview of the key features and technical methods underlying sSNAPPY, which will be illustrated by a case study involving the application of sSNAPPY and cell-type deconvolution of bulk RNA-seq data derived from primary malignant breast tumour tissues cultured ex vivo.

Identifying cellular interactions and communication in multiplexed in situ imaging data through cell state analysis

Sourish Iyengar

University of Sydney

At the heart of biological processes are a diverse collection of cells, with each cell playing its own unique role in the formation of the overall biological system. The behaviour of these cells can be described by hard-wired characteristics, but their functions can also dynamically change based on their environmental context, leading to cells transitioning into different states. Identifying how cells communicate and interact with each other can aid in understanding the drivers catalysing a change in a cell's state. This interplay between cell types could be key to navigating the complex biological landscape, diverse tissue functions, and patient responses to disease. Using spatial single-cell omics data, we have developed a computational method to identify and quantify changes to a cell's state to untangle the underlying mechanisms that drive heterogeneous tissue functions. We model how markers of a cell's state change with spatial proximity to other cell types. We propose this method as an initial framework to explore how the structure and function of different cell types may be altered by the agents they are surrounded with. Furthermore, we explore how changes in cell states can correlate with a patient's clinical outcomes. This is used to highlight the importance of cell state analysis in understanding the responses that drive disease status and patient survival.

Environment-dependent trajectory inference (ENTRAIN) to determine the extra-cellular signals that specify cell fate

Wunna Kyaw

Garvan Institute of Medical Research

Cell fate is commonly studied by profiling the gene expression of single cells to infer developmental trajectories based on expression similarity, RNA velocity, or statistical mechanical approaches. However, current approaches do not recover external signals from the microenvironmental niche that drive a differentiation trajectory. Here, we address this issue by presenting a computational method (Entrain) that unites trajectory inference methods, including RNA velocity, with ligand regulatory networks to infer extracellular modulators. By integrating trajectories and velocities with cell-cell communication, Entrain predicts driver ligands responsible for observed dynamics and decomposes trajectories into environmentally governed components and cell-intrinsic components. Further, Entrain quantifies the degree to which the niche is responsible for the observed trajectory dynamics, improving on existing methods for cell-cell communication inference that rely solely on differential expression of ligand-receptor genes in pre-defined clusters. We validate our approach on single-cell bone marrow and embryonic neurogenesis datasets to recapitulate known environmental drivers of cell fate commitment in haematopoietic, mesenchymal, and neurogenic lineages. We anticipate this method will help elucidate the driving interactions between developing cells and the governing niches which shape cell fate.

Multi-omic Profiling Uncovers Molecular Controls in Early Cerebral Brain Organoid Differentiation

Carissa Chen

Children's Medical Research Institute

Defining molecular programs that orchestrate human brain development is essential for uncovering the complexity behind neurodevelopment and the pathogenesis of neurological conditions. Given the difficulties in accessing embryonic and fetal brain tissues, the differentiation of human embryonic stem cell (hESC)-derived embryoid bodies (EBs) to three-dimensional brain organoids has made it possible to recapitulate this developmental process in vitro and provide a unique opportunity to investigate human brain development and disease. Here, we generate a molecular map of the phosphoproteome, proteome, and transcriptome of EBs differentiating towards cerebral brain organoids. We uncovered time-dependent key signalling events and downstream regulons in driving the neural lineage specification. Our comparative analysis of regulons demonstrates the fidelity of organoids to emulate embryonic brain formation. Finally, we validate AKT signalling as a key driver of neural differentiation. Our data provides a comprehensive resource to gain insight into the molecular control in human embryonic brain development.

