Computational Genomics Abstracts

CG1: Ancient DNA analyses in the Pacific - implications for human settlement history

Matisoo-Smith, E.A., Gosling, A., Collins, C.

Department of Anatomy and Allan Wilson Centre, University of Otago, Dunedin, NZ.

We have recently sequenced several complete ancient mitochondrial genomes from Pacific populations which are providing surprising results regarding population histories including high levels of variation and evidence of population replacements. These data will be presented and the implications for reconstructing population histories in the region discussed.

CG2: Understanding the evolution of island gigantism - a genomic approach

Knapp, M.[1], Prost, S.[2], Haile, J.[3], Scofield, P.[4], Bunce, M.[5], Gilbert, M.T.P.[6]

[1]Department of Anatomy, University of Otago, Dunedin, NZ, [2]Department of Integrative Biology, University of California, Berkeley, USA, [3]School of Archaeology, University of Oxford, UK, [4]Canterbury Museum, NZ, [5]Department of Environment and Agriculture, Curtin University, Australia, [6]Centre for GeoGenetics, Natural History Museum of Denmark, University of Copenhagen, Denmark.

Evolution on islands has produced extraordinary forms of adaptations including flightless birds, such as the New Zealand Kakapo, and dog sized elephants, such as the extinct Sicilian Pygmy Elephant. Understanding the genetic basis of such environmental adaptation is of fundamental importance for the study of evolutionary biology. Among the most noticeable environmental adaptations is island gigantism. While this phenomenon is well known[1], we know little about the genetic processes involved in the evolution of island gigantism. One of the most spectacular examples of island gigantism is New Zealand's extinct Haast's Eagle (*Harpagornis moorei*). This iconic representative of New Zealand's unique bird fauna descended from a lineage of very small eagles and, after arriving in New Zealand, rapidly evolved into the largest known raptor in the world. Based on genetic divergence, the age of the most recent common ancestor of Haast's Eagle and its two closest relatives, the Australasian Little Eagle (*Hieraaetus morphnoides*) and the Northern Hemisphere Booted Eagle (*H. pennatus*) was estimated to be between 0.7 and 1.8 million years[2], yet Haast's Eagle is as much as 15 times the size of these two closest relatives. This suggests that, in adapting to the New Zealand environment, Haast's Eagle went through the largest size increase (relative to time) known of any vertebrate species in the world. Using complete genome data from Haast's Eagle and its closest relatives we aim to identify the functional genomic basis of this rapid size increase.

1. Meiri, S., Cooper, N. and A. Purvis (2008). *The island rule: made to be broken?* Proceedings of the Royal Society B-Biological Sciences. 275:141-8.
2. Bunce, M., Szulkin, M., Lerner, H.R.L., Barnes, I., Shapiro, B., Cooper, A. and R.N. Holdaway (2005). *Ancient DNA provides new insights into the evolutionary history of New Zealand's extinct giant eagle.* PLoS Biology. 3:44-6.

CG3: Sex and speciation: Did sex chromosome turnover and rearrangement trigger mammalian divergence?

Jennifer A. Marshall Graves[1,2,3]

[1]School of Life Science, La Trobe University; [2]Research School of Biology, Australian National University; [3]Institute of Applied Ecology, University of Canberra, Australia

Humans and other therian mammals share a sex chromosome pair composed of a highly conserved X and a small Y chromosome, which progressively degenerated and specialised. The male-dominant *SRY* gene on the Y diverged from its X-borne partner *SOX3* about 190 MYA, the time that therians diverged from prototherian mammals (monotremes such as platypus). In reptiles, and even monotremes, this XY pair is represented by autosomes. Platypus and echidna share a bizarre $X_1X_2X_3X_4X_5Y_1Y_2Y_3Y_4Y_5$ system with homology to the bird ZW, suggesting that a mammalian ancestor may have had a bird-like ZW system.

I propose that the evolution of the sex determining *SRY* gene and definition of a novel XY chromosome pair in therian mammals ~190 MYA imposed a reproductive barrier with the ancestral population of mammal-like reptiles, and triggered the speciation event that led to the evolution of therian mammals. Meanwhile translocation with autosomes stabilised the ancestral system in monotremes. Hybrids between animals with the ancient system and the new *SRY* system, or the translocated monotreme system, would have had high frequencies of sex reversal, intersex development, and infertility, promoting divergence of prototherian and therian mammals. I also propose that more recently (~166 MYA), Robertsonian fusion of the therian XY pair with an autosome posed a reproductive barrier that promoted divergence of eutherian (placental) mammals and marsupials.

Several modern rodent lineages have variant sex determining systems (including complete loss of the Y), suggesting that rodents are undergoing a new explosion of speciation driven by Y chromosome degradation and sex chromosome turnover.

This theory conflicts with the long-prevailing paradigm that speciation results from accumulation of small mutational differences in isolated populations, but receives support from the many groups of fish and reptiles in which closely related species have different sex determination mechanisms, and interspecies hybrids are infertile.

CG4: Using genomics to manage adaptive potential in threatened populations

Santure, A.W.[1], Lee, K.D.[1], Ewen, J.G.[2], Brekke, P.[2]

[1]School of Biological Sciences, University of Auckland, Auckland, NZ, [2]Institute of Zoology, Zoological Society of London, London, UK.

For endangered species, one of the most promising applications of genomics is to understand the genetic basis of adaptation in the wild. For example, knowledge of genetic variants that increase or decrease aspects of fitness will be an important tool to help decide which individuals should be translocated to found new populations. This will maximise evolutionary potential of the new population while reducing risk from deleterious variants. In this talk, I will give an overview of our work to date in developing a genomic toolkit in the threatened New Zealand hihi (stitchbird; *Notiomystis cincta*), and the work we have planned to understand the genetic basis of important traits in the population. While discovering genes of large effect would help us understand selection and evolution in the wild, we believe it is likely that most traits will be influenced by many genes of small effect. However, borrowing from genomic prediction, we predict that we can use whole genome information to accurately identify high value individuals for breeding programs and translocations, thus maximising the evolutionary potential of threatened species.

CG5: Intrinsically disordered protein and the origins of complex multicellular organisms

A. Keith Dunker[1], Karl J. Niklas[2], .Sarah E. Bondos[3], Fei Huang[1], Christopher J. Oldfield[1]

[1]Indiana University School of Medicine, Center for Computational Biology and Bioinformatics, Indianapolis, IN, USA; [2]Cornell University School of Integrative Plant Science, Plant Biology Section, Ithaca, NY, USA; [3]Texas A&M Health Science Center, Department of Molecular and Cellular Medicine, College Station, TX, USA.

DNA sequencing has enabled the widespread construction of phylogenetic trees, revealing that *multicellular* organisms evolved independently from unicellular ancestors about 25 times among prokaryotes and eukaryotes. Multicellular organisms can be classified as simple, in which all of the cells are in direct contact with the surrounding milieu, or complex, in which some cells are completely surrounded by other cells. Current phylogenetic trees indicate that *complex multicellular* organisms evolved independently from unicellular ancestors about 10 times, and only among the eukaryotes, including once for animals, twice each for green, red, and brown algae, and thrice for fungi.

