VIBE/ICBG 2024
ABSTRACT BOOK

Table of Contents

Schedule

Conference Schedule – VIBE/ICBG 2024

Time	Thursday, 5 December 2024	
10:00-10:30	**Registration & Coffee/Tea**	
10:30-11:00	**Registration & Opening Address**	
11:00-12:00	**Keynote**	
	Dr Davide Cirillo	*Synthetic Data in Biomedical Research: From Knowledge Discovery to Knowledge Generation*
12:00-13:00	**Session 1:** Methods & Algorithms	
	Michael Lynch	*Supervised multimodal demultiplexing outperforms conventional demultiplexing of scRNAseq*
	Seyed Aghil Hooshmand	*Advancing Cancer Research through Integration of Liquid Biopsy Data in cBioPortal: Insights from the All-Ireland Cancer Liquid Biopsies Consortium (CLuB)*
	Yezhao Zhong	*Adverse Drug Reaction Profile Prediction: Denoising, Signal Enhancement and Missing Row Imputation*
13:00-14:00	**Lunch & Coffee/Tea**	
14:00-15:00	**Session 2**: Cancer Genomics	
	Roofiya Koya	*Mechanistic insights into long noncoding RNAs through high resolution analysis of tumour mutations.*
	Aideen McCabe	*Understanding Ovarian Cancer: Lessons Learned from Cells, Apps and Patients*
	Hannah Nyarko	*Assessing the spatial patterns of immune and stromal cells for prognosis in early-stage ER+/HER2- Breast Cancer*
	Micheál Ó Dálaigh	*Kinnex Long-Read Resequencing of Acute Myeloid Leukaemia Single-Cell RNA Samples for Improved Detection of Malignant Genomic Alterations*
15:00-16:00	**Keynote**	
	Dr Elisabeth Bik	*Errors and Misconduct in Biomedical Research*
16:00-16:30	**Lightning Talks & Gold Sponsor**	
	Ahmad Alkhan	*Deep Learning approach for detecting and segmenting Perineural Invasion in Colon, Prostate, and Pancreatic cancers*
	Stefanus Bernard	*Bioinformatics Refinement of CRISPR-Cas9 Knockout Screens Reveals Additional Genes Modulating Cellular Responses to CDC7 Inhibitors*
	Catherine Higgins	*Clustering imbalanced functional data - enhancing the clustering accuracy of time-course gene expression data*
	Pouya Motienoparvar	*A genetic network integrates regulation of the vegetative-reproductive phase transition in Arabidopsis thaliana*
16:30-17:30	**Poster Session & Sponsors**	
	Presenting: Odd Posters	
17:30	**Social Event**	

Time	Friday, 6 December 2024	
09:30-10:30	**Keynote**	
	Prof Karen Miga	*The Human Pangenome Project: Creating a Reference that Better Represents Human Global Genetic Diversity*
10:30-11:00	**Lightning Talks**	
	Elle Loughran	*Factors in the Development of Extreme Ploidy States in Cancer*
	Mariagiovanna Pais	*Exploring endocrine disrupting pathways using knowledge graph and network biology.*
	Anna Großbach	*Mapping Genetic Determinants of DNA Methylation Across Early Development*
	Emma Corley	*Associating Mood Symptom Severity with Subcortical Brain Volumes in Bipolar Disorder and Major Depressive Disorder using an Item Response Theory Model*
	Metin Yazar	*Unravelling Resistance Mechanisms to Synthetic Lethal Therapies in Cancer Through Protein-Protein Interaction Networks*
11:00-12:00	**Poster Session & Sponsors**	
	Presenting: Even Posters	
12:00-13:00	**Session 3:** Population Genetics & Molecular Evolution	
	Olivier Dennler	*Evaluating Sequence and Structural Similarity Metrics for Predicting Shared Paralog Functions*
	Sophie Matthews	*Variable gene copy number in cancer-related pathways is associated with cancer prevalence across mammals*
	Maria Eleonora Rossi	*Independent origins of spicules reconcile the evolutionary history of sponges (Porifera)*
	James McInerney	*panGPT: An AI transformer for generating large pangenome models.*
13:00-14:00	**Lunch & Coffee/Tea**	
14:00-15:00	**Session 4:** Metagenomics & Pathogen Surveillance	
	Kate Ryan	*Genomic identification and characterisation of novel bacterial species from Space Craft Assembly Clean Rooms*
	Emmet Campbell	*In silico phage-bacteria infection networks (PBINs) of Streptococcus suis reveal co-evolution patterns between host and prophage*
	Anna Tumeo	*Validation of a field-deployable automated DNA extraction system as a tool for assessment of microbial diversity in marine ecosystems*
	John Paul Wilkins	*Secrets in the Sewers - Revealing the hidden diversity of SARS-CoV-2 in Northern Irish wastewater*
15:00-15:30	**Coffee/Tea Break**	
15:30-16:30	**Session 5:** Genomics of Health & Disease	
	Javier Villegas Salmerón	*Characterization of the mouse intra-amygdala kainic acid model at single-cell resolution reveals cell-type specific contributions to epilepsy phenotype*
	Sophia Heneghan	*The impact of copy number variants in diagnosis and severity of autosomal dominant polycystic kidney disease*
	Nicole Glendinning	*DNA Methylation in Vulnerability to Opioid Use Disorder*
	Karen Guerrero Vazquez	*Predicting Age and Identifying Aging-Related Genes from Muscle Gene Expression Data*
16:30-17:00	**Awards & Close of Conference**	

Keynote talks

Dr Davide Cirillo: Synthetic Data in Biomedical Research - From Knowledge Discovery to Knowledge Generation

In the era of data-driven medicine, synthetic data generation (SDG) has emerged as a transformative approach paving the way from traditional knowledge discovery to active knowledge generation. At the Barcelona Supercomputing Center, the Machine Learning for Biomedical Research unit is advancing SDG across various data types (clinical, omics, and imaging data) to address critical challenges in healthcare. Through SDG, we unlock insights in data-scarce scenarios and forecast complex events in disease progression, illustrating the broad potential of SDG to drive early and personalized interventions. Our work prioritizes fidelity, utility, and privacy, ensuring that synthetic data is both robust for further modeling and valuable for clinical applications.

Dr Elizabeth Bik: Errors and Misconduct in Biomedical Research

Science builds upon science. Even after peer-review and publication, science papers could still contain images or other data of concern. If not addressed post-publication, papers containing incorrect or even falsified data could lead to wasted time and money spent by other researchers trying to reproduce those results. Elisabeth Bik is an image forensics detective who left her paid job in industry to search for and report biomedical articles that contain errors or data of concern. She has done a systematic scan of 20,000 papers in 40 journals and found that about 4% of these contained inappropriately duplicated images. In her talk, she will present several types of inappropriately duplicated images, how to report such problems and how journals and institutions handle such allegations. Finally, she will address the growing problems of 'paper mills', for-profit networks that produce and sell large amounts of low-quality or fake papers.