Developing a user interface for sharing federated genomic and phenotypic data using the Beacon v2 protocol

Ricky Nguyen

University of NSW

A major obstacle in human genomics research is the paucity of accessible, well-annotated genomic and phenotypic data. In an effort to bring disparate data sources together, the Children’s Cancer

Institute as part of the Human Genome Platforms Project (HGPP) has set out to build on existing standards such as the Global Alliance for Global Health's Beacon specification. A Beacon network is made up of federated data sources (beacons) that can be collectively searched with a single query. This alleviates both the need to move large data collections into a single repository and compromising their individual data security. Originally, the Beacon protocol (versions 0 and 1) only stored genomic variant data, made accessible via ELIXIR's Beacon network UI. The current version (v2) introduces new features such as phenotypic data and variant annotations (such as pathogenicity assessment), which increases the complexity and richness of the metadata. This new version of Beacon and the complexity of its data model necessitates a more sophisticated user interface. To address this, we have designed and implemented a UI to support Beacon v2. The purpose of the UI is to enable querying the full catalogue of data by all members of the Beacon network. In addition, this UI allows users to build cohorts by applying filters across seven common data entities that can be further adjusted to produce either simplified or nested queries as required. These filters have been guided by user stories gathered from the HGPP consortia and are customisable. Thus, we assert that our UI is a versatile and practical solution to extracting useful data from the v2 Beacon network. We aim to make this implementation open-source and welcome contributions to this effort.

Inference of ancestral protein properties in phylogenetic gene trees

Sam Davis

University of Queensland

Testing evolutionary hypotheses in functionally diverse protein families involves reconstruction of ancestral proteins by inference of their sequence, followed by laboratory characterisation. Such investigations can inform the evolutionary trajectories which shaped the family, as well as discover sequence and structural determinants which contribute to functional variation. Ancestors targeted for reconstruction are generally selected by applying pre-conceived ideas about the evolutionary history in a non-rigorous manner. As such, reconstruction is prone to producing ancestral proteins without the anticipated properties. This challenge may be mitigated by the ability to reliably predict an ancestor's properties prior to characterisation. An under-utilised source of information for such predictions is the phylogenetic topology itself. Tools which use phylogenetic trees as a basis for ancestral state prediction exist, however they generally incorporate continuous-state stochastic models and were designed for the inference of trait values in species trees. These enforce restrictive assumptions including fitness neutrality and constant evolutionary rates, and hence are unsuitable for inferring chemical or physical parameters of proteins which may be sensitive to discrete mutations. We developed a model for ancestral state inference to address these deficiencies. The model utilises discrete, latent states as a proxy for evolution of continuous-valued properties along branches of the tree. Latent states emit real property values via Gaussian features, the parameters of which can be learnt from known data. Inferred transitions between latent states along a particular branch can be considered alongside corresponding changes in sequence, which may allow for the prediction of causal relationships. The tool was evaluated for its ability to predict values of withheld experimental data from comprehensive ancestral reconstruction studies and curated datasets. It generally outperforms naïve baselines for a variety of property types. Utilisation of this tool in future reconstruction studies will likely facilitate a more targeted selection of candidate ancestors, leading to improvements in cost and time-efficiency.

Posters

Systematic construction of phylogenetic trees enriched with experimental properties

Sebastian Porras

School of Chemistry and Molecular Biosciences - The University of Queensland

Ancestral sequence reconstruction is quickly becoming an effective way for protein engineers to predict and resurrect ancient proteins for application in the bioeconomy. Expressing ancestral proteins in the lab is both time consuming and costly and new methods are needed to help guide decisions about which ancestors are most likely to have desired qualities. Data sets that are enriched with experimental results allow labs to test hypotheses about how properties change over time and help develop understanding of protein family evolution. Currently, researchers must build trees and search through literature to obtain such data sets. A method was developed to integrate information from the enzyme database BRENDA with UniProt annotations to systematically build phylogenetic trees that were highly annotated with experimental properties. BRENDA currently contains 8331 different enzyme groups that can contain hundreds of sequences across multiple domains of life. However, enzymes are grouped only on their activity, leading to non-homologous enzymes being placed together. The method seeks to filter out non-homologous enzymes and allow trees that are rich with experimentally validated data to be constructed. This method uses InterPro signatures assigned to each enzyme to determine similarity between enzymes within a BRENDA enzyme group. These signatures can represent protein families, domains or active sites and can help to determine what proteins are homologous. By applying a threshold score, enzymes can be grouped together to then construct trees. Manual verification of selected datasets revealed that trees constructed with this method conformed to prior expectations in terms of homologous relationships. The vast diversity of enzyme functions available means that a group of interest can be run through this pipeline easily from end to end to construct trees coupled with a wealth of experimental annotations in a reproducible manner.

