Bioinformatics Abstracts

B1: AI in healthcare: identifying and understanding superbugs

Wheeler, N.E.[1,2], Gardner, P.P.[3,4], Barquist, L.[5,6]

[1]Infection Genomics, Wellcome Sanger Institute, Hinxton, UK, [2]School of Biological Sciences, University of Canterbury, Christchurch, New Zealand, [3]Department of Biochemistry, University of Otago, Dunedin, New Zealand, [4]School of Biological Sciences, University of Canterbury, Christchurch, New Zealand, [5]Helmholtz Institute for RNA-based Infection Research, Wuerzburg, Germany, [6]Institute for Molecular Infection Biology, University of Wuerzburg, Wuerzburg, Germany.

Thanks to the dropping cost of whole genome sequencing technologies, and coordinated global surveillance efforts, we now have access to vast collections of genomic data from bacteria circulating in human populations. However, our ability to draw insights about why some of these bacteria cause severe disease, and which lineages might be problematic in the future is still limited. New methods are urgently required to effectively utilize these large resources of genomic data to provide functional insights into established and emerging pathogens. In this work, we have measured the burden of atypical mutations in protein coding genes across independently evolved *Salmonella enterica* lineages, and used these as input to train a machine learning algorithm to identify strains associated with extraintestinal disease. Members of the species fall along a continuum, from pathovars which cause gastrointestinal infection and low mortality, associated with a broad host-range, to those that cause invasive infection and high mortality, associated with a narrowed host range. Our machine learning algorithm was able to perfectly discriminate long-established gastrointestinal and invasive serovars of *Salmonella*. Additionally, it was able to identify recently emerged *Salmonella* Enteritidis and Typhimurium lineages associated with invasive disease in immunocompromised populations in sub-Saharan Africa, and within-host adaptation to invasive infection. We have dissected the architecture of the model to identify the genes that were most informative of phenotype, revealing a common theme of degradation of metabolic pathways in extraintestinal lineages. This approach identifies patterns of gene degradation and diversifying selection specific to invasive serovars that have been captured by more labour-intensive investigations, but can be readily scaled to larger analyses.

B2: *refynr2* and *Gentian*: a bioinformatic platform to enable automated precision cancer genomic medicine trials

John P Grady[1], Mark Pinese[2,3], Mandy Ballinger[2,3], Emily Collignon[2], John Simes[4], Anthony M Joshua[2,3], Marcel E Dinger[1,3], David M Thomas[2,3], Mark J Cowley[1,3]

[1]Kinghorn Centre for Clinical Genomics, Garvan Institute of Medical Research, Sydney, Australia, [2]Genomic Cancer Medicine Program, Garvan Institute of Medical Research, Sydney, Australia, [3]St Vincent's Clinical School, UNSW Sydney, Sydney, Australia, [4]NHMRC Clinical Trials Centre, University of Sydney, Sydney, Australia.

Identification of genetic alterations in patient tumours has the potential to substantially improve patient care, though more precise diagnosis and identification of optimal therapies. We established the Molecular Screening and Therapeutics (MoST) program to bring genomics-led therapeutics to patients with rare and advanced adult cancers. To date, we have screened and reported on over 600 patients, using several gene capture panels, currently utilising the Illumina TST170 gene targeted sequencing panel using DNA and RNA from tumour-only FFPE samples. To support precision cancer genomics trials, we have developed a modular and automated bioinformatics pipeline to enable hands-off analysis, interpretation, discussion, and reporting of patient genomic tumour profiles with genomically matched treatment and trial recommendations.

refynr2 is a flexible, modular, cloud-based analytical platform, which runs on DNAnexus, to identify SNVs, INDELs, CNVs, tumour mutation burden (TMB) and RNA fusions. To achieve hands-free automation, we overcame numerous bioinformatics challenges, including tumour purity estimation, germline/somatic variant classification, FFPE artefact removal, local ploidy estimation, and gene fusion detection. We developed *Gentian* to automatically make treatment and clinical trial recommendations, by integrating genomic, clinical, drug and trial data. A comprehensive tumour landscape report is prepared for rapid discussion at our molecular tumour board (MTB). *Gentian* also integrates with clinical patient management systems (Progeny, and JIRA for case tracking) for pre- and post-MTB patient management and automated reporting.

This system has been flexibly designed to integrate with other sources of genomic data (not restricted to gene capture panels), and other databases/patient management systems.

Furthermore, it has enabled us to upscale our patient throughput by an order of magnitude by reducing the human input from several hours per patient to minutes, whilst also standardising the analytics and facilitating cohort analysis and research.

Here we will present our bioinformatics research for enabling precision cancer genomic medicine trials.

B3: A Customisable Analysis Pipeline for Identification of Clinically Relevant Genetic Variants in Next Generation Sequencing Data

Benton, MC.[1,2]

[1] Institute of Environmental and Science Research (ESR), Kenepuru Science Centre, Porirua, New Zealand, [2] Institute of Health and Biomedical Innovation, Faculty of Health, Queensland University of Technology, Brisbane, Australia

The introduction of large scale next generation sequencing into clinical practice creats new challenges regarding the classification of potentially pathogenic genetic variants. Aside from the technical challenges involved (sufficient depth and breadth of coverage for target regions) there is the challenge of identification of known or novel pathogenic variants. As these variants are found among a host of common and rare polymorphisms, the identification of clinically relevant variants is time consuming and fraught with the potential for analysis induced false negatives. This is especially the case in larger datasets and further complicated in conditions or disorders where a multitude of genes or mutation sites within a gene may be responsible for symptoms or important for treatment recommendations.

