ABSTRACTS

GENOME 10K & GENOME SCIENCE

29 AUG - 1 SEPT 2017

NORWICH RESEARCH PARK, NORWICH, UK

Evolutionary Genomics 110

Conservation Genomics 124

KEYNOTE SPEAKERS

Dr Adam Phillippy: Towards the gapless assembly of complete vertebrate genomes

Computational and Statistics Branch, National Human Genome Research Institute, Maryland, US

A complete and accurate genome sequence forms the basis of all downstream genomic analyses. However, even the human reference genome remains imperfect, which affects the quality of experiments and can mask true genomic variations. For most other species, quality reference genomes do not exist. Long-read sequencing technologies from Pacific Biosciences and Oxford Nanopore have begun to correct this deficiency and have enabled the automated reconstruction of reference-quality genomes at relatively low cost. Further combination of these technologies with complementary scaffolding and phasing approaches such as chromatin conformation capture (Hi-C) may soon enable the complete reconstruction vertebrate haplotypes. I will review recent application of these approaches, and present a strategy for the automated assembly of hundreds of high-quality vertebrate reference genomes for the Genome10K project.

Prof Kathy Belov: Saving the Tasmanian devil from extinction

University of Sydney, AU

Kathy's research team have demonstrated that Tasmanian devils have extremely low levels of genetic diversity at the Major Histocompatibility Complex (MHC) providing an opportunity for Tasmanian Devil Facial Tumour Disease (DFTD), a rare contagious cancer, to spread through devil populations without encountering histocompatibility barriers. They continue this research by studying the relationship between MHC type and disease susceptibility in devil populations, as well as the impact of the emergence and evolution of DFTD strains using genomics technologies.

Prof Peter Holland: Homeobox genes and animal evolution: from duplication to divergence

University of Oxford, UK

To understand the evolution of animals, we must understand genomes and development. One of the most important discoveries in 20th century biology was the finding that widely different animal species use similar genes, such as homeobox genes, to build their embryos. But if the genes are conserved, why do animal species look so different? Does evolution subtly change the regulation of key genes, or change the number of genes, or change their protein coding sequences? Examples of all three routes have been revealed through comparative genomics, including some surprising examples of how evolution changed the number and function of homeobox genes in mammalian evolution.

Dr Hilary Burton: Genomics in healthcare: the challenges of complexity

Foundation for Genomics and Population Health, Cambridge, UK

Genomic technologies have greatly enhanced our understanding of health and disease. Sequencing has become cheaper and quicker, whilst our increasing ability to interpret the data using huge computer power and very big databases, means that genomic testing can now influence clinical decisions in many areas of medicine. Whilst new possibilities continue to escalate, moving from scientific research to tried, tested and routine healthcare is not straightforward.

In this presentation I will outline some of the many dimensions of genomics in healthcare including disease prevention, making a precise diagnosis in rare and more common diseases, choosing drug treatments and assessing reproductive risk. I will explore some of the challenges facing health systems, which arise in part from the complexity of genomic information and the fast-moving nature of the technologies, but also include organisational and professional challenges: for example, the regulatory and practical difficulties of sharing personal data in health systems, or the educational programmes required to ensure that all healthcare professionals can use genetic testing appropriately and safely in their practice.

As health systems face the demands of an ageing population, a constant stream of emerging technologies and raised public expectations, I will suggest that using genomics effectively can be part of the solution. Together with other biomedical and even digital technologies, it can enable a move towards more personalised healthcare and a shift from end-stage ‘rescue’ to prevention and earlier diagnosis.

INVITED SPEAKERS

Vertebrate Genomics

Alex Cagan: Comparative genomics of animal domestication

Wellcome Trust Sanger Institute, UK

The domestication of animal species was essential for the emergence of complex human societies. Despite its importance there is much about the domestication process that we still do not know. Domesticated species tend to share a suite of phenotypic traits referred to as the ‘domestication syndrome’. However, whether these phenotypic similarities are the result of convergence at the genetic level remains unclear. We generated whole-genome sequences from experimentally domesticated Norway rats and American mink, and identified genes and putatively functional variants that may underlie the phenotypic differences seen in the domesticated animals.

When we combine these data with whole-genome sequences from multiple pairs of domestic animals and their wild sister species we find biological pathways that appear to be recurrently affected by the domestication process across all domesticated animal species. One of these is the ErbB signalling pathway, involved in the development of the reproductive system and neural crest migration.

Plant Genomics

Ksenia Krasileva: Evolution of plant Immune receptors

Earlham Institute, UK

Understanding evolution of plant immunity is necessary to inform rational approaches for genetic control of plant diseases. The plant immune system is innate, encoded in the germline, yet plants are capable of recognizing diverse rapidly evolving pathogens. Availability of plant genomes plant species allowed us to elucidate evolutionary history of plant immune receptors of Nucleotide-Binding Leucine Rich Repeat class (NLRs) that provide genetic diversity to recognize pathogens and induce signaling cascade. We identified the 'core' and highly variable sub-clades of NLRs from across 60 plant species, including previously understudied monocots and uncovered sub-family clade expansions. A recent paradigm in NLR-based recognition of pathogens involves NLRs with exogenous gene fusions, called integrated domains (NLR-IDs) that can serve as baits for pathogen-derived effectors. We have shown that NLR-IDs are prevalent across flowering plants and identified their ID repertoires. We uncovered a clade of NLRs that is undergoing repeated independent integration events that produces diverse NLR fusions to other genes. This NLR clade is ancestral in grasses with members often found on syntenic chromosomes while integrated domains are exchanged from different genomic locations. Sequence analyses revealed that DNA transposition or ectopic recombination are most likely mechanisms of NLR-ID formation. The identification of a subclass of NLRs that is naturally adapted to new domain integration can inform biotechnological approaches for generating synthetic receptors with novel pathogen 'traps'.

Andrea Harper: Using Associative Transcriptomics to predict tolerance to ash dieback disease in European ash trees

University of York, UK

Associative Transcriptomics (AT) is a potent method, first developed in the crop plant Brassica napus, enabling rapid identification of gene sequence and expression markers associated with trait variation in diversity panels. It can be effective even when advanced genomic resources are unavailable, making it a valuable tool for studying traits in non-model species. Most recently, we applied AT to the problem of ash dieback disease, a fungal disease affecting ash trees which was first discovered in the UK in 2012.

Using a Danish ash diversity panel varying for susceptibility to the disease, we discovered expression-based markers that could be used to identify trees with high levels of tolerance to the disease. In addition, information about the genes in which the markers are located, is revealing clues to the mechanisms underlying the ability of some trees to tolerate the disease.

John Lees: Scalable pan-genome-wide association studies in bacteria

Wellcome Trust Sanger Institute, UK

Genome-wide association studies (GWAS) have long been a staple of human genetics. In the simplest case a population-matched cohort of unrelated individuals with and without a disease or trait is genotyped, and then every marker (SNP) is tested for association with the phenotype. The ease of design has allowed very large cohorts to be recruited to these studies, yielding excellent power for linking genotype to phenotype. With the recent availability of populations hundreds or thousands of sequenced bacterial isolates interest has developed in applying the same technique to relevant pathogen phenotypes such as drug resistance and invasive potential. However, the highly variable pan-genome and potentially confounding strong population structure of bacteria make GWAS difficult to apply in the same way. In this talk I will describe Sequence Element Enrichment analysis (SEER), a method we have published which overcomes these issues by using k-mers as a generalised sequence variant along with appropriate population structure corrections. SEER is freely available and scales to thousands of genomes, and has been used to discover variants affecting invasive potential of *S. pyogenes* and region specific patterns of *B. pseudomallei*. Finally, I will describe recent work which pushes the limits of GWAS, testing the contribution of rare and structural variants to bacterial phenotypes.

Gemma Langridge: Contaminant or infective agent? Re-classifying the staphylococci for modern medicine

University of East Anglia, UK

Gemma Langridge[1], Rebecca Clifford[1], Claire Hill[1], Emma Meader[1,2], Caroline Barker[2], Iain McNamara[2], Lisa Crossman[1,3,4] & John Wain[1]

[1] Medical Microbiology Research Laboratory, Norwich Medical School, UEA

[2] Norfolk & Norwich University Hospital, Norwich

[3] School of Biological Sciences, UEA

[4] SequenceAnalysis.co.uk, NRP Innovation Centre, Norwich Research Park, Norwich

In many hospital laboratories, non-*aureus* staphylococci (NAS) are the most common isolates in blood culture. Although *S. aureus* is considered a true pathogen, NAS is often categorised as a contaminant. However, NAS are an important cause of healthcare associated infections, particularly associated with indwelling medical devices, such as prosthetic joints. They are also a reservoir of antimicrobial resistance genes, with resistance to methicillin and other frequently used antibiotics on the rise.

To investigate the population structure of NAS, we are establishing a diverse collection, currently just over 400 isolates from clinical samples, healthy volunteers and animals. At the Norfolk and Norwich Hospital, the clinical microbiology laboratory identifies isolates to the nearest species match using the gold standard MALDI-TOF method; we have used both MALDI-TOF and Illumina whole genome sequencing to characterise around 300 isolates from the collection.

The lack of a large shared core in NAS directed us to a different approach, but to gain greater resolution over the single gene approach of 16S, we used the concatenated sequence of 16 ribosomal proteins to cluster the strains, resulting in 17 robust cluster (RC) groups. Overlaying the MALDI-TOF species names upon RC groups made it clear that the MALDI-TOF species designations do not necessarily follow the phylogeny. As a test case within NAS, we show that there is a significant phylogenetic distinction between "*S. saprophyticus*" strains isolated from urinary tract disease and those not causing disease.

Clustering of ribosomal protein sequences has revealed robust clades within *Staphylococcus* that provide the opportunity to generate a new, biologically sound definition of NAS.

Evolutionary Genomics

Emma Teeling: Growing old yet staying young: A genomic perspective on bats' extraordinary longevity

University College Dublin, IRE

Of all mammals, bat possess the most unique and peculiar adaptations that render them as excellent models to investigate the mechanisms of extended longevity and potentially halted senescence. Indeed, they are the longest-lived mammals relative to their body size, with the oldest bat caught being 41 years old, living approx. 9.8 times longer than expected. Bats defy the 'rate-of-living' theories that propose a positive correlation between body size and longevity as they use twice the energy as other species of considerable size, but live far longer. The mechanisms that bats use to avoid the negative physiological effects of their heightened metabolism and deal with an increased production of deleterious Reactive Oxygen Species (ROS) is not known, however it is suggested that they either prevent or repair ROS damage.

Bats also appear to have resistance to many viral diseases such as rabies, SARS and Ebola and have been shown to be reservoir species for a huge diversity of newly discovered viruses. This suggests that their innate immunity is different to other mammals, perhaps playing a role in their unexpected longevity. Here the potential genomic basis for their rare immunity and exceptional longevity is explored across multiple bat genomes and divergent 'ageing' related markers. A novel blood based population-level transcriptomics approach is developed to explore the molecular changes that occur in an ageing wild population of bats to uncover how bats 'age' so slowly compared with other mammals. This can provide a deeper understanding of the causal mechanisms of ageing, potentially uncovering the key molecular pathways that can be eventually modified to halt, alleviate and perhaps even reverse this process in man.

Clinical and Translational Genomics

Joris Veltman: *De novo* mutations in genetic disease

Institute of Genetic Medicine, Newcastle University, UK

Department of Human Genetics, Radboud University Medical Centre, Nijmegen, NL

How is it possible that severe early-onset disorders are mostly genetic in origin, even though the disorders are not inherited because of their effect on fitness? Genomic studies in patient-parent trios have recently indicated that most of these disorders are caused by de novo germline mutations, arising mostly in the paternal lineage.

In this presentation I will discuss our research on the causes and consequences of de novo mutations using novel genomic approaches. I will illustrate all of this using severe intellectual disability as a model, for which we are making rapid progress and now have the opportunity to provide medically relevant information to the majority of patients and families involved.

Matthew Hurles: Deciphering Developmental Disorders

Wellcome Trust Sanger Institute, UK

Children with severe, undiagnosed developmental disorders (DDs), including Intellectual Disabilities as well as multi-system congenital malformations, are enriched for damaging de novo mutations (DNMs) in developmentally important genes. Working with the clinical genetic services of the UK and Ireland we have exome sequenced 13,600 families. We have diagnosed thousands of children, by providing the information back to their clinicians. We've determined that 40-45% of these children have causal de novo mutations in protein-coding genes, and we've identified over 30 novel disorders so far. We've also determined that de novo mutations are also enriched in highly conserved regulatory elements that are active in fetal brain, but that these only account for a small minority of as yet undiagnosed patients.

Agricultural Genomics

Alan Archibald: Precision engineering for PRRSV resistance in pigs

The Roslin Institute, University of Edinburgh, UK

Porcine Reproductive and Respiratory Syndrome (PRRS) is arguably the most important infectious disease for the world-wide pig industry. The effects of PRRS include late-term abortions and stillbirths in sows and respiratory disease in piglets. The causative agent of the disease is the positive-strand RNA PRRS virus (PRRSV). PRRSV has a narrow host cell tropism, targeting cells of the monocyte/macrophage lineage. One of the host proteins involved in facilitating viral entry is CD163 which has been described as a fusion receptor for PRRSV. CD163 is expressed at high levels on the surface of macrophages, particularly in the respiratory system. The scavenger receptor cysteine-rich domain 5 (SRCR5) region of CD163 has been shown to interact with virus in vitro.

We used CRISPR/Cas9 gene editing technology to generate pigs with a deletion of the CD163 exon 7 which encodes the SRCR5 domain. Deletion of SRCR5 showed no adverse effects in pigs maintained under standard husbandry conditions with normal growth rates and complete blood counts observed. Pulmonary alveolar macrophages (PAMs) and peripheral blood monocytes (PBMCs) were isolated from the animals and assessed in vitro. Both PAMs and macrophages obtained from PBMCs by CSF1 stimulation (PMMs) show the characteristic differentiation and cell surface marker expression of macrophages of the respective origin.

Expression and correct folding of the SRCR5 deletion CD163 on the surface of macrophages and biological activity of the protein as hemoglobin-haptoglobin scavenger was confirmed. Both PAMs and PMMs were challenged with PRRSV genotype 1, subtypes 1, 2, and 3 and PMMs with PRRSV genotype 2. PAMs and PMMs from pigs homozygous for the CD163 exon 7 deletion showed complete resistance to viral infections assessed by replication. Confocal microscopy revealed the absence of replication structures in the SRCR5 CD163 deletion macrophages, indicating an inhibition of infection prior to gene expression, i.e. at entry/fusion or unpacking stages.

Nicola Patron: Engineering Plant Genomes for Farming and Pharming

Earlham Institute, UK

Synthetic biology applies engineering principles to biology for the construction of novel biological systems designed for useful purposes. It advocated for standards and foundational technologies to facilitate biological engineering. Defining standards for plants has enabled us to automate parallel DNA assembly at nanoscales, removing research bottlenecks and providing the international plant community access to reusable, interoperable, characterized, standard DNA parts. We are applying these principles to programmable genome engineering tools for multiplexed targeted mutagenesis and for the development of tunable, orthologous regulatory elements, synthetic transcription factors and genetic logic gates.

Conservation Genomics

Beth Shapiro: The genomic consequences of inbreeding in mountain lions, *Puma concolor*

University of California Santa Cruz, US

Beth Shapiro, Nedda Saremi, Megan Supple, Gemma Murray, Richard E. Green, Eduardo Eisirik and the puma genome sequencing consortium*

Human land-use changes, including deforestation and establishment of roads and highways, can obstruct natural dispersal and migration corridors, leading to population isolation and inbreeding. Among the most affected species in North America by human land-use changes is the mountain lion, *Puma concolor*. Once distributed across North America, mountain lions are today found only in southern Florida and the western part of the continent.

To explore the genomic consequences of increasing isolation between mountain lion populations, we sequenced and assembled a chromosome-scale de novo genome from a mountain lion from the Santa Cruz mountains, 36M, and generate high coverage resequencing data from mountain lions from populations across North America and Brazil. Using these data, we investigated the relative timing of onset and duration of inbreeding within potentially distinct mountain lion populations. North American mountain lions contain significantly less genomic diversity than Brazilian mountain lions, but show varying levels of inbreeding that does not correspond directly to present-day barriers between them. Finally, we explore the selective consequences of inbreeding on North American mountain lions, and identify genomic changes that may have evolved as a consequence of increased interaction with humans.

Developmental Biology

Kristin Tessmar-Raible: Genomic and transcriptomic approaches for the study of daily, monthly and seasonal timing

Max F. Perutz Laboratories, University of Vienna, AT

Life is controlled by multiple rhythms. While the interaction of the circadian clock with environmental stimuli is well documented, its relationship to endogenous oscillators with other periods, as well as natural timing variation between individuals of the same species is little understood.

The marine bristle worm *Platynereis dumerilii* harbors a light-entrained circadian, as well as a monthly (circalunar) clock. Our first studies suggest that the circalunar clock persists even when circadian clock function is disrupted as evidenced by the complete absence of molecular and behavioral circadian oscillatory patterns. However, the circalunar clock impacts on the circadian clock on two levels:

a) It regulates the level of a subset of core circadian clock genes.
b) In addition to its molecular input, we furthermore find that the circalunar clock changes the period and power of circadian behavior, although the period length of the daily transcriptional oscillations remains unaltered.

In order to study the molecular and cellular nature of its circalunar clock, as well as its interaction with the circadian clock, we have established transient and stable transgenesis, inducible specific cell ablations, chemical inhibitors, as well as TALEN-mediated genome engineering. We have been investigating the extent of transcript changes in the brain caused by the circalunar clock and compare these changes to other major conditions (sex determination, maturation) occurring during the life of the worm, as well as to the known extent of transcript changes caused by the circadian clock.

The marine midge *Clunio marinus* possesses a circadian clock, and in addition acquired a circalunar clock during the past 20.000 years. Strains of different geographic origins exhibit differences in their circalunar and circadian timing (“chronotypes”), which are genetically encoded and map to 3 quantitative trait loci (QTLs). We sequenced and assembled the 90Mbp genome of the midge and mapped the QTLs to the molecular map. Based on subsequent single nucleotide polymorphism (SNP) analyses differentially fixed in different timing strains, and molecular studies, we suggest that circadian chronotypes in Clunio are caused by activity variants in the enzyme CaMKII.

Given its evolutionary conservation and prominent role in the mammalian brain, it is tempting to speculate, that CaMKII could play a similar role in mammals, and could thus provide a molecular link between extreme chronotypes and frequently co-occurring neuropsychological diseases.

Andrea Münsterberg: Cellular dynamics and lineage specification in developing somites

University of East Anglia, UK

A fundamental process during both embryo development and stem cell differentiation is the control of cell lineage determination. In developing skeletal muscle, many of the diffusible signalling molecules, transcription factors and non-coding RNAs that contribute to this process have been identified. This has advanced our understanding of the molecular mechanisms underlying the control of cell fate choice. In vertebrate embryos, skeletal muscle is derived from paired somites. These are transient embryonic segments that also contain progenitors for other cell lineages of the musculoskeletal system, such as chondrocytes and axial tendon progenitors. In addition, some endothelial cells, adipocytes and brown fat cells are somite derived. We are developing approaches to examine the full complexity and molecular profiles of progenitor cells that are present in early and later stage somites. This will allow us to delineate molecularly distinct cell types, to define progenitors and lineage relationships, and to identify crucial pathways, hubs and markers for the lineages of the musculoskeletal system. In parallel, we use imaging approaches to assess cellular behaviours during somite maturation, a highly dynamic process that involves significant morphogenetic changes. A more detailed understanding of the key mechanisms and factors involved will be important for stem cell science, regenerative medicine and tissue engineering.

Microbial Communities

Mads Albertsen: Towards a fully populated tree of life

Aalborg University, DK

Small subunit (SSU) ribosomal RNA (rRNA) genes have been the standard phylogenetic markers for the study of microbial evolution and diversity for decades. However, the essential reference databases of full-length rRNA gene sequences are underpopulated, ecosystem skewed, and subject to primer bias; which hampers our ability to study the true diversity. In this talk, I will present out latest method development that combines poly(A)-tailing and reverse transcription of SSU rRNA molecules with synthetic long-read sequencing, to generate millions of high quality, full-length SSU rRNA sequences without primer bias. We applied the approach to complex samples from seven different ecosystems and obtained more than 1,000,000 SSU rRNA sequences from all domains of life. The novel diversity is overwhelming and include several potentially new archaeal phyla of the deeply branching Asgard Archaea, which are previously suggested to bridge the gap between prokaryotes and eukaryotes. This approach will allow expansion of the rRNA reference databases by orders of magnitude and will enable a comprehensive census of the tree of life. With a fully populated SSU tree of life, it will be possible to prioritize efforts towards making a fully populated genome tree of life. To demonstrate the progress with these efforts, I will also discuss our recent progress on extraction of complete (closed) genomes from metagenomes using high-throughput long-read Nanopore.

Lindsay Hall: Early life microbial communities

Quadram Institute, UK

The gut is home to an astonishingly diverse, dynamic, and populous ecosystem. This complex microbial community, termed the microbiota, is critical for host wellbeing. Disturbances in our microbiota, such as via caesarian sections and antibiotic exposure, can lead to increased susceptibility to pathogens, as well as atopic, and chronic inflammatory diseases. Bifidobacteria constitute a substantial proportion of the gut microbiota, particularly during early life and high-levels are associated with the development of mucosal defence. Currently there are many bifidobacterial species and strains with claimed health promoting or ‘probiotic’ attributes, however the mechanisms through which these strains reside within their host and exert benefits is far from complete. In this talk I will discuss the role of the gut microbiota with the host, focusing on the example of bifidobacteria in host colonisation, epithelial cell cross-talk, and pathogen protection.

Sequencing Technology and Developments

Aaron McKenna: Information and storage recovery using the diversity of second-generation sequencing technologies

University of Washington, US

Second-generation sequencing has been traditionally seen in terms of a key trade-off: a huge increase in information recovery at the cost of information fragmentation. Here we show that such weaknesses can be overcome by leveraging a series of inventive techniques developed by the field at large. First, we demonstrate that second-generation sequencing can be used to recover chromosomal level contiguity in the de novo genome assembly of a previously unsequenced Muridae species. In addition, we demonstrate it's utility in recovering the 'orthogonal genome': human engineered information storage within the genomes of single living cells, and it's application to tracing whole-organism lineage.

Genome Informatics

Doreen Ware: TBC

Cold Spring Harbor Laboratory, US

Population Genomics

Richard Durbin: Whole genome sequence studies of the Lake Malawi cichlid adaptive radiation

Wellcome Trust Sanger Institute, UK

The adaptive radiations of haplochromine cichlid fish in the East African great lakes provide paradigmatic systems to study the dynamics of species formation, and of natural and sexual selection. The most extensive radiation is in Lake Malawi, where in the last million years or so one or a few ancestral populations have given rise to a flock of more than 500 species, filling almost all piscine ecological niches in the lake.

Over the past few years we have collected with collaborators over 2500 samples and sequenced the whole genomes of over 300 fish from over 100 species of Lake Malawi cichlids. All species are genetically close, with pairwise divergence typically between 0.1 and 0.25%, compared to heterozygosity between 0.05 and 0.15%. In addition to extensive incomplete lineage sorting, we see strong signals of gene flow between clades at different levels in the radiation, based on PCA, F statistics and related methods. There appear to be several long chromosomal regions exhibiting unusual phylogeny, perhaps indicative of a role for large inversions in species separation.

At a finer scale, although for close species pairs Fst can be under 20%, we also see local spikes or “islands” of high differentiation that are statistically significant under simple models of population separation, suggestive of loci under selection. Finally, at a functional level, we see higher non-synonymous to synonymous differences between species in genes involved in retinal processing, the innate immune system, oxygen transport, and a number of other pathways.

Single Cell

Muzlifah Haniffa: Deconstructing the immune system using single cell technologies

Newcastle University, UK

Muzlifah has used functional genomics, comparative biology and more recently single cell RNA sequencing to study human mononuclear phagocytes. In this seminar, she will discuss the power and utility of single cell RNA sequencing to identify new dendritic cells, monocytes and progenitor cells relevant for immunotherapy.

Tamir Chandra: Understanding cellular heterogeneity in cellular senescence and ageing through single cell transcriptomics

MRC Human Genetics Unit, University of Edinburgh, UK

A key event in a healthy cell turning into a cancer cell is the activation of an oncogene. To prevent transforming to a cancer cell, the cell harbouring the oncogene activates a tumour suppressive programme, pushing itself into an irreversible growth arrest, called oncogene induced senescence (OIS). Everyone carries OIS cells, for example in the benign lesions (such as moles) that never progress to malignant cancer. Most of the time these lesions stably exist over decades, but sometimes individual cells escape and progress to cancer. What enables individual cells to turn malignant and how are they different from the cells around them? Here we present single cell transcriptomes of a time-course of human fibroblasts on their way to senescence after oncogene activation. Applying machine learning to order cells along a senescence trajectory, we find an unexpected bifurcation, leading to two distinct senescence endpoints. Each of these endpoints exclusively expresses sets of canonical senescence genes. Most importantly, one population failed to regulate key genes thought essential for the stability of the senescent state, leading to a scenario where the heterogeneity of the benign state might enable escape to malignancy.

Stephen Sansom: Transcript structures in the thymus: improvised or rehearsed?

Kennedy Institute of Rheumatology, University of Oxford, UK

Kathrin Jansen[1,2]*, Stefano Maio*[2]*, Annina Graedel*[2]*, Iain C. Macaulay*[3]*, Chris P. Ponting*[4]*, Georg A. Holländer*[2]*, Stephen N. Sansom*[1]

[1] Kennedy Institute of Rheumatology, University of Oxford, Oxford, UK

[2] Weatherall Institute of Molecular Medicine, University of Oxford, Oxford, UK

[3] Earlham Institute, Norwich Research Park, Norwich, UK

[4] MRC Institute of Genetics & Molecular Medicine, University of Edinburgh, Edinburgh, UK.

Epithelial cells of the thymus are remarkable for their ability to promiscuously express nearly all protein coding genes in order to assess the self-reactivity of developing T-cells. Such T-cells must also be able to tolerate the isoform specific epitopes that they will encounter as they monitor the various tissues of the body. Currently, the extent and fidelity of peripheral isoform representation in thymic epithelial cells is only poorly understood. We therefore used population and single-cell transcriptomics to compare transcript architectures between the thymus and peripheral tissues. These data also provide insights into the process by which the isoform repertoire of thymic epithelial cells is generated.