Given these multiple independent evolutionary lineages, we asked two questions: 1. Which molecular functions underpinned the evolution of multicellular organisms?; and, 2. Which of these molecular functions depend on intrinsically disordered proteins (IDPs, reviewed in [1])? Multicellular organisms require the advent of molecules for cellular adhesion, for cell-cell communication and for developmental programs. In addition, the developmental programs need to be regulated over space and time. Finally, each multicellular organism has cell-specific biochemistry. As for dependence on IDPs, we ranked Swiss Protein key-words with regard to their degree of association with predictions of protein structure or disorder. With a Z-sore of 18.8 compared to random-function proteins, “differentiation” was the biological process most strongly associated with IDP predictions. As expected from this result, large numbers of individual proteins associated with differentiation exhibit substantial regions of predicted disorder [2]. All five of the underpinning molecular functions for multicellularity were found to depend critically on IDP-based mechanisms [3].

These new findings necessitate a rethinking of the gene regulatory network models currently used to explain cellular differentiation and the evolution of complex multicellular organisms [4].

1. Oldfield CJ, Dunker AK.(2014) *Intrinsically disordered proteins and intrinsically disordered protein regions.* Annu Rev Biochem. 83: 553 – 584.
2. Xie H, Vucetic S, Iakoucheva LM, Oldfield CJ, Dunker AK, Uversky VN, Obradovic Z. (2007) *Functional anthology of intrinsic disorder. J Proteome Res. 6 (Three consecutive papers) 1. Biological processes and functions of proteins with long disordered regions 1882 – 1898; 2. Cellular components, domains, technical terms, developmental processes, and coding sequence diversities correlated with long disordered regions.*1899 – 1916; and 3. Ligands, post-translational modifications, and diseases associated with intrinsically disordered proteins.1917 – 1932..
3. Dunker AK, Bondos SE, Huang F, Oldfield CJ. (2015) *Intrinsically Disordered Proteins and Multicellular Organisms*. Seminars in Cell and Developmental Biology 37: 44 – 55.
4. Niklas KJ, Bondos SE, Dunker AK, Newman SA. (2015) *Rethinking gene regulatory networks in light of alternative splicing, intrinsically disordered protein domains, and post-translational modifications*. Front Cell Dev Biol. 3: 8.

CG6: Binding of fullerenes and nanotubes to the mechanosensitive channel of large conductance: A possible new class of antimicrobials

Hilder, T.A.[1], Ridone, P.[2], Nakayama, Y.[2], Martinac, B.[2,3], Chung, S.H.[1]

[1]Computational Biophysics Group, Australian National University, Canberra, Australia, [2]Molecular Cardiology and Biophysics Division, Victor Chang Cardiac Research Institute, Sydney, Australia, [3]St Vincent's Clinical School, University of New South Wales, Sydney, Australia.

Multi-drug resistance is becoming an increasing problem in the treatment of bacterial infections and diseases. For example, the re-emergence of tuberculosis is a serious threat that is spreading rapidly throughout the world. It has become increasingly important to develop new antibiotics to combat the rapidly emerging strains of multiple drug-resistant bacterial strains. A potential target for new antibiotics is the mechanosensitive channel of large conductance (MscL) which is highly conserved among prokaryotes.[1] MscL channels act as emergency release valves that open in response to membrane tension caused by hypo-osmotic shock. They relieve pressure and therefore prevent membrane damage. Evidence suggests that a pharmacological agent that can affect the gating of, or block the current through, MscL has significant potential as a new class of antimicrobial compounds[1] capable of targeting a range of pathogenic bacteria with minimal side-effects to infected patients.

Using molecular dynamics simulations we examine the binding of fullerenes and nanotubes to MscL and demonstrate that both are stable within the MscL pore.[2] We predict that fullerenes will attenuate the flow of ions through MscL by reducing the pore volume available to water and ions resulting in a slower recovery from osmotic downshock. However, nanotubes will prevent pore closure resulting in a permanently open pore, thus causing the cell to continually leak cytoplasmic contents. Both of these events may be sufficient to cause a bacterial cell to lyse. Moreover, we confirm experimentally that it is possible to attenuate the flow of ions through MscL using a C_{60}-γ cyclodextrin complex.[2]

1. Booth, I.R. and P. Blount (2012). *The MscS and MscL families of mechanosensitive channels act as microbial emergency release valves.* Journal of Bacteriology. 194: 4802-4809.
2. Hilder, T.A., P. Ridone, Y. Nakayama, B. Martinac and S.H. Chung (2014). *Binding of fullerenes and nanotubes to MscL.* Scientific Reports. 4: 5609.

CG7: Multiscale dynamics of the Dengue virus envelope

Marzinek, J.K.[1,2], Holbrook, D.A.[1], Huber, R.G.[1], Verma C.S.[1,2], Bond, P.J.[1, 2]

[1]Bioinformatics Institute (A*STAR), 30 Biopolis Str., #07-01 Matrix, Singapore 138671,
[2]National University of Singapore, Department of Biological Sciences, 14 Science Drive 4, Singapore 117543

Dengue is a pathogenic flavivirus, consisting of a nucleocapsid core surrounded by a lipid bilayer membrane plus anchored envelope (E) and membrane (M) proteins. Infection rearrangement of the E protein, exposing the unstructured fusion peptide (FP) for viral/host membrane fusion. Currently there are no available vaccines effective against all Dengue serotypes, due largely to the dynamic nature of the E protein. A multiscale simulation approach has thus been adopted to characterize the dynamics of the E protein in the context of fusion and viral architecture. We first report a systematic comparison of simulation methods and forcefields, constituting ~50 μs of sampling, to determine the complete conformational landscape of FP across all known flavivirus family members. The solvated peptide fold was significantly less ordered than in the crystallographic state, confirmed via circular dichroism and spectrofluorometric measurements, and the conformational landscape of a non-fusogenic mutant was perturbed compared to wild-type variants, emphasizing the importance of dynamics in therapeutic targeting. Subsequently, long- timescale FP/membrane assembly was investigated, helping to refine previously derived solution NMR data, and atomic-resolution free energy profiles calculated. FP binding affinities validated by fluorescence spectroscopy revealed functionally important differences between Dengue serotypes, and a reduced interaction (>50%) with pure zwitterionic membranes compared to physiologically realistic endosomal membranes, highlighting the importance of electrostatics in virus-host recognition. Finally, we reconstructed the entire Dengue viral envelope complex, containing 180 copies each of the M and E proteins within a ~40 nm lipid bilayer vesicle. Microsecond-timescale simulations revealed that the E protein induces extreme local curvature and lipid sorting, which may be important during membrane fusion. Calculation of theoretical, time-dependent density maps revealed almost perfect agreement with the equivalent density from low-resolution cryo-EM data. This is being further extended by integrating a geometrically plausible model of the viral core containing the entire viral genome.

CG8: A profile-based method for the detection of deleterious genetic variation

Nicole E. Wheeler[1], Lars Barquist[2], Paul P. Gardner[1]

[1] School of Biological Sciences, University of Canterbury, Christchurch, New Zealand, [2] Institute for Molecular Infection Biology, University of Würzberg, Würzberg, Germany

Comparative genomics investigations have yielded an abundance of information on the genetic diversity between organisms, however not all genetic variation has the same effect on fitness. We present a profile HMM-based method for assessing the functional significance of mutations in protein coding sequences. We demonstrate the ability of this method to distinguish between deleterious and neutral variation using protein mutagenesis data, and the application of the method to comparative analysis of bacterial genomes to scan for functionally significant genetic variation. We show that the method is able to detect functional degradation of genes associated with host adaptation in *Salmonella enterica* and that this approach offers a sensitive measure of the loss-of-function mutations that occur as a result of adaptation to a new niche.