Here I present the details of an approach I developed in a working NATA accredited DNA diagnostics laboratory, to rapidly facilitate the diagnosis of neurological conditions and to enable “pathogenic”, “likely pathogenic” and “variants of unknown significance” to be flagged for subsequent validation. This pipeline uses custom, curated gene lists to categorise variants into specific analysis tiers and to subcategorise them based on standard parameters to facilitate the rapid interrogation of potentially pathogenic variants by human operators. Numerous publicly available databases are used to assist with variant classification. Curation processes allow robust logging of parameters and database and software versions. A static report is automatically generated during the annotation and filtering step using RMarkdown. A Shiny-based web environment was implemented, providing a layer of interaction and visualisation not possible with static reporting. In practice this approach greatly improved the time to identification and validation of suspected variants, going from 4-6 weeks to 1-2 weeks. Overall, these tools provide a simple, customisable and entirely open source method to identify genetic variants that may be of clinical importance in a variety of genetically important conditions.

B4: Flexible Visualizations for Genome-Scale Genetics with HoloViews

Millar, T.R. [1,2] , Thomson, S.J. [1], McCallum, J.A. [1,2]

[1] Plant and Food Research, Lincoln, NZ, [2] Department of Biochemistry, University of Otago, Dunedin, NZ

Although there are dedicated standalone or server-based visualization tools for standard genome data formats, these have limited flexibility for *ad hoc* visualization and integration into experimental workflows. Statistical graphics packages have greater flexibility but can be limited by scalability and interactivity. The Python HoloViews project (http://holoviews.org) offers a powerful and flexible framework for scalable visualization that adapts well to genetic data and works well in Jupyter Notebooks. Importantly it allows use of multiple rendering engines (MatPlotLib, Bokeh and DataShader) to permit development of multi-layered scalable and interactive plots. We will illustrate use cases where we are employing these approaches for QC and exploration of whole-genome sequencing, genotyping and haplotyping in kiwifruit.

B5: Investigating a multi-species antibiotic-resistant infection using MinION sequencing

Jenny Draper[1], Xiaoyun Ren[1], Susan C. Morpeth[2], David Holland[2] and Helen Heffernan[3]

[1]Health and Environment Group, ESR, Wellington, NZ; [2]Infection Services, Middlemore Hospital, Counties Manukau District Health Board, Auckland, NZ; [3]Antimicrobial Reference Laboratory, ESR, Wellington, NZ

Carbapenemase-producing Enterobacterales (CPE) are enteric bacteria resistant to carbapenems, the antibiotic class considered the last line of defence against serious gram-negative bacterial infections. In New Zealand, infections with CPE are often associated with overseas travel. Their detection is of significant concern, as carbapenemase genes are carried on mobile elements which can easily spread under antibiotic pressure, and plasmid transmission has been implicated in hospital CPE outbreaks. Here we describe using MinION sequencing to investigate the case of a hospital patient infected with multiple species of CPE.

Four distinct carbapenemase-producing bacterial species were isolated from clinical and screening specimens from a single patient during a three-week period of hospitalisation. Standard Illumina sequencing revealed a common ~11 kb contig containing the *bla*$_{\text{NDM-1}}$ carbapenemase gene and a ~46-52 kb contig carrying the signature of IncA/C2, a highly efficient conjugative plasmid. This suggested that extensive transmission of a *bla*$_{\text{NDM-1}}$ carrying IncA/C2 plasmid between species had occurred within the patient. However, this could not be determined reliably from the incomplete Illumina short-read assemblies.

We therefore turned to Oxford Nanopore long-read sequencing, using a rapid barcoding kit on the same DNA preparations used for the Illumina libraries, and multiplexing up to 11 isolates on a single MinION flowcell. Even with the multiplexing and non-optimised DNA extraction, a simple hybrid assembly with UniCycler revealed that a novel 167 kb IncA/C2 plasmid had been transmitted between bacterial species within the patient. This plasmid carries multiple antibiotic resistance genes, including the *bla*$_{\text{NDM-1}}$ carbapenemase, thus explaining the multi-species CPE infection.

B6: Genomics Aotearoa: Aiming to improve the use of genomics in New Zealand

Dearden, P.K.[1]

[1]Genomics Aotearoa and Biochemistry Department, University of Otago. P.O. Box 56, Dunedin, New Zealand.

Genomics Aotearoa is an MBIE funded partnership between 3 Universities and 4 CRIs tasked to improve the use and uptake of genomics in New Zealand. Genomics Aotearoa aims to build capacity and capability in genomics and bioinformatics through exemplar projects that develop tools and technology to solve problems of importance to New Zealand. Genomics Aotearoa is developing both bioinformatics infrastructure, and skilled personnel to develop the use of genomics in health, environment and primary production. Many of the issues to be addressed are of importance to Māori, so Genomics Aotearoa will work in partnership with Māori to ensure benefit sharing and the development of trust.

This talk will present the principles of Genomics Aotearoa, introduce its current suite of funded activities and indicate upcoming opportunities for New Zealand researchers.

B7: Genomics Aotearoa – bioinformatics as underpinning infrastructure for genomics research in New Zealand

Black, M.A. [1]

[1]Department of Biochemistry, University of Otago, Dunedin, NZ

In early 2018, the New Zealand Government announced the funding of Genomics Aotearoa, a $35 million investment in genomics research over seven years, as a collaborative endeavour involving the University of Otago, University of Auckland, Massey University, AgResearch, Environmental and Science Research, Landcare Research, and Plant and Food Research. Genomics Aotearoa provides funding for exemplar projects across the broad fields of Primary Production, Environment and Health, as well as supporting the development of underpinning research infrastructure in bioinformatics and Vision Mātauranga.