ORAL PRESENTATIONS

Vertebrate Genomics

(O1/P61) Comparative whole-genome study of eleven Felidae species from six lineages

Gaik Tamazian[1], Ksenia Krasheninnikova[1], Pavel Dobrynin[1], Sergey Kliver[1], Aleksey Komissarov[1], Stephen O'Brien[1]

[1] Saint-Petersburg State University

The Felidae family represents a remarkable example of worldwide species radiation and adaptation to various environments. Here we present comparative analysis of whole genomes of eleven Felidae species - lion (*Panthera leo*), tiger (*Panthera tigris*), snow leopard (*Panthera uncia*), leopard (*Panthera pardus*), jaguar (*Panthera onca*), caracal (*Caracal caracal*), lynx (*Lynx lynx*), Asian leopard cat (*Prionailurus bengalensis*), fishing cat (*Prionailurus viverrinus*), puma (*Puma concolor*), cheetah (*Acinonyx jubatus*), and domestic cat (*Felis catus*) - that cover six lineages of the family (Panthera, Caracal, Lynx, Asian leopard cat, Puma, and Domestic cat). For each species, its whole-genome assembled sequence was assessed and annotated. The considered genomic features included genes, repeats, and variants. The structural alignment of the genomes was performed to identify homology and rearrangements between them. Homozygosity regions were determined based on single nucleotide variants called in the sequenced specimens. Differences and similarities between the annotated genomes are interpreted in terms of the evolutionary process that took place 10.8 million years ago and initiated branching from the last common Felid ancestor.

(O2/P52) A Cross-Species Bioinformatics and FISH approach to physical mapping of Mammalian Genomes

Rebecca Jennings[1], Rebecca O'Connor[1], Lucas Kiazim[1], Gothami Fonseka[2], Marta Farré[3], Joana Damas[3], Laura Buggioti[3], Denis M Larkin[3], Darren K Griffin[1]

[1] School of Biosciences University of Kent, Canterbury, Kent

[2] Cytocell Ltd, 3-4 Technopark Newmarket Road, Cambridge

[3] Department of Comparative Biomedical Sciences, Royal Veterinary College, University of London

To facilitate analysis of the whole genome, an unbroken length of contiguous DNA sequence, along the length of each chromosome is essential. Most de novo sequenced genomes exist as a series of scaffolds, and are therefore highly fragmented. This fragmentation limits their use in studies such as genome organisation, gene mapping, trait linkage and phylogenomics. To overcome these limitations, we recently presented a novel scaffold-to-chromosome anchoring technique that combines Reference-Assisted Chromosome Assembly (RACA) and fluorescence in situ hybridisation (FISH) to map the scaffolds of de novo sequenced avian genomes (Damas et al. 2016). To test this technique in non-avian species, we present preliminary work derived from a set of universal FISH probes, developed using BAC clones isolated from evolutionarily conserved sequences from the cattle (*Bos taurus*) genome. Using the cow as the reference genome, a selection of BACs were labelled and tested on multiple mammalian species using FISH to refine our selection criteria, with the ultimate goal of mapping scaffolds using these probes. Successful hybridisations were observed on all species tested, including human (*Homo sapiens*), pig (*Sus scrofa*), horse (*Equus ferus*), red lechwe (*Kobus leche*). The results generated provide preliminary evidence that our combined FISH and bioinformatics approach, as previously developed for avian species can also be applied to the mapping of genome scaffold assemblies in other orders, allowing comparative genomic research at a higher resolution than previously possible.

(O3/P65) Expansion of gene families and signatures of selection in the Australian marsupials

Will Nash[1], Wilfried Haerty[1]

[1] Earlham Institute, Norwich

The marsupials are thought to have diverged from other mammalian taxa around 160 million years ago. Since this time, Australian marsupials have undergone a unique radiation, lacking large predators, and needing to adapt to extreme dryness and specific diets. Gene family expansion and contraction has long been shown to be a unique process often associated with evolutionary innovations and ecological adaptation. The recent availability of a high-quality genome for the koala (*Phascorlarctos cinereus*), in addition to those of three other marsupials provides the opportunity to assess gene family dynamics within this unique lineage. In the koala, we recover signatures of expansion in over 1,000 gene families. The largest expansion was found within the CYP2C subfamily, representing two independent expansions in koala with 36 novel Cyp450 gene duplications. This is of interest as such proteins play essential roles in the metabolism of toxic compounds found in abundance in the koala's eucalyptus diet. An analysis of the conserved synteny of these genes also allowed their chromosomal placement, and showed the expansion within the marsupials to be of independent origin to a similar independent expansion found in rodents. Additionally, we analysed 1:1 orthologs across a tree of 9 species, for signatures of positive selection. In the koala, these genes enriched for GO terms associated with growth, muscular migration, sexual reproduction, and various responses to stress. Within Australidelphia we found a range of pathways to be enriched for positive selection, of particular interest the thyroid hormone synthesis pathway, important to marsupial development.

(O4) The tuatara genome project— Unlocking the genome of a living fossil

Neil Gemmell[1]

[1] Department of Anatomy, University of Otago, Dunedin 9054, New Zealand

The tuatara (*Sphenodon punctatus*) is an iconic and enigmatic terrestrial vertebrate, unique to New Zealand. Once widespread across the supercontinent of Gondwana, the tuatara, the only living member of an archaic reptilian order Rhynchocephalia (Sphenodontia) that last shared a common ancestor with other reptiles some 220-250 million years ago, is now only found on a small number of offshore Islands distributed around the coast of New Zealand. Through the efforts of a large international consortium, we have now sequenced, annotated, and analysed the 4.6-Gbp tuatara genome. In this presentation, I will highlight some of the challenges associated with sequencing this genome and the novel insights spanning genome architecture, sex determination, immunity, and homoeostasis that emerge from the genome of this important linchpin in vertebrate evolution. Last, the tuatara is a taonga, or special treasure, for Māori, and I will highlight the additional challenges, and rewards, of working in partnership with indigenous groups who have different cultural mindsets, albeit common goals.

Plant Genomics

(O5/P19) The evolution of photosynthetic efficiency

Steve Kelly[1]

[1] University of Oxford

Plant genes and genomes are composed of long strings of nucleotide monomers (A, C, G and T/U) that are built from metabolic precursors. The biosynthetic cost of each nucleotide differs in energetic and atomic requirements with different nucleotides requiring different quantities of energy and nitrogen atoms for their construction. Here I will discuss how natural variation in photosynthetic nitrogen use efficiency between plant species is a major determinant of genome wide GC content and genome wide patterns of codon bias. Specifically, plants that require more nitrogen per unit carbon fixation have correspondingly fewer nitrogen atoms in their genome and transcript sequences. Moreover, photosynthetic nitrogen use efficiency is the major determinant of genome-wide patterns of codon usage bias and genome-wide GC content in plants. I will also discuss what these findings mean in the context of land plant evolution past present and future.

(O6/P24) Designing multi-genome graphs for crop genomics and genetics: a wheat-centric view

Bernardo Clavijo[1]

[1] Earlham Institute

With a fairly complete genome assembly and annotation already published for Chinese Spring 42, and a set of anchored pseudomolecules from the IWGSC expected to be published soon, the wheat genome is now well characterised. Tools such as w2rap have made assembling the genome of wheat a reproducible analysis, and multiple genomes are already available for different wheat lines under pre-publication conditions. At the same time, new sequencing technologies provide longer-range data, such as linked reads (10x Genomics) and long reads (Nanopore, Pacbio). While these technologies can be directly applied to genome assembly, and are in part responsible for the advances in this field, they can also be used in novel ways to survey genetic variation. More importantly, in gene-rich regions of crop genomes, the scale gap between genetic resources and genomic long-range data is starting to narrow.

If these resources are integrated in informative analyses our understanding of genetics and genomics should enable a leap forward in breeding. But there are a number of limitations of current approaches to consider: most analyses are two-way (and reference-based) rather than multi-way, data integration is tackled too late, combining results rather than integrating analyses, and comparative analyses tend to be more exploratory and descriptive than quantitative.

We show here how appropriate design of multi-genome representations can enable better data integration, reducing the bias and increasing the power of comparative analyses. As an example, we apply this design to a preliminary comparison of 5 bread wheat lines, and show how new technologies provide valuable insights that can be better analysed within this framework.

(O7/P27) A novel species of human nasopharyngeal bacteria, distantly related to the avian pathogen *Ornithobacterium rhinotracheale*

Susannah Salter[1], Paul Scott[1], Paul Turner[2], Andrew Page[1], Julian Parkhill[1]

[1] Wellcome Trust Sanger Institute

[2] Cambodia-Oxford Medical Research Unit

Background: During an investigation into the nasopharyngeal microbiota of a cohort of children in Thailand, a large proportion of the data was found to belong to an unclassified taxon. It was present in 77% of samples, particularly those collected after 9 months of age. The standard 16S rRNA gene reference databases and public repositories could not shed any light on its identity, the closest match being the bird pathogen *Ornithobacterium rhinotracheale*. Subsequently we identified the same sequence in nasopharyngeal data from Australia, the Gambia, Kenya and Malawi. **Methods:** DNA extracted from nasopharyngeal swabs was randomly amplified and sequenced on the Miseq platform. Attempts were made to culture the bacterium, to visualise it in the original mixed sample using Gram staining and FISH, and to extract fixed cells using laser capture microdissection (LCM). **Results:** The novel genome was successfully assembled from metagenomic data. The bacterium is a member of the Flavobacteriaceae family, distantly related to *O. rhinotracheale*. With a POCP (percentage of conserved proteins) of 72% and a 16S rRNA gene identity of 93%, it is likely to represent a new genus.

(O8/P29) The population genetics of the ash dieback invasion of Europe highlights huge adaptive potential of the causal fungus, *Hymenoscyphus fraxineus*

Mark McMullan, Maryam Rafiqi, Gemy Kaithakottil, Bernardo Clavijo, Lorelei Bilham, Elizabeth Orton, Neil Hall, James K. M. Brown, David Swarbreck, Mark Blaxter, Allan Downie, Matthew D. Clark

A changing environment and accelerating international trade make pathogen spread an increasing concern. *Hymenoscyphus fraxineus* is the causal agent of ash dieback, a disease to which European common ash (*Fraxinus excelsior*) trees are highly susceptible (~5% partially resistant). The fungus invaded Europe around 20 years ago and, moving across continents and hosts from Asia to Europe. We have assembled and annotated a draft of the *H. fraxineus* genome which approaches chromosome scale. By resequencing 58 isolates of *H. fraxineus* from across its native (Asian) and invasive (European) ranges we find a tight bottleneck impacts pathogen genetic diversity across Europe but a signal of adaptive diversity remains in key host interaction genes (effectors). We use genetic diversity at Core Eukaryotic Genes to show that the European population was founded by two divergent haploid individuals and that divergence between these haplotypes represents the ‘shadow’ of a large source population. The signal of this European source population shows that it harboured as much genetic diversity as our native sample and that subsequent introduction would greatly increase adaptive potential and the pathogen’s threat.

(O9) A reference-free whole-genome alignment of hundreds of mammalian genomes

Joel Armstrong[1]

[1] UC Santa Cruz

The number of published vertebrate genomes is increasing rapidly. Each individual new assembly is useful to researchers studying that particular species, but when incorporated into a whole-genome alignment, the alignment can provide valuable information about the evolution of related genomes. This alignment can enable the transfer of information between related species. Our whole-genome alignment program, Cactus, has proven to be successful at aligning vertebrate-sized genomes without any reference bias and reconstructing ancestral genomes down to the base-level, providing insight into genome evolution. However, the whole-genome alignment problem becomes substantially more difficult to solve at the scale of hundreds of genomes, especially if they are of varying quality. To scale a reference-free whole-genome alignment to hundreds or thousands of genomes requires new alignment techniques, which we recently implemented. We present these improvements, in addition to preliminary results from an alignment of hundreds of new mammalian genomes as part of the 200 Mammals project. This is the most ambitious alignment project to date, both in terms of the number of genomes as well as total sequence size. We discuss the accuracy and completeness of our ancestral genome reconstructions and our goals for using the alignment to study mammalian evolution.

(O10) The 200 Mammals Genome Project: Understanding Evolutionary Conservation at Single Base Resolution

Elinor Karlsson[1,2], Jeremy Johnson[1], Diane Genereux[1], Jason Turner-Maier[1], Eva Muren[3], Voichita Marinescu[3], Joel Armstrong[4], Benedict Paten[4], Oliver Ryder[5], Harris Lewin[6], Kerstin Lindblad-Toh[1,3], Bruce Birren[1]

[1] Broad Institute

[2] U Mass Med

[3] Uppsala University

[4] University of California, Santa Cruz

[5] San Diego Zoo Institute of Conservation Research

[6] University of California, Davis

Evolutionary constraint is among the most powerful markers of genome function, critical for identifying functionally important variation. The 200 Mammals Project aims to achieve single base pair resolution of conservation across eutherian mammals. Using the cost-effective DISCOVAR de novo approach, we have sequenced 137 new genomes, yielding contig-level assemblies with high per-base accuracy well suited for detecting sequence conservation. In addition, we aim to scaffold, using Dovetail, at least one species from each of the eutherian orders. To date, 7 species from four orders are completed, with an average scaffold N50 of 19.5 Mb. Species were selected for inclusion based both on evolutionary branch length and community interest, with priority given to species with ancestral or unusual mammalian traits. For each genome, we generated one Illumina sequencing library and one lane of sequencing (2x250bp reads), and then assembled each genome using DISCOVAR de novo software. Critically, this process requires just 1 ug of low molecular weight DNA, allowing us to include species with small body sizes, like the bumblebee bat, and critically endangered species such as the northern white rhinoceros, Stephen’s kangaroo rat, and the solenodon. Our initial analysis of a 26-way alignment of ungulates and carnivores, using the Cactus7 reference-independent aligner, shows we can identify conserved sites and regions of accelerated evolution, and resolve functional variants. The full 200 mammal alignment will throw new light onto mammalian evolution, and how genomic variation impacts health.

(O11) Whole genome duplication and the evolution of salmonid fish: the state-of the art

Daniel Macqueen[1]

[1] University of Aberdeen

The common ancestor to all salmonid fishes underwent a spontaneous whole genome duplication (WGD) event (hereafter: ‘Ss4R’) around 88-103 Ma, remodelling constraints shaping genome structure and functional evolution. My talk will summarize our current understanding of Ss4R’s role in salmonid evolution, along with its relevance for other vertebrate WGDs. I will highlight the mechanistic and functional importance of rediploidization (cytological transition from a tetraploid to diploid genome), a process that was delayed after Ss4R and has yet to be fully resolved even in extant species. Consequently, many speciation events occurred before rediploidization was completed in large genomic regions, allowing thousands of duplicated genes to diverge independently in different salmonid lineages, including at the functional level. This model of evolution has been coined ‘LORe’ by my group (Robertson et al. Genome Biol. In press) and has general relevance for our understanding of the role played by duplicated genes in evolution and adaptation. My talk will frame the evolutionary significance of Ss4R in light of salmonid diversification patterns, as an example of the so-called ‘radiation-time lag’ model: I will argue here that WGD and LORe potentially promoted adaptations allowing an iconic life-history strategy to evolve, involving migration between freshwater and the marine environment. I will also describe plans for a new international research initiative called 'Functional Annotation of all Salmonid Genomes' (or 'FAASG'), which include comparative evolutionary analyses of several new salmonid genomes that are currently being sequenced and annotated.

(O12/P68) The evolution of social chromosomes in fire ants

Yannick Wurm[1], Rodrigo Pracana[1], Eckart Stolle[1], John Wang[2], Oksana Riba-Grognuz[3], Mingkwan Nippittwattanaphon[4], Dewayne Shoemaker[5], Laurent Keller[6]

[1] Queen Mary University of London

[2] Academia Sinica

[3] Université de Lausanne

[4] Kasetsart University

[5] USDA

[6] University of Lausanne

Variation in social behaviour is common, yet little is known about the genetic architectures underpinning its evolution. A rare exception is in the invasive red fire ant *Solenopsis invicta*. We recently demonstrated that alternative variants of a supergene region determine whether a colony will have exactly one or up to dozens of queens. The two variants of this region are carried by a pair of 'social chromosomes', SB and Sb, which resemble a pair of sex chromosomes. Recombination is suppressed between the two chromosomes in the supergene region. While the X-like SB can recombine with itself in SB/SB queens, recombination is effectively absent in the Y-like Sb because Sb/Sb queens die before reproducing. We use population genomic, phylogenomic and long-molecule sequencing approaches to understand the evolutionary history and the contrasting evolutionary forces affecting social chromosome variants. We identify large chromosomal rearrangements that are likely responsible for suppressed recombination in the social chromosome. Additionally, we show that the Sb haplotype of the supergene region has >600-fold less nucleotide diversity than the rest of the genome, indicating that a recent selective sweep has specifically affected Sb. Our findings are consistent with theoretical predictions regarding the importance of supergenes in evolution, suggesting that selection for reduced recombination between particular favorable allelic combinations may underlie additional phenotypic changes including other major social transitions.

Clinical and Translational Genomics

(O13) Nucleosome positioning as a cell memory in cancer transitions

Vladimir Teif[1]

[1] University of Essex

Nucleosome positioning is recognised as an important regulator of gene expression in normal and diseased cells. It is determined by several processes including the DNA sequence-dependent histone affinity landscape, active ATP-dependent chromatin remodelling, competition with transcription factors, chemical modifications of DNA and histones, and statistical positioning near genomic boundaries. Here I will provide an overview of our projects where nucleosome positioning is being investigated from the point of view of the analysis of cancer onset and progression. In the project conducted by the CancerEpiSys consortium we focused on nucleosome repositioning in B cells from patients with chronic lymphocytic leukaemia (CLL). We have shown that about 1% of nucleosomes reproducibly change their positions in cancer patients versus healthy individuals. A particularly important class of nucleosomes gained at promoters and enhancers marks the B-cell receptor signalling pathway specific for this cancer. Importantly, nucleosome positioning changes allow predicting cancer predisposition which is not yet evident from the changes of gene expression. In another project supported by the Wellcome Trust we are applying a similar concept to nucleosome repositioning in unrelated solid cancers, using paired cancer/normal tissue samples from the patients with glioblastoma and breast cancer. In this talk I will explain our current understanding of the role of nucleosome positioning as a cell memory in cancer transitions.

(O14) Good or bad sequencing data? Setting a benchmark for the quality of diagnostic NGS in the lab

Weronika Gutowska-Ding[1], JooWook Ahn[2], Kim Brugger[3], Jonathan Coxhead[4], Kate Thompson[5], Chris Boustred[6], Stephen Abbs[7], Erika Souche[8], Paul Westwood[9], Bauke Ylstra[10], Sander Stegmann[11], Graham Taylor[12], Farrah Khawaja[13], Zandra Deans[13], Simon Patton[1]

[1] EMQN

[2] Guy's and St Thomas' NHS Foundation Trust

[3] University of Cambridge

[4] Newcastle Biomedical Research Centre

[5] Oxford Medical Genetics Laboratories

[6] Queen Mary University of London

[7] Cambridge University Hospitals NHS Foundation Trust

[8] UZ Leuven

[9] NHS Greater Glasgow & Clyde

[10] VU University Medical Center Amsterdam

[11] MUMC Clinical Genetics Maastricht

[12] King's Collage London

[13] UK NEQAS

Next Generation Sequencing (NGS) is increasingly introduced into clinical genetics laboratories. The huge amount of data generated by NGS cannot be duplicated by alternative methods for laboratories to internally validate all results, therefore external quality assessment (EQA) of generated data is required. The European Molecular Genetics Quality Network (EMQN) and the UK National External Quality Assessment Service (UKNEQAS) for Molecular Genetics have developed a joint EQA scheme for NGS, with the aims to: (a) assess and improve quality; (b) enable laboratories to benchmark their NGS service against others and against best practice; (c) work towards consistency of reporting clinical results generated by NGS; and (d) contribute towards defining best practice. The objectives for developing an NGS EQA were to make it generic (testing context and technology independent) and to provide users with a broad range of quality indicators on their NGS data. This task has required the development of sophisticated tools for the integration and benchmarking of NGS data. So far, four pilot schemes have been run, with the two latest EQAs divided into Germline and Somatic schemes in order to address their different NGS challenges. The number of participant laboratories has grown from 30 in 2013, to 303 in 2016. The results obtained enable clinical diagnostic labs to start addressing the quality of their NGS testing. These technology-specific NGS EQAs will play an important role in enabling labs to benchmark this new technology, assess the accuracy of data and facilitate high quality reporting for patient benefit.

Agricultural Genomics

(O15/P38) Downregulation of immune genes in quail in response to H5N1 infection

Katrina Morris[1], Jacqueline Smith[1], Angela Danner[2], Robert Webster[2], David Burt[1]

[1] Roslin Institute, The University of Edinburgh

[2] St. Jude Children's Research Hospital

Highly pathogenic influenza A viruses (HPAI), such as H5N1, are responsible for enormous economic losses in the poultry industry and pose a serious threat to public health. Quail is a popular domestic poultry species raised for meat and eggs in Asia and Europe. While quails can survive infection with Low Pathogenic influenza viruses (LPAI), they experience high mortality when infected with strains of HPAI. Quails may play a key role as an intermediate host in evolution of avian influenza. While aquatic reservoir species such as duck are resistant to most HPAI strains and act as natural reservoirs, quails and chickens are highly susceptible. To better understand the effect this disease has on quails we performed differential analysis of gene expression in quails infected with LPAI an HPAI. We compare this to previous findings in ducks and chickens. We found that quails have a robust immune response to infection by LPAI, while they show dysregulation of the immune response after infection with HPAI, and this may explain their susceptibility to this disease. Genes associated with apoptosis were downregulated and quails did not show strong upregulation of IFITM proteins, which are thought to be key to HPAI tolerance in ducks. Many antiviral and innate immune genes, including those involved in antigen recognition, immune system activation, and anti-viral responses were downregulated in lung. This study provides crucial data that can be used to understand the differing response of bird species to avian influenza, which will be critical for managing and mitigating these diseases.

(O16) Pan-genome assembly of population haplotypes provides a comprehensive solution to common obstacles in modern breeding

Gil Ronen[1]

[1] NRGene

Much of the World's food production, in both livestock and crops, relies on modern breeding programmes which increasingly apply advanced genomic tools to achieve genetic gains, cost reduction, and development of new varieties and breeds. The increased usage of "genomic selection" and other approaches utilizing genetic markers is an important part of this change. However, often the use of genetic markers is limited due to the lack of high quality genome assemblies and genetic maps, or reliance on a non-ideal reference genome sequence.

NRGene's technology offers a comprehensive solution to analyzing a wide variety of food-producing organisms and detecting the most useful set of genomic markers to use.

NRGene's proprietary de-novo assembly software (DeNovoMAGIC) uses a unique pan-genome creation technology which creates a haplotype database that captures the full scope of the genomic diversity within a population. The haplotype database consists of very dense dominant sub-sequences distinctly related to a haplotype at any given position, which are used instead of SNP markers. On this basis, any low coverage genotype method (e.g. GBS or SNP array) can be used to impute back the haplotype of any given sample. The haplotype database is a "live system" that can be updated to account for new introduced germplasm in a time and cost-effective manner.

The above approach and its application to realistic breeding scenarios will be demonstrated in maize while detailing the cost benefit of the approach for different breeding entities.

Conservation Genomics

(O17) Conservation genomics of the pink pigeon

Camilla Ryan, Lawrence Percival-Alwyn, Mohammed Albeshr, Kevin Tyler, Ian Barnes, Carl Jones, Diana Bell, Cock van Oosterhout, **Matthew D. Clark**

Mauritius is a beautiful island in the Indian Ocean, but it's also infamous for humans driving the Dodo (a large flightless pigeon) to extinction. Other Mauritian species are still threatened, including the pink pigeon whose population collapsed to <20 in the 1970s, and again to just 9 in the 1990s. Due to the efforts of the Mauritius Wildlife Trust the population has now recovered to ~400, but it's still endangered. As well as predators and habitat loss, the pink pigeon suffers from inbreeding depression and susceptibility to pathogens. We are using some of the latest sequencing technologies to examine changes in pink pigeon's genetic diversity by comparing historic DNA with modern samples. We currently have a high quality draft genome, RAD-seq genotyping data for about half the world population and whole genome resequencing data for 8 historic samples. We intend to use this information to suggest captive breeding and reintroduction strategies that would increase the pink pigeon's genetic diversity. This strategy could contribute to the long-term survival of pink pigeons by reversing the negative impacts of inbreeding depression and increasing their resistance to pathogens like *Trichomonas gallinae*.

(O18/P84) Novel genome assembly approach contributes to natural history and conservation of the Hispaniolan solenodon, *Solenodon paradoxus*

Kirill Grigorev, Sergey Kliver, Pavel Dobrynin, Alexey Komissarov, Walter Wolfsberger, Ksenia Krashennikova, Audrey J. Majeske, Agostinho Antunes, Alfred L. Roca, Stephen J. O'Brien, Juan Carlos Martinez-Cruzado, **Taras K. Oleksyk**

Solenodons are insectivores living on the Caribbean islands, with few surviving related taxa. The genus occupies one of the most ancient branches among the placental mammals. The history, unique biology and adaptations of these enigmatic venomous species, can be greatly advanced given the availability of genome data, but the whole genome assembly for solenodons has never been previously performed, partially due to the difficulty in obtaining samples from the field. Because island isolation likely resulted in extreme homozygosity in *S. paradoxus* genomes, and tested several assembly strategies for performance with genetically impoverished species' genomes For Hispaniolan solenodon, *Solenodon paradoxus*, the string-graph based assembly strategy seems a better choice for the homozygous genomes, which is often a hallmark of endemic or endangered species. A consensus reference genome was assembled, annotated for genes, repeats, variable microsatellite loci and other genomic variants, sequencing 5 individuals from the southern subspecies (*S. p. woodi*) and one sequence of the northern subspecies (*S. p. paradoxus*). Genomic features acr were characterized and annotated, with a specific emphasis on the venomous genes. Phylogenetic positioning and selection signatures were inferred based on 4,416 single copy orthologs from 11 other mammals. Patterns of SNP variation allowed us to infer demography, which indicated a subspecies split of Hispaniolan solenodon, *Solenodon paradoxus* at least 100 Kya.