CG9: A dirt-powered generator

Cui, A.[1], Xiao, Z.[2],Korchina, V.[3], Stamenova, E.[4], Aiden, E.S.[5], Aiden, A.P.[5,6]

[1]Department of Biological and Biomedical Sciences, Yale University, New Haven, CT, USA, [2]Demyst Data, New York, NY, USA, [3]University of Houston, Houston, TX, USA, [4]Broad Institute, Cambridge, MA, USA, [5]Department of Molecular and Human Genetics, Baylor College of Medicine, Houston, TX, USA, [6]Department of Bioengineering, Rice University, Houston, TX, USA

Despite the fact that some clean energy options nearly 2 billion people utilize burning kerosene for lighting. We worked on optimizing low-cost microbial fuel cells (MFCs) to try and bridge this gap. We developed a modular MFC 'lego' that can be produced from locally available materials in sub-Saharan Africa and can be connected to produce larger generators. We noted that MFCs exhibit variability in their power output despite identical construction and operating conditions. We posited that the variability was due to stochastic differences in the microbial communities that grew up at the anode. We did whole genome sequencing and identified that the powerful cells had unusually high frequency of Pseudomonas species. Pseudomonas produce oxidizing agents, such as pyocyanins and phenazines that can serve as electron transporters. We hypothesized that introducing exogenous oxidizing agents would serve to increase the power output of MFCs. In order to maintain our low-cost, locally available materials model, we selected blood meal which is rich in Fe(II) and a byproduct of animal slaughter. As compared to controls of other agricultural byproducts, introduction of blood meal increased the power output by over 100 fold. We were able to use these devices to power an LED light and charge a Nokia 1100 cell phone.

CG10: Mapping shared, genome-wide markers in a New Zealand Māori population to understand a childhood liver disease

Sophia Cameron-Christie[1], Helen Evans[3], David Markie[2], Rick Tankard[5], Katherine Smith[4], Melanie Bahlo[4], Stephen Robertson[1]

[1]Departments of Women's and Children's Health, and [2]Pathology, University of Otago, Dunedin, New Zealand. [3]Starship Children's Hospital, Auckland, New Zealand, [4]Bioinformatics Division, The Walter and Elizabeth Hall Institute, Melbourne, Australia [5]Department of Medical Biology, University of Melbourne, Melbourne, VIC, Australia

Identifying susceptibility variants for multifactorial human traits can be complicated depending on the understanding of the genetic background on which it arises. This is especially difficult in small, under-studied populations without readily-available control cohorts. In such populations, low-penetrance variants in complex disease can be challenging to identify when normal variation and allele frequencies are uncharacterised. We present an investigation of biliary atresia (BA), a usually sporadic malformation of the biliary tree. Worldwide, BA leads to half of all paediatric liver transplants, and is always fatal without major surgical intervention. The causes of BA remain unknown; autoimmune processes and genetic background may both play a part. In Māori and Polynesian populations the incidence is elevated three-fold compared to Europeans. We have identified a small Māori iwi exhibiting an extremely elevated incidence of BA (1:100–300). To circumvent some of the problems in studying complex traits in Māori we have adopted a non-parametric, family based approach to localise a presumptive genetic factor conferring this susceptibility. Further, commercial SNP microarrays are not optimised to be informative in this population, so we have integrated rare variation from Whole Genome Sequencing (WGS) data. Assuming a single, segregating susceptibility factor contributes to BA in our family cohort, we have used the software Beagle, Germline and in-house methods to examine haplotype sharing between affected pairs across this family. We then construct maps of long, Identical-By-State (IBS) segments across the collective genomes of affected individuals without reliance on pre-existing assumptions about allele frequencies, a definitive inheritance model or exclusion of unidentified phenocopies. This allows resolution of shared haplotypes to define candidate regions in which a BA susceptibility variant could lie.

CG11: Whole Forest Modelling: Reconstructing the past, present and future performance of trees with big data.

Telfer, E.[1] Pont, D.[2] Dash, J.[2] Macdonald, L.[2] Dungey, H.[1] and Moore, J.[3]

[1]Forest Genetics, Scion, Rotorua, NZ, [2]Forest Industry Informatics, Scion, Rotorua, NZ, [3]Forest Systems, Scion, Rotorua, NZ.

In recent years, we have undertaken sequencing and assembly of the conifer mega-genome for radiata pine; a huge 25 Gbase genome with large multi-gene families and an extensive repeatome. This project sits alongside an industry aligned programme to generate and develop SNP genotyping resources for both genomic selection (prediction of future performance) and pedigree reconstruction (elucidation of parentage from within mixed seedlots). Genomic selection predictions rely on the quality of phenotypic data with which the model is developed and phenotypic traits can vary for an individual genotype between diverse environments.

The innovative tools of “precision forestry” are enabling whole forest phenotyping; a long way from the small numbers of manual measurements of the past. Combinations of data from terrestrial and airborne laser scanning, along with imagery from earth observation satellites and aircraft have become a big data reality. Methods are in development to identify and phenotype individual trees from a range of these remote sensing data sources. Extensive geospatial and meteorological data, sourced from satellites to microscale x-rays of the soil, are enabling us to characterise our environment in unprecedented detail and underpin our ability to understand global climate systems and the complex soil profiles and metabiomes that make up a forest. At this convergence of genomic, phenomic and environmental “big data” are opportunities to learn from each of the individual disciplines and incorporate the extraordinary power of modern computing capacity and machine learning. In an approach dubbed “forestomics” we intend to develop synergistic approaches to transform forest-wide “big data” into “big knowledge” about whole forest systems.

Our goal is to understand the relationships between, genotype, phenotype and environment, to the extent that we can predict and enhance the growth and resilience of whole forests. In this paper we present an overview of the opportunities for whole forest modelling, where a 25 Gbase genome in 27 million pieces may no longer be our greatest computational challenge.

CG12: Use Your Words: Executable Explanations of Computational Genomics with Docker and Jupyter

John McCallum[1,2], Hoang Nguyen[1,2], Ben Warren[2], Andrew Catanach[2] and Roy Storey[2]

[1]Department of Biochemistry, University of Otago, Dunedin, NZ, [2] Plant & Food Research, NZ.

The technological challenges of generating and processing ever-larger data streams appear to have displaced our attention from our responsibility as scientists to distil these into resources or accounts of our work others can use. We will describe our experiences in the use of Docker containerization and Jupyter notebooks to simplify configuration, analysis, documentation and exchange of complex informatics workflows as support for reproducible research. Although containerisation has been embraced as a means to organise complex workflows, we will describe how it may be used to provide a customised environment for authoring multi-kernel Jupyter (iPython) notebooks for small to large projects. Our use-cases emphasise large scale genetic analysis in plant genomes: CNV detection and marker design from whole genome sequencing in kiwifruit, and GWAS in wheat. The combination of Docker for configuration of the computational environment, Jupyter for documenting analyses, and Github for source control of these, provides a reproducible means to perform analyses and authoring in the same way as on desktop, cluster or cloud. We will illustrate how this way of working reduces barriers to collaboration and simplifies the exploration and explanation of new computational methods.