This talk will provide an overview of the Bioinformatics component of Genomics Aotearoa, focusing on the goals and operational structure, as well as outlining the computational resources that are being put in place, and the training and community building opportunities that are being developed.

B8: The Primary Production theme within Genomics Aotearoa

Brauning R [1] , Chagne D [2] , Buckley T [3] , Garrick D [4] , Lee M [5]

[1] AgResearch, Dunedin, New Zealand, [2] Plant and Food, New Zealand, [3] Landcare Research, New Zealand, [4] Massey University, New Zealand, [5] University of Otago, New Zealand

In New Zealand, to support conservation and primary production, we need high quality genomes of particular species, many of which are large, very complex, repetitive, or highly variable. We also need effective pipelines for the production of these high quality genomes. The project “High Quality Genomes”, co-led by David Chagne (Plant and Food) and Thomas Buckley (Landcare Research) addresses this initially for epichloë, ryegrass, blueberry, ship rat, pinus, and a range of species of conservational interest.

Our biological industries rely on genetic improvement to increase profitability, to remain competitive and to mitigate risk. We need to better use the data we have, and develop new technologies to gather new data sources and integrate them into breeding value prediction. Tools that deal with the statistical challenge need to be made available to the wider primary industries to ensure all the primary production industries can cost-effectively adopt genomic selection in their own circumstances.

The project “Better breeding values”, co-led by Dorian Garrick (Massey University) and Michael Lee (University of Otago) addresses this initially for cattle and sheep.

B9: Genomics Aotearoa and the Environmental metagenomics subprogram: Pilot study in the Puarenga stream

Waite, D.W.[1]; Tee, H.S.-[1]; Stott, M.[2]; Wood, S.[3]; Lee, C.[4]; Weaver, L.[5]; Weir, B.-[1,6]; Lear, G.[1]; Handley, K.M.-[1]

[1]School of Biological Sciences, University of Auckland, Auckland, NZ, [2]School of Biological Sciences, University of Canterbury, Christchurch, NZ, [3]Cawthron Institute, Nelson, NZ, [4]Biological Sciences, University of Waikato, Hamilton, NZ, [5]ESR, Christchurch, NZ, [6]Landcare Research, Auckland, NZ

The Genomics Aotearoa initiative was established in 2017 with the aim of supporting genomics research and improving bioinformatics and genomic research capacities in New Zealand, in order to facilitate advancements in health, conservation, and environmental research. The Environmental Metagenomics subprogram of Genomics Aotearoa is tasked with recovering high quality draft genomes of uncultivated microorganisms from the mixed microbial communities in the environment, and with developing capability in metagenomics. Metagenomic analysis represents a major advancement in microbial research, providing a means to bypass the traditional cultivation bottleneck and directly access microbial communities within an ecosystem. Ecological and evolutionary hypotheses gained from genomic insights can them be tested through other culture-independent methods.

The Puarenga Stream in Rotorua has been selected as the sampling site for this subprogram, with the goals of (1) using recovered genomes to predict the lifestyles of stream microorganisms across a spectrum of environmental conditions; (2) determining how selective forces and spatial separation influence genome composition; and (3) using genomic data to perform cross-kingdom (microbial-eukaryotic) analyses of stream communities. The team will use longitudinal and temporal gradients to examine the effects of water quality changes due to nutrient and thermal spring inputs on the microbial populations within the stream. Our final goal is to develop methods, and distribute our knowledge and experience to other New Zealand researchers through hosted workshops and software dissemination platforms.

B10: Cultural Considerations When Working with the Genomes of Taonga Species

Benjamin Iwikau Te Aika[1]

[1] Genomics Aotearoa, University of Otago, Dunedin, NZ

Taonga species and the use of taonga species will be discussed as widely as possible in the context of cultural considerations. Taonga species include taonga systems and need not be limited to native species, historical heritage, conservation or other limiting terms of reference.

Māori proprietary rights to fauna and flora remains an unresolved issue and is a matter of continued contention, despite the determination of the Wai262 report.

The Māori world view considers the continuance of pre-existing authority over natural resources. Therefore, the sequencing of material and conversion to data without fully informed consent to specifically extinguish pre-existing rights of Māori may not support the disassociation of pre-existing authority with data. Such sequencing projects are, therefore, challenging to Māori, and may be illegitimate when considered against the international human rights framework.

The notions such as “public domain” or the “commons” for Māori and indigenous people has been experienced as a precursor for alienation, annexation and acquisition. This pattern is also usually followed with corporate acquisition of any valuable commercial elements. The “public” or “commons” in knowledge also follows this pattern. The Convention of Biological Diversity now provides some support for indigenous people with stronger provisions in relation to the role indigenous peoples play in bio-diversity.

Māori peoples have a long history of an inclusive and magnanimous culture when presented with a mutually beneficial approach. The challenge for researchers will be in a shift in the culture of the professional researcher to one where best practice research models are normalised. The challenge for funding institutions will be in the design of research contracts that allow for the time-frame required for proper consultation and co-development with communities. The challenge for Māori communities is in being prepared for the tsunami which has arrived in the form of the data economy.

B11: Pinus radiata genome assembly

Sturrock, S.[1,2] Telfer, E.J.[1], Winkworth, R.C.[3], Graham, N.J.[1], Macdonald, L.J.[1], Frickey, T.[1], Wilcox, P.L.[4], and Dungey, H.S.[1]

[1]Scion, Rotorua, NZ, [2]Real Time Genomics, Hamilton, NZ, [3]Institute of Fundamental Sciences, Massey University, Palmerston North, NZ, [4]Department of Mathematics and Statistics, University of Otago, Dunedin, NZ.