(O19/P83) Characterisation of koala lactation genes using a combined transcriptomic, proteomic and genomic approach

Katrina Morris[1], Denis O'Meally[2], Xiaomin Song[3], Mark Molloy[3], Adam Polkinghorne[4], Katherine Belov[2]

[1] Roslin Institute, The University of Edinburgh

[2] The University of Sydney

[3] Macquarie University

[4] University of the Sunshine Coast

Marsupials are distinct from placental mammals in that offspring are born undeveloped and immunologically naïve, and compounds in the milk are critical for their immune protection. Additionally, due to their extended lactation period, the composition of milk varies dramatically throughout lactation in marsupials. Koalas (*Phascolarctos cinereus*) are an iconic Australian species that are increasingly threatened by disease. We used a mammary transcriptome, two milk proteomes and the koala genome to comprehensively characterise the protein components of koala milk, and investigated lactation gene families in the koala genome, with a focus on immune constituents. The most abundant proteins included several lipocalins, including β-lactoglobulin. We discovered that milk-expressed lipocalins, many of which likely serve nutritional roles, are found in a highly duplicated cluster of genes in the koala genome; their diversity perhaps linked to the long lactation period of marsupials. In the mammary transcriptome, 851 immune transcripts were expressed, including immunoglobulins and complement components. We identified many abundant antimicrobial peptides in koala milk. Additionally, we identified two novel marsupial-milk-specific proteins (VELP and MM1), which are closely linked in the koala genome, and based on homology and synteny with human antimicrobial genes, are likely to have antimicrobial functions. We also identified highly-abundant koala endogenous-retrovirus sequences, identifying a potential transmission route from mother to young. Characterising the immune components of milk is key to understanding protection of marsupial young, and the novel immune compounds identified may have applications in clinical research.

(O20) Genomic approaches to identification and preservation of wild tilapia species and unique genetic resources

Antonia Ford[1], Martin Genner[2], Federica Di Palma[3], Wilfried Haerty[3], Asilatu Schechonge[4], Benjamin Ngatunga[4], George Turner1

[1] Bangor University

[2] University of Bristol

[3] Earlham Institute

[4] Tanzania Fisheries Research Institute

Introduction: Tilapia cichlid fishes, particularly the genus Oreochromis, are a mainstay of tropical aquaculture. Future strain enhancement may benefit from availability of wild genetic resources, which have previously been used to enhance growth and environmental tolerance, control sex ratios, and introduce resistance to disease. Across East Africa native tilapia species are frequently threatened by invasive species, which have been introduced for aquaculture. A major threat stems from the propensity of tilapia species to hybridize, leading to significant loss of diversity. Here, we aim to characterize wild populations of Oreochromis across Tanzania to identify untouched populations for conservation priority, study evolutionary history of native species, and investigate the signature of ancient and recent hybridization events on the genome. **Methods and Results:** Alongside parallel work to investigate environmental adaptation (using de novo genome assembly), we employ population-level low coverage whole-genome sequencing to investigate patterns of admixture and selection in Tanzanian tilapia species. We focus on the interaction of exotic species *O. niloticus* and *O. leucostictus* (introduced into several regions from Lake Victoria) with native species including *O. urolepis* and *O. shiranus*. We find evidence of hybridization in aquaculture stocks, and in several natural water bodies, with hybrids persisting in the wild including F2 and backcrossed individuals. **Conclusions:** The spread of introduced tilapia species poses a threat to native species via ecological competition and hybridization. Several native species are endemic, and many exhibit traits of interest to aquaculture, so preserving native biodiversity is paramount from perspectives of conservation and genetic resource availability.

Developmental Biology

(O21) Enhancers and the convergent evolution of limb reduction in squamates

Carlos R. Infante[1]

[1] The University of Arizona

Among tetrapods, the lineage with the greatest number of independent reductions in limb length is the squamates (lizards and snakes). The independent evolution of a snake-like body form from a limbed ancestor has occurred at least 26 times within the group, the most notable instance being the snakes. Although limb loss has played a prominent role in squamate diversification, little is known about the genetic and developmental mechanisms that contributed to the evolution of this striking phenotype. For example, it is unknown whether the same genetic mechanisms were used repeatedly or if many different pathways have been involved in limb loss. To answer these questions, I focus on the evolution of cis-regulatory elements (enhancers) that control gene expression in the developing limbs using a combination of functional and comparative genomics. Using chromatin-immunoprecipitation followed by high-throughput DNA sequencing (ChIP-seq), I have identified thousands of active regulatory elements in the limbs and genital appendages of the lizard Anolis carolinensis. Using this information, I have compared the conservation of limb enhancers in the genomes of limbless squamates. These analyses reveal a striking conservation of limb regulatory elements in species that have lost limbs, possibly because of pleiotropic effects in other tissues. Future work will determine whether there is functional conservation of these regulatory elements across the multiple limbless lineages.

(O22/P40) Using long reads to understand small RNAs

Dominik Handler[1]

[1] IMBA - Institute of Molecular Biotechnology GmbH

Preserving genome integrity is of core importance for an organism's fitness. Especially important is the genome of reproductive cells. Selfish genetic elements such as transposable elements are particularly active in these cells in order to multiply in a population. Various defense systems against transposable elements have evolved in eukaryotes. Prominent among these are small RNA pathways that act as programmable silencing systems. The central small RNA pathway is the piRNA pathway, which is conserved from sponges to human. Similar to the microRNA pathway, a small RNA molecule bound to an effector protein acts at the heart of the pathway. Unlike miRNAs, piRNAs are processed from long single-stranded precursors transcripts. These originate from genomic regions called piRNA clusters that are densely populated by diverse transposon remnants. Some clusters are believed to be transcribed as single transcription units of up to 300-400kb in length. Understanding the biology of piRNA clusters is hampered by their repetitive sequence nature. Indeed, piRNA clusters are not properly assembled in genome sequencing projects. Long read sequencing technologies opened up new possibilities to study piRNA cluster biology. We set out to de-novo assemble the genome of a Drosophila melanogaster cell line that is widely used. Our results indicate that piRNA clusters can indeed be assembled from Nanopore reads without any prior genome sequence information. We are also planning to use the MinIon to sequence the long cluster transcripts emanating from clusters in order to learn how these long precursors RNAs are transcribed and how they get post-transcriptionally processed.

Microbial Communities

(O23) DESMAN: a new tool for De novo Extraction of Strains from MetAgeNomes

Christopher Quince[1]

[1] University of Warwick

Background We introduce DESMAN for De novo Extraction of Strains from MetAgeNomes. Metagenome sequencing generates short reads from throughout the genomes of a microbial community. Increasingly large, multi-sample metagenomes, stratified in space and time are being generated from communities with thousands of species. Repeats result in fragmentary co-assemblies with potentially millions of contigs. Contigs can be binned into metagenome assembled genomes (MAGs) but strain level variation will remain. DESMAN identifies variants on core genes, then uses co-occurrence across samples to link variants into strain sequences and abundance profiles. These strain profiles are then searched for on non-core genes to determine the accessory genes present in each strain.

Results We validated DESMAN on a synthetic twenty genome community with 64 samples. We could resolve the five *E. coli* strains present with 99.58% accuracy across core gene variable sites and their gene complement with 95.7% accuracy. Similarly, on real fecal metagenomes from the 2011 *E. coli* (STEC) O104:H4 outbreak, the outbreak strain was reconstructed with 99.8% core sequence accuracy. To mimic environmental applications we then tested DESMAN on a more complex 210 genome mock with 50 multi-strain species across 96 samples. Of the 25 multi-strain species that were binned DESMAN resolved 34 of their 67 strains exactly and 53 were within 5 SNPs of their closest matching reference, with just three false positive strains. Application to 32 MAGs from the TARA Oceans microbiome revealed that strain variation is endemic with (29/32 = 90.6%) of MAGs exhibiting strain variation. In many cases (57.5%) these strains were significantly correlated with geographic region. There was also a negative correlation between genome length and number of haplotypes in a MAG and evidence for more rapid change in gene complement in small genome organisms.

Conclusions DESMAN will provide a provide a powerful tool for *de novo* resolution of fine-scale variation in microbial communities. It is available as open source software from https://github.com/chrisquince/DESMAN.

(O24) Hansel and Gretel: A fairy tale of recovering haplotypes from metagenomes with a happy ending

Sam Nicholls[1]

[1] Department of Computer Science, Aberystwyth University

The diversity of microbial communities represent an untapped biotechnological resource for biomining, biorefining and synthetic biology. Revealing this information requires the recovery of the exact sequence of DNA bases (or "haplotype") that constitute functional isoforms of the genes on every individual present. This is a computationally difficult problem without a current solution, complicated by the requirement for environmental sequencing approaches (metagenomics).

Haplotypes are identified by their unique combination of DNA variants. However, standard approaches for working with metagenomic data require simplifications that violate assumptions in the process of identifying such variation. Furthermore, current haplotyping methods lack objective mechanisms for choosing between alternative haplotype reconstructions from microbial communities.

To address this, we have developed a novel probabilistic method for reconstructing haplotypes from complex microbial communities and propose the "metahaplome" as a definition for the set of haplotypes for any particular genomic region of interest within a metagenomic dataset.

Implemented in the twin software tools Hansel and Gretel, the algorithm performs incremental probabilistic haplotype recovery using Naive Bayes, from raw reads aligned to a pseudo-reference (such as a metagenomic assembly).

Our method is capable of reconstructing and ranking the haplotypes with the maximum likelihoods from metagenomic datasets without a priori knowledge or making assumptions of the distribution or number of variants. Additionally, the algorithm is robust to sequencing and alignment error and requires no altering or discarding observed variation, using all available evidence from the reads.

We validate our method using synthetic metahaplomes constructed from sets of real genes, and demonstrate its capability using metagenomic data from a complex HIV-1 strain mix. The results show that the likelihood framework can recover cryptic functional isoforms of genes with 100% accuracy, from microbial communities.

(O25) Linked-Reads enable efficient *de novo*, diploid assembly

Deanna M. Church[1], Stephen Williams[1], Claudia Catalanotti[1], Nikka Keivanfar[1], Jill Herschleb[1], Vijay Kuman[1], Preyas Shah[1], Neil Weisenfeld[1], Michael Schnall-Levin[1], David Jaffe[1]

[1] 10x Genomics

The determination of a reference sequence for the human genome fundamentally changed the way we approach studying human health and development. An important lesson from the past decade of research is that generating a single haploid consensus assembly for diploid organisms can lead to assembly errors and limited representation of biologically important sequences. Reconstruction of accurate, individual haplotypes provides a more complete picture of a genome. However, haplotype reconstruction of diploid genomes using cost effective, accurate short reads remains challenging. We describe a novel approach for the de novo assembly of individual mammalian genomes, requiring only small amounts (0.5-1.25 ng) of input DNA to construct a single library. We have developed a high-throughput microfluidic system for partitioning high-molecular weight DNA. Unique barcodes are applied within each partition, allowing for the retention of long-range information using short read sequencing, creating a data type called Linked-Reads. The Supernova™ Assembler takes advantage of Linked-Reads to perform de novo diploid assembly. Heterozygosity within the sample, coupled with molecular barcodes, allows for the separation of scaffolds into their distinct haplotypes, referred to as phase-blocks. We demonstrate the performance of this process on seven human genomes of diverse ethnic origin and validate the accuracy of the phase information using orthogonal data. We also show performance on diverse non-human genomes including hummingbird, dog and olive fly. Recent updates in the Supernova algorithms allow for assembling small genomes (<1Mb) using Linked-Reads. To demonstrate these new features, we constructed a diploid assembly for the ~500Mb Flame Grape genome.

(O26/P1) Novel approach to chromosome-level mapping of avian genomes doubles the number of assemblies

Rebecca O'Connor[1], Joana Damas[2], Marta Farré[2], Henry Martell[1], Lucas Kiazim[1], Rebecca Jennings[1], Anjali Mandawala[3], Sunitha Joseph[1], Katie Fowler[3], Eden Slack[2], Emily Allanson[2], Denis M. Larkin[2], Darren K. Griffin[1]

[1] School of Biosciences, University of Kent, Canterbury

[2] Department of Comparative Biomedical Sciences, Royal Veterinary College, University of London, London

[3] Canterbury Christchurch University, Canterbury, Kent

The ultimate aim of a genome assembly is to create a contiguous length of sequence from the p- to q-terminus of each chromosome. Most assemblies are however, highly fragmented, limiting their use in investigations into genomic organisation and mapping, trait linkage and phylogenomics. In order to overcome these limitations, we developed a novel scaffold-to-chromosome anchoring method combining reference-assisted chromosome assembly (RACA) and fluorescence in situ hybridisation (FISH) to position scaffolds from de novo assemblies with N50 > 1-2Mb on chromosomes. Using RACA, scaffolds were ordered and orientated into 'predicted chromosome fragments' (PCFs) against a reference and outgroup genome. PCFs were verified using PCR prior to mapping with FISH. A universal set of FISH probes developed through the selection of conserved regions were used to map PCFs of the peregrine falcon (*Falco peregrinus*) and the pigeon (*Columba livia*) genomes, improving the N50 of both seven-fold to 87% and 84% of the genome respectively, as well as identifying intra and interchromosomal rearrangements. Here we report the mapping of three additional genomes mapped using this method: ostrich (*Struthio camelus*), saker falcon (*Falco cherrug*) and budgerigar (*Melopsittacus undulatus*), illustrating the universal application of this method to avian genomes and doubling the number of chromosomally mapped genomes. Our combined Zoo-FISH and bioinformatics approach permits comparative genomic research at a higher resolution than previously described and opens up new avenues of investigation into genome karyotype evolution and the role of chromosome rearrangements in adaptation and phenotypic diversity in birds.

(O27) Scaling up the generation of reference quality genomes across a range of vertebrate diversity

Iliana Bista[1], Milan Malinsky[2], Michelle Smith[1], Dirk - Dominik Dolle[1], Karen Oliver[1], Marcus Klarqvist[1], Hannes Svardal[1], Shane McCarthy[1], Kerstin Howe[1], Eric Miska[3], Richard Durbin[1]

[1] Wellcome Trust Sanger Institute

[2] Zoological Institute, University of Basel, Switzerland

[3] Gurdon Institute, University of Cambridge

At the Wellcome Trust Sanger Institute we are working on scaling up sequencing and assembly of vertebrates at reference quality to support the Vertebrate Genomes Project (VGP) in association with the Genome 10K Project. We are targeting 50-100 species from three main groups of vertebrates, including several fish groups, the caecilian amphibians, and various rodent species. Currently the main focus is on sequencing fish groups, specifically: members of the Nototheniodinidae (Antarctic fish), members of the Cichlidae family (Haplochromine radiation), strains of zebrafish (*Danio rerio*) and related Cyprinidae, and species of the anabantoid group (gourami). Furthermore, we are evaluating a range of sequencing technologies, including PacBio, Oxford Nanopore, 10X Genomics, BioNano and Illumina for generating reference genome quality data. Our ultimate aim is to achieve assemblies with >1Mb contig N50, >10Mb scaffold N50 and >90% DNA assignment to chromosomes, while exploring novel contig scaffolding approaches, e.g. linkage disequilibrium from population variation data to order and orient contigs. Through collaboration with the European Bioinformatics Institute (EBI), data will be deposited in relevant archives with sufficient gene annotation (Ensembl). We will present preliminary results from a cichlid (*Astatotilapia calliptera*), strains of zebrafish (*Danio rerio*), and the grasshopper mouse (*Onychomys torridus*). Within 2017 we aim to provide genome data for additional species, including *Gouania willdenowi* (Blunt-snouted clingfish), *Erpetoichthys calabaricus* (reedfish), *Mastacembelus armatus* (tire track eel), and *Acomys russatus* (golden spiny mouse). This initiative will provide a valuable resource of genome data to the community, useful for in depth investigations of evolutionary relationships of vertebrates.

(O28/P2) Comparative Annotation Toolkit (CAT) - simultaneous annotation of related genomes using a high quality reference

Ian Fiddes[1],

[1] UC Santa Cruz

The recent introductions of low-cost, long-read and read-set sequencing technologies coupled with intense efforts to develop efficient algorithms have made high-quality de novo sequence assembly a realistic proposition. The result is an explosion of new, ultra contiguous genome assemblies. To compare these genomes we need robust methods for genome annotation. We describe the Comparative Annotation Toolkit (CAT), which provides a flexible way to leverage annotations in one species combined with whole genome alignments to simultaneously annotate entire clades, providing orthologous gene information. CAT also performs ab-initio gene prediction, allowing detection of gene family expansion and contraction. When given full length cDNA sequencing data, CAT can also predict novel isoforms. CAT can produce high quality annotation sets at a wide ranging of phylogenetic distances, from mouse-rat to human-human. We show that CAT can be used to improve annotations on the rat genome, annotate the primate clade, and annotate personal human genomes, discovering novel structural variants and greatly improving the ability to perform cross-species RNA expression experiments.

(O29) High Throughput Genomics Enabled by NEBNext Ultra II FS

Lesley Shirley[1]

[1] Wellcome Trust Sanger Institute

The DNA Pipelines team at The Wellcome Trust Sanger Institute (WTSI) process up to 16 000 DNA samples each month for whole genome (WGS) or targeted sequencing on the Illumina platform. These processing lines are underpinned by fully automated workflows that are designed to produce high quality DNA libraries from a wide range of sources. However, we have recently met strong demand for high quality WGS data from laser captured microdissection (LCM) biopsy material in which only a few hundred cells (~1 ng) are captured. Test data indicated that our existing LC processes could not produce high quality DNA libraries from such small amounts of available DNA. We therefore developed an entirely new LC workflow, which is underpinned by the NEBNext Ultra II Fragmentation System (FS) and an overview of this work will be presented. The benefits of the NEBNext Ultra II FS also led us to explore its widespread use for DNA library construction, in particular for PCR-free DNA library generation.

(O30/P15) Genome-wide characterization of RNA processing event dependencies

Colin Dewey[1], Matthew Amodio[1]

[1] University of Wisconsin-Madison

The current paradigm for annotating genes in eukaryotic genomes involves listing, for each loci, a set of full-length transcript structures. Each such set represents our current knowledge regarding the possible RNA processing events (e.g., transcription start sites, end sites, and splice events) associated with a gene, as well as the co-occurrence of these events within individual transcripts. This paradigm is implicitly predicated on the assumption that there exist strong dependencies or associations between processing events undergone by individual transcripts, because otherwise we would need only to catalog each gene's possible RNA processing events. However, this assumption has not been rigorously tested, in part because of a dearth of long-read RNA sequencing data, which are necessary for identifying or pairs or larger subsets of RNA processing events that co-occur along a single transcript. Here we present our ongoing work to test the extent to which the assumption of strong RNA processing event dependencies holds, if at all. Using recently generated data from long-read RNA-seq technologies and new short-read protocols that link the 5' and 3' ends of single transcripts, we have tested for dependencies between the start and end sites of transcripts, as well as between pairs of splice events. For the cell types from which we have such data, our results indicate that genes exhibiting RNA processing event dependencies are in the minority, and that dependencies, when found, are generally weak. These results have important implications for the future of gene annotation.

(O31/P18) Sequence alignment using optical correlation

Daniel Mapleson[1]

[1] Earlham Institute

Sequence alignment is an integral part of many bioinformatics pipelines. The progressive increase in data generated by next generation sequence technologies has driven the development of faster alignment algorithms, yet process times and availability of computing resources are still a limiting factor. Furthermore, the costs of configuring, running and cooling high performance computing (HPC) systems is a significant financial and logistical consideration. Optical computing has long been heralded as a solution to the limitations of traditional silicon-based computing technology. Optalysys in partnership with the Earlham Institute (EI), is developing a revolutionary, patented, sequence alignment technology based on an established diffractive optical approach that uses low-power laser light in place of electricity as the processing medium. This approach is inherently parallel, allowing for increased processing capacity that scales according to component properties such as resolution and frame rates, and the number of optical co-processors coupled together. We predict that this will lead to improvements in processing time with up to 95% drop in energy usage in some scenarios. Here we show some initial results from our system, codenamed Genesys, indicating improved read placement accuracy compared with BWA in a genomic mapping scenario. We go on to discuss how GENESYS could be adapted to other alignment tasks, and how it can leverage the increased sensitivity delivered by the optical correlator to perform alignments of more distantly related sequences, making it suitable for applications where BLAST is currently used.

(O32) Chromosome assemblies with Oxford Nanopore sequencing

John Davey[1]

[1] University of York

It is now possible to assemble complete chromosomes of small to medium-sized genomes with Oxford Nanopore sequencing. Typical MinION runs are now producing many gigabases of sequence with hundreds of times coverage of small genomes in reads longer than 30 kb, and with raw assemblies at 99% accuracy. I will report on several assemblies in progress, particularly that of *Galdieria sulphuraria*, an extremophile red alga. This organism has over 50 chromosomes, although the precise number is not yet known. We have assembled over half the genome in complete chromosomes, complete with telomeres and complex 10kb subtelomeric repeats. The remaining half of the genome is more challenging, as it appears that pairs of chromosomes share common regions up to 100 kb long. These features of the genome, previously unknown even based on high quality traditional assembly methods, have only been observable with nanopore sequencing. I will discuss what can be achieved with current genome assembly tools, and where new tools may be required to take advantage of high coverage with very long reads.

(O33/P12) gEVAL, a web-based browser to help you evaluate and assess the state of your assembly

William Chow[1], Kerstin Howe[1], Richard Durbin[1]

[1] Wellcome Trust Sanger Institute

Genome sequencing methods and assembly strategies are continuously improving, yet there are still challenges in reaching reference level quality. Established length-based metrics can be used to quickly assess an assembly but can often be misleading, and the real assembly problems are often only detected after aligning additional data and observing discordance. Here we present gEVAL (http://geval.sanger.ac.uk), a web based genome browser that integrates multiple data types to provide a one stop shop for the evaluation and improvement of assemblies. gEVAL features a variety of datasets including genome/optical maps, clone end sequences and transcript sequence as well as whole genome alignments to other assemblies of the same species (including 19 human assemblies). Locating and assessing troublesome regions is aided by the browsers bespoke assembly issue lists and the colour coding of data tracks. This not only provides ease of navigation to regions of interests, but the ability to facilitate developing assembly improvement strategies for use in curation efforts. gEVAL has been used by many projects across many species including: · the Genome Reference Consortium's (GRC) genome curation of the Human, Mouse and Zebrafish references · the draft assemblies of 16 mice strains as part of the Mouse Genomes Project, · the Swine Sequencing Consortium and the International Chicken Genome Consortium for the improvement and release of their respective reference genomes. More recently, the gEVAL team has been involved in the Vertebrate Genome Project at the Sanger Institute, a project to sequence hundreds of fish, rodents and caecilians.

(O34) Full-length Transcript (Iso-Seq) Profiling for Improved Genome Annotations

Jonas Korlach[1], Olivier Fedrigo[2], Jacqueline Mountcastle[2], Ting Hon[1], Tyson A. Clark[1], Sarah B. Kingan[1], Erich D. Jarvis[2]

[1] Pacific Biosciences, Menlo Park, CA;

[2] The Rockefeller University, New York, NY

Incomplete annotation of genomes represents a major impediment to understanding biological processes, functional differences between species, and their evolutionary mechanisms. Often, genes that are large, embedded within duplicated genomic regions, or associated with repeats are difficult to study by short-read expression profiling and assembly. In addition, most genes in eukaryotic organisms produce alternatively spliced isoforms, broadening the diversity of proteins encoded by the genome, which are also difficult to resolve with short-read methods.

In contrast, long Single Molecule, Real-Time (SMRT) Sequencing reads are able to directly sequence full-length transcripts without the need of assembly and imputation. This includes the PacBio isoform sequencing (Iso-Seq) application which is capable of directly generating full-length sequences for transcripts up to 10 kb in length. Here we demonstrate the application of the Iso-Seq method on several animal species, its utility for providing a higher quality annotation of their corresponding genome, and providing insights into alternative splice isoforms, alternative promoters and polyadenylation sites, as well as non-coding RNA. With the improved full-length transcript and gene models, it is also possible to reassess short-read RNAseq datasets to quantify expression data more accurately. The full-length transcript data can thus be integrated into new reference genomes being assembled with long-read sequencing to provide a more complete understanding of the organism's biology, and of differences between phylogenetically related species.

(O35/P88) Natural selection shaped the rise and fall of passenger pigeon genomic diversity

Gemma Murray[1], André Soares[1], Beth Shapiro[1]

[1] University of California, Santa Cruz

The extinct passenger pigeon was once the most abundant bird in North America, numbering between 3 and 5 billion individuals prior to its 19th century decline. Passenger pigeons were highly mobile, they bred in large social colonies, and their population lacked clear geographic structure. This suggests that their effective population size (Ne) may have been exceptionally large. Genome sequences from passenger pigeons therefore provide a rare opportunity to explore the evolutionary consequences of a large Ne. To this end, we performed comparative analyses of nuclear and mitochondrial genomes from passenger pigeons and band-tailed pigeons, the closest living relatives of passenger pigeons. We found that while the passenger pigeon Ne was large and stable for thousands of years prior to their extinction, the species had surprisingly low genetic diversity. We found that while this large Ne allowed for both a higher rate of adaptive evolution and more efficient selective constraint in passenger pigeons, the highly variable recombination landscape of bird genomes combined with the impact of selection on linked sites reduced diversity across the passenger pigeon genome by more than 60%. Our results demonstrate the combined effect of very large Ne and a highly variable recombination landscape on genetic diversity and the efficacy of selection: large Ne increases diversity and the efficacy of selection, while linkage in low recombination regions reduces diversity and constrains selection. For passenger pigeons, this meant a greater capacity for rapid adaptation when their population was large, but potentially a lower resilience to a rapid population decline.

(O36/P87) Determinants of the efficacy of natural selection on coding and noncoding variability in two passerine species

Padraic Corcoran[1], Toni Gossmann[1], Henry Barton[1], Jon Slate[1], Kai Zeng[1]

[1] University of Sheffield

Understanding the relative importance of natural selection and genetic drift in determining patterns of molecular evolution is a long-standing goal in population genetics. Theory predicts that selection should be more effective when the effective population size (Ne) is larger, and that the efficacy of selection should correlate positively with recombination rate. Here, we analysed the genomes of 10 great tits from Europe and 10 zebra finches from Australia. Nucleotide diversity at 4-fold sites indicates that the zebra finch has a 2.83-fold larger Ne. The proportion of 0-fold substitutions fixed by positive selection (α) is high in both species (great tit 48%; zebra finch 64%) and is significantly higher in zebra finches. To control for the confounding effects of GC-biased gene conversion (gBGC), we estimated α using only changes between A and T nucleotides and G and C nucleotides, and found reduced α estimates in both species (great tit 22%; zebra finch 53%), in agreement with the predictions of a theoretical model we describe herein. We present the first estimates in birds for α in the untranslated regions (great tit 19%; zebra finch 42%), supporting a substantial role for adaptive changes. Finally, although purifying selection is stronger in high-recombination regions, we obtained mixed evidence for α increasing with recombination rate, especially after accounting for gBGC, consistent with predictions of the new model. These results highlight that controlling for gBGC is essential for accurately quantifying selection and that our understanding of what determines the efficacy of selection is incomplete.

