G1: Finishing (small) eukaryotic genomes

Cox, M.P.[1], Winter, D.J.[1], Ganley, A.R.D.[2], Young, C.A.[3]
[1]School of Fundamental Sciences, Massey University, Palmerston North, NZ, [2]School of Biological Sciences, University of Auckland, Auckland, NZ, [3]Noble Research Institute, LLC, Ardmore, Oklahoma, USA.

Sequencing the genome is increasingly a standard first step in any study of a new non-model organism. Tens of thousands of species have now had their genomes sequenced. Except, of course, that they haven't. Almost every genome sequence is a draft – an unordered collection of hundreds or more fragments with unknowable levels of missingness. This talk discusses finished genome sequences – chromosome assemblies, telomere-to-telomere, no Ns, no gaps. We consider how small eukaryote genomes can be finished, where finished genomes can be useful, and why you may often not want to bother. These scenarios are illustrated by several finished fungal genomes.

G2: Revealing the hidden diversity of viruses in fish using meta-transcriptomics

Jemma L. Geoghegan
Department of Biological Sciences, Macquarie University, Sydney.

The total assemblage of viruses on earth is termed the 'virosphere', the vast majority of which is undescribed. It is futile to attempt to predict future emergence events until there has been a dramatic increase in the number of viruses identified and sequenced. Hitherto, this task was greatly hampered by a strong sampling bias toward pathogens that cause disease in humans or in economically important animals and plants. This is set to change with the development of whole transcriptome (RNA) shotgun sequencing. Using 'meta-transcriptomics' I provide a comprehensive view of viruses in aquatic hosts as a powerful system. Viruses are abundant in marine and freshwater environments. Crucially, fish likely harbour a greater diversity of viruses than any other class of vertebrate, and all virus families previously thought to only infect mammals have recently been found in bony fish. These viruses are the likely progenitors of viruses that infect a broad range of hosts, including mammals, amphibians and birds.

G3: The impact of the genetic background on gene deletion phenotypes in *Saccharomyces cerevisiae*

Bede P. Busby[1,2] *, Marco Galardini[2*], Cristina Viéitez[1], Alistair S. Dunham[2], Athanasios Typas[1] and Pedro Beltrao[2]
[1] European Molecular Biology Laboratory, Genome Biology Unit, Heidelberg, Germany,
[2]European Molecular Biology Laboratory, European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, UK, *authors contributed equally

Genetic functions have been investigated via chemical genomics in various species, however, these are generally limited to one strain of that particular species. Analysis of different isolates can provide orthogonal information to that provided from classical genetics, particularly for intra-species phenotypic diversity. In this study we used *S. cerevisiae* genome-wide deletion collections in four genetic backgrounds (S288C, Y55, YPS606, and UWOP587-2421), in order to compare their chemo-genomic profiles, across different environmental stresses. Conditional essentiality shows limited conservation across the strains at the single gene level. In addition, we also compared the growth of 900 *S. cerevisiae* natural-isolates across the same conditions. This also resulted in significant phenotypic differences between strains. Therefore, we conclude that gene function and conditional essentiality are strongly dependent on genetic background.

G4: Genome Assembly Strategies for Endemic New Zealand Species

Ann Mc Cartney[1], Joseph Guhlin[2], Alana Alexander[3], Anna Santure[4], David Chagne[5], Thomas Buckley[1].

[1] Systematics Team, Manaaki Whenua, 231, Morrin Road, Saint Johns Auckland.
[2] Department of Biochemistry, School of Biomedical Sciences, University of Otago, Cumberland Street, Dunedin.
[3] Gemmel Lab, Department of Anatomy, University of Otago, Dunedin.
[4] Commerce A, Room 113, University of Auckland, Symonds Street, Auckland.
[5] Plant and Food Research, Batchelar Road, Fitzherbert, Palmerston North.

Taonga species are those that have a special cultural significance amongst Māori in Aotearoa. As part of Genomics Aotearoa, a high-quality genomes project has been established alongside Māori to generate pipelines for the assembly of endemic and taonga species in New Zealand. These pipelines are specifically targeted at key species that are on the verge of extinction, treasured by Māori, key players in the primary production industry (kiwifruit), a significant threat to biosecurity within New Zealand or are considered genomically complex i.e. are abnormally large, have higher ploidy levels (blueberry), are highly repetitive or heterozygous. These species have been sequenced using a variety of NGS platforms, namely HiSeq/NovaSeq, ONT MinIon and Promethion, PacBio, Chromium 10X and Hi-C sequencing. The utilisation and hybridization of data from these technologies has already resulted in the generation of high quality genomes for invasive wasp species, Polynesian rat, manuka, rewarewa, 15 stick insect species, blueberry, hihi, robin and 125+ kakapo genomes to name but a few. A hybrid assembly pipeline was constructed and utilised to assemble a reference genome for the Phasmatodea clade, Clitarchus hookeri as well the low coverage assembly and phylogenetic analyses of genomes from a further 12 other stick insect species found in geographically distinct regions across Aotearoa to investigate the genomic basis of thermal tolerance and parthenogenesis.

G5: Ancient balancing selection maintains incompatible versions of a conserved metabolic pathway in yeast

Boocock, J[1], Bloom J.S. [1], Kruglyak L. [1]
[1] Department of Human Genetics, University of California, Los Angeles, USA.

Accumulation of genetic incompatibilities can lead to speciation, but there are few detailed examples of this process in action. We discovered that the classical galactose pathway of the yeast *Saccharomyces cerevisiae* exists in two incompatible states maintained by ancient balancing selection. In a mapping study of growth on galactose, we identified a genetic interaction among three unlinked loci in crosses involving the strain CBS2888. All three loci contained components of the galactose pathway (*GAL2, GAL1/10/7,* and *PGM1*). We noted that the CBS2888 galactose alleles were highly diverged from the reference strain. We engineered strains carrying all eight allelic combinations at these loci and showed that the reference version of *PGM1* is incompatible with the diverged alleles of the other genes. The diverged alleles appear to have split from those of most other *S. cerevisiae* strains before the birth of the *Saccharomyces sensu stricto* species cluster ~10-20 MYA. Globally, the diverged alleles are found in isolates from galactose-rich environments, such as cheese, kefir, and milk. Population genetics analysis of linked neutral sites revealed a strong signature of ancient balancing selection. Strains with the diverged alleles grew faster in galactose, but slower in glucose, revealing a trade-off on which this balancing selection may act. Our work shows that balancing selection on alternative versions of a metabolic process can maintain incompatible but functional unlinked alleles for millions of years.