Prof Karen Miga: The Human Pangenome Project - Creating a Reference that Better Represents Human Global Genetic Diversity

The initial Human Genome Project was a landmark achievement, serving as an essential resource for basic and clinical science, as well as for understanding human history, for over two decades. However, it is in need of an upgrade due to missing data, inaccurately assembled regions, and its inability to fully represent and identify sequence variants equitably. A single reference map, regardless of its completeness, cannot encapsulate the variation across the human population, leading to biases and ultimately inequity in genomic studies. Recognizing this limitation, the new initiative known as the Human Pangenome Project aims to deliver hundreds of highly accurate and complete genomes. This effort intends to define all bases of each chromosome from telomere to telomere (T2T), ensuring a broader representation of common variants across the human species. Achieving these goals will require the rise of new tools and technology standards for complete genome assemblies and pangenomics, which will have broad and lasting impact on genomic research.

Short talks

Cancer Genomics

6: Mechanistic insights into long noncoding RNAs through high resolution analysis of tumour mutations

Roofiya koya[1], Sunandini Ramnarayanan[1], Rory Johnson[1], Hugo Guillen[2]

[1]University College Dublin, Dublin, Ireland. [2]University of Bern, Bern, Switzerland

Background/Introduction: Long non-coding RNAs (lncRNAs) are extensively involved in diverse regulatory roles and have numerous disease associations, particularly in cancer. Our lab recently demonstrated that tumour mutations can impact lncRNA activity to increase cancer cell fitness. lncRNAs exert their bioactivity through interactions with other biomolecules in the cell, including DNA, RNA, and proteins. These interactions are encoded in lncRNA primary sequence in the form of sequence motifs and structures, termed 'elements'. To date, the elements of just a handful of lncRNAs have been elucidated.

Materials/Methods: This study posits that tumour single nucleotide variants (SNVs) can impact lncRNA activity by altering functional elements activity, and hence to gain mechanistic insights into cancer progression. We introduce Exinator-EL (Element), a pipeline designed to detect lncRNA elements with a significantly elevated mutational burden. Exinator-EL systematically estimates the mutational burden for each element type, with respect to a modified background estimation (non-element regions) and determining significance through Fisher's test. It accounts for mutational signatures with a trinucleotide-aware empirical significance estimation.

Results: We apply Exinator-EL to a catalogue of 270 classes of elements within 122 mutated cancer driver lncRNAs identified using cohorts from the Genomics England (GEL) resource. Our analysis reveals a significantly higher burden of mutations within the elements of these cancer-driver lncRNAs compared to non-cancer-driver lncRNAs, with elements Conserved RNA Structures and RNA Binding Protein Sites showing a two- to five-fold enrichment in mutations. These findings suggest that mutations in these elements alter lncRNA functionality, leading to abnormal cancer cell fitness.

Discussion/Conclusion: This comprehensive map of mutated lncRNA elements on both pan-cancer and individual cancer cohorts lays the groundwork for testable mechanistic hypotheses, advancing our understanding of how mutations affect these elements in both biological and clinical contexts.

30: Understanding Ovarian Cancer: Lessons Learned from Cells, Apps and Patients

Aideen McCabe[1,2], Oza Zaheed[1,2], Gerard P Quinn[3,4], Suneil Jain[4,5], Micheál Ó Dálaigh[6,7,2], Ross G Murphy[8,4], Simon S McDade[4,3], Kellie Dean[1]

[1]School of Biochemistry and Cell Biology, University College Cork, Cork, Ireland. [2]The SFI Centre for Research Training in Genomics Data Science, Galway, Ireland. [3]BlokBio, Ormeau Labs, Belfast, United Kingdom. [4]The Patrick G Johnston Centre for Cancer Research, Queen's University Belfast, Belfast, United Kingdom. [5]Department of Clinical Oncology, Northern Ireland Cancer Centre, Belfast Health and Social Care Trust, Belfast, United Kingdom. [6]School of Biological and Chemical Sciences, University of Galway, Galway, Ireland. [7]School of Mathematical and Statistical Sciences, Galway, Ireland. [8]The Centre for Genomic Medicine, Ulster University, Coleraine, United Kingdom

Background: Epithelial ovarian cancer (EOC) is a highly heterogeneous disease responsible for over 200,000 deaths worldwide each year. EOC is subdivided into five diverse histological subtypes, of which high-grade serous ovarian carcinoma (HGSOC) is the most common and aggressive. Effectively modelling these cancer subtypes in cell lines, classifying the molecular subtypes of HGSOC, and actively involving patients in the research process are essential for advancing our understanding of the disease.

Methodology: Using non-negative matrix factorization (NMF), we clustered 56 cell lines into five clusters to reflect histological subtype and assessed their similarity to primary tumours through gene expression, mutational, and copy-number data. Additionally, we developed *classifieRov*, an R Shiny app that assigns HGSOC tumours to four molecular subtypes, making transcriptional classifications accessible to non-bioinformaticians. Finally, we organized Public and Patient Involvement (PPI) workshops with people affected by ovarian cancer to align our research with patient priorities and improve the practical application of diagnostic tools.

Results: NMF and molecular analysis revealed five distinct clusters of EOC cell lines, reflecting histological subtype and identifying suitable models for HGSOC, clear cell, and mucinous carcinomas. ClassifieRov, a tool designed for non-bioinformaticians, integrates transcriptional classification of HGSOC tumours with data on cellular composition and transcription factor activity. Additionally, patient and public involvement (PPI) workshops provided crucial patient perspectives that guided the creation of a patient information leaflet, required for follow-up studies using clinical samples.

Discussion: This study integrates cellular, computational, and patient-driven approaches to advance our understanding of EOC. Accurate cell line models, accessible molecular classification tools, and patient engagement are vital for improving personalized treatment and diagnostic strategies in ovarian cancer.