CG13: Performance evaluation and benchmarking of differential DNA methylation analysis methods

McCormick, H. M.[1,2], Giannoulatou, E.[1,2], Hickey, P. F.[3], Cropley, J. E. [1,2] and Suter, C. M. [1,2]

[1]Victor Chang Cardiac Research Institute, Sydney, NSW, Australia, [2]University of New South Wales, Sydney, NSW, Australia, [3]Walter and Eliza Hall Institute of Medical Research, Parkville, Victoria, Australia

DNA methylation is one of the most widely used markers for the study of epigenetic contributions to phenotypic variation and disease. There are several methods for analyzing genome-wide DNA methylation data in common use, but there has been no rigorous evaluation of their performance. We have performed a systematic assessment and comparison of four packages: MethySig, methylKit, eDMR and DSS, using an empirical dataset of 12 reduced representation bisulphite sequencing libraries (6 test, 6 control). Surprisingly, we observed very low concordance among these commonly used model-based and binomial test-based approaches: using equivalent pre-processing and filtering parameters for each method, we found that the four methods identified significant differentially methylated cytosines at a concordance rate of less than 1%. Similarly low levels of concordance were observed with identification of differentially methylated regions using tiled data. Our study highlights the need for systematic approaches to reliable differential methylation analysis via data simulation. This concept of simulation will be discussed in the context of the growing implementation of epigenomic data in human medicine.

CG14: How the genome folds: now inside the loop

Erez Lieberman Aiden

The Centre for Genome Architecture; Department of Molecular and Human Genetics, Baylor College of Medicine; Department of Computer Science, Rice University

The human genome is over 2 meters long, but must fold up to fit inside the nucleus of a cell. How does it fold? We use in situ Hi-C to probe the 3D architecture of genomes, constructing haploid and diploid maps of nine cell types. The densest, in human lymphoblastoid cells, contains 4.9 billion contacts, achieving 1 kb resolution. We find that genomes are partitioned into contact domains (median length, 185 kb), which are associated with distinct patterns of histone marks and segregate into six subcompartments. We identify ~10,000 loops. These loops frequently link promoters and enhancers, correlate with gene activation, and show conservation across cell types and species. Loop anchors typically occur at domain boundaries and bind CTCF. CTCF sites at loop anchors occur predominantly (>90%) in a convergent orientation, with the asymmetric motifs 'facing' one another. The inactive X chromosome splits into two massive domains and contains large loops anchored at CTCF-binding repeats.

CG15: Optimisation of sequencing effort for relatedness estimation using genotyping-by-sequencing

Dodds, K. G., McEwan, J.C., Brauning, R., Clarke, S.M.

AgResearch, Invermay Agricultural Centre, Private Bag 50034, Mosgiel 9053, NZ.

Genotype-based relatedness estimation underlies pedigree reconstruction, genomic selection and some association mapping methods. Genotyping-by-sequencing provides genotypes from sequencing data. Molecular biology techniques allow the use of genome subsets and for multiple individuals to be combined in a sequencing run, allowing flexibility in the number of single nucleotide polymorphisms (SNPs) assayed and read depth at each SNP position. SNP depth affects the ability to correctly call genotypes, as only one of the alleles might be observed, with low depth.

We use simulation to find the optimal depth for estimating relatedness at a fixed total sequencing effort (number of SNPs times mean depth), assuming random sampling of SNPs and alleles. Relationship estimation used a method we have recently developed which accounts for read depth for each genotype, including zero reads (missing). We also investigate the effect of estimating allele frequencies. The optimal depth was defined to be the depth that gave the lowest standard deviations of relatedness estimates, across sets of individuals of the same relationship type. We found that the optimal depth was around 3 for estimating relatedness between individuals, and around 8 for estimating self-relatedness. The optima were flat, so that depths of 1-5 and 5-10 were close to optimal for between-individual and self-relatedness, respectively. There were small biases in the relatedness estimates at very low (≤0.5) mean depths. These were corrected using the true allele frequencies, indicating that allele frequency estimates based on low total numbers of reads can influence relatedness estimation.

CG16: Tumour Transcriptomes: Going beyond expression profiling

Davidson N.M.[1], Majewski, I.J.[2], Ekert, P.[1], Oshlack, A.[1]

[1]Murdoch Childrens Research Insititute, Parkville, Australia, [2]Walter and Eliza Hall Institute, Parkville, Australia

Genomic instability is a hallmark of cancer with structural rearrangements, chromosomal fusions and mutations all common occurrences. Sequencing the transcriptome using RNA-seq contains all the information about sequences and expression levels of the transcripts in the samples. However, the majority of analysis methods focus on summarising data into expression levels at either the gene or transcript level and often ignore the sequence information contained in the data. Here I will present our work on looking at sequences and mutations contained in the cancer transcriptome. I will begin by describing JAFFA, a tool we developed for detecting fusion genes in RNA-seq data. I will then move on to talking about our analysis methods for detecting of other transcriptional mutations commonly missed in RNA-seq data.

CG17: Phylogeny driven approaches to the study of microbial diversity

Eisen J.A.

University of California, Davis, USA. Department of Evolution and Ecology; Department of Medical Microbiology and Immunology, UC Davis Genome Center.

The rapid decreases in cost and difficulty of DNA sequencing has revolutionized studies of microbial diversity. In most cases, the tools and approaches used to analyze such data focus on similarity based metrics. I will present examples of how understanding the history behind taxa, genes and sequences can greatly improve studies of microbial diversity (e.g., in genome analyses, metagenomics, biogeography, and reference genome selection). In particular I will focus on phylogeny and phylogenetic approaches and how they can be used to help sort through massive sequence data sets and also improve planning and design of projects.

CG18: Gene expression noise is selected for and facilitates the evolution of novel regulatory interactions

Silander O.K.[1,2], Wolf, L. [2], van Nimwegen, E.[2]

[1]institute of Natural and Mathematical Sciences, Massey University, Auckland, New Zealand, [2]Biozentrum, University of Basel, Basel, Switzerland.

Gene expression levels fluctuate from cell to cell in clonal populations, a phenomenon called expression noise. Previous work has shown that certain cellular functions and pathways are enriched for genes exhibiting high or low levels of expression noise. However, it is not known how natural selection has affected the noise levels of these genes - for example whether it has acted to increase or decrease noise levels in specific functional categories. To address this question, we used laboratory evolution to evolve synthetic functional promoters. We compared the expression noise exhibited by these synthetic promoters to that exhibited by native E. coli promoters, and found that a substantial fraction of native promoters exhibit considerably higher noise. This suggests that many native promoters have experienced selection pressures that have increased their noise level. This observation is explained by recognizing that transcriptional regulators necessarily affect both mean expression and noise in expression. Explicit modelling shows that the unavoidable interaction between these two factors greatly facilitates the evolution of novel transcriptional regulation.