Conifers are abundant and economically important, with large, complex genomes that are technically challenging to assemble. The haploid radiata pine genome consists of approximately 25 billion base pairs – similar in size to that of other conifers [1] and approximately 8-10 times that of the human genome. Using a combination of high throughput sequencing technologies and bioinformatics workflows we have assembled a draft genome of *Pinus radiata*. Of this genome, approximately 1.5 – 2% is likely to be expressed genes [2]. Preliminary pairwise comparisons indicate that 74% [3] of the putatively expressed genes align between the radiata and loblolly pine [4]. The remaining 98-98.5% of the genome is composed of non-coding DNA; including non-expressed regulatory elements, and complex repetitive elements.

The assembly of the radiata pine genome has been confounded by genome size and the extent of the repetitive elements. This has meant heavy computational demands and necessitated the purchase of a dedicated super computer. Additionally, we have encountered software problems due to the volume of data to be assembled. Methods for assembling very large genomes are still in active development and a variety of assemblers designed to handle this type of genome were evaluated. A workflow that combines the most effective tools has emerged. We present here the preliminary draft of the *Pinus radiata* genome.

1. Ahuja MR, Neale DB: *Evolution of genome size in conifers.* Silvae Genetica 2005, 54(3):126-137.
2. Wegrzyn JL, Liechty JD, Stevens KA, Wu LS, Loopstra CA, Vasquez-Gross HA, Dougherty WM, Lin BY, Zieve JJ, Martínez-García PJ et al: *Unique features of the loblolly pine (Pinus taeda L.) megagenome revealed through sequence annotation.* Genetics 2014, 196(3):891-909.
3. Telfer E, Graham, N, Macdonald, L, Sturrock, S, Wilcox PL, Stanbra, L.: *An approach to variant discovery for conifer transcriptome sequencing* (2018). Under revision for PLoS ONE.
4. Zimin A, Stevens KA, Crepeau MW, Holtz-Morris A, Koriabine M, Marçais G, Puiu D, Roberts M, Wegrzyn JL, de Jong PJ et al: *Sequencing and Assembly of the 22-Gb Loblolly Pine Genome.* Genetics 2014, 196(3):875-890.

B12: Why NeSI is attending a Bioinformatics meeting

Lohani, N.[1]

[1] New Zealand eScience Infrastructure

The New Zealand eScience Infrastructure (NeSI) is the national provider of advanced research computing services, tasked with growing the capability of NZ research. We provide access to high performance compute for researchers, from a variety of research organisations. The majority of projects being run on our systems hail from the bioinformatics domain.

In this presentation we will provide you with a run-down on who we are, how our platforms are set-up for bioinformatics research and how you can access NeSI.

B13: Doing it in the dark: Methods for analysing and visualising transcriptomes without a reference genome

Davidson, N. M.[1,2]

[1]Murdoch Children's Research Institute, Parkville, Victoria, Australia, [2]School of Biosciences, University of Melbourne, Victoria, Australia.

Next-generation sequencing of RNA, RNA-seq, is a powerful technology for studying various aspects of the transcriptome including differential gene expression, differential splicing, gene discovery and identification of sequence variants. Significantly, RNA-Seq allows the transcriptome of non-model organisms to be generated without the need for a reference genome or annotation. Instead, sequences of expressed transcripts are reconstructed by *de novo* assembling the data. However, analysis and visualisation of this data beyond assembly for expression analysis remains challenging. In addition assembled transcriptomes may be inaccurate and the procedure to go from a *de novo* assembled transcriptome to scientific results has been unclear.

In this talk I will present a collection of methods we have developed to enable transcriptome analysis and visualisation in non-model organisms. In brief, our methods allow a multitude of standard analytical tools and techniques to be applied across all species, by converting a *de novo* assembled transcriptome into a pseudo reference genome. I will also show how the analysis of species where a reference genome is available, but incomplete, is improved because reference and assembled transcripts can be easily combined. In species such as chicken and sheep, we find hundreds of new genes using these techniques. Finally, I will discuss the exciting application of our methods to cancer. The cancer genome is riddled with variants and structural rearrangements so that it is no longer a perfect match to the human reference and cancer data can therefore benefit from *de novo* assembly approaches.

B14: From forensic genotype to forensic phenotype

McNevin, D.[1], Cheung, E.Y.Y.[2], Phillips, C.[3]

[1]Centre for Forensic Science, School of Mathematical & Physical Sciences, Faculty of Science, University of Technology Sydney, [2]National Centre for Forensic Studies, Faculty of Science, University of Canberra, [3]Instituto de Ciencias Forenses "Luís Concheiro", Universidade de Santiago de Compostela.

Until recently, forensic DNA profiling has been restricted to the use of identity informative markers, including high heterozygosity microsatellites and single nucleotide polymorphisms (SNPs), with the aim of differentiating between individuals. This approach requires a questioned profile to be compared with a reference profile and can result in a match (inclusion) or a non-match (exclusion). In many instances, there is no match: no database hit or no suspects. In these cases, forensic scientists are now looking to produce "intelligence" about the unknown donor of a DNA profile. This can take the form of biogeographical ancestry (BGA) and/or externally visible characteristics (EVCs) with the ultimate (but perhaps long term) goal being a molecular photofit of the DNA donor.