39: Assessing the spatial patterns of immune and stromal cells for prognosis in early-stage ER+/HER2- Breast Cancer

Hannah Nyarko[1,2], Zak Kinsella[1], Daria Kalinska-Lysiak[1], Niamh Connolly[1], Jochen Prehn[1], John Crown[3], Catherine Kelly[4], William Gallagher[5], Darran O'Connor[1]

[1]Royal College of Surgeons in Ireland, Dublin, Ireland. [2]SFI Centre for Research Training in Genomics Data Science, Dublin, Ireland. [3]St. Vincent's Private Hospital, Dublin, Ireland. [4]Mater Private Network, Dublin, Ireland. [5]University College Dublin, Dublin, Ireland

Background/Introduction: Oestrogen Receptor (ER)+ breast cancers are known to have a relatively favourable prognosis. However, ~10% of patients remain at risk of late recurrence. Prognostic tests, such as the Oncotype DX 21-gene recurrence score (RS), have been designed to predict late recurrence to inform treatment decisions, nonetheless, a fraction of these predictions tend to be erroneous. The spatial distribution of immune cells in the TME has been shown to influence breast cancer prognosis, yet this is not considered in current ER+ breast cancer prognostic tools. This study aimed to assess the spatial distribution of immune and stromal cell phenotypes in ER+ Breast Cancers and its potential as a prognostic factor in ER+/Her2- breast cancer.

Materials/Methods: Using the Nanostring GeoMx DSP whole transcriptome workflow and Illumina Next-Seq, FFPE tissues from 410 early-stage ER+/Her2- breast cancer patients were spatially profiled. Regions of interest were selected guided by panCK staining and segmented into tumour (panCK+) and TME (panCK-) regions.

Results: Results indicated a high infiltration of lymphocytes and macrophages in the TME, while tumour regions were mainly enriched in neutrophils. Patients with elevated fibroblast and lymphocytes in the TME were also found to show better survival outcomes.

Discussion/Conclusion: This study revealed distinct TME cell patterns that could impact prognosis in ER+ breast cancer, and these findings concur with our previous IHC analyses. Incorporating this characteristic into current prognostic models may improve risk stratification, to accurately and safely personalise therapy.

77: Kinnex Long-Read Resequencing of Acute Myeloid Leukaemia Single-Cell RNA Samples Improves Detection of Malignant Genomic Alterations

Micheál Ó Dálaigh[1,2,3], Jacopo Umberto Verga[1,2,3], Pilib Ó Broin[2,3], Eva Szegezdi[1,3]

[1]School of Biological and Chemical Sciences, University of Galway, Galway, Ireland. [2]School of Mathematical and Statistical Sciences, University of Galway, Galway, Ireland. [3]The SFI Centre for Research Training in Genomics Data Science, Galway, Ireland

Introduction: 10x Genomics single-cell RNA sequencing (scRNA-seq) uses short-read sequencing, ultimately biasing transcript coverage towards the mRNA capture site. This limits the detectability of genomic alterations (SNVs, indels, fusion gene junctions) in regions distal to the capture site. PacBio Kinnex scRNA-seq utilises long-read sequencing to cover the whole transcript and is compatible with existing 10x cDNA libraries. Here, we investigate the ability of Kinnex to identify the malignant cell population in heterogeneous AML samples previously sequenced with the 10x 3' workflow, based on malignant genomic alterations.

Methods: 4 patient samples harbouring a range of different AML alterations (*TET2/DNMT3A* point mutations, NPM1 insertions, and a *DEK::NUP214* fusion) were resequenced with the Kinnex scRNA-seq kit and processed with the single-cell Iso-Seq analysis pipeline. CTAT-Mutations and pbfusion were used for variant calling and fusion gene identification respectively. VarTrix was used to assign identified CTAT-Mutations variants to individual cells.

Results: Kinnex recovered 29% fewer cells across the 4 samples. For cells which were sequenced with both technologies, CTAT-Mutations identified, on average, 63% more single-nucleotide cancer-relevant variants in the Kinnex data than the 10x.
On average, the Kinnex data covered AML mutation hotspot sites in *NRAS, TET2, FLT3, DNMT3A,* and *NPM1* in 1.84x more cells; e.g. in the *NPM1* mutated sample, Kinnex data revealed ~ 3x more (1000 vs. 280) mutant *NPM1* cells than 10x. The Kinnex data also identified the *DEK::NUP214* fusion which was not detectable in the corresponding 10x data.

Discussion/Conclusion: Despite the lower number of cells recovered, overall, the Kinnex data identified more of the malignant population by improving coverage of AML mutation hotspots located distal to the 3' of relevant transcripts. Kinnex scRNA-seq data displays superior sensitivity in detecting multiple genomic alterations, highlighting the benefits of using long-read sequencing to characterise cancer samples in single-cell studies.

Population Genetics & Molecular Evolution

12: Evaluating Sequence and Structural Similarity Metrics for Predicting Shared Paralog Functions

Olivier Dennler[1,2,3], Colm Ryan[1,2,3]

[1]School of Medicine, University College Dublin, Dublin, Ireland. [2]School of Computer Science, University College Dublin, Dublin, Ireland. [3]Conway Institute, University College Dublin, Dublin, Ireland

Background. Gene duplication is the primary source of new genes, resulting in most genes having identifiable paralogs. Over time, paralog pairs may diverge in some respects but many retain the ability to perform the same functional role. Protein sequence identity is often used as a proxy for functional similarity and can predict shared functions between paralogs as revealed by synthetic lethal experiments. However, the advent of alternative protein representations, including embeddings from protein language models (PLMs) and predicted structures from AlphaFold, raises the possibility that alternative similarity metrics could better capture functional similarity between paralogs.

Methods. Here, using two species (budding yeast and human) and two different definitions of shared functionality (shared protein-protein interactions, synthetic lethality) we evaluated a variety of alternative similarity metrics. We modelled this as a binary classification problem – can similarity metrics derived from sequence, structure, and PLM embeddings effectively identify pairs of paralogs where both paralogs share functions?

Results. For some tasks, predicted structural similarity or PLM similarity outperform sequence identity, but more importantly these similarity metrics are not redundant with sequence identity, i.e. combining them with sequence identity leads to improved predictions of shared functionality. By adding contextual features, representing similarity to homologous proteins within and across species, we can significantly enhance our predictions of shared paralog functionality.

Conclusion. Overall, our results suggest that alternative similarity metrics capture complementary aspects of functional similarity beyond sequence identity alone.

22: Variable gene copy number in cancer-related pathways is associated with cancer prevalence across mammals

Sophie Matthews[1,2], Vahid Nikoonejad Fard[3], Marc Tollis[3,4], Cathal Seoighe[1,2]

[1]University of Galway, Galway, Ireland. [2]The SFI Centre for Research Training in Genomics Data Science, Galway, Ireland. [3]Northern Arizona University, Arizona, USA. [4]Arizona Cancer Evolution Center, Arizona State University, Arizona, USA

Background/Introduction: Cancer is a disease of multicellularity, observed across the tree of life. In principle, animals with larger body sizes and longer lifespans should be at increased risk of developing cancer. However, there is no strong relationship between these traits and cancer across mammals. Previous studies have proposed that increased copy number of cancer-related genes may enhance the robustness of cancer suppression pathways in long-lived mammals, but these studies have not extended beyond known cancer-related genes.