CG19: The emergence and evolution of the kiwifruit pathogen *Pseudomonas syringae* pv. *actinidiae*

Honour McCann

NZ Institute for Advanced Study, Massey University

The origins of agricultural diseases are linked with plant domestication; opportunities to witness disease emergence are therefore limited as most domestication events occurred centuries ago. Kiwifruit (Actinidia spp.) is an exception as it is one of the few plants domesticated in the 20th century. Bleeding canker disease caused by *Pseudomonas syringae* pathovar *actinidiae* (Psa) was reported soon after the expansion of kiwifruit cultivation in Asia. Disease outbreaks over the last three decades were caused by distinct lineages of Psa, which exhibited surprising diversity in virulence factor repertoires as well as evidence of between-clade recombination. The existence of a genetically diverse and recombining source population of Psa was predicted, from which stochastic sampling events followed by selection in agricultural environments for host specialization occurred. Subsequent sequencing and analysis of a small set of Chinese outbreak isolates showed these exhibit far greater diversity than all other outbreak strains. The recently domesticated kiwifruit is native to China and wild populations are often located in areas where kiwifruit cultivation also occurs, suggesting Psa may have emerged from a non-agricultural source population. Both wild and cultivated kiwifruit were sampled across eight provinces of China. Endophytic bacteria were isolated from both asymptomatic and symptomatic kiwifruit hosts. The diversity of kiwifruit-associated *P. syringae* is being characterised in order to identify whether these patterns vary according to cultivation and disease status. A population genomics approach will then be applied to identify the location of and extent of diversity within the source population of Psa, pinpoint the origin of the recent outbreak, and investigate the epidemiological spread and evolutionary processes leading to its emergence.

CG20: Inference of genomes from metagenomes: solving informatics problems in the wet lab

Darling, A.E.[1], Liu, M.[1], DeMaere, M.Z.[1], Beitel, C.W.[2], Eisen, J.A.[2]

[1]ithree institute, University of Technology Sydney, Sydney, AU, [2]Genome Center, University of California-Davis, Davis, CA, USA

Microorganisms play crucial roles in nearly every ecosystem, from the human body to the soil and sea. Metagenomics and related high throughput DNA sequencing techniques, have given us a first glimpse of our invisible and unculturable friends, leading to the recent discovery of dozens of new phyla[1]. These organisms have never been grown in the lab, yet their genomes have been recovered by applying new analysis techniques like correlated abundance binning to existing metagenomic data.

Meanwhile, population genetics and genomics have proven to be powerful tools to understand the ecological and evolutionary forces governing organisms. To date, it has been very difficult to apply population genomics approaches to microbes due to limitations inherent in DNA sequencing technology. I will describe recent work by my group and others to develop novel sequencing technologies and corresponding statistical methods to enable population metagenomics. We propose a solution to the computational problem of metagenome binning that addresses the root of the problem in the wet lab. By adapting the Hi-C and related 3C techniques to metagenomics, crucial information on the spatial organisation of sample DNA can be captured[2]. These techniques provide rich information that could support studies of population dynamics and gene flow in natural microbial communities. This new data type gives rise to several computational challenges that remain to be addressed.

1. Brown, C. T., L. A. Hug, B.C. Thomas, I. Sharon, C. J. Castelle, A. Singh, M. J. Wilkins, K. C. Wrighton, K. H. Williams, and J. F. Banfield. *Unusual biology across a group comprising more than 15% of domain Bacteria*. Nature. 2015 Jul 9;523(7559):208-11.
2. Beitel, C. W., L. Froenicke, J. M. Lang, I. F. Korf, R. W. Michelmore, J. A. Eisen, and A. E. Darling. *Strain- and plasmid-level deconvolution of a synthetic metagenome by sequencing proximity ligation products.* PeerJ. 2014 May 27;2:e415.

CG21: Crosstalk in bacterial and archaeal RNA regulatory circuits is selected against

Sinan U. Umu[1,2], Anthony M. Poole[1,2,3], Renwick Dobson[1,2] and Paul P. Gardner[1,2]

[1]School of Biological Sciences, University of Canterbury, Christchurch, New Zealand.
[2]Biomolecular Interaction Centre, University of Canterbury, Christchurch, New Zealand.
[3]Allan Wilson Centre for Molecular Ecology & Evolution, Massey University, Palmerston North, New Zealand.

RNAs play a significant role in various cellular functions such as translation, transcription, gene regulation, protein export and genome defence. While RNA-RNA interactions are generally thought to be functionally important, unfavourable crosstalk interactions may also occur. Despite a growing recognition of the importance of RNA-RNA mediated regulation, no study has yet evaluated the significance of unfavourable interactions between RNAs. We propose that, if crosstalk is selectively disadvantageous, this may be detectable as an 'avoidance signal': RNA-RNA interactions leading to crosstalk should be significantly underrepresented in prokaryotic genomes. We have tested this in silico using 1582 bacterial and 118 archaeal genomes together with bacterial gene expression and GFP reporter gene expression data. Our results show that crosstalk avoidance is a widespread phenomenon in bacteria and archaea. In addition, interaction avoidance level has a direct effect on protein abundances which makes it an important factor in gene optimization problem. We outline a number of models that account for our observations.

CG22: From genomics to effectors: how *Phytophthora* pathogens cause disease

Brett M. Tyler[1], Hua Wise[1,2], Ryan Anderson[2], Qunqing Wang[1,3], Danyu Shen[1,3], Wenwu Ye[3]; Brent Kronmiller[1], Nik Grunwald[4], Fredrik Heyman[5], Rays Jiang[6], Pieter van West[7], Carsten Russ[6], Richard Hamelin[8], Yuanchao Wang[3], John McDowell[2], Noel Chen[8], Mingying Gao[8]

[1]Oregon State University, Corvallis, USA; [2]Virginia Tech, Blacksburg, Virginia, USA; [3]Nanjing Agricultural University, Nanjing, China; [4]USDA-ARS HRL, Corvallis, Oregon, USA; [5]Swedish University of Agricultural Sciences, Sweden; [6]Broad Institute, Cambridge, Massachusetts, USA; [7]University of Aberdeen, Scotland; [8]University of British Columbia, Vancouver, Canada; [9]BGI Americas, Davis, California, USA

Oomycetes are fungus-like organisms that belong to the kingdom Stramenopila, which includes photosynthetic algae such as diatoms and kelp. Many oomycetes have evolved the ability to be destructive pathogens of plants while others infect animals. They have evolved pathogenicity independently of fungal pathogens, but have acquired many similar functions by convergent evolution and, in some cases, by horizontal gene transfer from fungi. Oomycete plant pathogens, including *Phytophthora* and *Pythium* species and many downy mildews, cause billions of dollars of damage to crops, forestry and ornamental plantings each year. Oomycete animal pathogens target mainly aquatic hosts and cause serious damage in aquaculture. We have generated draft genome sequences for the soybean pathogen *Phytophthora sojae*[1], the forest pathogen *P. ramorum*[1], the *Arabidopsis* pathogen *Hyaloperonospora arabidopsidis*[2] and the fish pathogen *Saprolegnia parasitica*[3]. We have also compared the genome sequences of the four major genotypes of *P. sojae*[4].

Comparisons among these genome sequences using data mining techniques have identified large numbers of rapidly evolving genes, including toxin and effector genes, that are likely involved in the interaction with hosts[5]. Plant and animal pathogens exhibit evolved very distinct repertoires of expanded virulence genes. Plant pathogen genomes encode a family of effectors, the RXLR effectors, that is particularly large (130-550 per genome) and rapidly evolving[5]. The effector proteins enter host cells by binding cell surface phosphatidylinositol-3-phosphate, a property that they share with some fungal pathogens of plants and humans[5].