G6: Hologenomics for conservation: a first test of utility using New Zealand’s endemic Māui and Hector’s dolphins

Alexander, A.[1], Constantine, R.[2], Roe, W.[3], Baker, C.S.[4], Heimeier, D.[5], Hamner, R.M.[6], Steel, D.[4], Jeunen, G.-J.[1], Gemmell, N.[1]
[1]Department of Anatomy, University of Otago, Dunedin, New Zealand, [2]School of Biological Sciences, University of Auckland, Auckland, New Zealand, [3]School of Veterinary Science, Massey University, Palmerston North, New Zealand, [4]Marine Mammal Institute, Oregon State University, Newport, Oregon, USA, [5]The Pirbright Institute, Surrey, UK, [6]Division of Fish and Wildlife, Rota, Commonwealth of the Northern Mariana Islands

In the face of the current biodiversity crisis, conservation genetics has been quick to adopt the latest genomic approaches to improve detection of demographic bottlenecks, inbreeding, selection leading to local adaptation, and the spread of maladapted genes. A new approach that has not yet been used in conservation genetics is hologenomics, where the genome of a host organism is considered in tandem with the microbes that naturally occur throughout the host’s body (the microbiome). The New Zealand endemic Māui dolphin (*Cephalorhynchus hectori maui*), numbering only ~63 individuals, and Hector’s dolphin (*C. h. hectori*), thought to number a total of ~24,000 individuals but subdivided into several regional populations, provide a compelling test case of the utility of hologenomics, as significant uncertainty remains about: (i) the potential for genetic interchange between these two subspecies; (ii) changes in their population sizes through time, (iii) the risks of inbreeding within each subspecies, and (iv) individual susceptibility to pathogens e.g. toxoplasmosis. Using an extensive database of individual Hector’s and Māui dolphins collected throughout their ranges, we are sequencing both host and microbial genomes to elucidate the hologenomic basis of disease susceptibility; assess the presence (or absence) of defining hologenomic characteristics between these subspecies; and use the increased resolution of the hologenome to describe more precisely anthropogenic impacts, such as fisheries related mortality, on population sizes and connectivity. This will provide a first ‘proof-of-concept’ with the potential to revolutionise conservation biology on a global scale, as well as improve management of New Zealand’s only endemic cetacean.

G7: Understanding genomic data management – an indigenous perspective of “data protection” and “data sovereignty.”

Ben Te Aika: Vision Mātauranga, Genomics Aotearoa, Kati Urihia, Kati Matamata, Kati Wairaki, Kati Mamoe, Waitaha, Ngati Mutunga, Ngati Tama, Te Ati Awa, Ngati Ruanui.

“What should culturally appropriate genomic data protection look like for indigenous species, data sovereignty?”
Unbundling the myriad of views is a sizable commitment. Understanding the ambiguity in these non-Māori terms such as “culturally appropriate protection” and “data sovereignty” is also a considerable task. Both these terms are discussed and compared to terms more consistent with mātauraka Māori. We consider terms such as “mana Māori” and various traditional expressions offer an improved cultural context and match when translated into English.

Understanding what may be considered “culturally appropriate” is important for Genomics Aotearoa. Quality engagement practices with takata whenua is crucial in reaching this understanding. Understanding the barriers to Māori participation in the science is key to facilitating an environment which fosters Māori exercising rakatirataka and kaitiakitaka. Maintaining the delicate relationship between the researcher and takata whenua is essential. Ensuring that biological whakapapa is maintained under the best cultural conditions possible. New standards for quality engagement in place, which assists the researcher to better integrate the desires of takata whenua with the science is important. A trusted environment is essential, where takata whenua exercise kaitiakitaka and rakatirataka over the biological whakapapa with the research engagement practices being at the highest standards possible. Responding effectively is the best possible solution in a rapidly evolving science and research environment.
“Appropriate protection” or enabling practices, may mean greater institutional ***responsibility*** in facilitating the nexus point where the researcher – takata whenua – data repository oversite develop a shared agreement on the use and storage of data.

G8: Raising eResearch Capability in New Zealand

Megan Guidry[2], Fabiana Kubke[1], Georgina Rae[2], and Nick Jones[2]

[1]University of Auckland, [2]New Zealand eScience Infrastructure

New Zealand eScience Infrastructure (NeSI) provides expertise and capability to researchers conducting computation and data intensive research in New Zealand. Within the training sector, our core purpose is to raise the computational capability of New Zealand research and, in turn, shrink the existing eResearch skills gap.

In this presentation, we will discuss our training efforts so far (both in terms of delivering training, but also cultivating the New Zealand training community) and reflect on the scale of the opportunity/challenge that we face. This will include a brief overview of the Kubke Report - *Growing researchers computational skills to meet future needs* - which reviewed NeSI's training activity and considered current size and demand for a digitally skilled workforce (https://auckland.figshare.com/articles/Growing_researchers_computational_skills_to_meet_future_needs/7482116). Additionally, we will discuss NeSI's position around collaboration, specifically, how we look to leverage a community of practice and partnering organizations when developing and delivering training.

Ultimately, training needs to be useful and relevant to those who need it. NeSI strives to be agile in its approach to training delivery and this presentation will conclude by noting what we are doing today to ensure our efforts are measurable, scalable, and community-focused.

G9: Genomics Aotearoa Training

Faya, N.
Genomics Aotearoa, University of Otago, Department of Biochemistry, Dunedin, New Zealand

A decrease in sequencing cost has seen a large amount of sequence data being generated in the last few years, leading to a paradigm shift from sequencing data generation to data analysis. Despite the ease of data generation, the same cannot be said for data analysis mainly due to fewer researchers with the bioinformatics skills necessary to analyze these datasets. Moreover, most data analysis tools are developed for use with the Linux command line and require use of high-performance computers, therefore there is need for hands-on data analysis training. Empowering researchers through hands-on training courses is the key to improve knowledge and understanding of bioinformatics approaches thereby easing the skills shortage.