Materials/Methods: In this study, we conducted a phylogenetic generalised least squares (PGLS) analysis to test for associations between copy number of all protein-coding genes and longevity, body size, and cancer prevalence across 94 species of mammals. In addition to investigating the copy number of individual genes, we tested sets of related genes for a relationship between the aggregated gene copy number of the set and these traits.

Results: We did not find strong evidence to support the hypothesis that adaptive changes in gene copy number contribute to the lack of correlation between cancer prevalence and body size or lifespan. However, we found several biological processes where aggregate copy number was associated with malignancy rate. The strongest association was for the gene set relating to transforming growth factor-beta (TGF-β), a cytokine that plays a role in cancer progression.

Discussion/Conclusion: Overall, this study provides a comprehensive evaluation of the role of gene copy number in adaptation to body size and lifespan and sheds light on the contribution of gene copy number to variation in cancer prevalence across mammals.

32: Independent origins of spicules reconcile the evolutionary history of sponges (Porifera)

Maria Eleonora Rossi[1,2], Nathan James Kenny[3,4], Mattia Giacomelli[2], Joseph N Keating[2], Sandra Alvarez-Carretero[2], Astrid Schuster[5], Paco Cárdenas[6,7], Sergi Taboada[8,9,4], Vasiliki Koutsuveli[4,10], Philip Donoghue[2], Ana Riesgo[8,4], Davide Pisani[2]

[1]School of Biological and Chemical Sciences, University of Galway, Galway, Ireland. [2]Bristol Palaeobiology Group, School of Earth Sciences, University of Bristol, Bristol, United Kingdom. [3]Department of Biochemistry Te Tari Matū Koiora, University of Otago, Dunedin, New Zealand. [4]Life Sciences Department, The Natural History Museum, London, United Kingdom. [5]Department of Biology, University of Southern Denmark, Odense, Denmark. [6]Museum of Evolution, Uppsala University, Uppsala, Sweden. [7]Pharmacognosy, Department of Medicinal Chemistry, Uppsala University, Uppsala, Sweden. [8]Department of Biodiversity and Evolutionary Biology, National Museum of Natural Sciences (CSIC), Madrid, Spain. [9]Departamento de Biodiversidad, Ecología y Evolución, Facultad de Ciencias, Universidad 19 Complutense de Madrid, Madrid, Spain. [10]Division of Marine Ecology, Marine Evolutionary Ecology, GEOMAR Helmholtz Centre for Ocean Research Kiel, Kiel, Germany

Introduction: Sponges (Porifera) are highly effective ecosystem engineers, playing a critical role in global biogeochemical cycles, and they have been closely linked to the evolution of Earth's environments.

However, determining the evolutionary history of sponges has posed challenges. Molecular divergence times have pointed to a Cryogenian or middle Ediacaran origin for sponges. This contrasts with the oldest unequivocal fossil evidence for sponges, consisting of silicified loose spicules dating to the latest Ediacaran, which leaves a 150-million-year gap in the sponge fossil record. The crucial question revolves around whether the common ancestor of sponges had spicules and if they were silicified. In this study, we employ an extensive phylogenomic dataset of sponges and revised fossil evidence to test hypotheses of spicule evolution.

Methods: We used phylogenomics, time calibration, and ancestral state reconstruction to address this question.

Results: Our analyses lead to a full revision of the evolutionary history of sponges. We present a new dated phylogeny, establishing that sponges originated in the middle Ediacaran rather than the Tonian period. Sponges are confirmed to include two main lineages: Silicea (Hexactinellida plus Demospongiae) and Calcarea plus Homoscleromorpha. Neither the last common sponge ancestor, nor the last common ancestors of Silicea, and Calcarea plus Homoscleromorpha possessed silicified spicules that independently evolved four times. Diversification rates analyses demonstrate that the origin of spicules did not align with species radiations.

Discussion: Taken together, our new molecular timescale, which substantially reduces the gap between fossil and molecular estimates for the origin of Porifera, and our conclusions that early sponges did not possess siliceous spicules, reconcile genomic and fossil evidence for the origin of sponges.

Metagenomics & Pathogen Surveillance

61: Genomic identification and characterisation of novel bacterial species from Space Craft Assembly Clean Rooms

Kate Ryan[1,2], Francesca McDonagh[1], Aneta Kovarova[1], Liam Burke[1], Matthew Dorman[3], Kasthuri Venkateswaran[4], Georgios Miliotis[1]

[1]Antimicrobial Resistance and Microbial Ecology (ARME) Group, School of Medicine, University of Galway, Galway, Ireland. [2]SFI Centre for Research Training in Genomics Data Science, Galway, Ireland. [3]School of Mathematical & Statistical Sciences, University of Galway, Galway, Ireland. [4]NASA Jet Propulsion Laboratory, California Institute of Technology, Pasadena, California, USA

Background/Introduction: Cleanrooms at spacecraft assembly facilities harbour unique microbial communities that adapt to their oligotrophic environment, shaped by selective pressures such as low organic material, high sanitation, controlled humidity, and restricted airflow. The study of extremophiles in these environments is essential, especially with upcoming missions to Mars and the Moon, as the potential for these microorganisms to survive on extraterrestrial surfaces raises concerns about forward contamination. These novel species may also possess unique biosynthetic mechanisms that contribute to their persistence in sterile conditions.

Materials/Methods: The sequences of 126 bacterial genomes from the Spacecraft Assembly Facility at the Jet Propulsion Laboratory (JPL) and the Mars Odyssey spacecraft at Kennedy Space Centre were investigated. Genus and species identification was performed using GTDB-Tk, Kraken, and subsequent Average Nucleotide Identity (ANI) and Average Amino Acid Identity (AAI) analyses. Strains with less than 95% ANI to any validly published type strain were considered potentially novel. We identified biosynthetic gene clusters (BGCs) using AntiSMASH to assess biotechnological potential and stress resistance.

Results:18 potentially novel species across 13 genera were identified. The newly identified species contains several biosynthetic gene clusters (BGCs), including those responsible for ectoine synthesis, which is associated with resistance to osmotic stress, high salinity, and survival in dry conditions. Furthermore, terpenes, betalactones and siderophores have been identified, which enhance microbial community adaptation to harsh environmental conditions. Additionally, the presence of *agrB-like* and *agrD* genes which together support biofilm formation were identified in a *Paenibacillus* species, which also carried a Type-IC CRISPR-Cas system.