Empirical tokenisation of genomic data and data-free deep-learning model evaluation

Tyrone Chen

Monash University

Machine learning is a useful technique which has been successfully applied in many fields, including the biological sciences. In image recognition, natural language processing (NLP) and other fields, many libraries and high-level workflows exist to streamline data processing. To our surprise, no analogue for this exists in the biological sciences for machine learning. Therefore, we created a software package (ziran) which meets this need of a high-level software library with a biology focus. We allow bioinformatics researchers to perform machine learning in biological sequences using state-of-the-art methods. While there is no substitute for the domain-specific knowledge required to interpret outcomes, the package is intended to reduce the barrier for entry for biologists and bioinformaticians in adopting machine learning in their research, and will be the first of its kind in the biological domain. We demonstrate use cases using multiple different strategies. By passing in two categories of DNA sequence collections, such as sequences enriched for regulatory features vs

sequence depleted for regulatory features, we successfully tuned and trained multiple natural language processing models on different data representations for a two-category classification. The process is heavily abstracted, where data transformations and representations are handled internally with minimal user input. However, for users who want to specify their own parameters, low-level tuning is also available. In our package, minimal preprocessing is required as biological sequence data is input directly as fasta or fastq file(s). The user can select from multiple combinations of state-of-the art (a) NLP preprocessing methods and (b) machine learning classifiers of choice to apply to their sequence data, comparing their resulting performances. Configurable hyperparameter sweeps are possible, and the best parameter configuration will automatically be selected by default. Metrics corresponding to the algorithm of interest - for example precision, recall, accuracy, f1, class membership and feature importance scores - are logged during both sweeps and runs using user-specified configurations. In addition, these metrics are tracked and output for the user, including real-time visualisations and plots for increased interpretability. We intend to continuously extend our library with new features over time, iteratively adding new methods as they become available.

Benchmarking long-read RNA-sequencing analysis tools using in silico mixtures

Xueyi Dong

Walter and Eliza Hall Institute of Medical Research

The growing popularity of long-read sequencing technology in transcriptomic studies has led to the development of customised algorithms for isoform identification and workflows for differential expression analysis. However, the current lack of benchmark datasets with inbuilt ground-truth makes it challenging to compare the performance of the different methods available. Here, we present a benchmark experiment using two human lung adenocarcinoma cell lines that were each profiled in triplicate together with synthetic, spliced, spike-in RNAs (“sequins”). Samples were deeply sequenced on both Illumina short-read and Oxford Nanopore Technologies long-read platforms. Alongside the ground-truth available via the sequins spike-ins, we created in silico mixture samples by combining reads from the two cell lines in known proportions to allow performance assessment in the absence of true positives or true negatives. Our results show that, StringTie2 and bambu outperformed other tools from the 6 isoform detection tools tested, DESeq2, edgeR and limma-voom were best amongst the 5 differential transcript expression tools tested and there was no clear front-runner for performing differential transcript usage analysis between the 5 tools compared, which suggests further methods development is needed for this application.