(O37) Copy number variation in the Atlantic salmon (*Salmo salar*) genome

Alicia C Bertolotti[1], Torfinn Nome[2], Simen R. Sandve[2], Samuel A. M. Martin[1], Sigbjørn Lien[2], Daniel J Macqueen[1]

[1] Institute of Biological and Environmental Sciences, University of Aberdeen, Scotland

[2] Centre for Integrative Genetics (CIGENE), Norwegian University of Life Sciences, Norway

A high-quality reference genome was recently published for the Atlantic salmon (*Salmo salar*), a species of considerable economic, cultural and scientific importance. This resource has opened up new opportunities to understand the role of the genome in population-level variation. There is a heavy current focus on single nucleotide polymorphisms, and larger structural genomic variation has not yet been characterized. Copy number variations (CNVs) represent duplicated or deleted regions of chromosomal DNA (>1Kb) that vary in copy number among individuals. CNVs frequently overlap functional genes and have been widely implicated in phenotypic variation of agricultural, evolutionary and clinical relevance. We are characterising the CNV landscape of Atlantic salmon, focusing on populations that broadly span its natural distribution. Using an array of bioinformatics tools with 10-20x coverage whole-genome re-sequencing data, we have mapped high-confidence CNVs throughout the genome of >500 individuals. Analysis of this encompassing dataset is ongoing, but we already know that at least 2.3% of the genome is CNV, with demonstrable regional variation linked to a salmonid-specific whole genome duplication (WGD) event that occurred 88-103 Ma. On average, 1,198 CNVs were identified per individual, with 50% overlapping annotated protein-coding genes. These gene-overlapping CNVs are markedly enriched for duplications and frequently span multiple genes. Our ongoing studies are investigating population variation in CNVs and testing whether functional redundancy and duplicate gene retention post-WGD has influenced CNV retention and evolution.

(O38) Genome-wide signatures of local adaptation in SNP loci and proteins of stonefly populations along a latitudinal gradient in Japan

Maribet Gamboa[1], Kozo Watanabe[1]

[1] Ehime University

Local adaptation plays a key role in determining genetic variation of natural populations. We combined genomic (Double Digest Restriction Associated DNA, ddRAD) and proteomics (protein expression analysis) approaches to explorer the genome wide associations of adaptive loci and environmental variables, focusing on stream stonefly populations. Seven stonefly species were collected from four regions in Japan (Matsuyama, Gifu, Sendai and Sapporo) along a nationwide latitudinal gradient (i.e., from north to south) with varying climatic conditions. The ddRAD analysis using 56 individuals among the 7 species generated a total of 247,580 loci with one single nucleotide polymorphism (SNP), with an average of 19,279 SNP loci per species. Outlier analysis found 7802 candidate SNP loci presumably under natural selection for all species, with 236-1927 candidate SNP loci per species. Correlation analysis of 8 environmental variables and allele frequency of the SNP loci found that altitude is a major environmental factor determining the spatial population genetic structure for all species. The proteomics analysis of 80 individuals for the 7 species by MALDI TOF/TOF yielded total of 446 proteins. Differential expression analysis of the identified proteins revealed that warmer regions caused up-regulation of metabolic proteins and down-regulation of proteins related to cold environmental stress, photoperiod and mating. Oxygen-related proteins and energy production proteins were up-regulated in the coldest and in the highest altitude regions. Our result overall showed high effectiveness of both genomics and proteomics approaches in understanding genomic adaptation and biological functions associated to local adaptation of stonefly populations along a climatic gradient.

POSTERS

Sequencing Technology and Developments

(P3) From RNA to Full-Length Transcripts: The Iso-Seq Method for Transcriptome Analysis and Genome Annotation

Michelle Vierra[1], Emily Hatas[1], Sarah Kingan[1], Ting Hon[1], Elizabeth Tseng[1], Tyson Clark[1]

[1] Pacific Biosciences

A single gene may code for a surprising number of proteins, each with a distinct biological function. This is especially true in higher organisms. Short-read RNA sequencing (RNA-seq) works by breaking up transcript isoforms into smaller pieces and bioinformatically reassembling them, leaving opportunity for misassembly or incomplete capture of the full diversity of isoforms from genes of interest. The PacBio Isoform Sequencing (Iso-Seq™) method employs long reads to sequence transcript isoforms from the 5' end to their poly-A tails - eliminating the need for transcript reconstruction and inference. These long reads result in complete, unambiguous information about alternatively spliced exons, transcriptional start sites, and poly-adenylation sites. This allows for the characterization of the full complement of isoforms within targeted genes, or across an entire transcriptome. The PacBio Sequel System generates hundreds of thousands of long and highly accurate single-molecule reads per SMRT Cell providing an opportunity to generate high-quality, cost effective genome annotation using the Iso-Seq method. Here we present the workflow from sample, through library prep, sequencing, and genome annotation using the latest protocols for the Sequel System. We will also share results of recent genome annotation work using the Iso-Seq method in comparison with short-read RNA-seq.

(P4) Improving enzymatic DNA fragmentation for NGS library construction

Fiona Stewart[1], Lynne Apone[1], Vaishnavi Panchapakesa[1], Karen Duggan[1], Chen Song[1], Timur Shtatland[1], Brad Langhorst[1], Christine Sumner[1], Pingfang Liu[1], Eileen Dimalanta[1], Theodore Davis[1]

[1] New England Biolabs, Inc. 240 County Road, Ipswich, MA 01938, USA.

The use of Next Generation Sequencing (NGS) data has been instrumental in advancing our understanding of human genetics, identifying the molecular events that contribute to human disease, and supporting drug development targeted towards precision medicine. Continued advancement relies on overcoming the limitations and bottlenecks associated with NGS. In this work, we have focused on NGS library preparation, where the requirement for expensive equipment and numerous steps can lead to sample loss, errors, and limited throughput. Specifically, we have developed a novel enzymatic DNA fragmentation reagent and have integrated this into the library prep workflow such that fragmentation is combined with end repair and dA-tailing in a single step. Integrating these reactions eliminates the need for costly equipment to shear DNA and reduces the number of sample transfers and losses. Adaptor ligation is also carried out in the same tube, after which a single cleanup step is performed. For low input samples, PCR amplification is performed prior to sequencing.

This method is compatible with a broad range of DNA inputs and insert sizes. Libraries generated using this streamlined method with inputs ranging from 500 pg to 500 ng of intact DNA show no significant difference in coverage uniformity or sequence quality metrics, compared to libraries generated with mechanically sheared DNA. Similarly, libraries generated to contain insert sizes that range from 150bp to 1kb display no significant difference in sequence quality from each other or from those generated with mechanically sheared DNA. Finally, this streamlined method generates libraries of substantially higher yields than those generated using mechanically fragmented DNA, allowing the use of lower DNA inputs and fewer PCR cycles.

Further, generation of larger DNA fragments enzymatically has utility for technologies including sequencing methods from Oxford Nanopore Technologies and Pacific Biosciences. We are evaluating additional novel enzymatic DNA fragmentation reagents for this application.

The ability to prepare high quality NGS libraries from intact DNA without the need for costly equipment and numerous cleanup or liquid transfer steps substantially reduces the time, cost and errors associated with library construction. In addition, these advances will enable greater use and adoption of NGS technologies in clinical and diagnostic settings.

Genome Informatics

(P5) The European Variation Archive

Cristina Yenyxe Gonzalez Garcia[1], Diego Poggioli[1], Gary Saunders[1], Jagadeesan Kandasamy[1], Jose Miguel Mut Lopez[1], Pablo Arce Garcia[1], Tom Smith[1], Sundararaman Venkataraman[1], Thomas Keane[1]

[1] European Bioinformatics Institute (EMBL-EBI)

The European Variation Archive (EVA, https://www.ebi.ac.uk/eva) is a primary open repository for archiving, accessioning, and distributing genome variation including single nucleotide variants, short insertion and deletions (indels), and larger structural variants (SVs) in any species. Since launching in 2014, the EVA and sister project DGVa have archived 520 million unique variants across 232 studies, containing 322,587 samples across 27 species, and submitted from 14 countries. A key function of the EVA as a long term data archive is to provide standard format, stable identifiers so that discovered variants and alleles can be referenced in publications, cross-linked between databases and integrated with successive reference genome builds. The EVA currently peers with the NCBI-based dbSNP and dbVar databases to form a worldwide network for exchanging and brokering submissions. From 2017, issuing and maintaining locus identifiers will be also divided by taxonomy: dbSNP will be responsible for all human locus accessioning, and EVA responsible for all non-human ones. Other services to researchers include: standard variant annotation, calculation of population statistics, and an intuitive browser to query and view variants from studies or across an entire species. The EVA currently offers a comprehensive REST API to query and export data. The API is species agnostic and is already extensively used by translational and species-specific resources including Ensembl Genomes, Open Targets, WheatIS and the 1000 Sheep Genomes Project.

(P6) The Human Ageing Genomic Resources

Daniel Thornton[1]

[1] University of Liverpool

The Human Ageing Genomic Resources (HAGR) are a collection of databases and tools designed to help researchers study the genetics of human ageing using bioinformatics approaches such as functional genomics, network analyses, systems biology and evolutionary analyses. HAGR originally comprised of AnAge, a curated database of ageing in animals, and GenAge, a database of genes affecting human longevity. Since then we have expanded HAGR to include LongevityMap, which comprises a comprehensive list of genetic variants that affect longevity, and GenDR, a database of genes affected during dietary calorie restriction in model organisms. We present the release of two new manually-curated databases to HAGR to help further elucidate the genetics of ageing: DrugAge and CellAge. DrugAge is a database of compounds shown to extend lifespan in model organisms. DrugAge advances ageing genomics by presenting drug-gene interactions specific to ageing. Combining life extension properties of compounds with drug-gene interaction data enables understanding of ageing genes and pathways through techniques such as gene functional enrichment. CellAge is a database of genes involved in cell senescence identified in human cells in vitro. Cell senescence is believed by many in the field to be a key driver of ageing, therefore in providing a repository of senescence genes, we move closer to providing potential genomic mechanisms of ageing. In summary, we present the key recent updates to HAGR, including CellAge and DrugAge, together with results from our most recent analysis of the data.

(P7) CyVerse UK: cyberinfrastructure to share bioinformatics data and applications

Alice Minotto[1], Erik Van Den Bergh[2], Annemarie Eckes[1], Robert P. Davey[1]

[1] Earlham Institute, Norwich Research Park, Norwich, NR47UZ

[2] EMBL-EBI, Wellcome Genome Campus, Hinxton, Cambridgeshire, CB101SD

CyVerse aims to provide a large scale cyberinfrastructure for bioinformatics, enabling robust reproducibility and collaborative analysis through data sharing and application versioning. The project goals are to overcome some of the pitfalls that are commonly experienced in bioinformatics and to provide research groups with readily available computational infrastructure for heavyweight analyses.

Despite the global reach of the parent CyVerse project, CyVerse UK provides geographically advantageous access to CyVerse infrastructure in Europe. Bioinformatics applications are registered through the Agave API on dedicated storage and execution systems that live as virtual machines on the EI cluster. We employ the Docker container virtualization system to package anything from single applications up to whole analysis environments, making them independent from the underlying computing architecture.

The CyVerse Data Store also allows users to share data at any time of their research, either with the whole community or with specific collaborators. Files can be stored and jobs submitted both via the CLI and a number of web interfaces.

CyVerse systems will be integrated with other projects, such as COPO, Galaxy and the Wheat Information System, to share and transfer datasets.

CyVerse promotes data sharing and accessibility by providing robust storage and compute infrastructures, and recording user specified metadata to promote FAIR data. The same principles of openness and reproducibility are applied to all analyses through metadata association and application versioning. Open tutorials and documentation are publicly available, as is most of the code: we aim to engage users and developers to contribute their apps and experience.

(P8) G-Anchor: a novel approach for whole genome comparative mapping utilising evolutionary conserved DNA sequences

Vasileios Panagiotis Lenis[1]

[1] IBERS, Aberystwyth University, Wales, UK

Abstract Background Cross-species whole-genome sequence alignment is a critical first step for genome comparative analyses ranging from the detection of sequence variants to studies of chromosome evolution. Animal genomes are large and complex, making their alignment a computationally intense process often requiring access to expensive high performance computing systems. With hundreds of sequenced animal genomes now available from multiple genome projects there is a need for tools that are capable of quickly and efficiently anchoring (or mapping) an animal genome to another species reference genome, without the need for the extensive computational resources used by traditional alignment software. **Results** Here we introduce G-Anchor. G-Anchor is a pipeline that utilizes highly conserved DNA sequences as anchors to rapidly map scaffolds of a de novo assembled genome to chromosome assemblies of a reference species. Our results demonstrate that G-Anchor is capable of successfully mapping a mammalian genome to a phylogenetically related reference species genome using a desktop or laptop computer within days, or sometimes a few hours, and with comparable accuracy to that achieved by LASTZ: thus making whole genome comparisons accessible to researchers with limited computational resources. **Conclusions** G-Anchor is a ready-to-use tool for anchoring a pair of mammalian genomes. It should also be suitable for use with large genomes that contain a significant fraction of evolutionally conserved DNA sequences, and that are not highly repetitive, polypoid or excessively fragmented. G-Anchor is not a substitute for whole-genome aligning software but can be used for fast and accurate initial genome comparisons.

(P9) Genome-wide identification of miRNAs and lncRNAs in *Cajanus cajan*

Ranjit Bahadur[1], Chandran Nithin[2], Amal Thomas[2], Jolly Basak[3]

[1] Indian Institute of Technology Kharagpur, 721302

[2] Department of Biotechnology, Indian Institute of Technology Kharagpur, Kharagpur-721302, India.

[3] Department of Biotechnology, Visva-Bharati, Santiniketan-731235, India

Non-coding RNAs (ncRNAs) are important players in the post transcriptional regulation of gene expression (PTGR). On one hand, microRNAs (miRNAs) are an abundant class of small ncRNAs (~22nt long) that negatively regulate gene expression at the levels of messenger RNAs stability and translation inhibition, on the other hand, long ncRNAs (lncRNAs) are a large and diverse class of transcribed non-protein coding RNA molecules (> 200nt) that play both up-regulatory as well as down-regulatory role at the transcriptional level. *Cajanus cajan*, a leguminosae pulse crop grown in tropical and subtropical areas of the world, is a source of high value protein to vegetarians or very poor populations globally. Hence, genome-wide identification of miRNAs and lncRNAs in *C. cajan* is extremely important to understand their role in PTGR with a possible implication to generate improve variety of crops. We have identified 616 mature miRNAs in *C. cajan* belonging to 118 families, of which 578 are novel and not reported in MirBase21. A total of 1373 target sequences were identified for 180 miRNAs. Of these, 298 targets were characterized at the protein level. Besides, we have also predicted 3919 lncRNAs. Additionally, we have identified 87 of the predicted lncRNAs to be targeted by 66 miRNAs. miRNA and lncRNAs in plants are known to control a variety of traits including yield, quality and stress tolerance. Owing to its agricultural importance and medicinal value, the identified miRNA, lncRNA and their targets in *C. cajan* may be useful for genome editing to improve better quality crop.

(P10) Gene expression of 29 immune cell types to estimate their proportions from mixed blood data

Gianni Monaco[1], Bernett Lee[2], Weili Xu[2], Leon Hwang[2], Michael Poidinger[2], Anis Larbi[2], João Pedro de Magalhães[1]

[1] University of Liverpool

[2] A*STAR

The cell composition of the hematopoietic tissue is highly heterogeneous and the characterization of all the immune cell types has a high impact on the treatment of diseases and increasing life expectancy. Deconvolution is the bioinformatics approach to define the proportions of specific cell types from transcriptomic data of a heterogeneous sample. Several deconvolution algorithms have already been proposed, nevertheless, there is still no consensus on the optimal methodology as well as on which cell types are more suitable for this approach. We used transcriptomic and flow cytometry data on 29 immune cell types and peripheral mononuclear blood cells (PMBC) of Singaporean individuals to validate the performance of two popular deconvolution methods based on basic linear regression (LLS) and support vector regression (CIBERSORT). Firstly, the transcriptomic data were used to estimate the cell-type proportions with the two methods. Secondly, the estimated proportions were compared with the real proportions calculated with flow cytometry. Regarding the methodologies, CIBERSORT gave better estimations as support vector regression is more robust to multicollinearity and noise. Regarding the cell types instead, we obtained good proportion estimation for almost all the immune cells of the innate system. Among the cells of the adaptive immune system, we found less agreeable results only for the memory subsets. In conclusion, we believe that deconvolving blood expression data is a promising approach that is still in its early stages, but with further studies could become widely adopted.

(P11) Campylobacter concisus pan-genomics: Elucidating the phylogeny and phenotypes through long & short read technologies

Matthew Gemmell[1,2], Susan Berry[3], Indrani Mukhopadhya[3], Richard Hansen[3,4], Hans Nielsen[5], Henrik Nielsen[5], Georgina Hold[3]

[1] Centre for Genome Enabled Biology and Medicine, School of Medicine, Medical Sciences and Nutrition, University of Aberdeen, Aberdeen, AB25 2ZD, U.K.

[2] Centre for Genomic research, University of Liverpool, Liverpool, UK.

[3] GI Research Group, School of Medicine, Medical Sciences and Nutrition, University of Aberdeen, Aberdeen, AB25 2ZD, U.K.

[4] Royal Hospital for Children, 1345 Govan Road, Glasgow, G51 4TF

[5] Department of Infectious Diseases, Aalborg University Hospital, PO Box 365, DK9100 Aalborg, Denmark

Campylobacter concisus is an oral commensal, which has historically been linked to gingivitis and periodontitis but more recently has been linked with gastrointestinal diseases including inflammatory bowel disease and gastroenteritis. This study aims to generate robust genome sequence data for a number of clinical *C. concisus* strains and compare the genomics of different phenotypes through pan-genomic analysis. 53 *C. concisus* isolates from patients with various gastrointestinal disease presentations were sequenced with the Illumina MiSeq. Of these 53, 2 were additionally sequenced with the PacBio RSII. Genome assembly, scaffolding and gap filling was carried out on the sequence data. The genome assemblies, along with 37 other publicly available *C. concisus*, were annotated and various virulence factors, including antibiotic resistance genes, were detected. Pangenome analysis with roary was then carried out. Phylogenetic analysis found the isolates formed into two main groups/genomospecies (GS). Clustering did not occur based on disease presentation or body site. Pangenome analysis found little difference in the core genome of the two groups. Pangenome analysis detected 28,465 genes within the *C. concisus* pangenome, 424 of these being core genes. Comparing the pangenome of *C. concisus* to that of Campylobacter found 10 core genes were unique to *C. concisus*. Our findings indicate that *C. concisus* strains are phenotypically and genetically diverse, and suggest the genomes of this bacterium contain modifications in secretion systems that may play an important role in their virulence potential.

(P13) Panthera: pipeline for prediction of protein-coding genes in large genomes

Sergei Kliver[1], Aleksey Komissarov[1], Gaik Tamazian[1], Stephen O’Brien[1]

[1] Dobzhansky Center for Genome Bioinformatics, Saint-Petersburg State University

Proper prediction of protein-coding genes in genomes plays a crucial role for further comparative analysis. Unfortunately, there is still no instrument for gene prediction in large genomes that can produce high quality gene models using only genome sequence, so extrinsic sources are of great value. However, pipelines using external support, such as MAKER, are known for issues with merged and fragmented genes. Also this tools don't provide enough metrics for quality estimation of gene models. We present new protein-coding gene prediction pipeline based on extrinsic sources for gene existence and de novo predictions by AUGUSTUS. However, de novo method is used only for gene fragment unification and extension. Further, gene models are checked by hits of corresponding proteins to Pfam and SwissProt databases. Finally, refinement of gene models is performed by comparison with homologous genes of related species to reduce number of fragmented and merged genes. Our gene prediction pipeline was developed to minimize number of incorrect gene models and ease estimation of their quality. So statistic and quality reports are provided for every stage of gene prediction that allows easy detection of possible issues with input data. This work was funded from RSF grant 17-14-01138.

(P14) KrATER – user-friendly tool for k-mer analysis

Sergei Kliver[1], Gaik Tamazian[1], Vladimir Brukhin[1], Stephen O'Brien[1], Aleksey Komissarov[1]

[1] Dobzhansky Center for Genome Bioinformatics, Saint-Petersburg State University

K-mer based approaches are widely used in bioinformatic areas related to DNA/RNA sequences. Analysis of k-mer distribution is crucial step in quality control of raw reads, error correction and genome or transcriptome assembly. However, all of tools developed for analysis of k-mer distribution are very difficult and sometimes impossible even to install or fail to perform in complicated cases. We present KrATER - K-mer Analysis Tool Easy to Run, which is user friendly for both installation and run. This tool have no assumptions about pattern of k-mer distributions that is important for complicated cases common for hybrid or highly heterozygous genomes. KrATER draws publication-ready plots in both logarithmic and linear scales. KrATER is available at the Python Package Index (PyPI, https://pypi.python.org) and GitHub (https://github.com/mahajrod/KrATER). KrATER was developed to fill the lack of simple, easy for installation and run tool in k-mer analysis area. KrATER can be used for estimation of genome sizes, quality control and initial analysis of reads, comparison of libraries, estimation of error correction efficiency, publication preparation and development of new k-mer based approaches. This work was funded from RFBR grant 16-54-21014 and SPbSU grant 1.52.1647.2016

(P16) Discovery and visualisation of homologous genes and gene families using Galaxy

Anil S. Thanki[1], Nicola Soranzo[1], Wilfried Haerty[1], Robert P. Davey[1]

[1] Earlham Institute, Norwich Research Park, Norwich NR4 7UZ, UK

The phylogenetic information inferred from the study of homologous genes helps us to understand the evolution of gene families and plays a vital role in finding ancestral gene duplication events and in identifying regions that are under positive selection within species.

The Ensembl GeneTrees pipeline generates gene trees based on coding sequences and provides details about exon conservation, and is used in the Ensembl Compara project to discover homologous gene families. Since expertise is required to configure and run the pipeline via the command-line, we created GeneSeqToFamily, an open-source Galaxy workflow based on Ensembl GeneTrees. GeneSeqToFamily helps users to run potentially large-scale gene family analyses without requiring the command-line while still allowing tool parameters, configurations, and the tools themselves to be modified.

At present, we are using this workflow on a set of vertebrate genomes, with some analyses comprising more than 13000 gene families. Gene families discovered with GeneSeqToFamily can be visualised using the Aequatus.js interactive tool, integrated within Galaxy as a visualisation plugin.

Aequatus.js is a JavaScript library developed as a part of Aequatus project which provides an in-depth view of gene structure across gene families, with various options to render and filter visualisations.

We are also working on integrating protein domain information from SMART (a Simple Modular Architecture Research Tool) to complement discovered gene families, and the incorporation of PantherDB into the workflow for validation of families.

Availability: https://github.com/TGAC/earlham-galaxytools

(P17) Grassroots Infrastructure: An interoperable data repository for plant science

Simon Tyrrell[1], **Xingdong Bian**[1], Robert P. Davey[1]

[1] Earlham Institute, Norwich Research Park, Norwich, NR4 7UZ, United Kingdom

Integrative research requires extensive multi-level approaches to enrich and expose data and workflows so that informatics infrastructures can process them effectively. The Grassroots Infrastructure is developed at the Earlham Institute (EI) to consolidate data and analyses, facilitating consistent approaches to generating, processing and disseminating public datasets in the plant sciences. Its lightweight reusable software stack comprises: an iRODS data management layer to provide structure to unstructured filesystems, with Elasticsearch-indexed metadata and Davrods-exposed WebDAV APIs; interfaces to interact with local or cloud-based analysis platforms; an Apache web server layer to deliver content and provide access to public programmatic interfaces; services such as: BLAST search on multiple databases across different sites; a mapping tool showing pathogen samples with temporal and spatial data. It can be run locally or packaged in virtual containers and deployed on a variety of hardware thus representing a decentralised system, allowing information generators to retain control over their resources but allowing interconnected resources to access each other consistently. Grassroots represents EI's contribution to the Wheat Initiative Wheat Information System (WheatIS) project, formalising the infrastructure as the federated UK WheatIS node involving partners from the University of Bristol, the European Bioinformatics Institute, Rothamsted Research, and the John Innes Centre.

We are currently working on standardised APIs such as the Breeding API (BrAPI) and schemas such as Frictionless Data and BioSchemas to enable greater interoperability with a variety of existing services, and integration with data analysis platforms such as CyVerse and Galaxy.

(P20) Time course RNAseq experiments in wheat to analyse response to heat stress and fertilizer input

Asier Gonzalez Uriarte[1], David Hughes[1], Peter Buchner[1], Matthew Audley[1], Keywan Hassani-Pak[1]

[1] Rothamsted Research

Population growth and climate change pose a major threat to worldwide food security. A key agricultural challenge will be to maintain and even increase the yield in much warmer conditions while decreasing fertilizer input. Hence, it is crucial that we gain insights into how heat stress and nutrient availability affect molecular mechanisms that control the yield of key crops such as wheat. Here we describe the analysis of two RNA-seq time course experiments in hexaploid wheat conducted at Rothamsted Research. The first study intends to elucidate the molecular basis of pollen infertility due to heat stress by sampling four time points at meiosis. The second experiment investigates the role of nitrogen on the remobilization of nutrients from the leaves to the head and senescence development, a process that is prolonged and spans a time frame of weeks. We combine two packages to identify differentially expressed genes, i) edgeR supports the analysis of factorial experiments and ii) maSigPro deals with the peculiarities of time course data. We also present a Shiny application that we are developing to enable interactive differential gene expression analysis and clustering; and KnetMiner (http://knetminer.rothamsted.ac.uk/) as a tool to help with the biological interpretation of the results.

(P21) Identifying Sex Determination Loci in the Highly Heterozygous White Guinea Yam (*Dioscorea rotundata* Poir.)