We have used transcriptional profiling to identify *P. sojae* effector genes that play an active role in promoting infection. Combining this information with mathematical modeling and high throughput functional screens for plant defense suppression has revealed an elite subset of effectors comprising around 10% of the repertoire appears to be responsible for most of the contribution of this family to virulence[4]. Many family members are individually indispensible for virulence, while other effectors exhibit transscriptional variation among strains. Recently, we have sequenced the genomes of many additional *Phytophthora* species with the ultimate goal of sequencing the genomes of all genus members. Preliminary results from species in clade 7, which contains *P. sojae*, reveal a core set of effectors that are shared between narrow and broad host range species, and are under purifying selection.

1. Tyler, B. M. *et al.* (2006). *Phytophthora genome sequences uncover evolutionary origins and mechanisms of pathogenesis*. Science **313**, 1261-1266
2. Baxter, L. *et al.* (2010) *Signatures of adaptation to obligate biotrophy in the Hyaloperonospora arabidopsidis genome*. Science **330**, 1549-1551
3. Jiang, R. H. Y. *et al.* (2013). *Distinctive expansion of potential virulence genes in the genome of the oomycete fish pathogen Saprolegnia parasitica*. PLoS Genetics **9**, e1003272
4. Wang, Q. *et al.* (2011) *Transcriptional programming and functional interactions within the Phytophthora sojae RXLR effector repertoire*. Plant Cell **23**, 2064-2086
5. Jiang, R. H. Y. & Tyler, B. M. (2012) *Mechanisms and evolution of virulence in oomycetes*. Ann. Rev. Phytopath. **50**, 295-318

CG23: "Small zoospores in a large pond": Genomics of *Phytophthora* tree pathogens in New Zealand

Rebecca McDougal[1], David Studholme[2], Christine Sambles[2], Murray Grant[2], Dan Jones[3], Emily Telfer[1], Lucy Macdonald[1], Nari Williams[1].

[1]Scion, Rotorua, New Zealand, [2]University of Exeter, Exeter, UK, [3] NZGL, Auckland, New Zealand.

The emergence of arboreal *Phytophthora* diseases is increasing worldwide. These diseases present many challenges due to their impact on long-lived hosts, difficulty to control and the biosecurity challenges they pose. In New Zealand there has been a long association of *Phytophthora* disease in forest trees[1], for example *Phytophthora cinnamomi* and *Phytophthora cactorum* in *Pinus radiata* forest nurseries and shelterbelts. However, recently new *Phytophthora* diseases of trees have emerged internationally. In New Zealand these include red needle cast in *P. radiata* caused by *Phytophthora pluvialis, Phytophthora kernoviae* also causing disease in *P. radiata*, and kauri dieback in *Agathis australis* caused by *Phytophthora agathadicida*. In addition, many species of *Phytophthora* are associated with disease in trees although they are not the primary pathogen, such as *Phytophthora multivora*, which is commonly isolated along with other *Phytophthora* species from soils under diseased kauri. Using a systems biology approach, including genomics and transcriptomics, we are looking to understand the host-pathogen interactions of these *Phytophthora* species in three model hosts. Investigation of the foliar and root infection systems for each model will provide fundamental knowledge on the mechanisms of infection. Genome sequencing of 12 *Phytophthora* isolates has been performed and comparative analysis of effector repertoire is underway. Transcriptomic analysis with needle samples from an infection time-series, using susceptible and tolerant genotypes of *P. radiata* and two strains each of *P. pluvialis* and *P. kernoviae,* will indicate which genes are differentially expressed in both the host and pathogens during infection.

1. Scott, P., & Williams, N. (2014). *Phytophthora diseases in New Zealand forests.* NZ Journal of Forestry, 59(2), 14-21.

CG24: Phytophthoras attacking trees: population genomics, evolutionary mechanisms and genome diversity

A. L. Dale, N. Feau and R. C. Hamelin

Department of Forest and Conservation Sciences, Faculty of Forestry, The University of British Columbia, Vancouver, BC, Canada

Phytophthora species attacking angiosperms and conifers have gained notoriety with outbreaks of *Phytophthora ramorum, P. lateralis, P. alni* and *P. pinifolia*. We have sequenced the genomes of six Phytophthoras involved in tree attacks and have conducted extensive re-sequencing of multiple lineages of two species, *P. ramorum* and *P. lateralis*. We have discovered a broad range of within and among lineage diversity, including homozygous and heterozygous single nucleotide polymorphisms (SNP) among individuals within lineages. We examined the patterns of heterozygosity and hypothesized that alternate mechanisms such as mitotic recombination drive evolution in these asexual lineages resulting in gene gain or loss, and these events are driven by genome structure including gene density and transposable element content. To detect genomic signatures of recombination and adaptation, and to characterize inter-lineage differences, we sequenced and compared 107 genomes from the four lineages of *P. ramorum*. We observed loss of heterozygosity in 23 isolates in all four lineages, affecting over 100 scaffolds. In the NA1 lineage, one scaffold, containing a large cluster of genes encoding necrosis-inducing enzymes, 18 putative effectors and several carbohydrate active enzymes, is almost entirely homozygous in all isolates. Eight EU1 isolates are also homozygous in half of the same scaffold. The high frequency of this conversion in two lineages suggests that it could result in an advantageous genotype. Lineage specific regions dominated by transposon-like elements, and also composed of pathogenicity related genes define inter-lineage diversity. Mitotic recombination could therefore be a prominent driver of evolution in *P. ramorum*. They could result in expression of recessive alleles, loss of host-recognized effectors, or in mating type switches that could lead to sexual recombination in populations.

CG25: Population genomic tools and inferences to characterize emergence in the plant pathogen genus *Phytophthora*

Niklaus J. Grünwald

Horticultural Crops Research Unit, USDA Agricultural Research Service, Corvallis, OR, USA

Phytophthora pathogens continue to emerge and reemerge globally. The recent availability of whole genome sequences for *P. sojae, P. ramorum* and *P. infestans* as well as several draft genomes of other taxa have provided novel tools for inferring the evolutionary history and dissecting processes involved in emergence of new linages, strains or species of *Phytophthora*. This talk will assess what we have learned about the genetic and evolutionary mechanisms that explain patterns of emergence. What is the importance of effectors such as RxLR and crinkler genes? Which processes such as migration mutation, gene flow, hybridization, and recombination are important in shaping patterns observed for the genus? What have we learned from extensively studied model pathogens such as *P. ramorum* and *P. infestans* relative to other *Phytophthora* species? Many questions remain unresolved but the recent population genetic and genomic tools have already provided novel insights into specific processes involved in emergence. All of these approaches are helpful in developing informed disease management systems.

CG26: Investigating essential protein families in family Enterobacteriaceae using transposon mutagenesis

Ashari Ghomi, F.[1], Gardner, P.P.[1,2], Barquist, L.[3]

[1]School of Biological Sciences, University of Canterbury, Christchurch, New Zealand, [2]Biomolecular Interaction Centre, University of Canterbury, Christchurch, New Zealand, [3]Institute for Molecular Infection Biology, University of Würzburg, Würzburg, Germany

Many genes have been identified thanks to advances in sequencing technology and genome annotation methods. However, not all of these genes are of the same importance for the survival and reproduction of these organisms. Transposon insertion experiments are a useful method for investigating gene essentiality in a variety of organisms. This method uses a large-scale transposon approach to generate a population of mutants, these are left to grow in competitive conditions. Only the fit mutants can survive this process. After a number of generations, the number and positions of insertion sites are identified using next-generation sequencing. The genes that are free of insertions are likely to be essential for the survival of the organism.