Quantifying native RNA integrity from nanopore sequencing

Aditya Sethi

John Curtin School of Medical Research, Australian National University

Nanopore direct RNA sequencing (DRS) unlocks the ability to sequence full-length transcripts in their native state and thus capture complex transcription, splicing and RNA modification patterns with single-molecule and single-nucleotide resolution. Direct RNA sequencing relies on sequencing RNA molecules in the 3’ to 5’ direction, meaning that any breaks in an RNA strand result in sequencing

termination. Given the vulnerability of RNA samples to chemical and enzymatic degradation, understanding how sample RNA integrity interacts with DRS to incur possible transcript coverage biases is a major consideration for DRS applications. Here, we show that systematic 3’ coverage bias (progressively diminishing 5'-directed coverage) is typical to DRS experiments across all classes of coding and noncoding transcripts, stemming from limited RNA integrity. We demonstrate that transcript coverage steadily decreases from endogenous or exogenous 3’ poly(A) tails, often resulting in drastic coverage differences between transcript 5’ and 3’ regions. Using an in vitro model of unbiased RNA degradation, we show that differential transcript coverage patterns as observed in DRS drives obscured and false discovery of differential gene expression, alternative splicing and RNA modification patterns. More generally, we show that DRS 3’ bias limit the application of DRS to study RNA features distal to 3' ends by > 1,000 nucleotides. To assess and controllably compensate these biases, we present nanograd, a direct RNA quality control tool to quantify transcript-specific 3’ bias and overall sample RNA integrity from DRS data. Nanograd estimates per-transcript sequence coverage bias and provides a single Direct Integrity Number (DIN) to summarise RNA integrity, enabling seamless and standardised quality comparison across experiments, libraries and flowcells. DIN is straightforward to use and provides coverage correction necessary to identify key features of RNA such as splice junctions, structural feature marks, and chemical RNA modifications in a bias-aware manner.

Single-cell RNA-seq analysis of amniotic fluid: a unique approach for studying foetal development

Calandra Grima

Peter MacCallum Cancer Centre

Studying live human foetal cells presents challenges to researchers, particularly with ethical constraints on the age of human foetuses grown in vitro. Single-cell RNA-seq (scRNA-seq) analysis of amniotic fluid samples may provide insight into live foetal development by profiling the cell types present and gene expression profile of these. This technique could provide unique insight into human prenatal developmental as well as the variety of structural and genetic abnormalities that are detected before birth. As a proof of principle, scRNA-seq was performed on amniotic fluid samples collected via amniocentesis from two second-trimester pregnancies using the 10X system. Using PCR, one pregnancy was confirmed to be infected with cytomegalovirus (CMV) and the other was CMV negative. Amniotic fluid contains high levels of dead cells and debris, however, across the two samples, we identified 1882 intact cells in the dataset after removal of empty droplets and damaged cells. Cell type annotation was performed in several ways using both supervised and unsupervised methods. We found that using Azimuth (supervised) with its inbuilt foetal reference gave annotations that mapped well onto the clusters produced by unsupervised methods. Marker gene analysis and Gene Ontology (GO) analysis was performed on clusters to give further insights into the detected cell types and validate annotations. Squamous epithelial cells originating from the lung and stomach were the most abundant cell types, followed by other epithelial cell types, myeloid cells and lymphoid cells. Interestingly, the CMV positive sample contained a lower proportion of megakaryocytes than the CMV negative sample, which may be related to the inhibition of megakaryocyte proliferation and differentiation and encouragement of megakaryocyte apoptosis by CMV as reported previously. The marker gene and GO analyses produced results that were concordant with the respective epithelial, immune, or other lineage of each cluster. Despite the

challenges of working with such a unique dataset and the need for further optimisation of laboratory protocols, this study presents proof-of-concept results for the utility of scRNA-seq in the examination of live foetal cells from amniotic fluid. These results give insight into the cell types that are available to collect and study from amniotic fluid which may hold unique clinical and research applications."

Disparities in spatially variable gene calling highlight the need for benchmarking methods in spatially resolved transcriptomics

Natalie Charitakis

Murdoch Children's Research Institute

Introduction

Spatially resolved transcriptomics (SRT) is a novel, disruptive technology set to push the boundaries of exploring gene regulatory networks while providing both spatial and temporal resolution. It is expected to be exponentially adopted by the transcriptomics community after being named Nature's Method of the Year in 2020. Despite this, analysis packages for SRT datasets are still in their infancy, with a clear forerunner yet to emerge. A comprehensive overview of the performance of commonly used packages on the same datasets is lacking and there is a need to determine their performance in correctly labelling spatially variable genes (SVGs).