Genomics Aotearoa (GA) is a collaborative platform established to ensure that New Zealand is internationally participating and leading in the fields of genomics and bioinformatics. One of GA's projects which is critical to genomics research is bioinformatics capability where bioinformatics tools and strategies needed to analyze information are provided. The bioinformatics capability project aims to address the increasing local demand for data analysis methods as well as training. The concept is: develop material/pipelines that can be accessed by everyone and travel to offer hand-on bioinformatics workshop. Post-doctoral researchers with strong bioinformatics background have been brought on board to develop open-source and reusable data analysis material and pipelines to benefit the genomics research community. A training coordinator has also been hired to assist with the coordination and organization of NZ-wide training workshops and interact with GA researchers and other relevant individuals to determine training needs and priorities. At this stage development of introductory, intermediate and advanced bioinformatics training material for genomics researchers is underway. Together with our partner, NeSI, coordination and delivery of data science and bioinformatics training workshops has already begun around the country.

G10: Discussion session: genomic data sovereignty

Ben Te Aika and Mik Black
Genomics Aotearoa, University of Otago, Department of Biochemistry, Dunedin, New Zealand

G11: NeSI and Genomics: Catering to Domain-Specific High Performance Computing needs

Senanayake, D.S.[1], Scott, C.[1], Healey, M.[1], Maxwell, P[1]

The advancement of sequencing technologies and decreasing cost is responsible in creating an avalanche of genomics data across multiple sub-domains. This data deluge demands an interdisciplinary approach to face the associated challenges such as data storage, security, scalability, data integration, and parallel or high-performance computing solutions for analysis. Ability to deliver solutions to these needs will result in converting highly granular and unstructured genomics data into real scientific insights, which will accelerate the advances being made in areas such as genomics assisted precision medicine and eukaryotic conservation programmes.

NeSI's multiple support teams and bioinformatics researchers have collaborated to address the challenges by various methods of optimisation of processing and analysing data. These methods may vary from a routine task such as utilisation of GPGPUs for accelerated base calling to deploying HPCGridRunner (HGR) to simplify high-throughput grid computing for Bioinformatics applications. For an instance, Trinity transcriptome reconstruction pipeline's phase 2 was transformed from an extremely inefficient single task to hundreds of thousands of commands that can be executed in parallel, each having independent inputs and outputs (embarrassingly parallel). All of these measures are encouraged to be used via the deployment of containers such as Singularity and HPC workflow management systems for better reproducibility.

G12: Simple combination of multiple variant callers to increase accuracy

Trevarton A.J.[1], Chang J.T.[2], Symmans W.F.[1]
[1]Department of Translational Molecular Pathology, The University of Texas MD Anderson Cancer Center, Texas, U.S.A., [2]Department of Integrative Biology and Pharmacology, The University of Texas Health Sciences Center, Texas, U.S.A.

Publications comparing variant caller algorithms present discordant results with contradictory rankings. Caller performances are inconsistent and wide ranging, and dependent upon input data, application, parameter settings, and evaluation metric. With no single variant caller emerging as a superior standard, combinations or ensembles of variant callers have appeared in the literature. In this study, a whole genome sequencing (WGS) somatic reference standard was used to derive some simple general principles to guide variant caller combination for cancer genome sequencing. Then, manually annotated variants called from the whole exome sequencing (WES) of a tumor were used to corroborate these general principles. Finally, we examined noise in targeted amplicon sequencing.

G13: Genomics Aotearoa: environmental genomics and the many different pathways to reconstruct microbial ecosystems and environmental adaptations

Handley, K.M.[1]; 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]
[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

Environmental genomics, metagenomics or community genomics, is a powerful approach for recovering the entire or dominant genetic content of an environment, whether that be the natural environment, built environment, or host associated. There are a multitude of approaches to achieve this. At one end of the spectrum metagenomics can be used to determine the bulk metabolic capacity of environments through large surveys using shallow unassembled reads. At the other higher-resolution end, the predicted lifestyles of individual microorganisms (bacteria, archaea, viruses and eukaryotes), and their unique functional roles in an ecosystem, can be determined through genome assembly. The aim of this project is two-fold. The first is to improve access to metagenomics through workflow documentation, training, and optimised approaches for detangling and analysing complex community genomes. The second is to illustrate the use of integrated metagenomics and metatranscriptomics to understand microbially driven ecological processes in a system encompassing multiple physical and chemical properties (freshwater-to-marine, water-to-sediment, clay-to-sand). All of these attributes represent ecological boundaries, and include different challenges associated with community complexity and nucleic acid recovery. The retrieval of genomic data across these boundaries also provides us with the opportunity to study how major chemical or physical boundaries impact microbial gene flow in the environment.

G14: Inferring cancer population history from a single-cell sequencing DNA methylation data

Moravec, J.C.[1], Bouckaert, R.[2], Diermeier, S.D.[3], Drummond, A.[2], Gavryushkin, A.[1]
[1]Department of Computer Science, University of Otago, Dunedin, New Zealand, [2]Centre for Computational Evolution, University of Auckland, Auckland, New Zealand, [3]Department of Biochemistry, University of Otago, Dunedin, New Zealand

DNA methylation is an epigenetic change in which cytosine is modified to methylcytosine. If this change occurs within a promoter region, this causes silencing of a particular gene. This process has been found to play an important role in carcinogenesis[1]. As DNA methylation changes are usually more frequent than genetic changes, they have also been a target of considerable interest in attempts to reconstruct the development and progression of cancer. However, most methods based on DNA methylation data do not use phylodynamic models. Instead, clustering approaches are used, which can reconstruct general patterns, but not a detailed model-based history of cancer spread.

Kingman's coalescent theory[2] allows quick and efficient reconstruction of population history but assumes a constant well-mixed population with no internal structure or migration. To account for this, coalescent theory can be extended to the structured coalescent[3] to allow population structure and migration. Recent progress in the structured coalescent approach enables efficient simulation of migration between multiple populations of interest[4]. However, current implementations allow only for constant population, which is unsuitable for modelling the spread of rapidly-growing populations of cancer cells.

This can be alleviated by extending the structured coalescent to allow nonparametric models of population size change through time, thus allowing sudden changes in the population dynamics to be uncovered, such as a metastatic development or the start of cancer treatment.

In this work, we extend the structured coalescent approach to estimate population changes during various stages of cancer development and introduce a novel substitution model for DNA methylation data that explicitly account for changes on a DNA level. We then demonstrate these methods by reconstructing a demographic history of cancer from a single-cell sequencing data and compare them to a standard approach that does not account for an internal structure of cancer cell populations within the body.