Muluneh Tamiru[1]*, Satoshi Natsume[1]*, Hiroki Takagi[1]*, **Benjamen White**[2]*, Hiroki Yaegashi1[*], Motoki Shimizu[1]*, Kentaro Yoshida[3], Aiko Uemura[1], Kaori Oikawa[1], Akira Abe[1], Naoya Urasaki[4], Hideo Matsumura[5], Pachakkil Babil[6], Shinsuke Yamanaka[7], Ryo Matsumoto[7], Satoru Muranaka[7], Gezahegn Girma[8], Antonio Lopez-Montes[8], Melaku Gedil[8], Ranjana Bhattacharjee[8], Michael Abberton[8], P. Lava Kumar[8], Ismail Rabbi[8], Mai Tsujimura[9], Toru Terachi[9], Wilfried Haerty[2], Manuel Corpas[2], Sophien Kamoun[10], Günter Kahl[11], Hiroko Takagi[7], Robert Asiedu[8], and Ryohei Terauchi[1,12]

[1] Iwate Biotechnology Research Center, Kitakami, Japan

[2] The Earlham Institute, Norwich, UK

[3] Kobe University, Kobe, Japan

[4] Okinawa Agricultural Research Center, Naha, Japan

[5] Shinshu University, Nagano, Japan

[6] Tokyo University of Agriculture, Tokyo, Japan

[7] Japan International Research Center for Agricultural Sciences, Tsukuba, Japan

[8] International Institute of Tropical Agriculture, Ibadan, Nigeria

[9] Kyoto Sangyo University, Kyoto, Japan

[10] The Sainsbury Laboratory, Norwich, UK

[11] University of Frankfurt, Frankfurt, Germany

[12] Kyoto University, Kyoto, Japan

White Guinea yam (*Dioscorea rotundata* Poir.) is a staple crop of great agricultural, cultural and economic significance to Africa, the Americas, the Caribbean, South Pacific and Asia. While demand for yam in sub-Saharan Africa is high, there is continuing decline in production due to pests, reduced soil fertility and disease. Despite the importance of this crop, there are limited genomics resources available for yam that could facilitate breeding initiatives, nor comprehensive phylogenetic or evolutionary studies. The breeding of the white Guinea yam is further impeded by its heterozygosity, long growth cycle, erratic flowering times, and dioecious (both female and male individuals) nature. The latter of which is a rare trait found in only 5 - 6% of angiosperms. To accelerate Guinea yam marker-assisted breeding, as part of an international collaboration, we have sequenced the 594 Mb genome, produced a chromosome anchored assembly and predicted a total of 26,198 genes. Phylogenetic analysis of 2,381 conserved genes has revealed Dioscorea not to form a monophyletic clade with the Poales, Arecales or Zingiberales, indicating an early divergence from the latter taxa in monocotyledons. Most importantly whole genome re-sequencing of bulked segregant F1 progeny, segregating for male and female individuals, has led to the identification of a genomic region and candidate genes associated with female heterogametic (male=ZZ, female=ZW) sex determination. The genome and identification of sex loci in Guinea yam will serve as an invaluable resource for genome-assisted breeding in yam and presents a unique opportunity to study the evolution of sex in monocots.

(P22) NLR diversity and evolution in exotic monocots

E.L. Baggs[1] , P. Bailey[1] , K.V. Krasileva[1]

[1] Earlham Institute, Norwich, UK

Devastating crop diseases can be prevented through activation of the plant immune system. Nucleotide Binding Leucine Rich Repeat (NLR) proteins are a type of plant immune receptor, identifiable by the presence of an NB-ARC domain and responsible for the recognition of intracellular pathogen molecules. The number of NLRs varies from approximately 190 to 1,000 in the *Poaceae* family alone. Although for many *Poaceae* species the NLR gene family is well understood this is not the case for other economically important monocot crops such as banana, yam and pineapple. We utilised 7 available genomes and bioinformatic methods to further our understanding of the NLR gene families' composition and evolution across divergent monocot families.

We identified over 900 NLRs across 7 monocot species after further developing a bioinformatic pipeline initiated in our laboratory. Using multi-species phylogenies we characterized the ancestral monocot NLR repertoire as well as lineage specific expansions and contractions of the NLR gene family. The *Poaceae* lineage of monocots has the largest lineage specific expansion in comparison to other monocot families in our study. Simultaneously, we were able to identify orthogroups which have been conserved in all species studied.

Current work aims to identify selective pressures acting on the NLRs identified as conserved across monocots. The identification of evolutionary conserved NLRs in monocots provides a testable prediction for the minimal required plant immune system. To test this hypothesis, we plan to generate and phenotype knockouts of a sub-set of the conserved genes identified.

(P23) It’s a grass, grass, grass: Profiling plant gluten genes with targeted resequencing and bioinformatics

Christian Schudoma[1], Matt Clark[1], and Ksenia V. Krasileva[1]

[1] Earlham Institute, Norwich, UK

Gluten proteins are important in food production and human health. Rich in proline and glutamine amino acids, these proteins serve as nutrients for the developing plant, and represent a key component in bread making quality, giving bread its unique ability to rise. Infamously, ingestion of gluten can lead to Coeliac disease in genetically susceptible people.

A single wheat plant harbours more than a hundred gluten genes with extremely low complexity sequences, tandem repeats and many motif variations due to a high tolerance to mutations.

Recently, we have established a new technique, called GlutenSeq, which enables the targeted capture and sequencing of gluten genes from any biological material using either short or long read sequencing technologies. Combining GlutenSeq with a dedicated bioinformatics pipeline for assembly, annotation and analysis, we are able to a) detect and classify gluten and gluten-like genes from different plant species and breeder’s varieties and b) to annotate harmful Coeliac-triggering epitopes in the detected sequences.

With GlutenSeq we present a useful tool for profiling gluten genes. The protocol and bioinformatics pipeline could also be used monitor the food chain and identify the origins of gluten contamination.

(P25) Comparative genome analysis of novel probiotic microorganism

Chul Lee[1], Jihyun Yu[1], Sook Hee Yoon[1]

[1] Seoul National University

As probiotics play an important role in maintaining a healthy gut flora environment through antitoxin activity and inhibition of pathogen colonization. In this regard, there is a growing interest in the isolation of novel probiotics and their functional effect. We perform de novo assembly and genome analysis on the Lactobacillus novel strain. Based on the complete genome, comparative genome analysis between Lactobacillus strains is examined to predict strain specific genomic characteristics using in-silico analysis. In addition, evolutionary genetic analysis revealed that novel strain has potential ability as lactic acid bacteria against pathogen and oxalate level. These results indicate that the novel strain is a suitable candidate of probiotics. This study provided insight into the Lactobacillus species as well as confirmed the possibility of its utility as a candidate probiotics.

(P26) Novel genomics-led approaches to characterise viral diseases in Atlantic salmon

Michael Gallagher[1], Iveta Matejusova[2], Daniel Macqueen[1]

[1] University of Aberdeen

[2] Marine Science Scotland

Global farmed production of salmonid fishes is worth > £8 billion annually, accounting for ~15% of total traded farmed fish. However, a major bottleneck limiting growth of this industry is loss caused by infectious diseases, which can have devastating economic impacts. Viruses - which cause 20% of all known infectious diseases in aquaculture - are of particular concern, as few effective anti-viral therapeutics or preventative vaccines have been developed. For example, an outbreak of Infectious Salmon Anaemia virus (ISAV) in 2007-08 cost the Chilean salmon industry around $1 billion and reduced salmon production from 650,000 to 100,000 tonnes in just two years. Rapidly and accurately diagnosing such outbreaks will help control strategies by identifying the strains present and the pathogenicity of sequence variation. The current standard is still to sequence a few marker genes with key roles in viral function using the Sanger approach. My project is developing methods to routinely and affordably sequence whole genome sequences for problematic salmonid viruses using second (Illumina) and third generation (Oxford Nanopore) sequencing platforms. The goal is to enable rapid genome-wide analysis and diagnostics, and to facilitate more comprehensive implementation of molecular epidemiology for inferring transmission routes and linking sequence variation to pathogenicity. Applying such technologies within the aquaculture industry may ultimately help control the spread of devastating diseases and contribute to both economic and food security.

(P28) EuPathDB: integrating eukaryotic pathogen genomic data with advanced search capabilities and large-scale data analysis

Jane Pulman[1]

[1] The University of Liverpool

The Eukaryotic Pathogen Database (EuPathDB.org) Bioinformatics Resource Center provides online open access to over 200 organisms within Amoebazoa, Apicomplexa, Chromerida, Diplomonadida, Trichomonadida, Kinetoplastida and numerous phyla of oomycetes and fungi. In addition to genomes (>200) and annotation, EuPathDB integrates structured sample and clinical data, and a wide range of functional data types (>500 datasets) encompassing transcript and protein expression, sequence and structural variation, epigenomics, clinical and field isolates, metabolites and metabolic pathways and host-pathogen interactions. EuPathDB provides easy to use tools to mine the underlying datasets or users can carry out custom dataset analysis in the EuPathDB Galaxy instance. The Galaxy instance was introduced in 2016 on essentially all EuPathDB family sites, offering pre-loaded genomes, private data analysis and display, and data analysis sharing and export. The instance houses a large variety of bioinformatics tools to facilitate large-scale analysis without the need for programming expertise. It was developed in partnership with Globus Genomics (https://www.globus.or.genomics) and currently has several RNASeq pre-configured workflows with more workflows planned. Workflows can be imported for editing and users can also create custom workflows which along with results and datasets can be stored privately or shared with the wider community. We also offer the option to export BigWig files directly to EupathDB GBrowse from the Galaxy instance. For questions and suggestions email us at help@eupathdb.org. Author presents on behalf of the EuPathDB team. Supported by NIH HHSN272201400030C and the Welcome Trust grant WT108443MA

(P30) Comparative network-omics: linking genomics and network data to investigate host adaptation of *Salmonella enterica* strains

Marton Olbei[1,2], Padhmanand Sudhakar[1], David Fazekas[3], Jozsef Baranyi[2], Rob Kingsley[2], Tamas Korcsmaros[1,2]

[1] Earlham Institute

[2] Quadram Institute

[3] Eotvos Lorand University

Salmonella enterica serovars are one of the most common foodborne pathogens, causing up to 100 million cases per year. The gastroenteritis and other illnesses caused by them are responsible for 155.000 fatalities per year.

Most of the *Salmonella* serovars are generalists, capable of infecting a range of hosts, but some of them become host adapted, specializing on a small range of targets or just a single organism. In this research project the goal is to try and determine what molecular interactions cause these large differences in lifestyle in otherwise closely related serovars. We attempt to determine this through the analysis of the regulatory, metabolic and protein-protein interaction networks of five gastrointestinal and five extraintestinal *Salmonella enterica* strains from the SalmoNet database (http://SalmoNet.org). SalmoNet contains integrated, multi-layered networks of *Salmonella* strains, compiled from the relevant literature, primary and secondary databases, high throughput experiments and inferred connections from the commensal bacteria *Escherichia coli*.

The analysis of the networks is done with both a supervised method, which relies heavily on previously amassed information from literature and databases, and an unsupervised method, which deals directly with the data and statistics found in the SalmoNet networks. With these “comparative networkomics” approaches we are planning to pinpoint potentially interesting regulatory or metabolic pathways, protein-protein interactions that are important in terms of host adaptation. To give validity to any predictions made with these methods appropriate molecular biology experiments are going to be conducted that can prove or disprove the proposed functions.

Clinical and Translational Genomics

(P31) Towards Precision Medicine for the Treatment of Cystinuria

Henry Martell[1], Kathie Wong[2], Juan Martin[1], Ziyan Kassam[2], Kay Thomas[2], Mark Wass[1]

[1] The University of Kent

[2] Guy's and St Thomas' NHS Foundation Trust

Cystinuria is an inherited disease that results in the formation of cystine stones in the kidney. Two genes (SLC3A1 and SLC7A9) are known to be responsible for the disease. Variants of these two genes disrupt amino acid transport across the cell membrane, which leads to the build-up of relatively insoluble cystinine and the formation of stones. Assessment of the effects of each mutation is critical in order to provide tailored treatment options for patients, and the use of computational methods offers a viable solution. In previous work, we sequenced SLC3A1 and SLC7A9 in a cohort of patients from Guy's Hospital, London, UK. In this project, we used various computational methods to assess the effects of cystinuria associated mutations, utilising information on protein function and structure, evolutionary conservation and natural population variation of the two genes. Our aims were to understand 1) How each mutation alters transporter function 2) How much is each variant contributing to the cystinuria symptoms. We also analysed the ability of some methods to predict the phenotypes of individuals with cystinuria, based on their genotypes, and compared this to clinical data.

(P32) Nucleosome repositioning in cancer

Graeme J. Thorn[1], **Navid Shafiei**[1], Joshua Burton[1], Christian Grumaz[2], Yevhen Vainshtein[2], Kai Sohn[2], Paul Brennan[3], Elena Klenova[1], Alexander Kagansky[3], Vladimir B. Teif[1,*]

[1] University of Essex, Colchester, UK

[2] University of Edinburgh, Edinburgh, UK

[3] Fraunhofer Institute for Interfacial Engineering and Biotechnology IGB, Stuttgart, Germany.

[*] University of Essex, Wivenhoe Park, Colchester CO4 3SQ, UK. Email: vteif@essex.ac.uk

Nucleosome positioning is recognised as an important regulator of gene expression in normal and diseased cells. It is determined by several processes including the DNA sequence-dependent histone affinity landscape, active ATP-dependent chromatin remodelling, competition with transcription factors, chemical modifications of DNA and histones, and statistical positioning near genomic boundaries. Here I will provide an overview of our projects where nucleosome positioning is being investigated from the point of view of the analysis of cancer onset and progression. In the project conducted by the CancerEpiSys consortium we focused on nucleosome repositioning in B cells from patients with chronic lymphocytic leukaemia (CLL). We have shown that about 1% of nucleosomes reproducibly change their positions in cancer patients versus healthy individuals. A particularly important class of nucleosomes gained at promoters and enhancers marks the B-cell receptor signalling pathway specific for this cancer. Importantly, nucleosome positioning changes allow predicting cancer predisposition which is not yet evident from the changes of gene expression. In another project supported by the Wellcome Trust we are applying a similar concept to nucleosome repositioning in unrelated solid cancers, using paired cancer/normal tissue samples from the patients with glioblastoma and breast cancer. In this talk I will explain our current understanding of the role of nucleosome positioning as a cell memory in cancer transitions.

(P33) The evolution of microRNA regulation in the ErbB signalling network

Mohab Helmy[1], Antonio Marco[1]

[1] University of Essex

MicroRNAs, post-transcriptional regulatory molecules, have an important role in cell signalling. Cell signalling allows cells to communicate with their environments. Often, the signalling process begins by the binding of a ligand to a cell-surface receptor which in turn initiates a series of intracellular processes which leads to a cellular action, often modifying gene expression via transcription factors. The ErbB signalling network is one of the most widely studied pathways in biomedical research due to its involvement in multiple diseases, mainly cancer. Although there are hundreds of studies on the regulatory role of individual microRNAs in the ErbB network, a global picture of how microRNAs regulate the entire signalling pathway is missing. In this work, we used a comparative genomics approach to explore the regulatory relationships between different groups of microRNAs and the gene members of the ErbB signalling network among different species. Our results show that the strength of microRNA regulation is high at the receptors. The evolutionary analysis of gains and losses of microRNA target sites indicates that some specifically conserved microRNA target sites on receptors have been lost in rodents. This observation is consistent with changes in cell-cycle regulation associated to a change in life-history in these species. We conclude that microRNA-mediated regulation of the ErbB signalling network is more important in the receptors, and that selective pressures (both purifying and adaptive) on microRNA target sites is particularly strong at this level.

(P34) Implementation of a clinical NGS amplicon panel for the genetic diagnosis of clinical lung and colorectal tumour tissue

Amy Slater[1], Graham Taylor[2], Nicola Foot[2], Nahid Kamal[2], Guillermina Nickless[2], Aled Jones[2], Shu Ching Yau[2], Michael Neat2

[1] GSTT NHS

[2] Viapath

With the advent of precision medicine and increasing demand, genetic analysis of tumours has become a standard of care requirement for delivery of targeted therapies, such as EGFR tyrosine kinase inhibitors for treatment of non-small cell lung carcinoma (NSCLC) and colorectal cancer. Major challenges facing the genetic analysis of NSCLC and colorectal cancer include increasing numbers of clinically relevant genes requiring assessment and the often small size/ low tumour content of biopsies. However, our previous methods of analysis including HRM (High Resolution Melt analysis) and Sanger sequencing permitted investigation of only one amplicon per reaction. Here we report the successful validation of a multiplex PCR amplicon panel (Swift Biosciences Accel-Amplicon EGFR Pathway Panel; AL-51048) for analysis of NSCLC samples using as little as 10ng starting DNA, successfully detecting variants with allele frequencies as low as 1%. Implementation of a bioinformatics pipeline that provides reduction of complexity using Amplivar (by grouping reads into read hashes prior to alignment) and limits the risks of generating artefactual results frequently observed when sequencing amplicons from low mass FFPE samples. These include damaged DNA and early cycle PCR errors being reported as mutations, amplicon recombination during PCR, reads mapping off-target and sequencing errors. We have successfully validated this method on 150 previously genotyped samples (including clinical samples), achieving full concordance; enabling us to introduce a diagnostic test which genotypes multiple clinically-actionable loci and increases sensitivity for detecting mutations at low allele frequency in samples with reduced tumour content, replacing HRM and Sanger sequencing.

(P35) Nanopore long read sequencing for clinical diagnostics

Andrew Bond[1,2], Kezia Brown[2], Shu Ching Yau[2], Graham Taylor[2]

[1] Guy's and St Thomas' NHS Trust

[2] Viapath

Emerging long read sequencing technologies have the potential to identify disease-associated variation that is undetectable using standard sequencing methodologies. We have used the Oxford Nanopore MinION to perform long read sequencing on clinical samples to identify the breakpoints of structural variants, detect single nucleotide variants and perform haplotype phasing. We describe the multiplex analysis of barcoded BRCA1, BRCA2, SMN1, HLA and LDLR amplicons (3.6 to 16kb). We found that alignment with BWA MEM using the ont2d setting, followed by consensus variant calling using SAMtools mpileup, allowed accurate detection of Genome in a Bottle truth set variants in BRCA1/2 at 500x read depth with 2D (template + complement) reads. All 10 BRCA1/2 variants were identified in our clinical cases in the 1D data, although false positives were detected due to systematic (non-random) errors. Two LDLR deletions (3.3kb and 500bp) were characterised at the base pair level, with confirmatory Sanger analysis identifying Alu elements at the breakpoints of the 500bp deletion. Additionally, we performed haplotype phasing to differentiate pathogenic variants in 16kb SMN1 reads from non-pathogenic variants in SMN2 which shares >99% homology. Despite high error rates in the reads, BWA MEM was able to correctly differentiate >80% of SMN1 and SMN2 reads. We are currently analysing HLA data and investigating the new 1D^2 technology. We show that random error rates are tractable by consensus alignment and over-sequencing. Providing systematic errors are avoided, Nanopore sequencing can deliver unique tools for clinical use and point of care testing.

(P36) DNA repair increases sequencing accuracy without altering actual mutation frequency in clinical samples

Pingfang Liu[1], Chen Song[1], Lixin Chen[1], Laurence Ettwiller[1], Eileen T. Dimalanta[1], Theodore B. Davis[1], and Thomas C. Evans Jr. [1]

[1] New England Biolabs, Ipswich, MA 01938, USA

Targeted cancer therapy based on genomic alterations can be remarkably effective. Currently, cancer genome profiling using next generation sequencing (NGS) is routinely applied in cancer care to guide personalized treatment. The accuracy of this profiling directly impacts therapeutic choices and the outcomes of patient care. We show here that false positive variant reads are abundant and can account for a major fraction of identified low frequency somatic variations in publicly available datasets. Some of these false positive variants are originated from mutagenic DNA damage. We have further demonstrated that enzymatic DNA repair increases sequencing quality by lowering damage-induced background noise. As a result, enzymatic DNA repair has the potential to improve sequencing accuracy, avoiding incorrect somatic variant calls and consequently reducing incorrect diagnostic conclusions.

In addition, we have investigated whether enzymatic DNA repair introduces any bias to NGS libraries using analysis by Droplet DigitalTM PCR (ddPCRTM). DNA reference standards containing multiple common cancer mutations (Horizon Discovery, Inc.) were spiked into NA19240 genomic DNA at defined frequencies (0.25-2.5% quantified by ddPCR). Genotyping of the NA19240 gDNA ensured that they were free of any of the spiked-in mutations. After DNA repair and library preparation, mutation frequencies were quantified by ddPCR, and compared to the mutation levels in input DNA and control libraries without repair. ddPCR data showed no difference in mutation frequency for the spiked-in mutations between the control and repair groups.

(P37) NRGene’s Technology - Genome de novo Assembly and Beyond

David Macheto[1]

[1] NRGene

NRGene is a world leading genomics big data company, developing cutting-edge software and algorithms to reveal the complexity and diversity of plants, animals, and aquatic organisms for the most advanced and sophisticated research programs. Whole genome sequencing, using 180X coverage of Illumina reads and NRGene’s DeNovoMAGIC™-3.0 software, was used to successfully complete the full assembly of over 200 genomes in the last 2 years. Among these were highly-complex animal and plant genomes such as the Bovine Nelora genome (N50>39 Million bp, N90>8 Million bp) and the 17Gbp hexaploid bread wheat genome (N50>38 Million bp, N90>7 Million bp). Intensive, in-depth analyses by multiple specialists from highly reputed institutions, such as IWGSC, IPK, KDRI and more, have proven that the assemblies meet with the highest standards of quality and accuracy. NRGene offers many services beyond de novo genome assembly. With NRGene’s tools, it’s possible to obtain an analysis on the full genetic diversity of an organism, to perform genome-to-genome mapping, discover haplotype markers, or to order custom solutions for other challenges and research needs. NRGene’s services are consistently chosen by world leading researchers and breeding companies for high quality start-to-end genomics services and software.

(P39) Assessing Percentage Purity and Genetic Diversity between *O. niloticus* and *O. urolepis urolepis* using RAD Sequencing

Levinus Leonard Mapenzi[1,3], Dirk Jan de Koning[2], Aviti John Mmochi[3]

[1] The University of Dodoma, College of Natural and Mathematical Sciences, P. O. Box 338, Dodoma, Tanzania.

[2] Swedish University of Agricultural Sciences, Department of Animal Breeding and Genetics, Box 7023, 751 23 Uppsala, Sweden

[3] University of Dar es Salaam, Institute of Marine Sciences, P. O. Box 668, Mizingani Road, Zanzibar, Tanzania

Aquaculture undertakings in Tanzania are largely dependent on wild collection for “fingerlings” or seeds. This is because there is no any significant fingerlings production center for sustainable aquaculture in the county. The wild collected and reared fish may have high degree of hybridization and thus could not be the true breeds of species highly cultured in Tanzania. Reliance on wild collection is unsustainable for aquaculture development. In that regard, the present study intends to determine percentage purity and genetic diversity within and between *Oreochromis niloticus* and *Oreochromis urolepis urolepis* using RAD Sequencing. This follows a successful study on hybridization of the two species in which the hybrids presented better growth, survival and tolerance in plastic tanks’s saline coastal waters. Also, hybrids showed better growth performance at different stocking densities and dietary probiotics treatments than their parents. Therefore, it is crucial to determine hybrids parental lines’ genetic diversity which can be helpful in getting the starting founder population aiming at establishment of the breeding programme. This will enable selection for better growth performing traits for introgression in hybrids to establish production of fast hybrids growers hence improvement of the nutritional status and aquaculture production in Tanzania. This study is part of the ongoing PhD project with the title “Potential for Aquaculture of *Oreochromis niloticus* and *Oreochromis urolepis urolepis* Hybrids: Genetic Characterization and Effects of Salinity, Stocking Density and Dietary Probiotics on Growth Performance”.

(P41) Bioinformatics approaches for assessing the impact of temperature on eukaryotic phytoplankton

Kara Martin[1,2]

[1] Earlham institute

[2] UEA

Phytoplankton are an extremely diverse group of organisms that include eukaryotic and prokaryotic species and can be found inhabiting waters throughout the world. Marine phytoplankton are important for life on earth, as they are a major part of the marine food web, are photosynthetic {Armbrust, 2009} and contribute about half of atmospheric carbon dioxide fixation. There is evidence of a latitudinal gradient of optimal growth temperature across the ocean. Temperature drives phytoplankton diversity and global warming is predicted to reduce diversity in tropical regions {Mock and Medlin, 2012}. Temperature is also vital for metabolism, evolution and the ocean's biogeochemical cycles, but the effects of the changing temperature due to global warming is unknown for eukaryotic phytoplankton {Toseland et al., 2013}. Samples were collected from different latitudes from the Arctic Ocean and Arctic Ocean close to the Norway's coast down to South Atlantic Ocean Cape Town. In addition, environmental data at the time of sample collection was recorded, making it possible to analyse how diversity vary under different conditions of temperature and nutrient levels. To do this we developed a pipeline to analyse 18s rDNA and 16s rDNA data. In our results we observe that the diversity of phytoplankton are in agreement with the latitudinal gradient of optimal growth temperature that has been reported in literature {Mock and Medlin, 2012}. We investigated the co-occurrence relationship of eukaryotic and prokaryotic species. We found two networks, one network correlating with a warm environment and the other correlating with a cold environment.

(P42) Profiling Complex Population Genomes with Highly Accurate Single Molecule Reads: Cow Rumen Microbiomes

Michelle Vierra[1], Cheryl Heiner[1], Itai Sharon[2], Steve Oh[1], Alvaro G. Hernandez[3], Itzhak Mizrahi[4], Richard Hall[1]

[1] Pacific Biosciences

[2] Tei-Hai College

[3] University of Illinois at Urbana-Champagne

[4] Ben-Gurion University

Determining compositions and functional capabilities of complex populations is often challenging, especially for short-read sequencing technologies that do not uniquely identify organisms or genes. Long-read sequencing improves the resolution of these mixed communities, but adoption of this application has been limited due to concerns about throughput, cost and accuracy. The PacBio Sequel System generates hundreds of thousands of long and highly accurate single-molecule reads per SMRT Cell. We investigated how the Sequel System might increase understanding of metagenomic communities. In the past, focus was largely on taxonomic classification with 16S rRNA sequencing. Recent expansion to WGS enables functional profiling as well, with the ultimate goal of complete genome assemblies. Here we analyze the complex microbiomes in 5 cow rumen samples, for which Illumina WGS sequence data was also available. To maximize the PacBio single-molecule sequence accuracy, libraries of 2 to 3 kb were generated, allowing many polymerase passes per molecule. The resulting reads were filtered at predicted single-molecule accuracy levels up to 99.99%. Comparison of the community compositions of the 5 samples determined with PacBio sequence data versus Illumina WGS assemblies from the same set of samples indicate that rare organisms were often missed with Illumina. These results illustrate ways in which long accurate reads benefit analysis of complex communities.