Methods

To establish which of the current packages is most effective in identifying SVGs within datasets generated using the same technology, a combination of publicly available and simulated 10X Genomics Visium datasets generated from 8 fresh frozen and FFPE human tissue samples were analysed. I will be assessing the performance of SpatialDE, SPARK, SpaGCN, scGCO and Seurat as each uses a different mathematical model to identify SVGs. Separate simulated datasets with known true positives or randomised signal helps begin to elucidate package performance.

Results

In all datasets, most packages identify SVGs independent of results from other packages and overlap is minimal. The number of SVGs identified by each package differ substantially and differences are statistically significant. SpaGCN appears consistently the most conservative while SpatialDE consistently labels the greatest number of genes as SVGs. Many datasets have an underlying distribution of gene expression that does not fit the assumptions of many of the packages, which may be affecting performance.

Conclusion

Given the disparity of results, this comparison study highlights the need of further development of benchmarking methods to for identifying SVGs in SRT experiments. This should include methods for reliably simulating SRT data for method validation. "

Multiomics profiling of lung transplant recipients identifies molecular signatures linked to chronic lung allograft dysfunction

Giulia Iacono

Monash University

Long-term survival of lung transplant recipients remains limited by chronic lung allograft dysfunction (CLAD). Repeated injury to the allograft results in dysregulated wound repair that culminates in irreversible airway fibrosis, bronchial obstruction, and permanent lung function deterioration. There is an urgent need to discover CLAD biomarkers able to predict changes in the allograft before a decline in lung function, as well as understanding underlying CLAD mechanisms to improve the limited therapies available. Using a multi omics approach, we performed host transcriptomics, metabolomics, lipidomics profiling, and microbial rRNA sequencing on longitudinal broncho-alveolar lavages from 12 CLAD and 20 CLAD-free lung transplant recipients over 36 months post-transplant. CLAD patients displayed a unique gene expression profile characterised by the sustained upregulation of genes involved in the inflammatory response. Intriguingly, this pre-CLAD signature was detectable during the first year post-transplant, before any significant decline in lung function had occurred. Using machine learning and an elastic-net regularised generalised linear model, we evaluated the predictive potential of the top 30 DE genes, obtaining high discriminative power with area under the curve (AUC) values of 0.84. When compared against the CLAD-free group, CLAD patients showed altered surfactant lipid composition involving the increase of glycerophospholipids and ceramide lipid classes. This increase was correlated with an active decline in lung function, suggesting a mechanism responsible for CLAD progression. This peri-CLAD signature was exacerbated if the patient also had a concurrent infection. Our multi-omics approach identified important gene expression signatures preceding a decline in lung function in CLAD patients, contributing to biomarker discovery and better patient stratification. Metabolomic and lipid signatures contribute to a better understanding of the pathophysiological mechanisms underlying CLAD.

Theoretical prediction of potential molecular mechanisms of Simiao Pills against hyperuricemia: A computational modelling and screening study

Hong Li

RMIT University

Hyperuricemia is a pathological condition that serum uric acid elevates due to purine metabolism disorders. Simiao Pill (SMP) is a Chinese herbal formula applied to attenuate HUA with a long history of use. Although studies have reported the efficacy of SMP against HUA, the corresponding mechanistic elucidation remains inadequate. Here, we performed computational structural modeling and screening to predict the potential mechanisms of SMP against hyperuricemia. Our study included network pharmacology analysis, molecular docking, and chemical structure analysis based on physicochemical and pharmacokinetic characteristics. Overall, 403 SMP compounds and 15 potential targets for hyperuricemia were identified to produce 6,045 docking results with an average binding affinity of -7.081 kcal/mol. Three inflammatory pathways were mainly enriched, particularly the TNF signaling pathway, IL-17 signaling pathway, and the C-type lectin receptor signaling pathway. Among SMP compounds, terpenes flavonoids and alkaloids were predicted to have relatively strong binding affinity with Xanthine Dehydrogenase. In conclusion, SMP may exert its effect on treating hyperuricemia by multiple targets and pathways. Our study provided a novel perspective for exploring the mechanisms of action of herbal medicine in hyperuricemia. Future experimental studies could be conducted based on our computational results.