1. Cao, Y. (2015). *Environmental pollution and DNA methylation: carcinogenesis, clinical significance, and practical applications.* Frontiers of medicine, 9(3), 261-274.
2. Kingman, J. F. C. (1982). The coalescent. Stochastic processes and their applications, 13(3), 235-248.
3. Takahata, N. (1988). *The coalescent in two partially isolated diffusion populations.* Genetical Research*,* 52(3), 213-222.
4. Müller, N. F., Rasmussen, D., & Stadler, T. (2018). *MASCOT: parameter and state inference under the marginal structured coalescent approximation.* Bioinformatics, 34(22), 3843-3848.

G15: Shared regulatory pathways reveal novel genetic correlations between grip strength and neuromuscular disorders

Sreemol Gokuladhas[1], William Schierding[1], David Cameron-Smith[1], Melissa Wake[2], Emma Scotter[3,4], Justin O'Sullivan[1]

[1]Liggins Institute, University of Auckland, Auckland, New Zealand, [2]Murdoch Children's Research Institute, The University of Melbourne, Melbourne, Australia, [3]Department of Pharmacology and Clinical Pharmacology, University of Auckland, Auckland, New Zealand, [4]Centre for Brain Research, University of Auckland, Auckland, New Zealand.

Pathological muscle weakness can develop during the course of aging or due to a range of neuromuscular diseases. Genome-wide association studies have identified single nucleotide polymorphisms (SNPs) associated with grip strength (GS; an inverse measure of age-related decline of muscle strength) and neuromuscular disorders (multiple sclerosis (MS), myasthenia gravis (MG) and amyotrophic lateral sclerosis (ALS)). However, little is known about the functional roles of these SNPs and how or if they contribute to the comorbidity between muscle weakness caused by aging and that caused by neuromuscular disorders.

We combined chromatin interaction data (Hi- C)[1] and functional expression quantitative trait loci (eQTL)[2] data using a bioinformatics algorithm (CoDeS3D)[3] to identify genes that are spatially regulated by SNPs (*i.e*. eQTL- eGene pairs) associated with GS, MS, MG and ALS. Biological pathways enriched with these spatial eGenes were identified using pathway enrichment analyses and the drug-gene interaction database were used to identify those eGenes that have the property of druggability.

None of the eQTLs associated with GS, MG, MS and ALS were shared but we identified eGenes that are commonly regulated by these eQTLs. Particularly, GS eQTLs share eGenes with all three disease-associated eQTLs indicating shared gene regulatory mechanisms between GS and neuromuscular diseases causing muscle weakness. Furthermore, pathway analysis revealed 24 pathways shared between GS and MG, 70 pathways between GS and MS and 18 pathways between GS and ALS. Strikingly, three pathways: mTOR signaling pathway, axon guidance pathway, and alcoholism were identified to be shared by all four phenotypes. A few eGenes implicated in these shared pathways were identified as druggable. Collectively, these findings identify significant biological overlap between age- (GS) and disease-related (MG, MS and ALS) muscle weakness, demonstrating the utility of spatial genetic analysis for the identification of potential therapeutic targets and mechanisms underlying multimorbidity.

G16: The Humble Guinea Pig: Cute, Furry and Deserving of a Better Genome Assembly?

Miles C. Benton[1], David A. Eccles[2], Martin A. Smith[3], Max Berry[4], and Donia Macartney-Coxson[1].

[1]Human Genomics, Institute of Environmental and Science Research (ESR), Porirua, Wellington, New Zealand. [2]Malaghan Institute of Medical Research, Wellington, New Zealand. [3]Garvan Institute of Medical Research, Sydney, NSW, Australia. [4]Neonatal Intensive Care Unit, Wellington Hospital and Department of Paediatrics and Child Health, University of Otago, Wellington, New Zealand

For centuries the domestic Guinea pig has been used as an experimental animal, thus the term 'Guinea pig' was coined. Their utility as a platform to study whole system integrated physiology remains integral to advances in our understanding of health and disease. Unlike other laboratory species they share many key physiological similarities with humans, especially with respect to pregnancy and newborn development, and the regulation of fundamental cardio-metabolic, immune and inflammatory pathways.

The new wave of ultra-long-read next generation sequencing is making it a reality to resolve repeats and close gaps in genome assembly. While improvements in base-calling software for this data continue, higher error rates are still an issue. To this end the accuracy of more 'traditional' shorter read technology provides a useful means with which to error correct and infer accurate genomic variation.

We will provide a high-quality reference genome utilising hybrid-assembly of both Nanopore (PromethION) and Illumina (X10) data. We performed long-read sequencing on 3 outbred females to generate longer scaffolds, complementing this with Illumina sequencing on the same 3 females as well as 3 additional males. This short-read data will be used to polish the long-read scaffolds and generate a comprehensive map of genomic variation. An additional benefit to long-read sequencing is the detection of base modifications, i.e. cytosine methylation. Establishing the Guinea pig methylome will improved the framework for epigenetic research in this animal model.

We present our work towards a more complete reference genome for the laboratory Guinea pig and demonstrate the value of lower coverage long-read sequencing in improving both genome assembly contiguity and completeness. In addition we will profile our initial methylome findings, highlighting these alongside more traditional RRBS data. We believe these findings and datasets will be an invaluable molecular resource and further support the translational strength of this animal model.

G17: Hi-C detects novel structural variants in HL-60 and HL-60/S4 cell lines

Elsie C. Jacobson[1], Ralph S. Grand[1,2], Jo K. Perry[1], Mark H. Vickers[1], Ada L. Olins[3], Donald E. Olins[3], Justin M. O'Sullivan[1]

[1]Liggins Institute, University of Auckland, Auckland, New Zealand, [2]Friedrich Miescher Institute for Biomedical Research, Basel, Switzerland, [3]University of New England, Portland, ME, USA

Cancer cell lines often have large structural variants (SVs) that evolve over time. There are many reported differences in large scale SVs between HL-60 and HL-60/S4, two cell lines derived from the same acute myeloid leukemia sample. However, the stability and variability of inter- and intra-chromosomal structural variants between different sources of the same cell line is unknown. Hi-C and RNA-seq are often used to study gene regulation, but can also be exploited to study the linear genome architecture as well. The high contact frequency of linearly proximal loci can identify genomic rearrangements in Hi-C heatmaps, and RNA-seq read pairs spanning breakpoints can identify putative gene fusions. Using the two methods in combination can improve sensitivity and specificity of SV detection. hic_breakfinder and STAR-fusion were used to identify breakpoints in Hi-C and RNA-seq data respectively, and characterisation of complex and multi-chromosome SVs was performed by visual inspection of Hi-C heatmaps. Here, we used Hi-C and RNA-seq to identify and compare large SVs in HL-60 and HL-60/S4 cell lines. Comparisons with previously published karyotypes identified novel SVs in both cell lines. Hi-C was used to characterise the known expansion centered on the MYC locus. The MYC expansion was integrated into known locations in HL-60/S4, and a novel location (chr4) in HL-60. The HL-60 cell line has more within-line structural variation than the HL-60/S4 derivative cell line. Collectively we demonstrate the usefulness of Hi-C and with RNA-seq data for the identification and characterisation of SVs.