I will present an overview of the field of forensic phenotyping, with an emphasis on bioinformatics, from its beginnings at the turn of the century when applied to a Louisiana serial rapist and killer and the Madrid (11-M) train bombings, to the current time. While there is much hope that molecular photofitting is possible, progress is slow. Major limitations include the availability of representative reference databases for both BGA and EVCs; the existence of genetic clines rather than discontinuities; the realisation that most phenotypes are polygenic with many SNPs each contributing a small effect; and the significant contributions of epistasis, epigenetics and the environment to phenotype.

B15: Challenges and opportunities in transforming Forensic DNA genotyping to massively parallel sequencing technology

SallyAnn Harbison[1], Ryan England[1], Kate Stevenson[1], Jayshree Patel[1], Lisa Melia[1], Alex Liu[2], Nadine Dunn[1] and Rachel Fleming[1].

[1] Forensic Biology, Institute of Environmental Science and Research Ltd, Mt Albert Science Centre, 120 Mt Albert Road, Sandringham Auckland, NZ, [2] Forensic Science Programme, School of Chemical Sciences, University of Auckland, NZ.

Worldwide, forensic genotyping still relies on the analysis of short tandem repeats and capillary electrophoresis for forensic genotyping. However, a small number of laboratories have embraced next generation sequencing recognising the potential for increased investigative intelligence information and greater discrimination. This transformation is however, not without substantial challenges. In this presentation I will focus on the analysis of short tandem repeats, the mainstay of forensic genotyping. First I will outline the current landscape including statistical models used for data interpretation, the use of DNA profile databases and the quality assurance measures required. I will then describe the challenges faced by the forensic community in moving towards a sequenced based approach including nomenclature, bioinformatics, transformation of DNA Databases and method validation to ensure reliability and reproducibility. validation. Of course forensic samples from crime scenes also contain RNA, and at ESR we have been one of a very small number of laboratories using mRNA to identify the source of body fluids in case samples. I will present work describing the application of RNA Seq to solve some of the inherent challenges in identifying stable and specific RNA markers and the bioinformatics we have developed to enable use by forensic caseworkers with limited bioinformatics expertise.

B16: Machine learning methods in the analysis of forensic sequencing data

Yao-Yuan Alexander Liu[1], SallyAnn Harbison[2]

[1]Forensic Science Program, School of Chemical Sciences, University of Auckland, [2]Institute of Environmental Science and Research Limited, Private Bag 92021, Auckland 1142 New Zealand

The arrival of massively parallel sequencing platforms into forensic science reveals previously unseen complexity and variability of the genetic markers used, along with new computational problems in the analysis of the data.

Short tandem repeats (STRs), the current gold standard of forensic DNA typing, are highly variable and problematic to genotype in massively parallel sequencing data. In recent years, a variety of bioinformatic tools have been developed for the extraction of STR sequences from raw sequencing data. Each of these tools employ a distinctive method for STR detection, with the two most utilized approaches to identifying STRs in sequencing reads being the detection of known features such as repeat motifs and the flanking sequences adjacent to the repeats. Often these features are treated as discrete clues to the identity of the STR, which sometimes lead to the extraction of artefact alleles.

To take a more holistic approach, we employed machine learning methods to model the behaviour of different features present in STRs. Instead of manually specifying rules for STR sequence structure, such as motifs and their expected number of repeats, machine learning algorithms may discover features or rules from the training dataset that are more appropriate for the task.

B17: Striking loss of global DNA methylation levels and constitutive PD-L1 expression in melanoma

Chatterjee, Aniruddha [1], Rodger, Euan [1], Ahn, Antonio [1], Stockwell, Peter [2] , Parry, Matthew [3], Motwani, Jyoti [1], Gallagher, Stuart [4], Eccles, Michael [1] and Hersey, Peter [4]

[1] Department of Pathology, Dunedin, University of Otago, New Zealand, [2] Department of Biochemistry, Dunedin, University of Otago, New Zealand, [3] Department of Mathematics & Statistics, Dunedin, University of Otago, New Zealand, [4] Centenary Institute, Sydney, University of Sydney, Australia

One of the hallmarks of cancer (including melanoma) is its ability to evade the host immune system by up-regulating immune checkpoints. The programmed death-ligand 1 (PD-L1) receptor is one of most important immune checkpoints and is often upregulated in cancer cells. Recent evidence suggests that in melanoma, the patients with PD-L1 expression and absense of tumour infiltrating lymphocytes (TILs) (i.e. "constitutive PD-L1 or PD-$L1_{CON}$") show worse response rates and have a worse prognosis than patients with PD-L1 expression and the presence of TILs (i.e. "inducible PD-L1 or PD-$L1_{IND}$"). However, how PD-L1 expression is regulated in melanoma cells remains elusive. Understanding the mechanisms of how PD-L1 is regulated is important for predicting responses for anti-PD-L1 treatment and for developing new combinatorial therapies. We hypothesised that genes and transcription factors involved in pathways that normally regulate PD-L1 expression are either silenced (by hypermethylation) or activated (by hypomethylation) by epigenetic mechanisms in melanoma. As a result, expression or repression of PD-L1 in these cancer cells is determined by their epigenetic status.