Synergistic mechanisms of herbal compounds from Toujie Quwen Granules against the main protease of SARS-CoV-2: a structure-based multi-ligand molecular modelling study

Hong Li

RMIT University

The Chinese patent medicine, Toujie Quwen Granules (TQG), is known for its symptomatic relief of COVID-19 symptoms, inhibition of disease progression, and improvement in recovery rate. In accordance with known principles of traditional medicine, these effects are likely mediated by the synergy between herbal molecular components, although the specific ligands responsible are not yet known. Previous in silico studies suggest that main protease (Mpro) is a possible target for TQG. Mpro is considered a promising drug target due to its dissimilarity to human proteases and its crucial role in viral function. In the current study, we employed a novel strategy involving structure-based multi-ligand molecular modelling to investigate the synergistic effects across compounds in TQG. We first used molecular docking to identify molecular components of the formula which may inhibit Mpro. We found that HQA004 was the most favourable inhibitory ligand. We also identified a ligand from the second herbal component, CHA008, which may act to support the binding of the proposed HQA004 inhibitor. Molecular dynamics simulations were then performed to further elucidate the possible mechanism of inhibition by HQA004 and synergistic bioactivity conferred by CHA008. We found that HQA004 bound strongly at the active site, and that CHA008 enhances the contacts between HQA004 and Mpro. Furthermore, CHA008 also dynamically interacted at multiple sites and continues to enhance the stability of HQA004 despite diffusion to a distant site. We proposed that HQA004 acts as a possible inhibitor and CHA008 serves to enhance its effects via allosteric effects at two sites. By using the structure-based multi-ligand molecular modelling, we have demonstrated a real-time computational simulation of the synergistic interactions of two natural compounds, baicalein (HQA004) and cubebin (CHA008), and how they collaboratively inhibited the Mpro of SARS-CoV2. This study highlights the potential molecular mechanisms of synergistic effects between different herbs as a result of allosteric crosstalk between two ligands at a protein target.

Functional analysis of the stable phosphoproteome reveals cancer vulnerabilities

Di Xiao

The University of Sydney

The advance of mass spectrometry-based technologies enabled the profiling of the phosphoproteomes of a multitude of cell and tissue types. However, current research primarily focused on investigating the phosphorylation dynamics in specific cell types and experimental conditions, whereas the phosphorylation events that are common across cell/tissue types and stable regardless of experimental conditions are, so far, mostly ignored. Here, we developed a statistical framework to identify the stable phosphoproteome across 53 human phosphoproteomics datasets, covering 40 cell/tissue types and 194 conditions/treatments. We demonstrate that the stably phosphorylated sites (SPSs) identified from our statistical framework are evolutionarily conserved, functionally important and enriched in a range of core signaling and gene pathways. Particularly, we show that SPSs are highly enriched in the RNA splicing pathway, an essential cellular process in mammalian cells, and frequently disrupted by cancer mutations, suggesting a link between the dysregulation of RNA splicing and cancer development through mutations on SPSs.