G18: Epigenetic impact of cannabis use in the Christchurch Health and Development Study

John Pearson, Amy Osborne, Alex Noble, Martin Kennedy, John Horwood, Joseph Boden

The New Zealand Parliament recently passed a Medicinal Cannabis bill, effectively legalising the use of cannabis for terminally ill people, and is considering a referendum on law changes concerning recreational cannabis use in light of the recommendation to decriminalise personal use of cannabis. The Christchurch Health and Development study (CHDS), a birth cohort of 1265 Christchurch children born in 1977, provides evidence of harms caused by the prohibition of cannabis. However the CHDS also shows that cannabis use is associated with negative psychosocial outcomes, most strongly in youth (<18 years). There is clear evidence that environment factors, including the use of cannabis and tobacco alters epigenetic marks across the genome. Hence we explored the association of cannabis use with methylation in the DNA from purified peripheral blood of heavy long term cannabis smokers, long term cannabis and tobacco smokers and matched controls who smoked neither cannabis nor tobacco.

We confirmed the most differentially sites in tobacco and cannabis users were hypomethylation in the AHRR and F2RL3 genes (adjusted P $3x10^{-6}$, 0.002 respectively) replicating previous studies into the epigenetic effects of tobacco. Cannabis users who did not use tobacco had no evidence of differential methylation at any sites in these genes (all adjusted P>0.97). There was no differential methylation in cannabis only users at the epi-genome wide significance level ($P<10^{-8}$). However there were multiple sites differentially methylated at more than 1% beta value at a nominal significance of P<0.001 in 218 genes. Pathway analysis of these genes involved neuronal development and neural signalling.
We conclude that the effects of cannabis use on the mature human methylome differ from the effects of tobacco use. We did not identify any methylation loci associated with exclusive cannabis use at the epi-genome wide level of significance. However loci differentially methylated at nominal significance levels were enriched in pathways expressed in the brain, consistent with psychosocial phenotypes associated with cannabis use in the CHDS cohort. Further studies are required to elucidate the long term effects of cannabis use on the developing and mature human brain.

G19: A System-wide Approach to Monitor Responses to Synergistic BRAF and EGFR Inhibition in Colorectal Cancer Cells

Ressa A[1], Bosdriesz E[2], de Ligt J[3*], Mainardi S[2], Maddalo G[1], Prahallad A[2], Jager M[3], de la Fonteijne L[3], Fitzpatrick M[1], Groten S[1], Altelaar AFM[1], Bernards R[2], Cuppen E[3], Wessels L[4,5], Heck AJR[6].

[1]Biomolecular Mass Spectrometry and Proteomics Group, Utrecht Institute for Pharmaceutical Science, Utrecht University, Padualaan 8, 3584 CH Utrecht, The Netherlands. [2]Division of Molecular Carcinogenesis, Cancer Genomics Centre Netherlands, Oncode Institute, The Netherlands Cancer Institute, Plesmanlaan 121, 1066 CX Amsterdam, The Netherlands. [3]Center for Molecular Medicine and Cancer Genomics Netherlands, Division Biomedical Genetics, University Medical Center Utrecht, Utrecht University, Universiteitsweg 100, 3584 CG Utrecht, The Netherlands. * Current: ESR, Institute of Environmental Science and Research, Porirua, New Zealand. [4]Division of Molecular Carcinogenesis, Cancer Genomics Centre Netherlands, Oncode Institute, The Netherlands Cancer Institute, Plesmanlaan 121, 1066 CX Amsterdam, The Netherlands. [5]Department of EEMCS, Delft University of Technology, Mekelweg 4, 2628 CD Delft, The Netherlands. [6]From the Biomolecular Mass Spectrometry and Proteomics Group, Utrecht Institute for Pharmaceutical Science, Utrecht University, Padualaan 8, 3584 CH Utrecht, The Netherlands

Intrinsic and/or acquired resistance represents one of the great challenges in targeted cancer therapy. A deeper understanding of the molecular biology of cancer has resulted in more efficient strategies, where one or multiple drugs are adopted in novel therapies to tackle resistance. This beneficial effect of using combination treatments has also been observed in colorectal cancer patients harboring the BRAF(V600E) mutation, whereby dual inhibition of BRAF(V600E) and EGFR increases antitumor activity. Notwithstanding this success, it is not clear whether this combination treatment is the only or most effective treatment to block intrinsic resistance to BRAF inhibitors.
Here, we investigate molecular responses upon single and multi-target treatments, over time, using BRAF(V600E) mutant colorectal cancer cells as a model system. Through integration of transcriptomic, proteomic and phosphoproteomics data over time we obtain a comprehensive overview, revealing both known and novel responses. We primarily observe widespread up-regulation of receptor tyrosine kinases and metabolic pathways upon BRAF inhibition. The integrative approach allows the tracing of signals throughout the different omics levels over time and provides insights into how these cells can rapidly reprogram.

These findings point to mechanisms by which the drug-treated cells switch energy sources and enter a quiescent-like state as a defensive response, while additionally compensating for the MAPK pathway inhibition. These findings are important for this specific treatment but also demonstrates how omics technology can help grow our understanding of drug effects.

G20: A per-sample pathway analysis tool for DNA methylation data

Santana, A.F. [1,2], Benton, M.C.[1], Macartney-Coxson, D.[1], Black, M.A.[2]
[1]Human Genomics, Institute of Environmental Science and Research (ESR), Porirua, New Zealand, [2]Department of Biochemistry, University of Otago, Dunedin, New Zealand.