To address these questions, we have generated whole-genome scale DNA methylomes (using reduced representation bisulfite sequencing, RRBS) and transcriptomes (RNA-Seq) for 12 patient derived melanoma cell lines (six PD-$L1_{IND}$ and six PD-$L1_{CON}$). We discovered extensive global hypomethylation in the constitutive lines, particularly pronounced in intergenic repeat regions and gene bodies. A high proportion of hypomethylated regions exhbited dichotomous methylation patterns indicating a common regulatory mechanism between the inducible and constitutive groups. RNA-Sequening data indicated that the hypomethylated state of the PD-$L1_{CON}$ cells was correlated with higher upregulation of the differentially expressed genes at a global-scale and the upregulated genes were associated with several cancer hallmark properties. The upregulated genes exhibited expression signatures of viral mimicry and cytosolic sensing of dsRNA genes similar to what has been observed after DNA methyltransferase inhibitor (DNMTi) treatment in cancers. Further, our analysis reveals that expression of a subset of epigenetic regulator genes were strongly correlated with PD-L1 expression and global methylome status in PD-$L1_{CON}$ and PD-$L1_{IND}$ cell lines. Finally, we show inhibition of DNA methylation resulted in increased PD-L1 transcription in the PD-$L1_{IND}$ cell lines, demonstrating that global hypomethylation mediated changes regulate PD-L1 expression in melanoma. We believe these results are the first to show that DNA methylation levels play a role in regulating PD-L1 on melanoma and suggest they may have important implications for combined treatments targeting DNA methylation (DNMTi) and PD1/PD-L1 (anti-PD1 antibodies).

B18: Multi-omics Analysis in Primary Cell Culture Models Reveals the Genetic and Epigenetic Basis of Phenotypic Diversity within Tumors

Sixue Liu[1], Zuyu Yang[1,2], Xiaolan Xu[3], Xuemei Lu[1]

[1]Key Laboratory of Genomics and Precision Medicine, Beijing Institute of Genomics,Chinese Academy of Sciences, Beijing, China, [2]Invasive Pathogens Laboratory, Institute of Environmental Science and Research, Porirua, Wellington, NZ, [3]National Key Laboratory of Biomacromolecules, Institute of Biophysics, Chinese Academy of Sciences, Beijing, China.

Uncovering the functionally essential variations related to tumorigenesis and tumor progression from cancer genomics data is still challenging due to the genetic diversity between patients, and extensive inter- and intra-tumor heterogeneity at the genomic, epigenomic, and transcription levels. To rule out the effects of the differences in germline genetic background, in this study, we successfully established multiple primary cultures from the primary and recurrent tumors of a hepatocellular carcinomas patient. Multi-omics sequencing was performed for these heterogeneous cells that encompassing the diversity of tumor cells within this patient. Variations in the genome sequence, epigenetic modification, and gene expression were used to infer the relationships of these cell cultures. We observed the discrepancy between phylogenetic relationships revealed by single nucleotide variations (SNVs) and transcriptional/epigenomic profiles in cell cultures. Significantly overlapping between sample-specific mutated genes and differentially expressed genes (DEGs) were not observed, suggesting that most of the heterogeneous SNVs among the lineages within the patient were functionally insignificant. Moreover, copy number alterations (CNAs) and DNA methylation variation within gene bodies, rather than promoters, were significantly correlated to differences in gene expression among these cell cultures. Pathway analysis of CNA/DNA methylation-related genes indicated one cell clone from the recurrent tumor exhibited distinct cellular characteristics and tumorigenicity, which were confirmed by cell experiments both *in vitro* and *in vivo*. Our systematic analysis revealed that CNAs and epigenomic changes, rather than SNVs, are more likely to contribute to the phenotypic diversity among subpopulations in tumors, suggesting new therapeutic strategies targeting gene dosage and epigenetic modification should be considered in personalized cancer medicine. This culture model may be applied for further identifying the plausible determinant of cancer metastasis and relapse.

B19: Gene expression effects of sodium valproate in a serotonergic cell line

Sinha P.[1], Cree S.[1], Miller A.L.[1], Pearson J.F.[2], Kennedy M.A.[1]

[1] Department of Pathology and Carney Centre for Pharmacogenomics, University of Otago, Christchurch, NZ, [2]Biostatistics and Computational Biology Unit, University of Otago, Christchurch, NZ.

Sodium valproate (VPA) is a histone deacetylase (HDAC) inhibitor, widely prescribed in the treatment of bipolar disorder. Molecular and pharmacological studies have revealed some relevant properties and targets but the precise modes of action are not yet understood[1-2]. Some of its therapeutic effects are attributed to HDACi property. We have carried out a range of experiments to test the effects of VPA on gene expression in a cell line (RN46A), derived from the rat medullary raphe nucleus which represents a relevant model for the mood stabilising effects of VPA.

RNA-Seq analysis of cells treated with VPA (0.5mM) for 72 hrs induced widespread changes in expression of many genes, with >700 significantly upregulated and >400 downregulated. High correlation (r^2=0.75) was observed between RNA-Seq and NanoString for selected 23 genes. Clear validation was observed for *ADAM23, LSP1, MAOB, MMP13, PAK3, SERPINB2, SNAP91, WNT6,* and *ZCCHC12*.

We also explored the effects of the VPA analogue valpromide, and HDAC inhibitors including trichostatin A and CI994, on the selected genes. Expression of eight genes was modified by HDACi: *ZCCHC12* and *SHANK3* were upregulated by VPA and CI994 but downregulated by trichostatin A; *CDKN1C, MAOB, NGFR* and *WNT6* were upregulated by CI994 only, and *MMP13* and *VGF* were upregulated by TSA only.

We observed extensive gene expression changes in response to VPA with several genes involved in neuronal function or nervous system development. The complex regulatory effects of different HDACi or the non-HDACi VPA analogue valpromide on the selected genes suggest VPA can exert its regulatory effects via both HDACi-dependent and independent properties. Understanding the broader gene regulatory effects of VPA in this serotonergic cell model should provide insights into how this drug works, whether other HDACi compounds may have similar gene regulatory effects, and perhaps highlight molecular processes that may underlie regulation of mood.