We are analysing "Transposon-Directed Insertion-site Sequencing" (TraDIS) data [1] from 12 bacterial strains from the family Enterobacteriaceae, a scientifically and medically important clade including widely used model species such as *Escherichia coli* and *Salmonella enterica*. While a few studies have used transposon insertion experiments to perform comparisons between closely related strains, this data is generally not compared between species and genera. To study gene essentiality within an evolutionary framework, we have developed a profile hidden Markov model-based protein clustering tool to identify orthologous genes within these organisms. We are exploring the essentiality of ortholog clusters by defining three levels of essentiality: essential genes, non-essential genes, and beneficial gene losses in laboratory media. The results to-date suggest that genes with about one copy per genome are predicted to be mostly essential or non-essential by transposon insertion experiments. Moreover, genes with more than one copy per genome are predicted to be non-essential and ORFans are scattered among essential, non-essential and beneficial losses.

1. Barquist, L., Boinett, C. J. & Cain, A. K. Approaches to querying bacterial genomes with transposon-insertion sequencing. *RNA Biol.* **10,** 1161–1169 (2013).

CG27: Utilisation of genomic tools and techniques to further understand the genetics of the New Zealand Greenshell™ Mussel, *Perna canaliculus*

Rachael Ashby[1,2], Rudiger Brauning[1], Kim Rutherford[2], Tracey van Stijn[1], Hayley Baird[1], Chris Brown[2], Neil Gemmell[2], and Shannon Clarke[1]

[1] AgResearch, Mosgiel, [2]University of Otago, Dunedin

Perna canaliculus is a native New Zealand mollusc of economic importance to the New Zealand Aquaculture industry. The advent of Next Generation Sequencing (NGS) technologies and the advancement of computing power have enabled the generation of large volumes of genomic sequence data. This data can be used for *de novo* genome assembly and has resulted in the development of high throughput methods, such as Genotyping-By-Sequencing (GBS), for genotyping multiple individuals from a species cost effectively and efficiently. We present the *de novo* assembly of *P. canaliculus* and the development of a GBS pipeline to efficiently process GBS data for mussel breeding.

CG28: Predicting adaptive potential in Hihi (*Notiomystis cincta*) for conservation management

Miss. Kate Lee[1], Dr Anna Santure[1]

[1]University of Auckland, Auckland, New Zealand

Hihi, also called Stichbird (*Notiomystis cincta*), are an endemic New Zealand forest-dwelling bird which were last seen on the mainland in 1883. A Hihi recovery program has been operating since the 1980's, and the birds have been translocated to predator-free islands and fenced wildlife parks with increasing success. In particular the birds on Tiritiri Matangi Island in the Hauraki Gulf have been monitored through every breeding season since they were translocated there in 1995 and their pedigree has been established by a microsatellite genetics study. As part of the ongoing Hihi recovery program, we are developing a genomics toolkit to look at the genomic variation in the population. Our aim is to predict the adaptive potential of the population to inform future conservation work.

CG29: The Evolution of Poxvirus Immuno-modulatory Proteins

Elham Torbati [1], Kurt L. Krause [1], Chris M Brown [1]

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

Orthopoxviruses are double-stranded DNA viruses with linear genomes. Some have a board range of hosts but the others are species-specific (e.g smallpox/human). Orthopoxviruses consist of 9 fully sequenced reference species, one of these (smallpox) has been eradicated (176-224 kb). However, frequent outbreaks of orthopoxvirus diseases in animals and/or human indicate an ongoing dynamic between virus and host and the evolution of orthopoxviruses.

We analysed each gene of the full genomes of orthopoxviruses. As previously observed conserved genes are mostly located at central part of viral genome while there is variability at the ends. Some of the genes at extremities play roles in viral life in the host cells. In particular, immuno-modulatory proteins are a group of proteins that help the virus to evade host immune system using a number of different strategies - such as mimicking receptors or attaching to receptors. Interestingly, some of these immuno-modulatory proteins have similar structures but located in different locations (but mostly the termini of the genome). We targeted three of these proteins and aim to trace their origin using HMM, multiple alignment system and phlyogenetic methods. Based on the presence and absence of these genes, the level of identity, and also the recently published vaccinia ribosome foot printing map, we hypothesise that these proteins may arise from recombination, insertion and deletion resulting sometimes in new multifunctional proteins.

CG30: The mobilization and detection of TE insertions in a grapevine mutant-tagged population.

Thomson, S.J.[1], Lizamore, D.[2], Winefield, C.[2]

[1]The New Zealand Institute for Plant & Food Research LTD, Christchurch, NZ, [2]Department of Agriculture and Life Sciences, Lincoln University, NZ.

Plant genome sizes can vary quite significantly, by as much as 2000 fold, yet protein-coding genes are not highly variable. The transposable element (TE) fraction however does vary considerably, ranging between 3% and 85% generally following a linear correlation with genome size[1]. TEs are thought to be highly instrumental in shaping the genetic and epigenetic environments of eukaryotes as a result of their mobilisation and replication. The coding portion of the grapevine (Vitis vinifera) genome is dwarfed by an abundance of repetitive DNA sequences, and like many agricultural crops propagated through cutting, grapevine has been shown to accumulate TE-derived mutations which have led to valuable new phenotypes[2].

We have utilized the ability of TEs to mobilise and induce genetic and epigenetic modifications in order to generate a mutant-tagged population of Pinot Noir and Sauvignon Blanc grapevine cultivars. We have initiated a large-scale pipeline to regenerate a population of grape vines from totipotent somatic embryogenic callus cultures, some of which had been treated with environmental stressors, whereby the mobilome of grape has been activated. As part of a pilot study, we have generated whole-genome pool-seq of 20 regenerated plants and individual re-sequencing of a further 6 plants, and have identified a high degree of TE polymorphism in comparison with the parental material. Transposon mobility was not limited to stressed cultures and new insertions were attributed to a diverse range of TE types.

Here we describe the bioinformatic pipeline for detection and validation of TE insertions between regenerated Pinot Noir plantelets and parental material, and highlight the benefit of *in silico* validation.

1. Lee, S. I., & Kim, N. S. (2014). Transposable Elements and Genome Size Variations in Plants. *Genomics & Informatics*, *12*(3), 87–97. doi:10.5808/GI.2014.12.3.87
2. Kobayashi, S., Ishimaru, M., Ding, C. K., Yakushiji, H. and Goto, N.(2004). Retrotransposon-induced mutations in grape skin colour. *Science (80-.).* **304,** 982.

CG31: Deciphering the signals activating dendritic cells in skin allergy

Patel, D.[1], Cognard, E[1], Connor, L.[1], Eccles, D.[1] , Hilligan, K.[1], Ochiai, S., [1]Smith, A.[1], Tang, S.C.[1], McConnell, M.[1,2], Ronchese, F[1].

[1]Immune cell biology, Malaghan Institute of Medical Research, Wellington, NZ [2]School of Biological Sciences, Victoria University, Wellington, NZ.

Dendritic cells (DC) are sentinels of environmental stress. They recognise changes in our environment and activate innate immunity to drive long-lasting immune responses. There are several types of dendritic cells that reside in the skin including CD11b+, CD103+ and CD11b- CD103- CD326- DCs. The role of these different DC subsets in driving the allergic immune response is not well understood. To study the processes that drive the TH2 allergic immune response we use the parasite *Nippostrongylus brasiliensis* as a model of skin allergy. We have recently undertaken an RNA sequencing project to understand the signals that are driving the activation of these DC subsets. By comparing numerous bioinformatic resources including DAVID, InnateDB and Ingenuity pathway analysis to interrogate the data we have identified diverse transcriptional signatures across the dendritic cell subsets.