Pathway enrichment analysis tools play an important role in the understanding of diseases and biological processes. By detecting sets of differentially methylated CpG sites and/or differentially expressed genes, pathway analysis interrogates whether changes to members of a biological pathways are occurring by chance or not, indicating the possible biological relevance of these functional groupings. In the context of DNA methylation, statistical testing investigates which methylation sites are significantly altered across samples, traditionally using either a paired or case:control design. However, the nature of these tests makes per sample analyses infeasible, as most approaches require replication to allow the estimation of variability. Moreover, unlike gene expression data, methylation can skew the pathway analysis results if not properly handled, as CpG probes are generally unevenly distributed across genes. For example, there are genes with only one or two probes mapped, while others have hundreds: consequently the latter is more likely to exhibit differential methylation, and hence perturbed pathways. Here, we propose a novel pathway analysis tool for DNA methylation data, which enriches gene sets and analyses pathways disruption on a per-sample basis, and minimises the bias from CpG-to-gene mapping. This approach applies categorisation techniques to methylation signals (beta values) creating CpG sets, which are next converted into gene sets after adjusting for multi-mapping bias, and finally enriched for pathway membership. To assess statistical significance, a resampling step is then applied to test whether the enriched pathways could have been obtained by chance. Here we will demonstrate the use of this methodology on DNA methylation data obtained from adipose tissues.

G21: Machine Learning Identifies the Tissue-Specific Effects of SNPs Associated with the Risk of Developing Type 1 Diabetes

Ho, D.[1], Nyaga, D.[1], Schierding, W.[1], Wake, M.[2], Saffery, R.[2], Kempa-Liehr, A.[3], O'Sullivan, J.M.[1]
[1]Liggins Institute, The University of Auckland, Auckland, New Zealand
[2]Murdoch Children Research Institute, The University of Melbourne, Melbourne, Australia
[3]Department of Engineering Science, The University of Auckland, Auckland, New Zealand

The spatial organization of genomes emerges from the impact of all nuclear functions (*e.g.* transcription, repair, and replication) that are occurring within a cell or tissue. As such, understanding spatial organisation represents the key to understanding how the genome works and the impact of genetic variants (Single nucleotide polymorphisms (SNPs)) associated with diseases. Here we present a tissue-specific understanding of an individual's risk of developing type 1 diabetes (T1D) by integrating data on the genome organisation, transcription profile, and phenotype.

Genotype data for individuals with T1D and controls from the Wellcome Trust Case Control Consortium (WTCCC) were cleaned and imputed (1960 T1D cases and 2933 controls, each with 5,957,907 SNPs). T1D related SNPs from the Genome Wide Association Studies catalog ($p<10^{-5}$) were analysed (CoDeS3D) to identify the functional targets of the SNPs. Mann-Whitney test was used to subselect informative data features with the most significant causal relationships (FDR <= 0.05). A multivariate logistic regression (with L1 regularization and gradient descent optimization) was used to integrate information on the tissue-specific impacts of T1D-associated SNPs, trained on 80% of the WTCCC dataset and validated with the remaining 20%.

The regression model incorporated tissue-specific, spatial-regulatory disease patterns to classify T1D patients and controls (AUC = 0.77). The L1 regularization selected 290 data features as having informative relationships in the predictive model, including 50 SNPs, and 74 genes across 48 different tissue types. Of the 48 tissues, lung, testis, and thyroid were the most strong major organ contributors to the T1D disease risk. By contrast, the pancreas and adipose tissue contribute only modestly to the prediction model. Therefore, the tissue-specific contributions of T1D-associated SNPs to models of risk prediction suggest novel pathological pathways for the transition of risk to pathology in T1D.

G22: A Cloud Decoder Ring - Demystifying Cloud Jargon

Duckles, J.M[1]
[1]AbacusBio,Dunedin, NZ.

To be an effective data driven bioinformatician in the modern era, you need to have a good mental model for what various cloud technologies are and how they interoperate. In this talk we'll demystify buckets, virtual machines, EC2, S3, and other jargon terms used by commercial cloud providers. We'll give you a rubric and decoder-ring for understanding these marketing terms and technological jargon terms so you can begin to explore cloud technologies for your bioinformatics workflows. With a basic understanding of cloud technology and some simple computational budgeting you will be well on your way to test and evaluate if a cloud driven workflow might help accelerate your data intensive biology research. There will be a short presentation and an interactive session helping to demystify jargon terms. By attending this session, you should be able to have a mental picture for when and how cloud technologies might help you scale beyond your local compute resources and how to effectively manage costs and capabilities for scientific impact.

G23: Decoding the comorbidities between psychiatric disorders and cognitive functions

Golovina E.[1,2], Vickers M.H.[1], Erb C.D. [3], O'Sullivan J.M.[1,2]
[1] Liggins Institute, The University of Auckland, Auckland, NZ, [2] A Better Start National Science Challenge, Auckland, NZ, [3] School of Psychology, The University of Auckland, Auckland, NZ

Epidemiological studies have identified comorbidities for psychiatric disorders and cognitive phenotypes and suggest that common biological mechanisms may underlie these phenotypes. Single nucleotide polymorphisms (SNPs) represent common predisposing genetic factors. These can occur at regulatory sites and contribute to the observed comorbidities via their impact on gene expression and biological pathways. Genome-wide association studies (GWASs) have identified thousands of such SNPs associated with psychiatric disorders and cognitive functions. Recent studies have reported the regulatory effects of GWAS SNPs and their potential contribution to comorbidity. However, spatial regulatory networks linking psychiatric disorders as comorbid conditions and their associations with cognitive function remain largely unknown.

In the present study, we performed an integrated analysis of 2,893 GWAS SNPs ($p < 1\times10^{-6}$) associated with cognitive functions and five psychiatric disorders (i.e. attention-deficit hyperactivity disorder, anxiety, bipolar disorder, unipolar depression and schizophrenia) in the context of 3D genome organization, functional genomics, biological pathways and drug effects. We identified 45,269 tissue-specific short and long-range regulatory interactions that can contribute to the comorbidities seen in these phenotypes. The analysis revealed different levels of commonalities between psychiatric disorders and cognition - shared regulatory effects on genes and biological pathways. We found that combinatorial actions of regulatory SNPs on specific or common genes within the same neurotrophin signaling pathway may lead to disruptions in the signaling cascade and, consequently, to a series of cellular functions and events associated with both psychiatric and cognitive phenotypes. Drug interactions with spatially regulated genes uncovered potential pharmaceutical influences associated with these phenotypes.

Collectively, our results inform the extent of shared genetic influences, tissue-specific regulatory mechanisms and pathways that contribute to the comorbidities observed between the psychiatric and cognitive phenotypes. The identified drug interactions may be clinically relevant and represent potential opportunities for therapeutic drug repurposing to improve outcomes related to these conditions.