1. Chiu, C.T., et al. (2013). *Therapeutic potential of mood stabilizers lithium and valproic acid: beyond bipolar disorder*. Pharmacol Rev. 65: 105-42.
2. Atmaca, M. (2009). *Valproate and neuroprotective effects for bipolar disorder.* International review of psychiatry. 21: 410-3.
3. White, L., et al. (1994). *Distinct regulatory pathways control neurofilament expression and neurotransmitter synthesis in immortalized serotonergic neurons*. The Journal of Neuroscience 14, 6744–53.

B20: Investigating Repetitive Sequences in Ultra-long Nanopore Reads

Eccles, D.A.[1]

[1] Malaghan Institute of Medical Research, Wellington, NZ

Imagine standing in the middle of a dense forest, being overwhelmed by the diversity you see. You start walking and take pictures of what appears to be a world of randomness. Later on, you notice that many pictures have a vine that hangs in a particular way, with a little fern underneath.

This is similar to what I can see in DNA sequences. On the surface, there doesn't seem to be anything remarkably different between the sequence ACTCCAGCCTGGTGACAGAGTGAGA and the sequence ACTCCAGCTCACAGTCCTGTCGATG, yet the first appears 41 times, and the second only once in a DNA sequence 15,062,723 bases long.

I have created a tool to aid in the discovery of repetitive DNA patterns. The tool works by carrying out a kmer survey of a DNA sequence, and identifying repeat distances and abundances of the most repetitive kmer. Additional features include an analysis of repetitive reverse and complement patterns, which expose repetitive features of DNA that are less emphasised in most sequence mapping tools.

This tool is supplemented with graphical representations of DNA to explore DNA at a megabase scale. The graphical representations use a similar kmer-based system for displaying repeated features, compressing the entire sequence (regardless of length) into a single linear, semicircular, or circular "profile plot". This allows for very quick eyeball comparisons of DNA sequences at megabase scales based on infrequent, but repetitive, patterns in the sequences.

It is my hope that this tool, combined with the visualisations, will eventually lead to new discoveries about the large-scale structure of DNA, and an understanding of why these non-random patterns exist.

Summary of Abstracts for the Poster Session

No.	Title	Presenter	Institution
B21	Movement of genes, song, and morphology across the black-capped/Carolina chickadee hybrid zone in Missouri	Alana Alexander	University of Otago
B22	Meta-analysis of mRNAs loaded into exosomes suggests cell type specific RNA localisation elements	Harry Biggs	University of Otago
B23	Analysis of transcript splicing	Chun Shen Lim	University of Otago
B24	The Twilight Zone of Homology	Stephanie McGimpsey	University of Canterbury
B25	Prediction of the host domain for partial viral genomes from metagenomes	Thomas Nicholson	University of Otago
B26	Differentially methylated genes and pathway perturbation in adipose tissues from gastric bypass intervention	Alessandra Santana	Environmental Science and Research

B21: Movement of genes, song, and morphology across the black-capped/Carolina chickadee hybrid zone in Missouri

Alexander, A.[1,2], Robbins, M.[2], Holmes, J.[2,3], Moyle, R.[2,3], Peterson, A.T.[2,3]

[1]Department of Anatomy, University of Otago, Dunedin, NZ, [2]Biodiversity Institute, University of Kansas, Lawrence, Kansas, USA [3]Department of Ecology and Evolutionary Biology, University of Kansas, Lawrence, Kansas, USA.

The black-capped (*Poecile atricapillus*) and Carolina chickadee (*P. carolinensis*) contact zone occurs across a narrow latitudinal band in the USA, ranging from Kansas in the west to New York in the east. Genetic and morphological studies in Pennsylvania (PA) and Ohio (OH) demonstrate northward movement through time of the hybrid zone located within the contact zone, likely in response to climate change. In contrast, analysis of song data in Illinois suggested little movement of the hybrid zone in this area. Here, we focussed on characterising a temporal shift in the understudied western portion of the hybrid zone in Missouri (MO), using 67 samples from 1980 and 92 samples from 2016, including 3 black-capped and 2 Carolina chickadee reference samples obtained from outside the hybrid zone. We compared and contrasted patterns in the morphological, song and genetic (using 11,834 SNPs derived from ddRADseq) datasets for these birds. While the hybrid zone did appear to move northwest across our temporal sample based on genetic markers, it did not move at the rapid rate seen in other areas in the United States (e.g. PA, OH). We suggest this slower movement is due to differences in warming between areas: the temperature change between 1976-2016 in PA (where the contact zone has moved rapidly) is as much as 50% greater than in MO.

B22: Meta-analysis of mRNAs loaded into exosomes suggests cell type specific RNA localisation elements

Harry Biggs[1], Christopher Brown[1], Stephanie Hughes[1]

[1] Department of Biochemistry, University of Otago

Exosomes, also known as extracellular vesicles, have recently become a cellular communication mechanism of interest in many areas. These vesicles are capable of transmitting many biologically active molecules such as mRNA, miRNA and proteins. This is especially seen in cancers, where vesicles from tumours can drive a regular cell towards oncogenesis. As the contents of these vesicles are biologically active and have roles in cellular communication, it is suggestive of an exosome specific signal present for the targeting of contents to exosomes. Using publicly available microarray data, we set out to discover the mechanism behind the vesicle mRNA loading in eight different cell types.

Analysis included searches for overlapping differentially expressed mRNAs. The 3' UTRs were then analysed for protein binding sites, miRNA binding sites, other nucleotide motifs, and RNA secondary structures. These analyses revealed several binding sites, motifs and structures that were unique to the vesicles compared to the control cellular mRNA. There was also a distinct lack of overlap between the cell types, for example out of a total of 774 differentially localised mRNAs, only 18 had any overlap between different cell types. These data together suggest that rather than a single common loading mechanism there are in fact several mechanisms, each unique to that cell type.