(P43) TruePrime™, a unique primer-free MDA technology for single cell amplification with low bias and superior variant recovery

Florent Fordoxel[1], Angel Picher[2], Bettina Budeus[3], Luis Blanco[4], Armin Schneider[3]

[1] Expedeon, Cambridge, United Kingdom

[2] SYGNIS Biotech SLU, Madrid, Spain

[3] SYGNIS Bioscience GmbH & Co KG, Heidelberg, Germany

[4] Centro de Biología Molecular Severo Ochoa, Madrid, Spain

TruePrime™ is a novel MDA (multiple displacement amplification)-type DNA amplification method that utilizes a primase activity (TthPrimPol) rather than random hexamers to generate DNA-primers. TthPrimPol is a monomeric enzyme (34 kDa) that displays a potent primase activity, preferring dNTPs as substrates unlike conventional primases. This DNA primase activity shows a wide sequence specificity for template recognition. In this setup, TthPrimPol synthesizes the DNA primers needed for Phi29 DNA pol, which allows for the exponential amplification of genomic DNA. Key advantages of the TruePrime™ technology for amplification of single cell genomes include complete absence of primer artefacts, superior sensitivity down to the femtogram range, and an easy reaction workflow. Analyses on genomic DNA amplified from single Hek293 cells in comparison with non-amplified DNA, the commercially available MDA methods based on random synthetic primers and MALBAC, reveal superior breadth and evenness of genome coverage, high reproducibility, excellent single nucleotide variant (SNV) detection rates with low allelic dropout (ADO) and low chimera formation. Moreover, copy number variant (CNV) calling yields superior results than random primer-based MDA methods. The advantages of this method promise to facilitate and improve single cell genomic analysis.

(P44) Expression dynamics of HTLV-1 at the single-cell level

Jocelyn Turpin, **Anat Melamed**, Ashleigh Lister, Wilfried Haerty, Iain Macaulay, Charles R. M. Bangham

Human T-cell lymphotropic virus type 1 (HTLV-1) is a human retrovirus which causes a life-long persistent infection of T-cells. HTLV-1 infection is often considered latent, with viral propagation dominated by cell division of the host cell, generating expanded, long-lived clones of infected T cells; each clone defined by a unique viral integration position in the host genome.

At a population level, HTLV-1 is clearly not latent – strong, constitutive immune response against viral epitopes suggests the presence of recurrent viral expression. We postulate that the HTLV-1 provirus is expressed *in vivo* in frequent but intermittent bursts, with only a minority of cells expressing viral genes at a given time. The mechanisms regulating the timing and duration of expression in a given cell are not known.

This project aims to utilize single cell sequencing to study the causes and consequences of cell-to-cell heterogeneity in viral latency, in terms of expression of both viral and host genes.

As a proof of concept, we initially analysed single cell mRNA from T-cell clones isolated by limited dilution from two HTLV-1 infected individuals, including both cells from infected (carrying a provirus at a single integration site) clones and cells from an uninfected clone. We will present preliminary results on the gene expression unique to each cell population in order to dissect the inter- and intra-clone heterogeneity. We are now aiming to extend this to *ex vivo* polyclonal cell samples.

(P90) Single cell recombination screening in wheat pollen

Ashleigh Lister[1], Ned Peel[1], Azahara Martin[2], Lola Santome[2], Peter Shaw[2], Graham Moore[2], Matt Clark[1], Iain Macaulay[1]

[1] Earlham Institute, Norwich Research Park, Norwich

[2] John Innes Centre, Norwich Research Park, Norwich

Recombination is a process which takes place during meiosis, a system necessary during gametogenesis in all sexual organisms. The two sets of replicated parental chromosomes undergo crossover formation followed by two rounds of segregation, to produce haploid gametes ready for fertilisation. Variation in the offspring population, caused by meiotic recombination as well as random segregation, is highly important as it creates a fitness in the event of a change in environment or pathogen attack, as well as creating genetic diversity required in agricultural plant breeding.

We are developing a process to screen wheat meiocytes at the single cell level to investigate the extent of meiotic recombination. Using FACS, we isolate single meiocytes into individual wells of 96 well plates for subsequent whole genome amplification and sequencing library preparation. Here, we present preliminary data from shallow sequencing of individual meiocytes, exploring chromosome coverage and recombination events in pollen cells from hybrid plants.

Further development of this approach could generate a rapid low-cost/high-throughput method of scoring new hybrids, and to screen for treatments which affect meiotic recombination. Identifying hybrids or treatments have elevated recombination events would be a powerful tool for breeding or programmes and for future research.

(P45) Single-molecule real-time sequencing and long-range scaffolding improves the contiguity of the elephant shark genome assembly

Prashant Shingate[1], Nisha Pillai[1], Vydianathan Ravi[1], Byrappa Venkatesh[1]

[1] Comparative and Medical Genomics Lab, Institute of Molecular and Cell Biology, A*STAR, Biopolis, Singapore

The slow-evolving relatively small genome (~1 Gb) of the elephant shark (**Callorhinchus milii**), a member of the oldest extant group of jawed vertebrates (the cartilaginous fishes), is a valuable reference genome for comparative genomic studies of vertebrates. We have previously generated a draft genome assembly of the elephant shark based mainly on 454 sequences, with N50 contig and scaffold sizes of 46 kb and 4.5 Mb, respectively. This assembly, however, contains ~67,000 gaps and some important gene loci such as the MHC loci are fragmented. The advent of single-molecule real-time sequencing platforms such as PacBio has made it possible to achieve highly contiguous de novo assemblies. We have now generated a PacBio-only assembly of the elephant shark genome followed by scaffolding using Dovetail's Chicago library and Hi-C data. The N50 contig length of the PacBio assembly is 30-fold longer than that of the 454 assembly whereas the N50 scaffold length is 6-fold longer than the 454 assembly. Six of the PacBio assembly scaffolds account for half of the assembled genome indicating that they are likely to represent full-length chromosomes. The highly heterozygous genomes of outbred populations such as sharks and fishes with high content of dispersed tandem repeats pose major problems in generating highly contiguous PacBio assemblies such as those of birds and mammals. We have overcome these problems to some extent by developing in-house programs followed by manual curation that considerably improved the contiguity of PacBio contigs.

(P46) Scaling annotation of vertebrate genomes

Fergal Martin[1], Carlos García Girón[1], Leanne Haggerty[1], Thibaut Hourlier[1], Osagie Izuogu[1], Konstantinos Billis[1], Daniel Murphy[1], Rishi Nag[1], Paul Flicek[1], Bronwen Aken[1]

[1] European Molecular Biology Laboratory, European Bioinformatics Institute, Wellcome Genome Campus, Hinxton, CB10 1SD, UK

There are many efforts currently underway to expand the breath and depth of assembled vertebrate genomes. These range from haploid and diploid human assemblies, to strains of rodent species, to the Vertebrate Genome Project's goal of sequencing at least one individual from every vertebrate species. These genome assemblies will be a valuable reference resource for the life sciences. Additional complementary reference data sets are needed for each assembly, for example consistent gene annotation and comparative analyses that will enable meaningful scientific studies. Ensembl is a leading source of genome annotation for vertebrates, and our goal is to annotate the coming influx of assemblies. Over the past two years, we have significantly redeveloped our gene annotation system to allow us to scale. We can now produce consistent annotation on multiple assemblies in parallel. The new annotation system is highly automated, and is capable of producing a gene set in less than two weeks. The system builds on our extensive knowledge of genome annotation and uses species-specific transcriptome data (where available), protein-to-genome alignments, and mapping of annotation from a suitable high-quality reference annotation from within the clade of interest. We are currently trialing this system on rodent and primate genomes, using the mouse and human GENCODE reference gene sets.

(P47) Convergences of Avian Vocal Learning Clades on cAMP-based Vocal Learning Pathway – Not How Many, But Where

Chul Lee[1], Seoae Cho[2], Kyu-won Kim[3], Hong Jo Lee[1], Jae Yong Han[1], Osceola Whitney[4], Andreas Pfenning[5], Miriam V. Rivas[6], Erina Hara[7,8], Peter V. Lovell[9], Claudio V. Mello[9], Guojie Zhang[10,11], Dave W. Burt[12], Heebal Kim[1,2,13], Erich D. Jarvis[7,14]

[1] Seoul National University

[2] Cho & Kim genomics

[3] Kongju National University

[4] New Mexico State University

[5] Carnegie Mellon University

[6] Durham Technical Community College

[7] Howard Hughes Medical Institute

[8] Duke University Medical Center

[9] Oregon Health and Science University

[10] University of Copenhagen

[11] BGI

[12] Roslin Institute, University of Edinburgh

[13] Shinshu University

[14] Rockefeller University

Vocal learning is the ability to imitate vocalizations based on auditory experience. It is a homoplastic character state observed in polyphyletic lineages of animals such as songbirds, parrots, hummingbirds and human, but is not observed in most animals including chicken, duck, and chimpanzee. It has now become possible to perform proteome-wide molecular analyses across vocal learners and vocal non-learners with the recent expansion of avian genome data. Here we analyzed the whole genome of avian species including those belonging to the three vocal learning clades. We aimed to determine if behavior and neural convergence is associated with molecular convergence in polyphyletic avian vocal learners. We found molecular convergences are correlated to products of origin branch lengths. We discovered convergences/divergences ratio of vocal learners does not exceed other control sets in avian lineages, but illuminated the function of homoplastic genes specific to avian vocal learners was enriched for learning. Out of the learning genes, key candidate genes of vocal learning were identified by supports of multiple evidences: positive selection, fixed differences, gene expression profile in songnuclei of a vocal learner, and relationships with human diseases causing language disorder. Moreover, human also had unique substitutions different from non-human primates in the candidate genes. Our findings suggest a novel cAMP-based vocal learning pathway, indicating molecular homoplastic changes associated with a complex behavioral trait, vocal learning.

(P48) De novo genome and comparative evolutionary genomic analysis of the common lizard, *Zootoca vivipara*

Andrey Yurchenko[1], Hans Recknagel[1], **Kathryn Elmer**[1]

[1] IBAHCM, Univ. of Glasgow

Squamate reptiles are important ecological and evolutionary model taxa as they have evolved a dramatic breadth of adaptations, including limblessness, live-bearing and parthenogenic reproduction, broad climatic tolerance, and venom. However there are relatively few lineages with reference genomes available – a critical resource for identifying the genetic basis and evolutionary history of these fascinating traits. We are developing a high-quality annotated and oriented reference genome for the Eurasian common lizard, *Zootoca vivipara* (Lacertidae). This species has the most northerly distribution of any reptile and is reproductively bimodal, with some lineages being live-bearing and others egg-laying. The genome was sequenced with high-coverage Illumina short-reads and iteratively scaffolded with mate-pair libraries, Pacbio long reads, and RNA-seq data. Total assembly length is 1.44 Gbp (current scaffold N50 = 12.51 Mbp, 2.9% gaps). To orient scaffolds and arrange them into linkage groups, we genotyped mothers and clutches of *Z. vivipara* with ddRAD-seq for genetic mapping. Protein-coding gene annotation on the repeat-masked assembly was done by combining de-novo prediction with homology-based and RNA-seq evidence, yielding 15,257 genes. Benchmarking using single-copy orthologs to Tetrapoda revealed high completeness of the genome, with 96.8% of single-copy orthologs (including 90.2% completely assembled) being identified. Comparative genomic analysis of *Z. vivipara* was conducted against available Reptilia genomes to identify genome rearrangements, repeat dynamics, and analyze gene family expansions and contractions. As far as we are aware this is the first reference genome for a lacertid and therefore fills an important gap for genome information in this and related lineages.

(P49) Transcriptome analysis of metabolic bone symptom in Siamese Crocodile (*Crocodylus siamensis*)

Prapatsorn Areesirisuk[1], Soyoung Song[2], Aorarat Suntronpong[1], **Worapong Singchat**[1], Tanawut Srisuk[3], Narongrit Muangmai[4], Sasimanas Unajak[5], Yosapong Temsiripong[6], Surin Peyachoknagul[1], Chinae Thammarongtham[7], Seyoung Mun[2], Kyudong Han[2], Kornsorn Srikulnath[1]

[1] Laboratory of Animal Cytogenetics and Comparative Genomics, Department of Genetics, Faculty of Science, Kasetsart University, 50 Ngamwongwan, Chatuchak, Bangkok 10900, Thailand

[2] Department of Nanobiomedical Science & BK21 PLUS NBM Global Research Center for Regenerative Medicine, Dankook Univerisity, Cheonan 31116, Republic of Korea

[3] Pilot Plant Development and Training Institute, King Mongkut's University of Technology Thonburi, Bang Khun Thian, 49 Soi Thian Thale 25, Bang Khun Thian Chai Thale Rd, Tha Kham, Bangkok, 10150, Thailand

[4] Department of Fishery Biology, Faculty of Fisheries, Kasetsart University, 50 Ngamwongwan, Chatuchak, Bangkok 10900, Thailand

[5] Department of Biochemistry, Faculty of Science, Kasetsart University, 50 Ngamwongwan, Chatuchak, Bangkok 10900, Thailand

[6] Sriracha Moda Co., Ltd., Sriracha, Chonburi 20110, Thailand

[7] Biochemical Engineering and Pilot Plant Research and Development Laboratory, National Center for Genetic Engineering and Biotechnology (BIOTEC), King Mongkut's University of Technology Thonburi, Bang Khun Thian, 49 Soi Thian Thale 25, Bang Khun Thian Chai Thale Rd, Tha Kham, Bangkok, 10150, Thailand

Siamese crocodiles (*Crocodylus siamensis*) are economically important animals in the Thai agricultural industry. Recently, metabolic bone symptoms (osteoporosis, cretinism and crooked bone, generally in spines) have been found in captive crocodiles, negatively impacting crocodile industry products. Diagnosis of these symptoms has not been determined as either genetic disorder or microbial infection. Here, transcriptomic analysis was performed using the Illumina platform with one normal and three abnormal crocodiles. RNA sequences of the one normal and three abnormal crocodiles were generated as 5.93 – 7.28 Gb. DEGseq (TCC package) R program analysis showed that 293 genes were overexpressed and 4,109 genes were under-expressed. Notably, DHFR, LRRC8D, PSMB2, MASTL, FBSO22, GGT1, AMOTL2, MPDZ, TMEM72, PLEKHA3, and DRAM2 genes indicated significantly down regulated genes in severe symptoms of the three afflicted crocodiles. Gramma-glutamyltranspeptidase 1 (GGT1) plays a role in calcium dynamics in kidney proximal tubules as a calcium sensing receptor. The transcriptome analysis of normal and metabolic bone symptoms of crocodiles will provide a valuable resource for identifying genes and symptom origins for diagnosis and treatment to improve farming and breeding management. This data will also be useful for Siamese crocodile genome annotation which is currently ongoing.

(P50) Whole genome sequencing of Siamese crocodile, *Crocodylus siamensis*, to extensively investigate genome structure and rearrangement in the lineage of crocodiles

Kornsorn Srikulnath[1,2,3*] Prapatsorn Areesirisuk[1,2], Worapong Singchat[1,2], Aorarat Suntronpong[1,2], Tanawut Srisuk[4], Narongrit Muangmai[5], Sasimanas Unajak[6], Yosapong Temsiripong[7], Surin Peyachoknagul[1], Chinae Thammarongtham [8], Seyoung Mun[9], Kyudong Han[9]

[1] Laboratory of Animal Cytogenetics and Comparative Genomics, Department of Genetics, Faculty of Science, Kasetsart University, 50 Ngamwongwan, Chatuchak, Bangkok 10900, Thailand

[2] Animal Breeding and genetics consortium of Kasetsart University, 50 Ngamwongwan, Chatuchak, Bangkok 10900, Thailand

[3] Center for Advanced Studies in Tropical Natural Resources, National Research University-Kasetsart University, Thailand (CASTNAR, NRU-KU, Thailand), Kasetsart University, Bangkok, Thailand

[4] Pilot Plant Development and Training Institute, King Mongkut's University of Technology Thonburi, Bang Khun Thian, 49 Soi Thian Thale 25, Bang Khun Thian Chai Thale Rd, Tha Kham, Bangkok 10150, Thailand

[5] Department of Fishery Biology, Faculty of Fisheries, Kasetsart University, 50 Ngamwongwan, Chatuchak, Bangkok 10900, Thailand

[6] Department of Biochemistry, Faculty of Science, Kasetsart University, 50 Ngamwongwan, Chatuchak, Bangkok 10900, Thailand

[7] Sriracha Moda Co., Ltd., Sriracha, Chonburi 20110, Thailand

[8] Biochemical Engineering and Pilot Plant Research and Development Laboratory, National Center for Genetic Engineering and Biotechnology (BIOTEC), King Mongkut's University of Technology Thonburi, Bang Khun Thian, 49 Soi Thian Thale 25, Bang Khun Thian Chai Thale Rd, Tha Kham, Bangkok 10150, Thailand

[9] Department of Nanobiomedical Science & BK21 PLUS NBM Global Research Center for Regenerative Medicine, Dankook Univerisity, Cheonan 31116, Republic of Korea

[*] Corresponding author: kornsorn.s@ku.ac.th

A genome sequencing project concerning Siamese crocodiles (*Crocodylus siamensis*), economically important animals in the Thai agricultural industry, is currently ongoing to determine the relationship between genes, biological systems, and economic traits. Whole genome sequencing of Siamese crocodile was performed using the Illumina platform which generated 797,800,548 reads of 150-base paired-end sequences. After low quality read filtering, 769,892,900 clean reads were obtained from total bases of 113,390,323,500. The GC content of the sequence data was 45%. By comparison with the genome sequence of Chinese alligator (*Alligator sinensis*) using the Burrows-Wheeler Aligner mapping tool, 48.58% of the clean read were mapped with an average depth of 22x onto the Chinese alligator genome. Numbers of INDELs and SNPs were 7,677,890 and 118,263,137, respectively, in the mapped Siamese crocodile genome sequence. Single nucleotide variants between Chinese alligator and Siamese crocodile were also found in intergenic, intron, downstream, upstream, and exon regions as 53.14%, 36.26%, 4.30%, 4.12%, and 0.77%, respectively. Comparative genomics with *Alligator mississippiensis*, *Gavialis gangeticus* and *Crocodylus porosus* will investigate different genome structure and rearrangement in the lineage of crocodiles. A comprehensive genome map for Siamese crocodile will provide a valuable resource for identifying genes with economic traits in captive crocodiles to improve farming and conservation methods, and address many biological questions in diverse vertebrates.

Keywords: Siamese crocodile, Chinese alligator, whole genome sequence

(P51) The genome sequence of Danube salmon (*Hucho hucho*): a key lineage for understanding salmonid fish evolution

Manu Kumar Gundappa[1], Simen R. Sandve[2], David Hazlerigg[3], Jürgen Geist[4], Samuel A. M. Martin[5], Daniel J. Macqueen[5]

[1] Institute of Biological and Environmental Sciences,University of Aberdeen

[2] Centre for Integrative Genetics (CIGENE), Faculty of Biosciences, Norwegian University of Life Sciences, Ås NO- 1432, Norway

[3] Department of Arctic and Marine Biology, Faculty of BioSciences Fisheries & Economy, University of Tromsø, Norway

[4] Aquatic Systems Biology Unit, School of Life Sciences Weihenstephan, Technical University of Munich, Mühlenweg 22, D-85354 Freising, Germany

[5] Institute of Biological and Environmental Sciences, University of Aberdeen

Many salmonid species migrate from freshwater into the marine environment, a life-history strategy called anadromy. However, the ancestral state was a purely freshwater life-cycle, which several lineages have retained. The Danube salmon is an endangered member of one such freshwater clade, which is evolutionarily sister to a more well-known group (including Atlantic and Pacific salmon spp.) that evolved anadromy ancestrally. The Danube salmon thus holds a key phylogenetic position to reconstruct the genomic and evolutionary basis of adaptations underpinning anadromy. We are also trying to understand the role of a salmonid-specific whole genome duplication (WGD) in the diversification of these commercially and ecologically important fish. My presentation reports a high-quality, annotated genome for the Danube salmon. A haploid individual was sequenced to improve assembly contiguity by removing heterozygosity. Paired-end and mate-pair libraries were sequenced at ~80/40x respective coverage (2x250bp reads). K-mer analysis predicted a 2.3 Gb genome, in line with other sequenced salmonids. Contig assembly/scaffolding was done using the W2RAP assembler, which outperformed other tested pipelines. Annotation was performed using MAKER, run on trained ab initio gene predictions from SNAP, along with transcript evidence gained by sequencing ~450M 2x150bp RNAseq reads, a de-novo predicted library of repeats, and the UniProt database as protein evidence. The Danube salmon genome will be used for comparative analysis of different salmonid genomes, as part of the 'Functional Annotation of All Salmonid Genomes' (FAASG) initiative, including to characterize the evolution of duplicated genes retained from WGD and their role in the evolution of anadromy.

(P53) The genome of Southern Grasshopper Mouse (*Onychomys torridus*) genome as a model for studying aggressive behaviour

Jingtao Lilue[1], Laura Masullo[2], Dirk-Dominik Dolle[1], Shane McCarthy[1], Marco Tripodi[2], David Thybert[3], Thomas Keane[4]

[1] Wellcome trust Sanger Institute

[2] MRC Laboratory of Molecular Biology

[3] Earlham Institute

[4] European Bioinformatics Institute

The Grasshopper mouse (*Onychomys* of family Cricetidae) is a new world mouse genus endemic to North America, only distantly related to the common house mouse, *Mus musculus*. Grasshopper mice display several unique behavioural and morphological features, e.g. they are highly aggressive carnivores who stalk their prey in the manner of cats, and defend their territory by "howling" like wolves. The grasshopper mouse is known to prey on insects, centipedes, scorpions, snakes and even other mice, and has evolved immunity to venoms released by its prey. In this project, we are using third generation long read sequencing techniques (Pacbio Sequel sequencing, 10x genomics, and Bionano optical maps) to create a reference quality genome assembly of the Southern grasshopper mouse (*Onychomys torridus*), with 23 pairs of autosomes compared to 19 in house mouse. We will perform a comparative genomics analysis to study their aggressive behavior by comparing its genome with a set of non-aggressive rodent species such as house mouse, Chinese hamster, and deer mouse. In addition, we have generated RNA-Seq from a set of regions of the brain involved in aggression response for both laboratory mouse (non-aggressive) and Grasshopper mouse (aggressive) to aid in the identification of candidate genes. Based on these results, we will use CRISPR techniques in both species to assess functional effects of candidate mutations on the aggression phenotype.

(P54) Nonhuman Primate Annotation in Ensembl

Leanne Haggerty[1], Konstantinos Billis[1], Carlos Garcia Giron[1], Thibaut Hourlier[1], Osagie Izuogu[1], Daniel Murphy[1], Rishi Nag[1], Fergal Martin[1], Paul Flicek[1], Bronwen Aken[1]

[1] EMBL-EBI

In response to the multitudinous large-scale genome sequencing efforts, the Ensembl gene annotation system has been rebuilt to provide high-quality gene annotations in a matter of days. The approach, based on a combination of RNA-seq alignments, annotation projection via whole genome alignments and protein-to-genome alignments using selected UniProt proteins, allows us to perform clade-based annotations with consistency and efficiency. Owing to this, the annotated assemblies of several primates will become available in Ensembl (version e91). A major focus of genomic research is human genetics and its relevance to disease. Genomic analyses of nonhuman primates provide crucial information for understanding the evolutionary history of humans and to expound the genetic basis of human disease. In addition to the updated gene set, we will also ensure that all other nonhuman primate resources in Ensembl are up-to-date, including: genome variation resources, gene names, gene-trees and orthologues, cross-references to external databases including UniProt and RefSeq, vertebrate multiple whole genome alignments, and conserved and constrained elements. Where available, annotated genomes will include RNA-seq data, which can be viewed on the Ensembl genome browser. Data and tools to facilitate research on nonhuman primates will be accessible through our website (www.ensembl.org), REST API (http://rest.ensembl.org), Variant Effect Predictor (www.ensembl.org/Tools/VEP), BioMart (http://www.ensembl.org/biomart) and our public MySQL server (ensembldb.ensembl.org).

(P55) High quality genome assemblies for zebrafish and relatives

Kerstin Howe[1], William Chow[1], Joanna Collins[1], Gregory Concepcion[2], Sarah Pelan[1], Michael Quail[1], Michelle Smith[1], Glen Threadgold[1], James Torrance[1], Jonathan Wood[1], Shane McCarthy[1], Jason Chin[2], Gene Myers[3], Richard Durbin[1]

[1] Wellcome Trust Sanger Institute, Hinxton, Cambridge, UK

[2] Pacific Biosciences, Menlo Park, CA, USA

[3] Max Planck Institute of Molecular Cell Biology and Genetics, Dresden, Germany

The advent of long-range sequencing and mapping technologies is driving the generation of an ever-increasing range of high quality vertebrate genome assemblies in a time- and cost-effective manner. The Sanger Institute Vertebrate Genome Project (VGP) is taking advantage of the recent developments to produce high quality genome assemblies for diverse fish, rodents and caecilians. As part of the VGP, we are sequencing multiple *Danio rerio* strains and several additional species of the Danioninae subfamily to provide additional context to the reference genome of zebrafish, an important model organism. The project started in December 2016, and has so far seen 17 *Danio rerio* strains/populations, 7 additional Danio species and 2 other close relatives added to the sequencing pipeline, with samples being kindly provided by collaborators listed below. We are trialling a variety of sequencing and assembly technologies and quality check and improve the resulting assemblies using a multi-stage workflow that includes curation using the Genome Assembly Evaluation Browser gEVAL (geval.sanger.ac.uk). Particular attention is being paid to generating a haplotype-resolved high quality assembly of the SAT strain, a cross between double-haploid parents from Tuebingen and AB. These new genomes enable a powerful comparative approach to the study of zebrafish genetics, development and Danioninae evolution. Sample providers: Ralf Britz, Uwe Irion, Braedan McCluskey, Elisabeth Busch-Nentwich, Lukas Ruber, Bill Trevarrow, Zoltan Varga and Andrew Whiteley

(P56) Multiple mouse reference genomes defines subspecies specific haplotypes and novel coding sequences

Anthony Doran[1], Mouse Genomes Project[2],

[1] Wellcome Trust Sanger Institute

[2] The Mouse Genomes Project consortium

The Mouse Genomes Project has completed the first draft assembled genome sequences and strain specific gene annotation for twelve classical laboratory and four wild-derived inbred mouse strains (WSB/EiJ, CAST/EiJ, PWK/PhJ, and SPRET/EiJ). We used a hybrid approach for genome annotation, combining evidence from the mouse reference Gencode annotation and strain-specific transcript evidence (RNA-seq and PacBio cDNA), to identify novel strain-specific gene structures and alleles. As these strains are fully inbred, we used heterozygous SNP density as a marker for highly polymorphic loci, and found these loci to be enriched for genes related to immunity, olfaction and sensory function. We focus in particular on four immune related loci (IRG, Nlrp1, Schlafen and Raet1) containing novel sequence, coding alleles and diverse gene structures in the wild derived strains. In mouse, anthrax lethal toxin is currently the only known activator of Nlrp1 and genetic differences in Nlrp1b are linked to response sensitivity. For the first time, our data shows the striking allelic diversity in this locus, identifying new coding alleles shared by subsets of the susceptible and resistant strains. At another locus on Chr10, we identified novel strain-specific allelic combinations of H60 and Raet1 homologs that segregate with susceptibility to Aspergillus infection. Of particular note was the discovery of a previously unannotated rodent specific 138 exon gene on Chr11. Manual annotation extended this novel gene as a combination of the human genes EFCAB3 and EFCAB13 on human Chr17. The genome sequences and annotation can be viewed in the UCSC and Ensembl genome browsers.