Summary of Abstracts for the Poster Session

No.	Title	Presenter	Institution
G24	When can’t we use homology search tools to infer homology?	Stephanie McGimpsey	University of Otago
G25	Pairwise gene-gene interactions from RNAi perturbation screens: scalability and accuracy of recent machine learning tools	Kieran Elmes	University of Otago
G26	Highly accessible translation initiation sites are predictive of successful heterologous protein expression	Chun Shen Lim	University of Otago
G27	The use of computational modelling to investigate gene-gene interactions in the context of antibiotic resistance	Astra Heywood	University of Otago
G28	TIsigner: Controlling protein expression with designer translation initiation regions	Bikash Kumar Bhandari	University of Otago
G29	Differential Transcript Expression Analysis Using Nanopore Sequencing	David Eccles	Malaghan Institute of Medical Research
G30	Using machine learning algorithms to identify suitable traits for predicting non-coding RNA functionality	Helena Cooper	University of Otago

G24: When can’t we use homology search tools to infer homology?

McGimpsey, S.R.[1] & Gardner, P.P.[1]
[1]Department of Biochemistry, University of Otago, Dunedin School of Medicine, NZ

Homology search tools are important for inferring homology in the abundance of genomes currently sequenced. These tools utilise sequence similarity in order to assign a score between two sequences from which homology is inferred. The relationship between sequence similarity and homology can break down for certain levels of similarity. The zone of pairwise identity where a known pair of homologs has a 50% chance or less of being inferred as homologous based on the alignment score is called the twilight zone. The twilight zone for nucleotide homology has previously been calculated using databases that were small or contained bias. Therefore, the aim of this research was to calculate the twilight zone of nucleotide homology using a carefully designed database of homologous sequences.

A database of core ncRNA and mRNA genes from a large range of genus representative bacteria was generated, from which sequence pairs were chosen. The database was used to calculate where the twilight zone of nucleotide homology was for four different types of alignment algorithms; BLASTn, ggsearch, nhmmer and ssearch. The effect of G+C content and sequence length on the location of the twilight zone was also examined. The twilight zone was shown to be between 38-50% pairwise identity for all alignment algorithms tested. Both sequence length and G+C content shift the twilight zone for all four alignment algorithms. This research has shown that between 38-50% pairwise identity homology should not be inferred based only on the alignment score, as there is a greater chance of incorrectly inferring homology than correctly inferring homology. Furthermore, the analyses have shown that a parametric approach to database design is required to further balance the database used for the twilight zone calculation.

G25: Pairwise gene-gene interactions from RNAi perturbation screens: scalability and accuracy of recent machine learning tools

Kieran Elmes[1], Fabian Schmich[2], Ewa Szczurek[3], Niko Beerenwinkel[4,5], and Alex Gavryushkin[1]
[1] Biological Data Science Lab, Department of Computer Science, University of Otago, Dunedin, New Zealand [2] Roche, Munich, Bavaria, Germany [3] Institute of Informatics, University of Warsaw, Warsaw, Poland [4] Department of Biosystems Science and Engineering, ETH Zurich, Basel, Switzerland [5] SIB Swiss Institute of Bioinformatics, Basel, Switzerland

Targeting synthetic lethal pairwise genetic interactions (where suppressing either gene does no significant harm, but suppressing both is lethal) is a promising recent approach, with important application in cancer research. The number of potential pairwise interactions grows quadratically with the number of genes, making discovery in a lab impractical without first reducing the search space.

This has previously been done by only searching for interactions with genes in specific cell lines [3, 1]. We propose instead to exhaustively search for pairwise inter- actions in large-scale siRNA perturbation screens. In this paper, we simulated large siRNA perturbation screens based on data from real siRNA libraries. A selection of known interactions, main effects, and noise was added, which combine to produce an overall fitness score. We then compared the performance of two recent methods for discovering interactions, group lasso regularised regression (Glinternet) [2], and the xyz algorithm [4].

We evaluate the ability of each to correctly and quickly identify lethal pairs on sets of up to 4,000 genes. For each simulated data set, we tested varying signal-to-noise ratios, numbers of true interactions, numbers of observations per double knockdown, and effect sizes, considering all returned interactions or only those that deviated significantly from zero.

A separate set of simulations was performed to determine their ability to identify only strong negative interactions which did not obey the strong hierarchy assumption of the group lasso. These were used to compare the runtime of each algorithm.

We found that xyz suffers from a drop in accuracy as the size of the data set increases, while Glinternet is still able to identify the correct interactions. Glinternet does, however, take a prohibitively long time to run on large data sets. In particular, it does not scale to the approximately 20, 000 genes that would be required to analyse human protein-coding genes.

Work is also ongoing to produce a lasso implementation that is specifically designed for finding (strong) interactions on large perturbation screens, using multi-core ma- chines.

1. Christina Laufer et al. “Mapping genetic interactions in human cancer cells with RNAi and multiparametric phenotyping.” In: Nat Meth (2013).
2. Michael Lim and Trevor Hastie. “Learning interactions via hierarchical group- lasso regularization”. en. In: J. Comput. Graph. Stat. 24.3 (2015), pp. 627–654.
3. Michael Steckel et al. “Determination of synthetic lethal interactions in KRAS oncogene-dependent cancer cells reveals novel therapeutic targeting strategies.” In: Cell Res. (2012).
4. Gian-Andrea Thanei, Nicolai Meinshausen, and Rajen D Shah. “The xyz al- gorithm for fast interaction search in high-dimensional data”. In: (Oct. 2016). arXiv: 1610.05108 [stat.ML].

G26: Highly accessible translation initiation sites are predictive of successful heterologous protein expression

Chun Shen Lim[1], Bikash Kumar Bhandari[1], Paul P. Gardner[1,2]
[1]Department of Biochemistry, School of Biomedical Sciences, University of Otago, Dunedin, NZ, [2]Biomolecular Interaction Centre, University of Canterbury, Christchurch, NZ.

Recombinant protein production in microbial systems is well-established, yet half of these experiments have failed in the expression phase. Failures are expected for 'difficult-to-express' proteins, but for others, codon bias, mRNA folding, avoidance, and G+C content have been suggested to explain observed levels of protein expression. However, determining which of these is the strongest predictor is still an active area of research. We used an ensemble average of energy model for RNA to show that the accessibility of translation initiation sites outperforms other features in predicting the outcomes of 11,430 experiments of recombinant protein production in *Escherichia coli*. We developed TIsigner and showed that synonymous codon changes within the first nine codons are sufficient to improve the accessibility of translation initiation sites. Our software produces scores for both input and optimised sequences, so that success/failure can be predicted and prevented by PCR cloning of optimised sequences.