Developing methodology to scan public databases using recount3 to identify potential fusion genes

Caitlin Page

Peter MacCallum Cancer Centre

Fusion genes, where genomic rearrangements bring together two separate genes to form one combined transcript, are common in cancer. Current fusion detection methods rely on identifying RNA-Seq reads containing both genes. These methods are computationally intensive, and rely on strict false positive filters for accuracy. Our project aims to develop a method to scan for fusions across large datasets by comparing changes in expression before the fusion breakpoint (expected low) and after (expected high).Cancer RNA-seq data from The Cancer Genome Atlas (TCGA) and non-cancer RNA-seq from the Genome Tissue Expression project has been uniformly processed and compiled within R package recount3 for downstream analysis. Using the count data provided, we selected 10 of the most commonly represented fusion genes in TCGA datasets to validate the accuracy of our fusion calling method on recount3 summarised exon and gene counts. We then investigated factors which may affect the accuracy of our method - including the position of the fusion within the second gene in the fusion, and the tumour purity of samples. Potential fusion samples are first identified using gene counts to look for over-expression, comparing distributions across cancer to select samples that are overexpressed. Changes in exon expression across the gene are then compared to identify samples with significant differences in expression between the start and end of the gene to indicate the point at which the fusion transcript begins. Samples with very low expression at the start of the gene and normal to high expression at the end of the gene are identified as potential fusions, allowing a subset of the data to be explored in greater detail.We hope to make our method a publicly available pipeline that would allow researchers to identify samples with potential fusions involving their gene of interest.

A computational model for gene regulatory networks in early kidney development

Adeline Trieu

Murdoch Children's Research Institute

Introduction

The interplay of key markers in gene regulatory networks underpin the robustness of early kidney development. However, the specifics of these interactions, how variability in gene expression translate to a final consistent phenotype, are poorly understood. Here, we aim to generate a computational model that recapitulates the early developmental trajectory of the kidney, in order to identify feedback and feedforward loops that characterise its robustness.

Method/Results

To capture the essence of the kidney developmental network from current literature, the model must be able to build complex interactions from the basis of a few key markers. Thus, we have found Boolean models to be well suited for these purposes. Boolean models provide a binary representation of a gene regulatory network whilst abstracting co-inhibitory and activatory interactions into an influence graph. Utilising this modelling language, we developed a new model for early kidney development and validated its accuracy with existing single-cell RNA-sequencing data.

Our constructed model identified pathways and stable states in which transcriptional factors may interact and form stages of kidney development.

Conclusion

We presented a novel Boolean network to recapitulate the gene regulatory networks underlying kidney development. Our gene regulatory network could mimic the early steps and requirements of early kidney development, resembling the patterns observed in published single-cell RNA-sequencing data.

Finding novel transcripts in bulk and single cell cancer transcriptomics datasets through data analysis of both short and long read RNA sequencing

Michael Nakai

Peter MacCallum Cancer Centre

RNA sequencing is a powerful method to probe the transcriptome of cancerous samples, which often exhibit markedly different transcriptomes to healthy cells due to high levels of mutation. These mutations can result in transcripts with large amounts of structural variation that can’t be mapped to the reference, and subsequently remain uncharacterised in downstream analyses. In addition, short read sequencing has limited utility in reconstructing complete transcripts, complicating potential novel transcript identification in the unmapped reads. Combining both short read and long read RNA-sequencing on samples has the potential to mitigate these weaknesses, allowing for use of the long reads as a sample-specific reference with quantification by short reads. and can therefore capture the formerly unmapped reads. Drawing from the SG-NEx bulk RNA-sequencing dataset generated from cancer cell lines, we mapped short reads taken from K562 cells to nanopore cDNA long reads using Minimap2 and found that mapping rates were higher when using the long reads as a reference. In addition, we found that 9.9% of the mapped short reads were uniquely mapping to the long read reference and not the reference genome, confirming that some short reads contain information uniquely found in the long reads that would be missed when mapping to the genome. Using this data as a starting point, I am developing a method to identify and quantify potential novel transcripts within the formerly unmapped sequences in bulk RNA-sequencing datasets. In addition, through use of a single-cell RNA-sequencing dataset looking at dendritic cells taken from ten C57BL6 mice, I will adapt this method to be further applicable to single-cell datasets."