(P57) The Oz Mammals Genomics (OMG) Initiative: Developing genomic resources for Australian mammals

Janine Deakin[1]

[1] University of Canberra

Australia has an incredibly diverse range of mammals, with species spanning all three major mammalian lineages. They are a treasure trove of interesting biology, enabling the evolution of mammalian-specific or lineage-specific traits to be uncovered. Unfortunately, Australia also has the highest mammal extinction rate in the world and many extant Australian mammals are listed as threatened. Consequently, we require a comprehensive understanding of the relationships of Australian mammals, including recently extinct species, to underpin studies of their evolution, as well as improve our understanding of extinction risk. It is now time that we capitalised on the advances in genome sequencing technology and bioinformatic analyses to enhance our understanding and conservation of Australia’s unique mammals. To facilitate the uptake of genomics for the conservation of Australia’s diverse and unique mammalian species, we have formed the Oz Mammals Genomics (OMG) consortium. The consortium consists of over 30 partners, including Australia’s natural history museums, university researchers and wildlife management agencies, with a one million dollar investment from BioPlatforms Australia, with co-investment from universities and museums nationwide. OMG aims to produce well-assembled marsupial genomes from representative species across the marsupial phylogeny, generate a comprehensive phylogenomic framework for Australian mammals (bats, rodents, marsupials) that includes recently extinct species, and generate population genomic datasets for threatened Australian species to inform conservation and management programs. This presentation will discuss our progress to date towards each of these three aims.

(P58) Tissue-specific enhancer and promoter evolution in mammals

Maša Roller[1], Ericca Stamper[2], Diego Villar[2], Aisling Redmond[2], Duncan T. Odom[2], Paul Flicek[1]

[1] European Molecular Biology Laboratory, European Bioinformatics Institute, Wellcome Genome Campus, Hinxton, Cambridge,CB10 1SD, UK

[2] University of Cambridge, Cancer Research UK Cambridge Institute, Robinson Way, Cambridge, CB2 0RE, UK

Gene expression is established through spatiotemporal coordination of regulatory elements including enhancers and promoters. Tissue-specific gene expression is strongly associated with enhancer function. Enhancers have already been shown to have a fast evolution in mammalian liver. Here, we investigate the evolution of enhancers and promoters in four different tissues of ten species of mammals by profiling H3K27 acetylation, H3K4 monomethylation, and H3K4 trimethylation. The four tissues - liver, muscle, brain and testes - provide in vivo snapshots of diverse regulatory patterns in reproductive and somatic tissues of different functions and developmental origins. The ten mammals cover primates, carnivores, artiodactyla, lagomorphs, and rodents. Within this framework we compare the rates of evolution both between mammalian linages and between tissues. We investigate whether species-specific regulatory elements are also more likely to be tissue-specific. These results provide important insights into tissue-specific mammalian regulatory evolution.

(P59) Virtual Genome Walking across the repeat-rich 30Gb axolotl genome

Teri Evans[1]

[1] University of Nottingham

Axolotl (*Ambystoma mexicanum*) is the salamander model species used to study basal vertebrate embryogenesis and limb regeneration. Although this key model species is widely used, certain experiments are impossible without a reference genome sequence. Unfortunately, all salamanders have undergone genome gigantism, such that the axolotl genome is approximately 30Gb spread across 14 chromosomes. The longest chromosome is estimated to be longer than the entire human genome combined. Furthermore the genome is repeat-rich, almost 15Gb is thought to be high copy repeats. As such, previous attempts to sequence the axolotl genome have proved unsuccessful, although 20X of short-read data was obtained, conventional assembly methods are simply unable to cope with the high level of complexity. We have developed a methodology that allows us to extend transcript sequences into genomic space by repetitively mapping and locally assembling genome reads. We have run this Virtual Genome Walking (VGW) program on over 20,000 axolotl transcripts, generating gene models which include intron and promoter sequence. We are now able to conclusively identify transcript variants, gene duplications and in a few cases, local synteny. We are making these genome scaffolds publically available, providing a unique resource for anyone using axolotls as a model system. Furthermore, VGW can be used to assemble genic regions for any large, repeat-rich genome that cannot be assembled using conventional techniques.

(P60) Genome-wide association studies to identify loci and variants associated with behavioral traits in dogs

Voichita D. Marinescu[1], Marcin Kierczak[1], Katarina Tengvall[1], Jagoda Jabłońska[1], Per Arvelius[2], Sharadha Sakthikumar[1], Fabiana Farias[1], Erling Strandberg[2], Erik Wilson[3], Åke Hedhammar[4], Kerstin Lindblad-Toh[1,5]

[1] Department of Medical Biochemistry and Microbiology and Science for Life Laboratory, Uppsala University, Uppsala, Sweden

[2] Department of Animal Breeding and Genetics, Swedish University of Agricultural Sciences, Uppsala, Sweden

[3] Swedish Armed Forces

[4] Department of Clinical Sciences, Swedish University of Agricultural Sciences, Uppsala, Sweden

[5] Broad Institute of MIT and Harvard, Boston, USA

The domestic dog (*Canis familiaris*) represents a promising genetic and genomic model for identifying the genes and variants underlying behavioral traits given that the many breeds available today are the result of a careful selection of desired traits including behavioral ones. Moreover, in Sweden, behavioral tests such as the Dog Mentality Assessment (DMA) that is used to evaluate the suitability of a dog for a desired purpose (e.g. service in the police forces) and quantifies several aggregated personality traits (aggressiveness, chase proneness, curiosity/fearlessness, playfulness, sociability) based on a number of standardized subtests, have become widely used also for privately owned dogs. As a result, it is now possible to perform genome-wide association studies (GWAS) for a large number of dogs across different types of breeds for which DMA results are known. In a GWAS mapping behavioral traits across a well-defined population of 466 Swedish German shepherd dogs, we found two loci that were genome-wide significantly associated with playfulness, that point towards genes involved in synaptic plasticity, axonal navigation and nervous system development. We have extended this study to include two additional populations of 160 Rottweilers and 273 Rough Collies that are popular breeds displaying a relatively large variation in behavior. This analysis is currently ongoing. By understanding the genetic variation associated with a particular behavioral trait in a given breed, we hope to identify markers that could be used for selection and to gain insights into the molecular mechanisms underlying canine behavior.

(P62) Chromosome organization and tissue specific gene expression patterns in chicken

David Martín-Gálvez[1], Duncan Odom[2], Paul Flicek[1]

[1] European Molecular Biology Laboratory, European Bioinformatics Institute

[2] University of Cambridge, Cancer Research UK - Cambridge Institute

The causes behind chromosomal organization and their effects on the evolutionary process are not fully understood yet. Birds are an interesting group to address such questions as they have a typical karyotype that is fairly conserved across the evolution of the group. The typical avian karyotype consists of 6-7 pairs of macrochromosomes, a pair of sex chromosomes and 30-32 pairs of microchromosomes. The existence of mechanisms preventing abrupt changes in the karyotype of birds have previously been suggested. One possibility is that the different types of chromosomes in birds provide the specific genomic environment to ensure an optimal level of expression necessary for the type and function of the genes that they contain. This could be especially relevant for those genes requiring similar regulation, such as tissue-specific genes and house-keeping genes. We address this question in chicken by comparing at the chromosomal level the tissue specificity (using Shannon entropy) of existing gene expression (RNA-seq) data from 21 tissues. Chicken macrochromosomes had smaller values of gene entropies (i.e. variable gene expression among tissues) and less tissues-specific genes than expected. In contrast, chicken microchromosomes show greater gene entropies (i.e. uniform gene expression among tissues) and more house-keeping genes than expected. Sexual chromosomes were positively enriched with tissues-specific genes. Our results suggest a link between the morphology of chromosomes and regulation of the expression of their genes. We are now investigating possible mechanisms underpinning this link.

(P63) When less is more: the role of gene loss in phenotypic diversification

Filipe Castro[1], Mónica Lopes-Marques[1], Miguel Fonseca[1], Susana Barbosa[1], André Machado[1], Raquel Ruivo[1]

[1] CIIMAR-UPorto

Evolutionary biology seeks to unravel the link between the phenotypic diversity and the genetic makeup of different lineages. These can include for example variable gene complements, or subtle/drastic mutations in the coding region of genes. The vertebrate sub-phylum comprises a considerable number of lineages with tantalizing morphological and physiological adaptations to particular environments. In many cases, the evolution of novel traits (e.g. teeth, hair, taste, digestion) has been clearly linked to gene duplication, with the retention of descendent gene copies leading to the emergence of novel roles. Much less explored is the role of gene loss (pseudogenization) and the ensuing phenotypic consequences. Here, I will present a few examples combining comparative genomics with functional assays to show that gene loss has been a prolific event underscoring adaptation. This approach puts into an evolutionary context the structural and functional dynamics of vertebrate genomes.

(P64) *Acomys* genomes project: A comparative genomic framework for evolutionary and biomedical studies

David Wright[1], Dirk-Dominik Dolle[2], Shane McCarthy[2], John Muturu Kimani[3], Ashley Seifert[4], Noga Kronfeld-Schor[5], Thomas Keane[6], David Thybert[1]

[1] Earlham institute, Norwich Research Park, Norwich UK

[2] Sanger Institute, Welcome Trust Genome Campus, Hinxton, Cambridge, UK

[3] Department of Veterinary Anatomy, University of Nairobi, Chiromo Campus, Nairobi, Kenya

[4] Department of Biology, University of Kentucky, Lexington, United States.

[5] Department of Zoology Faculty of Life Sciences, Tel-Aviv University, Tel Aviv, Israel

[6] European Bioinformatics Institute, Welcome Trust Genome Campus, Hinxton, Cambridge, UK

The *Acomys* or spiny mouse, are rodents from the family of Muridae. These close relatives to the laboratory mouse carry extreme phenotypes of high interest in evolutionary or biomedical studies. For instance, the desert adapted species *Acomys russatus* has the possibility to drink sea water and still maintaining kidney function. It can contract type 2 diabetes and become obese when kept in captivity. *Acomys cahirinus* is the only rodent for which menstruation has been described, making it the only laboratory rodent that can be used to study menstrual disorder in women. *Acomys* are also the only known mammals able to regenerate tissues and making them the closely human related organism that can be used as a model for regenerative medicine. In order to study these phenotype at the genomic level, we undertook to sequence chromosome level assembly of the genome of four *Acomys* species (*A. cahirinus, A. russatus, A. kempi, A. percivali*) and one close related outgroup (*Psammomys obesus*). We combined Pacbio and 10x long reads technologies with Illumina short reads to optimise contiguity. The comparison of these genomes and the analysis of the lineage and species-specific loci within the Acomys genus will help to find the genomic loci associated to these phenotypes and understand the evolutionary mechanisms involved in shaping them.

Evolutionary Genomics

(P66) Gross genomic reconstruction of evolutionary events suggests that the genome organisation of dinosaurs closely resembled that of modern birds

Rebecca E. O'Connor[1], Michael N. Romanov[2], Paul Barrett[3], Marta Farré[4], Joana Damas[4], Malcolm Ferguson-Smith[5], Nicole Valenzuela[6], Denis M. Larkin[4], **Darren Griffin**[2]

[1] University of Kent

[2] School of Biosciences, University of Kent, Canterbury

[3] Department of Earth Sciences, Natural History Museum, Cromwell Road, London SW7 5BD

[4] Department of Comparative Biomedical Sciences, Royal Veterinary College, University of London, London

[5] Department of Veterinary Medicine, Cambridge University, Cambridge

[6] Department of Ecology, Evolution, and Organismal Biology, Iowa State University

Non-avian dinosaurs remain subjects of intense biological enquiry while pervading popular culture and the creative arts. While organismal studies focus primarily on their morphology, relationships, likely behaviour, and ecology there have been few academic studies that have made extensive extrapolations about the nature of non-avian dinosaur genome structure prior to the emergence of modern birds. In this study, we used multiple avian whole genome sequences assembled at a chromosomal level, to reconstruct the most likely gross genome organization of the overall genome structure of the diapsid ancestor and reconstruct the sequence of inter and intrachromosomal events that most likely occurred along the Archosauromorpha-Archosauria-Avemetatarsalia-Dinosauria-Theropoda-Maniraptora-Avialae lineage from the lepidosauromorph-archosauromorph divergence ~275 mya through to extant neornithine birds. Using a combined bioinformatics and molecular cytogenetics approach, our results suggest that the ‘typical avian karyotype’ of ~2n=80 was probably established around the time that the first dinosaurs and pterosaurs emerged 240–245 mya with the gross karyotypic structure remaining largely intact inter-chromosomally to the present day (with rare exceptions). Further analysis of the evolutionary breakpoint regions in the archosauromorph and diapsid reconstructed genomes suggest that chromosomal evolutionary breakpoint regions are enriched for gene ontology terms associated with chromatin organisation, consistent with recent studies in rodents. We propose therefore that the overall genome organization and evolution of dinosaur chromosomes (inclusive of the avian radiation) had deeper origins than previously appreciated and was a major contributing factor to the morphology, physiology, ecology, evolutionary change, and ultimately survival, of this fascinating group of animals.

(P67) Comparative BAC Mapping of Macrochromosomes from Nine Avian Species Reveals Strong Chromosome Homology Over 98 Million Years of Evolution

Lucas Gem Kiazim[1], Rebecca O'Connor[1], Rebecca Jennings[1], Joana Damas[2], Marta Farré[2], Denis M. Larkin[2], Darren K. Griffin[1]

[1] School of Biosciences, University of Kent

[2] Department of Comparative Biomedical Sciences, Royal Veterinary College

Despite technological advances in genome sequencing as well as a reduction in sequencing costs, the majority of avian species do not have a sequenced genome. For these species, comparative BAC mapping using fluorescence in situ hybridisation (FISH) provides the opportunity to trace chromosome evolution through the mapping of genomic rearrangements between species. Using the chicken (*Gallus gallus*) genome as a reference, the mapping of individual BAC clones to the chromosomes of multiple avian species allows for the identification of fusions, duplications, inversions, and deletions, all of which contribute to the chromosomal changes that influence speciation. In order to map the macrochromosomes of multiple avian species, 74 selected BAC clones isolated from evolutionary conserved sequences from the chicken genome were hybridised to metaphase spreads of the blackbird (*Turdus merula*), canary (*Serinus canaria*), chestnut manikin (*Lonchura castaneothorax*), collared dove (*Streptopelia decaocto*), Eurasian woodcock (*Scolopax rusticola*), guinea fowl (*Numida meleagris*), houbara (*Chlamydotis undulata*), duck (*Anas platyrhynchos*), and pigeon (*Columba livia*). Mapping of chromosomes 1-9 and Z from the nine different avian species revealed strong chromosome homology despite 47-98 million years of evolutionary divergence. Chromosomal rearrangements were predominantly intra-chromosomal, with inter-chromosomal rearrangements being identified in selected species. Our study demonstrates a new comprehensive approach to tracing evolutionary relationships of multiple distantly related bird species and provides new insight into the nature of avian genomes and genomic stability.

(P69) The functional landscape of the fission yeast genome

Daniel Jeffares[1], Leanne Grech[2], Jurg Bahler[2]

[1] University of York

[2] University College London

A proportion of the non-protein-coding section of genomes contributes to cell function and to fitness. It is difficult to identify or quantify what this proportion is using bioinformatics approaches, comparative genomics or population genomics. Locating the specific functional non-coding elements is even more challenging, particularly when we require ‘function’ to mean ‘important to the cell’ rather than the ENCODE definition, ‘has some output’. Genome-scale transposon mutagenesis/selection experiments (TRADIS/Tn-Seq) provide a powerful complimentary method to locate functional elements with high accuracy, producing exquisite descriptions of functional elements bacterial genomes. We show that this method can describe the more complex functional elements of a eukaryote genome. We generated multiple high-density transposon mutagenesis libraries in fission yeast (*S. pombe*). We generated a total of 26 million insertions, an insertion density of 1 insertion site/13 nucleotides of the genome, essentially saturating the non-lethal sites. Insertion data were analysed using a customised hidden Markov model that categories the functional significance of each position in the genome accounting for insertion count, density and insertion biases. The results are consistent with current gene annotations, comparative genomics and population genomics, indicating that we can generate a statistically robust map of all the functional landscape of this genome. Results show that 91% of this yeasts genome is functionally significant, and that 81% of the non-protein-coding genome is functionally significant. The picture that emerges is that the genome of fission yeast is densely packed with undescribed non-coding functional elements.

(P70) ‘Multilayer’ FISH and interphase mapping to detect intra-chromosomal rearrangements

Jacob Ward[1], Darren Griffin[1], Rebecca O'Connor[1]

[1] School of Biosciences, University of Kent, Canterbury, UK

BAC mapping of evolutionary intra-chromosomal rearrangements is typically impeded by three factors: The unavailability of probes that hybridise effectively to multiple species, the capricious nature of multicolour hybridisation strategies and the need to prepare metaphases from actively dividing cells. Methodologies that can be used to map such rearrangements in multiple species and without the need for actively dividing cells are thus needed. We have recently developed a set of BAC clones that hybridise to the majority of avian species by judicious selection based on sequence homology. Here we report the development of a FISH protocol that involves a series hybridization layers to achieve the effect of a multi-colour strategy. The nature of nuclear organization in avian cells is such that the macrochromosomes locate toward the periphery of the nucleus in chromosome territories reminiscent of metaphase chromosomes. Here we demonstrate, with our multi-layer strategy, that inter-chromosomal rearrangement can be detected at both metaphase and interphase in five avian species. In preliminary experiments we have concentrated on Galliformes but provide proof of principle for a strategy that may be extended to other avian species, and eventually diagnostic uses for humans. Without the need to prepare chromosomes, a wider range of species (e.g. when tissue is hard to come by) may be examined for gross genomic rearrangement to gain insight into the mechanisms of chromosome evolution.

(P71) Transposon-driven, species-specific binding of insulator protein CTCF in two closely-related mice species

Dhoyazan Azazi[1], Christine Feig[2], Duncan Odom[2], Paul Flicek[1]

[1] European Bioinformatics Institute (EMBL-EBI), Hinxton, Cambridge

[2] Cancer Research UK Cambridge Institute, Cambridge, United Kingdom

CTCF is a genome-wide chromatin organiser protein with several roles in genome regulation. The evolution of CTCF binding sites and their influence on transcriptional regulatory activity has previously been studied in the mammalian lineage across species separated by approximately 5-200 million years. The insertion of novel CTCF binding sites has been shown to result from the action of the B2-B4 family of transposable elements. The exact mechanisms and functional impact of the appearance of novel binding sites are not yet fully understood. Here we investigated highly active, expanding repeats that have rapidly remodelled CTCF binding, and thus chromatin architecture and transcription, in two mouse subspecies, separated by one million years of evolutionary divergence: *Mus musculus domesticus* (C57BL/6J) and *Mus musculus castaneus*. We explicitly compared repeat content, evolution, and lineage loss/gain. Our results show that the evolution of CTCF binding in a limited evolutionary time-scale is strongly governed by the introduction of new sites through the action of the B2-B4 family of transposable elements independently in each lineage. This species-specific, repeat- driven expansion of CTCF binding does not predate speciation. Our results show that some of these sites may have acquired transcriptional regulatory function as illustrated by conservation of binding across several tissues in *M. musculus*. Nonetheless, the pattern of CTCF species-specific binding in terms of proximity to nearest transcription start sites and/or topologically-associated domains (TADs) is largely similar to deeply conserved CTCF sites.

(P72) Micro-evolution of CTCF binding sites helps maintain the integrity of topologically associating domains

Elsa Kentepozidou[1], Christine Feig[2], Maša Roller[1], Duncan Odom[2], Paul Flicek[1]

[1] EMBL - European Bioinformatics Institute

[2] Cancer Research UK Cambridge Institute

Topologically associating domains (TADs) are fundamental units of the 3D structural organisation of eukaryotic genomes. Their formation is an essential component of transcriptional regulation. Remarkably, it has been shown that TADs are highly conserved across species. Despite their importance and conservation, the mechanisms underlying their formation, as well as their evolution, remain largely unclear. Previous research has indicated that, among others, binding of the CCCTC-binding factor (CTCF) is an important player in TAD formation, though not sufficient to demarcate TAD boundaries alone. Moreover, it has been shown that TAD boundaries tend to overlap with CTCF binding sites that are conserved across species. This implies a potential interplay between CTCF binding site evolution and the maintenance of TADs. Aiming to elucidate the role of CTCF binding in the evolution of TADs, we investigate the extent of CTCF binding conservation at TAD boundaries in five mouse species. Our preliminary results indicate that, although most TAD-boundary-associated CTCF sites are conserved across the studied species, there is also a number of CTCF sites that are characterised by slight topological shifts around TAD boundaries. We believe this demonstrates that dynamic micro-turnover of CTCF binding sites within a short genomic range around TAD boundaries is one evolutionary mechanism to insure TAD stability.

(P73) Assembly and analysis of transcriptomes of the Zebrafish: characterisation of transcripts associated with learning and memory

Abril Izquierdo[1]

[1] The University of Nottingham

Synapses are central in the functioning of the brain. Yet, its mutations are the cause of more than 130 neurological alterations. It is therefore fundamental the mapping of gene expression in the human brain and characterized its proteome. For this intention, the zebrafish has emerged as a key model for researching the vertebrate gene function in neuroscience. Recently, it was characterized the proteome and ultrastructure of zebrafish synapses. Noteworthy in this study is the absence of complexity of the post synaptic density proteome compared that of the mammalian, even though that the zebrafish underwent a teleost-specific genome duplication. To adequately provide a richer understanding of neurological diseases in humans, it is essential to investigate the magnitude of which zebrafish genes have mammalian orthologous. To explore the presence of genes known to be essential in learning and memory, we sequenced and generated a de novo transcriptome assembly from three main regions of the zebrafish brain; the optic lobe, olfactory lobe and hindbrain, together with four zebrafish whole brains. Transcripts enriched and specific to each tissue was determined. Along with the analysis of which mouse genes present in the brain, synaptosome and post synaptic density have zebrafish orthologous. Methods: RNA was extracted from zebrafish brain tissues, and sent for sequencing using the Illumina NextSeq500 sequencing platform. Reads for each tissue were assembled into transcripts using Trinity. High quality transcripts were identified using Transrate. Transcriptomes were merged and isoform abundance estimated using HISAT and Stringtie. Orthology and functional annotation was achieved using dammit.

(P74) African Trypanosome comparative genomics with the focus on *Trypanosoma vivax*

Ali Abbas[1]

[1] University of Liverpool/Institute of Integrative biology

African animal trypanosomiasis known as “nagana” is a chronic disease causes high economic losses due to the low productivity of infected animals and fatal outcome. This illness mainly caused by *T.vivax*, *T. congolense* and *T. brucei*. In general, these species have digenic lifecycle, insect stage and mammalian stage with phenotypic differences and variances in the life cycle. Such variations might reflect modifications on the genome level among those parasites. Standard reference genome assembly of *T. brucei* strain TRUE927 (Tb927) is available. However, the current draft genome sequence of *T. vivax* based on Sanger sequencing is highly fragmented (>12,000 contigs). In order to achieve comparative genomic study, more contiguated genome assembly is needed. PACBIO SMRT sequencing was adopted to sequence *T. vivax* strain Y486 gDNA. ~ 6Gb of data were generated from eleven SMRT cells which assembled into ~770 contigs using HGAP assembler version 2 with total assembly size of ~67Mbp, the max contig size was ~2.5Mbp and N50 contigs’ length of ~ 261Kbp. The preliminary results revealed that large contigs have regions of synteny to more than one Tb927 chromosomes which make it difficult to infer distinctive eleven mega-base chromosomes based on Tb927 reference assembly using ABACAS version 1. Further data analysis unravelled putative large inter- chromosomal rearrangements that affected *T. vivax* genome in comparison to the Tb927 genome. Our initial data showed for the first-time predicted large structural variations on the chromosomal level that could have an impact on the speciation and life cycle of this parasite.

(P75) Phylogenomic analysis of kangaroos identify strong incomplete lineage sorting and introgression signals

Maria Nilsson Janke[1], Yichen Zheng[1], **Vikas Kumar**[1], Matthew Phillips[2], Axel Janke[1]

[1] Senckenberg Museum

[2] Queensland University of Technology

The iconic Australasian kangaroos and wallabies (genera Macropus and Wallabia) represent a successful marsupial radiation. However, the evolutionary relationship and timing of kangaroo evolution is controversial. We sequenced and analyzed the genomes of nine of the 13 species to investigate the evolutionary cause of the conflicting trees. A multi-locus coalescent analysis using ~14,900 genome fragments, each 10 kilo base pairs long, could significantly resolve the species relationships between and among the sister-genera Macropus and Wallabia. The phylogenomic approach reconstructed the swamp wallaby (Wallabia) as nested inside Macropus, making this genus paraphyletic. However, the genomic analyses show that the swamp wallaby genome includes conflicting phylogenetic signals. We interpret this as at least one introgression event between the ancestor of the genus Wallabia and a now extinct ghost lineage. Additional phylogenetic signals are found in the swamp wallaby genome that are caused by incomplete lineage sorting and or introgression, but statistical methods are not able to disentangle the two processes. In addition, the relationships inside the Macropus sub-genus (M. (Notamacropus)) represent a hard polytomy. Thus, the relationship between tammar, red-necked, agile and parma wallabies remains unresolvable even with whole-genome data. Even if most methods resolve bifurcating trees from genomic data, hard polytomies, incomplete lineage sorting and introgression complicate the interpretation of the phylogeny and in addition the taxonomy.

(P76) The “insightful” repertoire of nuclear receptors in the Ecdysozoa *Priapulus caudatus*

Elza Fonseca[1], Raquel Ruivo[1], Mónica Lopes-Marques[1], Miguel Fonseca[1], Miguel Santos[1], **Filipe Castro**[1]

[1] CIIMAR-UPorto

The homeostatic coordination of biological functions such as development or reproduction crucially depends on the action of numerous transcription factors. Among these, Nuclear Receptors (NRs) are the most abundant and unusual in Metazoan genomes. NR monomers, homodimers or heterodimers, triggered by ligand binding, selectively modulate transcription upon recognition of specific DNA responsive elements, in the promoter region of target genes. The emergence of full genome sequences brought, in recent years, a radical change to our understanding of NR gene family diversification with two clear waves of NR duplication in the ancestor of Bilateria and Vertebrates. Moreover, a significant aspect of NR biology is their activation or inhibition by man-made chemicals. Why some animal species are more sensitive than others to man-made chemicals is a major scientific conundrum of the Anthropocene. To address such a broad question, we must consider genome diversity at an ecosystem scale. Here, we investigate the full collection of NRs in the Ecdysozoa *Priapulus caudatus*. Our findings, illustrate the power of sequencing technologies and comparative genomics to infer contrasting patterns of gene duplication and loss and their wider impact in the Anthropocene Epoch.