Discussion/Conclusion: Cleanrooms foster novel extremophiles that may resist extreme conditions and produce bioactive compounds. As space missions to bodies like Mars and the Moon increase, understanding these extremophiles is crucial for preventing contamination, protecting extraterrestrial environments, and ensuring the safety of future human settlers.

67: *In silico* phage-bacteria infection networks (PBINs) of *Streptococcus suis* reveal co-evolution patterns between host and prophage

Emmet Campbell[1], Timofey Skvortsov[1], Nicholas Dimonaco[1,2], Lucy Dillon[1], Christopher Creevey[1]

[1]Queen's University Belfast, Belfast, United Kingdom. [2]Aberystwyth University, Aberystwyth, United Kingdom

Background/Introduction: Phage-bacteria infection networks (PBINs) and evaluation of nestedness or modularity is guides understanding of host interactions, which can help the identification of therapeutic phages. In silico methods may help expedite this process, though they face challenges from limited in vitro data and early-stage host prediction methods. This study aims to simulate PBINs by using prophage prediction and k-mer-based comparisons to group bacteria and prophages into communities, approximating PBIN structures.

Materials/Methods: We analyzed 2,119 *Streptococcus suis* genomes downloaded from BV-BRC with specific filters (WGS, Complete, Good). *De novo* gene annotations were created with Prokka, anti-phage genes detected with PADLOC, and prophage regions were identified using geNomad. Core gene alignments and phylogenetic trees were generated with Roary and FastTree, while prophage proteomic trees were created using ViPTree. Sourmash was used for k-mer comparisons, and custom python scripts used for community formation of genomes and PBIN creation.

Results: Iterative edge weight removal generated communities with larger core genomes and reduced total pangenome compared to the whole species. *S. suis* phylogeny aligned with bacterial communities, prophage community presence, and phage-defense genes. Prophages within communities clustered on the proteomic tree based on community, and communities showed similar genome size.

Discussion/Conclusion: Patterns of prophage community presence/absence and phage-defense genes aligning with host phylogeny suggest co-evolutionary dynamics of detected prophages. These may reflect resistance, superinfection exclusion, or susceptibility, though further study is needed to distinguish these patterns from cases where hosts and phages have not interacted. These networks may facilitate the prediction of host susceptibility/phage infection potential through insertion of target isolates into host PBINs.

71: Validation of a field-deployable automated DNA extraction system as a tool for assessment of microbial diversity in marine ecosystems

Georgios Miliotis[1], Diego Jiménez[2], Tahira Jamil[2], Junia Schultz[2], Niketan Patel[2], Lila Aldakhee[2], Nicholas Kontis[2], Francisca García[2], Helena Villela[2], Gustavo Duarte[2], Adam Barno[2], Andy Page[3], Season Wong[4], Adam Kabza[4], Alexandre Putra[2], Park Changsook[2], Angel Angelov[2], Patrick Driguez[2], Raquel Peixoto[2], Stefan Green[5], Anna Tumeo[1], Scott Tighe[6], Alexandre Rosado[2], Kasthuri Venkateswaran[7]

[1]University of Galway, Galway, Ireland. [2]KAUST, Thuwal, Saudi Arabia. [3]Innovaprep, Drexel, USA. [4]AI Biosciences, Texas, USA. [5]RUSH, Chicago, USA. [6]University of Vermont, Vermont, USA. [7]NASA-JPL California Institute of Technology, Pasadena, USA

Background: Microbial monitoring of environmental ecosystems is susceptible to biases from sample storage, DNA extraction methods, contaminants, sequencing technology, and computational analyses, potentially skewing biological interpretations. This study aimed to validate a field-deployable automated nucleic acid extraction system, xTitan, coupled with full-length 16S rRNA gene and shotgun metagenomic sequencing, to assess microbial diversity and composition in marine-derived samples, including corals, mangrove sediments, and seawater. By minimizing processing-associated biases, this approach seeks to enhance the recovery of broader microbial diversity. We compared DNA-extraction using xTitan in the field with in-lab extractions via xTitan and a commercial Qiagen kit.

Results: Both xTitan and Qiagen methods showed similar alpha diversity metrics and prokaryotic DNA proportions. However, DNA contaminants ("kitomes") significantly influenced bacterial composition profiles in coral samples across all extraction methods. The xTitan system generally increased alpha diversity in most environmental samples. Extraction strategies affected bacterial community structures in corals (*Pocillopora verrucosa*), mangrove sediments, and seawater. In mangrove sediments, in-field xTitan extraction resulted in significant differences in beta diversity and functional profiles compared to in-lab xTitan. Hundreds of amplicon sequence variants (ASVs) were differentially abundant when DNA was extracted on-site, including a higher abundance of *Endozoicomonas acroporae* in corals and decreased Cyanobacteria in seawater after short-term ice transportation. In mangrove sediments, Balneolaceae taxa were consistently abundant across methods.

Conclusions: The xTitan system provided consistent diversity results comparable to Qiagen when DNA was extracted in-lab. However, significant differences in beta diversity and taxonomic profiles emerged when extraction was performed in-field versus in-lab, particularly in mangrove sediments and seawater. These findings underscore the importance of on-site DNA extraction to accurately detect marine-associated microbes sensitive to storage and transportation conditions. This study validates the xTitan system for marine-derived samples, highlighting the impacts of DNA extraction methods, contaminants, and sample handling on microbial community analyses.

48: Secrets in the Sewers - Revealing the hidden diversity of SARS-CoV-2 in Northern Irish wastewater

John-Paul Wilkins, Francesco Rubino, Andrew J. Lee, Marina I. Reyne, Evan Troendle, Timofey Skvortsov, Deirdre F. Gilpin, Jennifer M. McKinley, David Simpson, John W. McGrath, Connor G.G. Bamford, Christopher J. Creevey

Queen's University Belfast, Belfast, United Kingdom

Background/Introduction: Whole-genome sequencing of Sars-CoV-2 from symptomatic individuals has tracked viral evolution in Northern Ireland as part of a UK-wide initiative. Expanding sequencing surveillance to wastewater, covering over 60% of NI's population, enables detection of viral sequences from asymptomatic and pre-symptomatic individuals, offering a broader view of transmission. By integrating clinical and wastewater data, we identify previously undetected 'cryptic' variants—sequences that, without this approach, might be misclassified or dismissed as noise by using standard SNP-based databases.