B23: Analysis of transcript splicing

Chun Shen Lim[1], Chris M Brown[1]

[1] Department of Biochemistry, School of Biomedical Sciences, University of Otago, Dunedin, New Zealand

Analysis of transcript splicing requires accurate genome annotation. For well-annotated genomes, in-depth analysis such as completeness of splicing can be done on RNA-seq data. However, this is challenging for poorly annotated genomes. To overcome this problem, we suggest to first re-annotate the splice sites using ReSplicer, which is based on orthology. For *ab initio* detection of splice sites, we proposed a combination of two-pass mapping and transcriptome assembly using STAR and StringTie, respectively. Furthermore, the strength of splice sites could be predicted using tools such as MaxEntScan

B24: The Twilight Zone of Homology

Stephanie McGimpsey[1,2,3] and Paul P Gardner[1,2,3]

[1] School of Biological Sciences, University of Canterbury, Christchurch, NZ, [2] Biomolecular Interaction Centre, University of Canterbury, Christchurch, NZ, [3] Department of Biochemistry, University of Otago, New Zealand

Sequence similarity is often used to infer homology between sequences but the relationship between the two breaks down as the number of similar residues reduces. Therefore the alignment score for a certain interval of pairwise similarity can no longer be used as a way to determine homology without other evidence. This interval is the Twilight Zone; the area of pairwise similarity where homology prediction accuracy by alignment score alone is less then 50% accurate. This zone has been well characterised for protein alignments, but less comprehensive work has been done on the nucleotide level. A unique database has been curated that aims to minimise bias through phylogenetic and gene diversity as well as ensuring each pairwise identity has a representative sample of the data. This was to improve upon earlier work where researchers used databases designed for other projects which contained flaws that affected the calculation of the nucleotide alignment homology Twilight Zone. Current estimations of the Twilight zone have its upper bound between 50 - 60% which is much higher then for protein alignments (20 - 35%) and higher then anticipated (based on probability of similarity by chance given a 4 letter alphabet).

This work is on-going and has no definitive results yet.

B25: Prediction of the host domain for partial viral genomes from metagenomes

Thomas Nicholson[1,2] and Chris Brown[1,2]

[1] Department of Biochemistry, University of Otago, PO Box 56, Dunedin 9054, New Zealand,
[2] Genetics Otago, University of Otago, New Zealand.

Diverse viruses are integral parts of biological systems. New technologies including metagenomic analysis has allowed many more viral sequences to be found. Its is expected that virome metagenomes contain a mix of sequences that may be from viruses infecting eukaryotic, bacterial or archaeal hosts. However, one challenging part of viral metagenomics has been identifying the likely host from a partial viral sequence. Archaeal viruses are apparently the least common, and their sequences are underrepresented in databases. They are not easily detected in viromes, we address this problem here. A set of Hidden Markov Models was made for all viral proteins from DNA viruses with known hosts in each of the three domains of cellular life (Archaea, Bacteria, Eukarya). These models were analysed to identify those found specifically in viruses from one domain. In a test set of viral proteomes from viruses with known hosts, multiple models matched proteins and predicted the correct host. By using these 24 199 models we were able to predict the domain for polycistronic sequences from metagenomes. Most predictions were for Bacteria or Eukarya hosts, but the program also enabled prediction of Archaeal viruses from metagenomic sequences.

B26: Differentially methylated genes and pathway perturbation in adipose tissues from gastric bypass intervention

Santana, A.F.[1,2], Macartney-Coxson, D.[1], Benton, M.C.[1], Black, M.A.[2]

[1]Biomarkers Group, Environmental Health, Institute of Environmental Science and Research (ESR), Wellington, New Zealand, [2]Department of Biochemistry, University of Otago, Dunedin, New Zealand.

DNA methylation is a key player in modulating gene expression, regulating changes in phenotype without altering the underlying DNA sequence. This epigenetic mechanism had been shown to relate to a wide range of pathologies, including: obesity; type-two diabetes; cardiovascular disease and cancer. The dynamic nature of epigenetic signatures in obesity can be demonstrated by studies examining DNA methylation before and after an intervention process, such as diet, exercise and/or bariatric surgery. Several such studies have related changes in DNA methylation to obesity traits, including BMI, body percentage fat and waist circumference. Methylation has been measured in a range of tissues, from sperm cells, blood and muscle, to adipocytes. Adipose is the main tissue for energy storage and, moreover, is related to endocrine functions, which can be impacted by obesity. Therefore, studying methylation profiles in adipose tissues could provide insight into the relationship between epigenetic mechanisms (i.e. DNA methylation) and obesity, a condition which affects over 650 million people around the world, and almost a third of the New Zealand population.

We have performed a preliminary analysis of epigenomic data from 15 women who underwent gastric bypass surgery, with DNA methylation assessed pre and post weight loss. Biological pathways exhibiting enrichment for methylation changes (20% or above) included Extracellular Matrix Organization (11 women), Developmental Biology (10 women), Signal Transduction (9 women), Muscle Contraction (7 women), Cell-Cell Communication (6 women), Homeostasis (6 women) and Immune System (5 women). Our work utilised a novel analytic approach, where changes in gene-related methylation were used to explore potential pathway perturbation, on a per-individual basis. The results present a pathway-centric view of the impact of differential methylation in response to bariatric surgery, and provide additional insight into the mechanisms associated with weight loss.