CG32: Combining Tn-seq with comparative genomics identifies proteins uniquely essential in *Shigella flexneri*

Nikki Freed[1,2], Dirk Bumanan[1], and Olin Silander[2,3,4]

[1]Infection Biology, Biozentrum, University of Basel, Basel, Switzerland, [2]Institute of Natural and Mathematical Sciences, Massey University, Auckland, New Zealand, [3]Computational and Systems Biology, Biozentrum, University of Basel, Basel, Switzerland, [4]corresponding author

Background

Protein functions that are essential for the growth of bacterial pathogens provide promising targets for antibacterial treatment. This is especially true if those functions are uniquely essential for the pathogen, as this might allow the development of targeted antibiotics, i.e. those that disrupt essential functions only for the pathogenic bacteria.

Here we present the results of a Tn-seq experiment designed to detect essential protein coding genes in *Shigella flexneri* 2a 2457T on a genome-wide scale. Our results suggest that 471 protein-coding genes in this organism are critical for cellular growth in rich media. Comparing this set of essential genes (the essential gene complement) with their orthologues in the closely related organism *Escherichia coli* K12 BW25113 revealed a significant number of genes that are essential in *Shigella* but not in *E. coli,* suggesting that the functional correspondence of these proteins had changed. Notably, we also identified a set of functionally related genes that are essential in *Shigella* but which have no orthologues in *E. coli*.

We found an extreme bias in proteins that have evolved to provide essential functions, with many proteins essential in *Shigella* but not *E. coli,* but with none (or very few) being essential in *E. coli* but not *Shigella*. We also identify a set- of genes involved in nucleotide biosynthesis that are essential in *Shigella*, but which lack orthologues in *E. coli.* Consequently, the data presented here suggest that the essential gene complement can quickly become organism specific, especially for pathogenic organisms whose genomes might have reduced robustness in their metabolic capacity (e.g. functional redundancy), or a reduced numbers of protein coding genes. These results thus open the possibility of developing antibiotic treatments that target differentially essential genes, which may exist even between very closely related strains of bacteria.

CG33: Designing powerful RNA-seq experiments

Todd, E.V.[1], Black, M.[2], Gemmell, N.J.[1]

[1]Department of Anatomy, University of Otago, Dunedin, NZ, [2]Department of Biochemistry, University of Otago, Dunedin, NZ.

RNA sequencing (RNA-seq) is a powerful tool in transcriptomics that enables whole transcriptome analysis *de novo*, and is quickly replacing microarrays for differential gene expression (DE) studies in non-model organisms. However, utilising powerful technology is not the same as having statistical power to address a research question. Technological advances do not eliminate biological variation, or circumvent the need for biological replication in establishing statistical significance of DE tests. Real-world budget constraints necessitate a trade-off between sequencing depth and sample size in RNA-seq experimental design. However, the field lacks clear experimental design guidelines and many published studies utilise few (2-3) biological replicates. We synthesise recent progress in statistical power analysis and sample size calculation for RNA-seq experiments examining DE, and derive 'rules-of-thumb' for RNA-seq experimental design. Sequencing more replicates at lower depth achieves the greatest statistical power for DE analysis. Deeper sequencing gives diminishing returns on power once read depths are sufficient to overcome Poisson counting error (av. 10 mapped reads per gene). Additional biological replicates have a far greater impact on power, although required sample sizes depend critically on the biological variance and fold change of the data: datasets with higher biological variance or smaller fold changes require larger sample sizes to overcome higher measurement error. With modest replication (3-5 replicates per condition), only the very largest expression differences (≥ 3 fold) are detectable with high statistical confidence (≥ 80% power), unless biological variance is very low (i.e. typical of cell lines and inbred animal strains). Detecting subtle expression differences (< 2 fold change) requires very large sample sizes (> 10 replicates per condition). More powerful (well-replicated) study designs remain prohibitively expensive with current technology, especially for med-high variance datasets. We suggest a pilot sequencing approach in order to estimate power, and establish the feasibility of larger experimental designs

CG34: The transcriptomic shift accompanying the photoperiod induced acceleration of flowering in *Medicago truncatula*

Thomson, G.[1], Tsai, P.[2], Putterill, J.[1]

[1]School of Biological Sciences, The University of Auckland, Private Bag 92019, Auckland Mail Centre, Auckland 1142, New Zealand, [2]Bioinformatics Institute, The University of Auckland, Private Bag 92019, Auckland Mail Centre, Auckland 1142, New Zealand

Seasonal variation in the length of the day acts as an important cue regulating the life-cycle of many plant species, allowing them to predict changes in their environment and accurately align their flowering with conditions conducive for reproduction. In legumes, the third most diverse plant family on Earth, floral induction appears to differ from previously studied species. Using the model species *Medicago truncatula* this study utilises high-throughput RNA-Sequencing (RNA-Seq) to describe the global transcriptional change associated with a shift from short-day to long-day conditions. *M. truncatula* is known to accelerate its flowering time in response to both long-days and vernalisation (a period of winter cold). This allows it to remain in a vegetative state over winter and then flower in spring when it is warm and the days are long. The results present a broad picture of the altered patterns of gene expression associated with this acceleration and also compare the performance of five popular approaches to modeling differential gene expression not previously reported in any plant system. Ultimately this study contributes to the wider understanding of the molecular control of flowering and provides a valuable resource for future research in this area

.

Summary of Abstracts – Computational Genomics Poster Session

No.	Title	Presenter	Institutions
CG26	Investigating essential protein families in family Enterobacteriaceae using transposon mutagenesis	Fatemeh Ashari Ghomi	School of Biological Sciences, University of Canterbury, Christchurch, NZ
CG27	Utilisation of genomic tools and techniques to further understand the genetics of the New Zealand Greenshell™ Mussel, *Perna canaliculus*	Rachael Ashby[1,2]	1-AgResearch, Mosgiel, NZ 2-University of Otago, Dunedin, NZ
CG28	Predicting adaptive potential in Hihi (*Notiomystis cincta*) for conservation management	Kate Lee	The University of Auckland, Auckland, New Zealand
CG29	The Evolution of Poxvirus Immuno-modulatory Proteins	Elham Torbati	Department of Biochemistry, University of Otago, Dunedin, NZ
CG30	The mobilization and detection of TE insertions in a grapevine mutant-tagged population.	Susan Thomson	The New Zealand Institute for Plant & Food Research, Christchurch, NZ
CG31	Deciphering the signals activating dendritic cells in skin allergy	Deepa Patel	Malaghan Institute , Wellington, NZ
CG32	Combining Tn-seq with comparative genomics identifies proteins uniquely essential in *Shigella flexneri*	Nicole Freed[1,2], Olin Silander[1,2]	1-Massey University, Auckland, NZ, 2-University of Basel, Basel, Switzerland
CG33	Designing powerful RNA-seq experiments	Erica Todd	Department of Anatomy, University of Otago, Dunedin, NZ
CG34	The transcriptomic shift accompanying the photoperiod induced acceleration of flowering in *Medicago truncatula*	Geoffrey Thomson	School of Biological Sciences, The University of Auckland, Auckland, NZ