Materials/Methods: Wastewater samples from 31 treatment plants across Northern Ireland were sequenced using the ARTIC protocol. These mixed samples were then disentangled to allow reconstruction of individual viral spike protein haplotypes. A hierarchical classification system based on nucleotide sequence similarity linked 21,304 reconstructed haplotypes with 42,427 NI clinical sequences. Viral variants present in multiple wastewater samples but absent in the clinical dataset were identified for further analysis.

Results: Over 75% of reconstructed wastewater spike protein haplotypes clustered with clinically derived viral sequences. Notably, a subset of reoccurring haplotypes did not correspond to the clinical sampling. Temporal and geographic analyses provided robust evidence of these cryptic variants' circulation within the community, correlating their emergence and disappearance with transitions between dominant viral strains.

Discussion/Conclusion: Wastewater genomic sequencing provides real-time data on pathogen outbreaks across large populations at a fraction of the cost of clinical sequencing programs. Through the development of novel computational tools we overcome some of the unique challenges associated with wastewater sampling to reconstruct viral spike protein haplotypes. By integrating the classification of clinical and wastewater sequences, we unveiled the widespread circulation of multiple cryptic SARS-CoV-2 variants, providing new insights into pathogen evolution during critical moments in the pandemic.

Genomics of Health & Disease

11: Characterization of the mouse intra-amygdala kainic acid model at single-cell resolution reveals cell-type specific contributions to epilepsy phenotype

Javier Villegas Salmeron[1,2,3], Albert Sanfeliu[1,3], Omar Mamad[1,3], Niamh M C Connolly[1,3], David Henshall[1,3]

[1]Royal College of Surgeons in Ireland, Dublin, Ireland. [2]The SFI Centre for Research Training in Genomics Data Science, Galway, Ireland. [3]Futureneuro, SFI Research Centre for Chronic and Rare Neurological Diseases, Dublin, Ireland

Introduction: Single-cell sequencing technologies are transforming our ability to resolve the cellular diversity of organs and the cell type-specific contributing factor to a broad range of diseases. Mouse models are critical tools to model the pathophysiology of epilepsies, as well as to evaluate potential novel therapeutics. The intra-amygdala kainic acid model in mice has emerged as one of the leading models of drug-resistant temporal lobe epilepsy (TLE), displaying unilateral neuropathology, frequent behavioral seizures and broad resistance to anti-seizure medicines. Additional insights into the pathomechanisms of epilepsy in this model may enhance its use and understanding. Here, we generated a single-nuclei RNA sequencing dataset from hippocampal tissue from the intra-amygdala kainic acid (KA) mouse model (C57Bl/6) to better understand this model.

Materials/Methods: A total of 14052 nuclei from 4 mice were sequenced using the 10X platform, epileptic mice treated with KA and mice treated with a PBS as controls. Count matrices were generated and analyzed using Seurat v5. After quality control, cell types were annotated using the MapMyCells software from the Allen Brain Atlas. Differential expression analysis was performed after pseudo-bulking the samples using the DESeq2 software.

Results: Annotation of the dataset identified all major transcriptomic subtypes of hippocampal cells and differential expression analysis between conditions revealed cell-subtype specific dysregulation in the epileptic mice, where the CA1/subiculum/CA3 cells and the dentate gyrus appeared to be the most affected areas in the model.

Discussion: In conclusion, we present the first single-cell data on the intra-amygdala kainic acid mouse model of TLE. These findings enhance our understanding of the cellular and molecular processes in this model and identify both shared and distinct features compared to human TLE.

16: The impact of copy number variants in diagnosis and severity of autosomal dominant polycystic kidney disease

Sophia Heneghan[1,2], Elhussein Elhassan[3,4], Hamidah Ghani[1,2,5], Kane Collins[1,2], Monika Sigg[6], Peter Conlon[3,4], Gianpiero Cavalleri[1,2,5], Katherine Benson[1]

[1]School of Pharmacy and Biomolecular Sciences, Royal College of Surgeons in Ireland, Dublin, Ireland. [2]SFI Centre for Research Training in Genomics Data Science, Dublin, Ireland. [3]Department of Nephrology and Transplantation, Beaumont Hospital, Dublin, Ireland. [4]Department of Medicine, Royal College of Surgeons in Ireland, Dublin, Ireland. [5]FutureNeuro SFI Research Centre, Royal College of Surgeons in Ireland, Dublin, Ireland. [6]BioMarin Pharmaceutical Inc., Novato, California, USA

Introduction: Autosomal dominant polycystic kidney disease (ADPKD) is characterized by cyst development and kidney enlargement, inducing progressive kidney failure. Mutations in *PKD1* or *PKD2* cause most cases, but some remain genetically unresolved and have variation in disease severity. Here, we assessed the contribution of copy number variants (CNVs) in 1) unresolved diagnoses and 2) disease severity.

Methods: Using ClinCNV, CNVs in cystic kidney genes were identified from next-generation sequencing (NGS) of 378 ADPKD patients. CNVs were validated from array-CGH of 5 samples and MLPA of 50 samples. Regression models assessed the association of the presence or burden (in kilobases) of CNVs to age at kidney failure, height-adjusted total kidney volume, and liver cysts as clinical outcomes.

Results: Disease-causing CNVs were identified in 13 individuals across 7 families, increasing the diagnostic yield from 86% to 89.5%. ClinCNV detected *PKD1*/*PKD2* CNVs with 100% specificity and 62.5% sensitivity compared to MLPA; array-CGH data indicated 83% sensitivity for regions with sufficient probe coverage.

After filtering, 5.7% (21/371 individuals) had an additional non-diagnostic cystic kidney gene CNV, all with a pathogenic PKD variant alongside the CNV. Regression models indicated an association of duplication burden in genes unrelated to cystic kidney disease to a mild but significantly lower likelihood of liver cysts.

Conclusions: ClinCNV effectively identifies disease-causing CNVs in ADPKD, suggesting its utility when initial NGS analysis is inconclusive. Regression analyses indicate a possible connection between non-cystic kidney gene duplications and absence of liver cysts. This study underscores the importance of bioinformatic tools in repurposing NGS for diagnostics in ADPKD.