(P77) Evolution of tissue-specific regulatory programs in cichlids

Tarang K. Mehta[1], Christopher Koch[2], Sara A. Knaack[3], Padhmanand Sudhakar[1], Luca Penso-Dolfin[1], Tomasz Wrzesinski[1], Will Nash[1], Tamas Korcsmaros[1], Wilfried Haerty[1], Sushmita Roy[2,3,4], Federica Di-Palma[1]

[1] Earlham Institute (EI), Norwich, UK

[2] Dept. of Biostatistics and Medical Informatics, UW Madison, Madison, USA

[3] Wisconsin Institute for Discovery (WID), Madison, USA

[4] Dept. of Computer Sciences, UW Madison, Madison, USA

In vertebrates, the East African cichlid radiations represent arguably the most dramatic examples of adaptive speciation. In the great lakes Victoria, Malawi and Tanganyika and within the last few million years, one or a few ancestral lineages of haplochromine cichlid fish have given rise to over 1500 species exhibiting an unprecedented diversity of morphological and ecological adaptations. Such explosive phenotypic diversification of East African cichlids is unparalleled among vertebrates and the low protein divergence between species implies the rapid evolution of regulatory regions and networks underlying the traits under selection.

Comparative functional genomics, transcriptomics and epigenomics are powerful tools to study the evolution of tissue and species divergence. We recently developed Arboretum, an algorithm to identify modules of co-expressed genes across multiple species in a phylogeny. By integrating inferred modules with nucleotide variation, predicted cis regulatory elements and miRNA profiles from five East African Cichlids, we investigated the evolution of tissue-specific gene regulation. Our analyses identified modules with tissue-specific patterns for which we reconstructed the evolutionary gene regulatory networks across the five cichlids species. We report striking cases of rapid network rewiring for genes known to be involved in traits under natural and/or sexual selection such as the visual systems, and more specifically the cone opsins (*sws2a* and *sws1*) responsible for colour vison of selected cichlid fishes. Our unique integrative approach that interrogates the evolution of regulatory networks allowed us to identify the rapid regulatory changes associated with certain traits under selection in cichlids.

(P78) Binding sites within long non-coding RNAs discriminate between RNA- and transcription-mediated mechanism

Tomasz Wrzesinski[1], Wilfried Haerty[1]

[1] Earlham Institute, Norwich Research Park, Colney Lane, Norwich, Norfolk, United Kingdom, NR4 7UZ

Tens of thousands of long non-coding RNAs (lncRNAs) have now been annotated in the human genome. They represent a highly heterogeneous class of transcribed elements with respect to their genomic position, molecular mechanisms, cellular localization and potential function. Only few lncRNAs has been experimentally characterized showing functions in dosage compensation, genomic imprinting and gene expression regulation. Despite active research, little is still known for most lncRNAs including the proportion that are biologically functional. Previous reports highlighted a dichotomy between lncRNAs, identifying loci whose function was solely conveyed by the act of transcription (Bendr) and loci with a RNA based function (Xist). We previously identified significant signals of purifying selection for splicing regulatory elements within a subset of lncRNAs supporting a RNA mediated mechanism.

Here we further investigate these roles in human aiming to identify the proportion of lncRNAs belonging to either or both classes. We tested the selective constraints acting on binding sites (AGO, TFs, RBP) within lncRNA exons, introns, as well as 1kb upstream and downstream sequences. LncRNAs with binding sites for RNA binding proteins are more highly expressed than those with either TF binding sites only or without any binding sites. We report increased conservation of the binding sites relative to matched sequences. Interestingly, lncRNAs with AGO sites are depleted within annotated enhancers whereas the opposite was found for lncRNAs with TFBSs. The joint analysis of purifying selection acting on functional elements within lncRNAs and of the genomic context help in distinguishing loci with different mechanisms of action.

(P79) The role of structural variation in Cichlid evolution

Luca Penso-Dolfin[1], Wilfried Haerty[1], Federica Di Palma[1]

[1] Earlham Institute, Norwich, UK

Cichlids represent a unique example of adaptive radiation, which gave rise to thousands of species in only a few million years. There is a great interest in gaining a better understanding about the link between genomic variation and the observed adaptive phenotypes.

In this study, we want to investigate the contribution of structural variants (SVs) in speciation events (through a reduction of gene flow) and adaptation to different ecological niches. We used paired-end libraries for five African Cichlids (*Astatotilapia burtoni*, *Metriaclima zebra*, *Neolamprologus brichardi*, *Pundamilia nyererei* and *Oreochromis niloticus*). Alignments of paired-end reads against the reference genome (*O. niloticus*) were used to identify putative rearrangements, locate chromosomal breakpoints, as well as classify these variants as deletion, duplication, inversion or translocation events.

Results will be used to infer the gain and loss evolution of SVs across the phylogeny, investigate their association with genomic features (copy number variants, SNPs, transposons, regulatory elements) and identify minimum conserved synteny blocks across species.

(P80) Differential gene regulation by microRNAs in the extremely young repeated parallel adaptive radiations of cichlid fish from Nicaragua

Paolo Franchini[1], Peiwen Xiong[1], Ralf Schneider[1], Carmelo Fruciano[1], Joost Woltering[1], Axel Meyer[1]

[1] Department of Biology, University of Konstanz, Universitätsstraße 10, 78457 Konstanz, Germany

Cichlid fishes are an ideal model system for studying the evolutionary mechanisms underlying rapid lineage divergence. Their adaptive radiations are textbook examples for explosive phenotypic diversification and sympatric speciation. Although gene regulation has been widely recognized to be an important mechanism that links diversification in gene function to speciation, so far the involvement of post-transcriptional regulation by miRNAs during speciation has received little attention. We investigated the potential importance of miRNA regulation in the repeated adaptive radiations of Midas cichlids (*Amphilophus spp.*) from Nicaraguan crater lakes. To this end, we sequenced miRNAs and mRNAs of five Midas species from embryos at 1-day post hatching. Comparing data from differentially expressed genes and miRNAs in each species pair, we identified several miRNAs and their potential target genes whose expression was significantly negatively correlated, thus providing candidate miRNA/gene functional pairs with a potential role in phenotypic diversification. The expression levels of these functional pairs were then validated using RT-qPCR on developing embryos and the expression domains of four selected miRNAs and several target genes investigated by in situ hybridization. Although these species are extremely young and shared common ancestors only a few thousand years ago, we found clear species-specific expression domains and, most interestingly, found that a novel Midas cichlid miRNA and its target gene were differentially expressed in the jaw area of a benthic and a limnetic species that have speciated sympatrically in the same crater lake. These results suggest that regulation by miRNAs might be an important and extremely fast evolving mechanism that contributed to the rapid phenotypic evolution of cichlid fishes more generally.

Conservation Genomics

(P81) Assembly and comparative analysis of parrot genomes from the Caribbean

Sofiia Kolchanova[1,2], Pavel Dobrynin[2], Klaus-Peter Koepfli[3], Stephanie Castro[4], Kirill Grigorev[4,5], Jonathan Foox[6], Juan Lorenzo Rodriguez Flores[5], Jafet Velez-Valentin[7], Stephen J. O'Brien[2,8], Juan Carlos Martinez Cruzado[4], Taras K. Oleksyk[4]

[1] University of Puerto Rico at Mayaguez

[2] Dobzhansky Center for Genome Bioinformatics, St. Petersburg State University

[3] Smithsonian Institution, Washington

[4] University of Puerto Rico at Mayaguez

[5] Weill Cornell Medical College, New York

[6] American Museum of Natural History, New York

[7] Puerto Rican Parrot Recovery Program, US FWS

[8] Halmos College of Natural Sciences and Oceanography at NOVA Southeastern University

Amazon parrots (*Amazona sp.*) that currently inhabit the Greater Antillean Islands of Cuba, Jamaica, Hispaniola and Puerto Rico, all descend from a common ancestor in Central America, and represent a fascinating model of island speciation. Since these species are very closely related, their genomes can be studied in direct genome comparisons. First, we assembled, annotated and compared complete mitochondrial genomes of five species to test existing island speciation hypotheses. Second, we created nuclear genome assemblies of parrot genomes starting with Puerto Rico's own critically endangered *Amazona vittata* using Illumina and PacBio technologies, improved genome assemblies improved with transcriptome-based scaffolding approach and then annotated. In addition, we sequenced one individual for each of the five Caribbean and central American species - *A. vittata*, *A. ventralis*, *A. leucocephala*, *A. agilis* and *A. collaria, A. albifrons* and *A. xantolora* - using paired reads and mate pairs of different insert sizes. For gene annotation, we used homology-based approach and pre-trained HMM gene models derived from existing genome annotations of other annotated birds (*Melopsittacus undulatus*, *Gallus gallus*, *Ficedula albicollis*, *Taeniopygia guttata*). Quality of the initial and improved assemblies was assessed with Quast, BUSCO and CEGMA. Inferences about genomic variation and population structure in these species will be to help adjust decisions in developing conservation strategies and to help the existing captive breeding programs. We have expanded our analysis to search for signatures of selection, time the speciation events, and look for gene families' expansion and contraction.

(P82) Evaluating the performance and replicability of fecal DNA targeted sequencing

Jéssica Hernández Rodríguez[1], Mimi Arandjelovic[2], Jack Lester[2], Cesare de Filippo[2], Antje Weihmann[3], Matthias Meyer[3], Samuel Angedakin[2], Ferran Casals[4], Arcadi Navarro[1], Linda Vigilant[2], Hjalmar S Kuhl[2], Kevin Langergraber[5], Christophe Boesch[2], David Hughes[6], Tomàs Marquès Bonet[1]

[1] Institute of Evolutionary Biology (Universitat Pompeu Fabra/CSIC), Ciencies Experimentals i de la Salut, Barcelona, Spain

[2] Department of Primatology, Max Planck Institute for Evolutionary Anthropology, Leipzig, Germany

[3] Department of Evolutionary Genetics, Max Planck Institute for Evolutionary Anthropology, Leipzig, Germany

[4] Genomics Core Facility, Departament de Ciències Experimentals i de la Salut, Universitat Pompeu Fabra, Barcelona, Spain

[5] School of Human Evolution & Social Change, Arizona State University, Tempe, USA

[6] MRC Integrative Epidemiology Unit, School of Social and Community Medicine, University of Bristol, Bristol, UK

Target capture technologies have risen in the past years, proving to be a very efficient tool for selectively sequencing regions of interest. These methods have also allowed the use of non-invasive samples such as feces (characterized by their low quantity and quality of endogenous DNA) in conservation genomics, evolution, and population genetics. Here we aim to test different protocols and strategies for exome capture with Roche SeqCap EZ Developer kit (57,5Mb). First, we captured a considerable pool of DNA libraries. Second, we assessed the influence of using more than one fecal sample, extract and/or library from the same individual on the molecular variability of the experiment. We validated our experiments with 18 chimpanzee fecal samples (9 from Kibale National Park, Uganda, and 9 from Loango National Park, Gabon). We have demonstrated that at least 16 libraries can be pooled and hybridized with ¼ diluted probes, obtaining a considerable number of SNPs for population genetic analyses. We also observe that library richness increases when using multiple libraries from the same extract or multiple extracts from the same sample. We conclude that repeated capture significantly decreases the proportion of off-target reads from 34.15% after one capture round to 7.83% after two capture rounds.

(P85) Genome wide analysis of drift and selection using historic and contemporary samples of the endangered Mauritius Pink Pigeon

Camilla Ryan[1,2], Lawrence Percival-Alwyn[2], Diana Bell[3], Ian Barnes[4], Carl Jones[5], Matthew Clark[2], Cock van Oosterhout[1]

[1] School of Environmental Sciences, University of East Anglia

[2] Earlham Institute

[3] School of Biological Sciences, University of East Anglia

[4] Department of Earth Sciences, Natural History Museum

[5] Durrell Wildlife Conservation Trust

The pink pigeon (*Nesoenas mayeri*) is an endangered species, endemic to the island of Mauritius, that experienced a severe bottleneck in the 1970's to less than 20 individuals. Today the population has recovered to just under 400 birds but still suffers from many threats such as predators, habitat loss, inbreeding depression and susceptibility to pathogens. We will be using a DNA capture approach alongside previously generated genomic resources to examine changes in pink pigeon's genetic diversity. We plan on analysing genetic variation pre- and post-bottleneck by comparing museum samples collected in the 19th Century with present-day birds, and studying genetic drift and selection using both forward-in-time simulations as well as coalescence theory. This will also allow us to investigate how the effective population size has changed over time and whether any changes in genetic diversity can be linked to specific threats. Such an approach will enable us to identify alleles or genotypes that are under balancing or positive selection. This could help inform genetic supplementation, ex situ breeding programs and contribute to the genetic rescue and long-term viability of the pink pigeon.

Population Genomics

(P86) Selection against microRNA target sites

Antonio Marco[1],

[1] University of Essex

MicroRNAs are powerful gene regulators that play an important role in the evolution of regulatory networks. At the population level, the study of polymorphisms allows the identification of microRNA target sites under positive or purifying selection. Mutations generate novel microRNA target sites, some of which may affect the expression of genes. Consequently, some transcripts will avoid target sites for specific, co-expressed, microRNAs. I have developed a model for microRNA target avoidance based on the comparison of allele frequency distributions in untranslated regions (UTRs). When applying this strategy to Drosophila, I detected that genes transmitted by the mother into the egg avoid target sites for microRNAs also deposited in the egg. That is, maternal genes avoid maternal microRNAs. Ongoing work in our lab indicates that selection against microRNA target sites is prevalent in human populations. Target avoidance may be a broad evolutionary phenomenon, resulting from gene regulatory conflicts.

(P89) Genome-Wide Analysis of the evolution of new species in a Tanzanian crater lake

Wilson IW, Tyers AM, Malinsky M, Svardal H, Miska A, Durbin R et al.

The African Great Lakes contain thousands of cichlid fishes comprising large scale adaptive radiations, the genomes of which are indicative of myriad speciation events, both historic and incipient. However, despite the affordable availability of whole genome sequencing (WGS) technologies, high levels of relatedness due to extensive hybridisation and introgression within each radiation limit our ability to study drivers of speciation in many cases. Lake Massoko - a small crater lake, north of Lake Malawi - offers a solution to this problem. Incipient sympatric speciation has recently been observed in this isolated habitat involving two ecomorphs of a haplochromine cichlid species related to members of the Lake Malawi radiation. A Genome-Wide Association Study of members of this population has been employed to enable us to understand how genomic divergence is driving speciation here in the presence of gene flow. We have performed WGS and 25 phenotypic measurements of 193 individuals from Lake Massoko. Variant calling and filtration has produced a set of 460,419 biallelic single nucleotide polymorphisms (SNP), which has been entered into association analyses incorporating univariate linear mixed models. Variants closely associated with phenotypic differences between ecomorphs will be validated through their inclusion on a SNP genotyping array, which will be tested against the genomes of a further 100 cichlids isolated from the crater lake. The Lake Massoko population provides an outstanding model for cichlid adaptive radiation, benefiting from its relative simplicity, whilst involving species derived from the mega-radiation of Lake Malawi, which continues to draw great interest.

SPONSORS SHOWCASE

Verne Global: Why Icelandic HPC is Bioinformatics’ best friend

Sarah Cossey[1], **Spencer Lamb**[2]

[1] Earlham Institute

[2] Verne Global

Cutting-edge, high-throughput DNA sequencing instruments generate large amounts of data, from a few hundred gigabytes to several terabytes per run. This output requires significant high performance computing effort, making the storage, processing, analysis and sharing of the data extremely challenging.

This presentation highlights how Earlham Institute is looking to address the increasing data-driven challenge by migrating a strategic, collaborative bioinformatics analysis platform to the Verne Global data center in Iceland via the National Research Education Network (NREN) providers Janet and NORDUnet.

In this project, a first for a UK research institution, Earlham Institute has been able to take advantage of Verne Global’s highly-optimised, secure, scalable and 100% renewably-powered, data center campus. Further, due to the geothermal and hydro-electric sources of power, plus the ability for ambient air cooling, due to Iceland’s temperate climate, Earlham Institute will be able to significantly reduce the carbon footprint of its HPC workloads.

The presentation will conclude with the findings from Earlham Institute’s first year of operations in Iceland and will provide recommendations for similar UK research institutes that are involved in high capacity, data-driven science.

New England BioLabs: NEBNext®: Optimised Workflows for NGS Library Preparation

Adam Peltan, PhD. Technical Application Specialist, New England Biolabs

Grounded in over 40 years of research, discovering and engineering enzymes for basic molecular biology, New England Biolabs have leveraged their expertise to develop highly efficient, easy to use library preparation kits for NGS. The new NEBNext® Ultra™ II DNA and RNA library preparation kits have been optimised at every step of the process in order to deliver high quality sequencing data from both low quality and high quality sample types, allowing streamlining of the number of sample prep workflows needed for your research. The talk will focus on RNA-seq library prep and reagents for 16s rDNA sequencing and metagenomics. We will also briefly touch on our new Ultra II FS Library Construction kit. For additional details please also see the talk from Lesley Shirley, Wellcome Trust Sanger Institute on Wednesday 3.15-3.30pm where she will describe how this novel library construction kit has enabled the Sanger Institute's high-throughput sequencing pipelines.

Intel: AI + Precision Medicine + Moore’s Law = The 21st Century virtuous cycle

Gaurav Kaul

As the life science industry increasingly moves towards genomic sequencing, massive volumes of data are being generated to understand genotypic variants, spread of diseases in populations and evolution itself. However, the gap in managing, exchanging and deriving insights from these different data types continues to be an industry challenge. Two developments can help solve this issue and have shown tremendous potential in other fields – Artificial Intelligence and Moore’s Law. By accelerating artificial intelligence using newer and sophisticated computer architectures thanks to continuing improvements due to Moore’s Law, the diversity of the data generated by genomics can be handled and truly predictive analytics done. This has potential to revolutionize how healthcare is delivered at point of care. In this talk, we will discuss some use cases where AI is being used in digital pathology, genomics and pharmaceutical industry with increasingly sophisticated computational techniques to improve healthcare for all. We will also discuss details of Intel’s new BigStack architecture for genomics and how it can leverage some of these technical advancements in service of genomics.

TTPLabtech: Automated low-volume liquid handling for cost-effective NGS library preparation and single cell genomics

Dr Klaus Hentrich

The rapid advance of NGS technologies has led to an increasing need for high-throughput, low-cost NGS library preparation. Assay miniaturisation with automated liquid handlers can potentially address these needs. Suitable instruments must deliver high accuracy at sub-microlitre volumes, combined with the ability to aspirate, dispense and mix, while avoiding cross-contamination.

Here we demonstrate how TTP Labtech's positive displacement pipetting technology enables automated and miniaturised NGS library preparation workflows. With mosquito® liquid handlers, Nextera® XT and NEBNext® Ultra™ library preparation for Illumina® sequencing have been established and validated at up to 25-fold reduced volumes compared with the manufacturers' standard protocols (1).

Furthermore, we present the use of mosquito to automate and miniaturise two different protocols for single cell RNA-seq with FACS-sorted cells in 384-well plates.

The novel CEL-seq2 (2) procedure, with first-strand cDNA synthesis in only 400 nL and subsequent pooled library preparation, provides a sensitive and cost-effective pipeline for 3'-end counting (3).

In the miniaturised SMART-seq2 (4) workflow, full-length cDNA is amplified in a 5 µl reaction, while Nextera XT library preparation routinely takes place in only 4 µl total volume.

These approaches provide a significant reduction in reaction volumes, hands-on time and cost; allow for reproducible high-throughput, parallel processing; and thus help maximise data output in a research lab or core facility setting.

References:

1) Mora-Castilla, S., et al. (2016). JALA 21, 557-567.

2) Hashimshony, T., et al., (2016). Genome Biol. 17, 77.

3) Herrtwich, L., et al. (2016). Cell 167, 1264-1280.

4) Picelli, S., et al. (2014). Nat. Protocols 9, 171-81.

10x Genomics: The Chromium System for Enabling High Resolution Biology

Deanna Church, Senior Director of Applications

Reconstructing individual genomes and understanding the impact on biology remains a significant challenge. While large numbers of genomes and transcriptomes have been sequenced, the resulting resolution of these data remains insufficient for many applications. Traditional reference based, short-read analysis of genomes provides an incomplete picture of individual genome architecture. Likewise, while traditional transcriptomics has provided many biological insights, higher resolution data will allow for new information to be obtained. We have developed a high-throughput solutions that addresses both areas.

For genomic applications, we partition limiting amounts of high molecular weight DNA such that unique bar codes can be added as part of library generation. This approach allows us to couple long-range information with high-throughput, accurate short read sequencing, generating a data type known as Linked-Reads. Coupling this novel datatype with new algorithms allows us to access a greater percentage of the genome as well as identify the full spectrum of variant types. Additionally, Linked-Reads enable de novo assembly with modest amounts of sequencing.

For transcriptomic applications, our microfluidics system partitions single cells and then barcodes their transcriptional content. This high resolution transcriptional profiling allows for the discrimination of discrete cell types from complex mixtures, allowing for the dissection of complex biological processes at high throughput. This opens up new applications for better discriminating immunological processes as well as understanding tumor micro-environment.

Eppendorf: Yield, Specificity and Inhibition of PCR: how to get better results!

Dr. Kay Körner

Although PCR is a well-established technology, optimizing certain variables of a PCR setup can still have great benefits. Avoiding inhibition or partial inhibition can increase the yield of a PCR. The same is true for optimized denaturation temperatures, which can additionally help with specificity. Last but not least, the reaction volume can be optimized to get better results.

DNAnexus: Creating Community in the Cloud

Mark Mooney, Alessandro Riccombeni, Andrew Carroll

As consortia and biological projects become larger and more global, maintaining “big data”, pipelines and searchable results in an accessible form is an expanding issue. The goal of these projects is to create and maintain an active and actionable environment for biologists and bioinformaticians to work and collaborate. The cloud seems like an ideal solution to this problem. The “what you want and when you want it” virtual hardware of the cloud allows data to be processed without capital investment. Approved pipelines can be created in the US and used in China. Consortia members can access approved tools, pipelines, and datasets generated by the consortia community. The DNAnexus platform is already the data backbone for a number of community programs, including eMERGE, CHARGE and the FDA precision medicine challenge sites. We will talk today about our role in supporting and enabling the Vertebrate Genomes Project. This consortium is implementing a range of tools to create pipelines for Genome Assembly. We will discuss how we interact with the cloud provider Amazon Web Services to access free cloud resources (storage and free data movement), how we coordinate with a range of sequencing partners to allow access to multiple data types for the “kitchen sink assembly”, and finally, how we hope to work with the research community on making this data freely available.

BioNano genomics: *Acomys* genome project: A comparative genomic framework for evolutionary and biomedical studies

Dr Thomas Keane[1]

[1] EMBL - EBI

The *Acomys* or spiny mouse, are rodents from the family of Muridae. These close relatives to the laboratory mouse carry extreme phenotypes of high interest in evolutionary or biomedical studies. For instance, the desert adapted species *Acomys russatus* has the possibility to drink sea water and still maintaining kidney function. It can contract type 2 diabetes and become obese when kept in captivity. *Acomys cahirinus* is the only rodent for which menstruation has been described, making it the only laboratory rodent that can be used to study menstrual disorder in women. However, the most spectacular phenotype of *Acomys* is their capacity to regenerate tissues making this genus the only mammal model to study regenerative medicine. In order to study these phenotype at the genomic level, we undertook to sequence chromosome level assembly of the genome of four Acomys species *(A. cahirinus, A. russatus, A. kempi, A. percivali*) and one close related outgroup (*Psammomys obesus*). We combined Pacbio, Bionanogenomics, and illumina to generate chromosome level assemblies with high nucleotide sequence accuracy. The analysis of the lineage and species-specific loci in the Acomys genus will help to find the genomic loci associated to these phenotypes and understand the evolutionary mechanism involved in shaping them.

Perkin Elmer: A Comparison of 16S Amplicons in Microbial Community Standards & Environmental Samples

Brad Hehli (Bradley.Hehli@PERKINELMER.COM)

In the evolving field of bacterial metagenomics, many researchers choose to use variable regions of the 16S rRNA gene to understand which bacteria are present and how many bacteria are present in a sample. The expanding number of next generation sequencing options and platforms provide many possibilities of 16S sequencing, from looking at one variable domain to combinations of variable domains. The variable regions of the 16S gene offer differential ability to discriminate bacteria from each other and environmental contaminants. Here, we show the performance of four different options of next generation 16S rRNA sequencing- 16S V1-V3, 16S V3-V4, 16S V4, and 16S V5-V6. Each region was tested on both a known community of bacteria and on environmental samples. The known communities show the relative representation and performance of each primer set. The environmental samples show the performance and challenges these primers sets face in samples similar to those of metagenomics researchers. The environmental samples were analyzed and classified using the One Codex pipeline and database.

PacBio: PacBio SMRT Sequencing on the Sequel System: higher throughput, lower cost, better science

Michelle Vierra

Join the leaders in long read sequencing to hear the latest updates on the Sequel System and to learn how scientists are employing Single Molecule, Real-Time (SMRT) Sequencing to gain a deeper understanding of evolution, diversity, and environmental interactions for all walks of life. With long reads, high consensus accuracy, uniform coverage, and simultaneous epigenetic characterization, SMRT Sequencing enables a complete view of genetic diversity in large and highly complex plant and animal genomes and their environment.

Illumina: Sequencing and Array-based methods for resolving genomic inquiry

Cindy Lawley[1]

[1] Illumina, Inc., 499 Illinois Street, San Francisco, CA 94158

The drop in price and simultaneous >10 fold increase in accuracy over traditional Sanger methods of sequencing have fueled creative ways to characterize genetic variation answer biological questions. *De Novo* sequencing using NGS is an early step toward understanding the genetic underpinnings of a plant or animal's function and its interaction with the environment. Generating publication-ready scaffolds can still be prohibitively expensive, especially in complex, repetitive and/or highly diverse genomes. Through strong partnerships, collaboration and consortia with innovators and early adopters, Illumina has been able to help drive forward lower cost solutions for building publication-ready scaffolded assemblies, without compromising sequencing accuracy. Sequencing sample prep and analysis methods are changing/innovating quickly. Here we share customer stories and newly published methods, including the key parameters to consider for research and industry partners implementing genomic tools for driving applied solutions.