G27: The use of computational modelling to investigate gene-gene interactions in the context of antibiotic resistance

Heywood, A.[1,2], Redmond, A.[2], Rehman, A.[1], Gavryushkin, A.[2], Lamont, I.[1]
[1]Department of Biochemistry, University of Otago, Dunedin, NZ, [2]Department of Computer Science, University of Otago, Dunedin, NZ

Antibiotic resistance is becoming increasingly prevalent in bacterial populations worldwide. One bacterium that has both intrinsic and acquired resistance is the opportunistic pathogen *Pseudomonas aeruginosa*. Constant exposure to antibiotics is the primary cause of acquired resistance in *P. aeruginosa.* Currently, there is an incomplete understanding of genetic mechanisms that contribute to antibiotic resistance. Using computational modelling we aim to infer the impact of gene-gene interactions that contribute to antibiotic resistance over and above what each gene accounts for individually. Whole genome sequencing (WGS) data was generated from highly antibiotic resistant laboratory evolved strains, clinical isolates, and environmental isolates with known antibiotic resistance profiles (Minimum Inhibitory Concentrations, MIC).

Firstly, laboratory strains were exposed to increasing concentrations of antibiotics and the resulting mutated genes became the bases for building an artificial resistance genotype. Our data included genomic profiles that had multiple (different) resistance measurements (MIC) and unknown resistance profiles for hypothetical genotypes. We fitted this data to multiple interaction models with the aim to infer gene-gene interactions. The models were compared using Akaike's Information Criterion (AIC), a measure of parsimony that rewards goodness of fit and penalises overfitting. To assess genomes that were found outside of the laboratory setting, 380 clinical and environmental genomes (with known MICs) were incorporated into the dataset. Increased genomic variation of the dataset led to additional filtering methods that elaborated on specific antibiotic resistance alleles.

Further testing and expansion of this model could lead to a tool, capable of quickly and effectively processing clinical isolates of *P. aeruginosa* to determine the antibiotic resistance profiles, and hence the best course of action for effective treatment.

G28: TIsigner: Controlling protein expression with designer translation initiation regions

Bikash Kumar Bhandari[1], Chun Shen Lim[2], Paul P. Gardner[1,2]

[1]Department of Biochemistry, School of Biomedical Sciences, University of Otago, Dunedin, NZ, [2]Biomolecular Interaction Centre, University of Canterbury, Christchurch, NZ.

It is known that mRNA levels do not dictate protein levels as strongly as expected. Several sequence features such as codon adaptation index, tRNA adaptation index, mRNA secondary structure and mRNA:ncRNA avoidance have been suggested to explain this discrepancy. Using 11,430 recombinant protein experiment dataset from Protein Structure Initiative (PSI:Biology), we found that the ‘opening energy’ of translation initiation sites exceeds the performance of all other features for the classification of recombinant protein expression success and failure. Furthermore, we found that the optimisation of just the first nine codons can reduce the opening energy of region around translation initiation. We have shown that this is often sufficient to improve the outcome of experiments.

Based on these findings, we have developed Translation Initiation coding region designer (Tisigner), which optimises a given mRNA sequence by doing synonymous changes on the first nine codons. Such optimised sequences can be used for PCR cloning. This is more economical than full gene synthesis. Our tool also predicts the probability of success of a given sequence and the optimised sequence. The optimised sequence can also be tailored to a specific outcome such as low, medium or high expression. The tool is available as a web service (https://tisigner.otago.ac.nz) as well as offline scripts (https://github.com/Gardner-BinfLab/Tisigner).

G29: Differential Transcript Expression Analysis Using Nanopore Sequencing

Eccles, D. A.[1], Grasso, C.[1], Berridge, M. V.[1]
[1]Cancer Cell Biology Group, Malaghan Institute of Medical Research, Wellington, NZ.

We used MinION long-read cDNA sequencing to determine the effect of mitochondrial DNA (mtDNA) deletion on nuclear gene expression in a mouse tumour model[1]. We have been sequencing cDNA on the MinION sequencer since December 2017, and have experienced substantial changes in the sequencing technology during that time. Based on our cDNA sequencing experience, we recommend the following practices:

1. Prepare cDNA from fresh samples
2. Do a length-based QC prior to sequencing
3. Aim for >20ng/µl DNA post-amplification
4. Load samples with poor QC separately
5. Aim for >1M sequenced reads per sample
6. Filter out reads with ambiguous barcode matches
7. Use the strand-switch primer sequence to determine read orientation
8. Map reads to the transcriptome for count matrix generation
9. Map reads to the genome for visual confirmation
10. Re-call reads with the latest available basecaller

Despite technology challenges, our results have been consistent enough across five separate runs that we have been able to make a number of observations about genes whose expression requires mtDNA:

- 4T1-ρ0 cells without mtDNA do not express mitochondrial transcripts
- Nuclear transcripts involved in mitochondrial respiratory complexes are transcribed normally when mtDNA is absent
- Expression of some genes involved in immune and stress responses require the presence of mtDNA

These results will be used to investigate the molecular mechanisms of mitochondria acquisition by tumour cells lacking mtDNA and the regulation of early tumour immune responses.

1. Tan, A. S., Baty, J. W., Dong, L. F., et al. (2015). Mitochondrial genome acquisition restores respiratory function and tumorigenic potential of cancer cells without mitochondrial DNA. Cell Metabolism. 21(1): 81-94.

G30: Using machine learning algorithms to identify suitable traits for predicting non-coding RNA functionality

Cooper, H.[1], Biggs, H.[1], Gardner, P[1],

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

There are multiple classes of non-coding RNAs (ncRNA), each of which has specific characteristics, sequence length and biological functions. As a result, the dysregulation or absence of ncRNA has been linked to numerous human diseases, particularly various forms of cancer. Predicting whether a ncRNA is functional using computational methods has proved to be more difficult in comparison to proteins and coding genes, with lack of conservation, poor annotations and repetitive sequences being some of the challenges that need to be overcome. In this study, human ncRNA from RNAcentral was used to create a machine-learnt prediction model by combining data from six categories of functional traits.

RNA:RNA interactions, population statistics, genomic copy number, transcription, sequence conservation and conservation of secondary structure were all examined. These categories then had a random forest analysis applied to both a test set and a larger group of randomly generated genomic regions as a negative control set.

It was found that the majority of the ncRNA in the dataset could have their functionality predicted correctly, with covariance and conservation predictors standing out as being significant.