19: DNA Methylation in Vulnerability to Opioid Use Disorder

Nicole Glendinning[1], Sarah Abdulmalek[1], Analia Kinen[1], James Mackle[1], Brittany Kuhn[2], Nazzareno Cannella[3], Laura Soverchia[3], Massimo Ubaldi[3], Leah Solberg Woods[4], Peter Kalivas[2], Roberto Ciccocioppo[3], Hui Wang[5], Dongjun Chung[6,7,8], Gary Hardiman[1,9]

[1]School of Biological Sciences, and Institute for Global Food Security, Queen's University Belfast, Belfast, United Kingdom. [2]Department of Neuroscience, Medical University of South Carolina, Charleston, USA. [3]School of Pharmacy, Center for Neuroscience, Pharmacology Unit, University of Camerino, Camerino, Italy. [4]Department of Internal Medicine, Wake Forest University, Winston-Salem, USA. [5]School of Electronics, Electrical Engineering and Computer Science, Queen's University Belfast, Belfast, United Kingdom. [6]The Interdisciplinary Ph.D. Program in Biostatistics, The Ohio State University, Columbus, USA. [7]Department of Biomedical Informatics, The Ohio State University, Columbus, USA. [8]Pelotonia Institute for Immuno-Oncology, The James Comprehensive Cancer Center, The Ohio State University, Columbus, USA. [9]Departments of Medicine and Public Health Sciences, Medical University of South Carolina, Charleston, USA

Background/Introduction: Opioid Use Disorder (OUD) is characterised by the habitual consumption of opioids such as heroin or morphine, despite substantial negative consequences associated with continued use. OUD represents a significant health and socio-economic concern both globally, as well as locally within Ireland and the UK. Although a combination of genetic, epigenetic, and environmental factors collaboratively influence vulnerability to OUD, current clinical risk prediction measures rely exclusively on the assessment of environmental and social influences.

Materials/Methods: The Hardiman lab group is involved in a multi-disciplinary collaborative study which has developed an outbred rodent model of heroin self-administration. Based on the results of behavioural tests, rodents can be stratified into two populations, those which are vulnerable to OUD, and those which are resilient. Blood methylation data from reduced representation bisulfite sequencing has been collected from vulnerable and resilient rodents both before, and after heroin exposure. Data was analysed using methyKit to identify differences in methylation patterns between vulnerable and resilient rodents.

Results: Analysis identified numerous differentially methylated cytosines (DMCs), and differentially methylated regions (DMRs) between vulnerable and resilient rodents. Many DMCs and DMRs were detected within pre-heroin samples, and localised within genes that have been previously associated with various psychiatric disorders.

Discussion/Conclusion: The identification of DMCs and DMRs highlight the presence of discernible epigenetic features that may be important in determining vulnerability to OUD, and has identified genes and pathways for further investigation. The fact differential methylation is detectable in samples that have not yet been exposed to heroin creates potential for findings to be applied to development of predictive epigenetic biomarkers for opioid vulnerability. Such markers would be highly beneficial in guiding clinicians to make informed decisions when prescribing opioids for pain management, as well as identification and timely application of prevention strategies for at risk individuals.

37: Predicting Age and Identifying Aging-Related Genes from Muscle Gene Expression Data

Karen Guerrero Vazquez[1], Qin Hong[2], Pilib Ó Broin[1], Katarzyna Goljanek-Whysall[1]

[1]University of Galway, Galway, Ireland. [2]Old Dominion University, Norfolk, USA

Background: Understanding the molecular mechanisms underlying aging is essential for developing targeted interventions to mitigate age-related diseases. While recent advancements have employed omics analyses to identify signature genes of various age-related conditions, skeletal muscle aging remains relatively underexplored.

Methods: We present a deep learning approach to predict the age of individuals based on gene expression data from human vastus lateralis, while identifying key genes associated with muscle aging. We identified genes implicated in well-known aging mechanisms such as inflammation, cell proliferation, and autophagy, intersecting with genes previously published as differentially expressed in sarcopenia.

Our data collection integrates expression data from 19 microarray projects and 12 RNA-seq experiments, totalling more than 900 samples from different ethnicities. The datasets were rigorously curated through pseudoalignment, sequence ID analysis, batch correction, and two-point normalization. Datasets were binned across three age groups: young (18-35 years old), middle-aged (35-65), and old (older than 65).

Results: Our results show that a set of 300 genes is sufficient to predict an individual's muscle age with a mean absolute error of 8.6. Our study represents the first comprehensive effort to predict age based on muscle tissue gene expression profiles at this scale that can be used to predict the biological age of the muscle.

Discussion: Our methodology overextends existing approaches by effectively generalizing across diverse data sources, ensuring independence from project-specific biases and ethnicities, enhancing applicability in real-world scenarios.

The genes identified in our study hold potential implications for understanding the molecular basis of aging-related processes in muscle. Furthermore, these findings provide a foundation for future research to develop targeted interventions for sarcopenia. In ongoing work, we plan to integrate these genes into a microRNA regulatory network model to identify potential therapeutic targets for sarcopenia, underscoring the translational relevance of our findings.

Methods & Algorithms

10: panGPT: An AI transformer for generating large pangenome models

JAMES MCINERNEY

Liverpool, Liverpool, United Kingdom

The increasing availability of microbial genome sequences has revolutionized our understanding of genetic diversity. However, leveraging this vast data for predictive modelling remains challenging. I present panGPT, an innovative AI transformer designed to generate and analyse large pangenome models. The model can be trained on tokenized pangenome datasets using a transformer-like architecture for next-token prediction. panGPT employs a tokenization approach, where individual genomes are expected to be broken into functional units representing small units of Darwinian selection (e.g., promoters, protein domains, terminators). panGPT implements a simple Transformer model with positional encoding that takes as input a pangenome file with each genome on a single line. Each gene name is separated by a space and gene names must be consistent across all genomes - i.e. homologs or orthologs all get the same name. Output from ROARY or PANAROO can be adapted to provide this kinds of input file. The LPM is then trained on these data, using the Transformers approach outlined in Vaswani et al., (2017). The supplementary program panPrompt can then be used to prompt the model with a string of input gene names, whereupon it will carry out "next token prediction" for a specified number of tokens.

35: Supervised multimodal demultiplexing outperforms conventional demultiplexing of scRNAseq

Michael P Lynch[1], Yufei Wang[2,3], Shannan Ho Sui[4], Laurent Gatto[5], Aedin C Culhane[1]

[1]School of Medicine, Limerick Digital Cancer Research Centre, Health Research Institute (HRI), University of Limerick, Limerick, Ireland. [2]Department of Cancer Immunology and Virology, Dana-Farber Cancer Institute, Boston, USA. [3]Harvard Medical School, Boston, USA. [4]Harvard T.H. Chan School of Public Health, Boston, USA. [5]Computational Biology and Bioinformatics Unit (CBIO), de Duve Institute, UCLouvain, Brussels, Belgium

Background: Multiplexing single-cell RNA sequencing (scRNAseq) allows simultaneous sequencing of multiple samples in a single scRNAseq experiment, greatly increasing sample throughput, which reduces cost and facilitates larger scale studies. Poor scRNAseq demultiplexing algorithm performance can result in low data yield, lost data and increased cost.

Materials/Methods: We investigated factors that impact scRNAseq demultiplexing performance of existing algorithms and developed a multimodal supervised algorithm and R package, demuxSNP, which leverages both cell hashing and genetic variation between individuals (SNPs). demuxSNP infers the genotypes of singlet and doublet clusters from the hashing data and uses these to predict hashing negative, uncertain or doublet cells using a nearest neighbour approach adapted for missing data. We benchmarked demuxSNP against existing hashing, genotype-free SNP and hybrid methods on simulated and real data from renal cell cancer.

Results: demuxSNP outperformed standalone hashing methods on low-quality hashing data benchmark, improved overall classification accuracy and allowed more high RNA quality cells to be recovered. Through varying simulated doublet rates, we showed that genotype-free SNP and hybrid methods that leverage them, were impacted by class size imbalance and doublet rate. demuxSNP's supervised approach was more robust to doublet rate in experiments with class size imbalance.

Discussion/Conclusion: demuxSNP uses hashing and SNP data to demultiplex datasets with low hashing quality where biological samples are genetically distinct. Unassigned, negative and uncertain cells with high RNA quality are recovered, making more cells available for analysis. demuxSNP is available as an R/Bioconductor package (https://doi.org/doi:10.18129/B9.bioc.demuxSNP).

References

Lynch MP, Wang Y, Ho Sui S, Gatto L, Culhane AC. demuxSNP: supervised demultiplexing single-cell RNA sequencing using cell hashing and SNPs. GigaScience (Accepted) doi: https://doi.org/10.1093/gigascience/giae090

62: Advancing Cancer Research through Integration of Liquid Biopsy Data in cBioPortal: Insights from the All-Ireland Cancer Liquid Biopsies Consortium (CLuB)

Seyed Aghil Hooshmand, Pilib Ó Broin

School of Mathematical and Statistical Sciences, University of Galway, Galway, Ireland

Background/Introduction: Liquid biopsy is a minimally invasive technique for analyzing tumor-derived biomarkers—such as circulating tumor cells (CTCs), extracellular vesicles (EVs), and circulating tumor DNA (ctDNA)—from biological fluids. This approach provides real-time insights into cancer progression, avoiding the complications of invasive tissue biopsies and enhancing personalized cancer care.

Materials/Methods: Through the HEA-funded All-Ireland Cancer Liquid Biopsies Consortium (CLuB), we are deploying and extending cBioPortal, an open-source web platform hosted on Amazon Web Services (AWS), for visualizing and analyzing cancer genomic data. cBioPortal integrates clinical and genomic information, improving data accessibility and functionality for the research community. To ensure data security, Keycloak identity and access management restricts sensitive data access to CLuB-approved users.

Results: The cBioPortal deployment within CLuB includes diverse datasets for lung, breast, and ovarian cancers. For lung cancer, clinical data from 136 patients (57 features) and genomic data from 315 samples (47 features) were available, including 228 CTC and 40 EV samples. Breast cancer data comprises clinical information from 48 patients (62 features), methylation data from 22 case and 18 control samples. Ovarian cancer data includes clinical data from 4 patients (109 features) and 4 samples (51 features). Analysis identified significant relationships between clinical outcomes and genomic profiles, showcasing cBioPortal's potential to accelerate cancer research.

Discussion/Conclusion: Integrating clinical and genomic datasets through cBioPortal enhances cancer research by uncovering relationships between clinical outcomes and molecular profiles. Future work will apply advanced computational techniques, including machine learning and network analysis, to identify novel biomarkers, supporting more precise diagnostics and prognostic assessments in oncology.

73: Adverse Drug Reaction Profile Prediction: Denoising, Signal Enhancement and Missing Row Imputation

Yezhao Zhong

University of Galway, Galway, Ireland

Background/Introduction: Adverse Drug Reactions (ADRs) cause significant risks to human health, making it essential to identify potential ADRs in early stage of drug development. However, this process is costly and time-consuming. Therefore, developing advanced computational methods to predict ADR profiles is significantly important.

Materials/Methods: We developed a series of approaches to enhance ADR profile prediction across three main strategies. First, to address noise in imbalanced ADR data, we proposed a novel hybrid method, Kernel Regression (KR) on V (VKR), which combines Non-negative Matrix Factorization (NMF) with KR on the drug-component matrix V derived from NMF. Second, we introduced Smoothed KR (SKR) to enhance signal detection for low-frequent ADRs. Finally, we developed a missing row imputation strategy to enrich drug databases by imputing missing row of features for non-overlapping drugs, increasing the dataset's breadth and predictive capability.

Results: VKR demonstrated superior performance over existing methods on both single-feature and integrated-feature datasets. SKR significantly improved prediction performance for low-frequent ADRs, outperforming other methods in this challenging category, but the performance across frequent ADRs was slightly improved among models. The extended dataset size further enhanced model performance with both single-feature and integrated-feature, indicating the benefit of the missing row imputation strategy.

Discussion/Conclusion: VKR effectively reduces noise introduced by imbalanced data and binary representation of drug-ADR data, while SKR addresses the gap in low-frequent ADR prediction, which is crucial for real-world applications. Current models often overlook low-frequent ADRs due to the dominance of frequent ADR signals, but capturing these low-frequent ADR profiles is essential. The limited overlap of drugs across feature databases significantly reduces usable training data, making the missing row imputation strategy a valuable addition for preserving critical drug information and improving predictive outcomes.

Lightning talks

7: Clustering imbalanced functional data - enhancing the clustering accuracy of time-course gene expression data

Catherine Higgins, Michelle Carey

University College Dublin, Dublin, Ireland

Background/Introduction: Functional data analysis studies realizations of a smooth continuum, recorded with some error, observed at discrete points in time. Taken together these observations can be regarded as a curve or function. Functional clustering techniques classify a sample of curves into homogeneous groups of curves, without prior knowledge of the actual underlying clustering structure.

Materials/Methods: The class imbalance problem, that is where the number of curves in one cluster is considerably more than the number of curves in another cluster, is a challenging issue. Some minor clusters are often ignored and/or misclassified into the major clusters, which leads to a poor clustering accuracy and a distortion of the true underlying cluster structure. Learning from imbalanced data has numerous real-world applications, such as detecting faults in industrial machinery and identifying genes affected by infection or disease. While the class imbalance problem has been extensively studied in the context of supervised classification, this problem has received limited attention in an unsupervised environment.

Results: We adapt the iterative hierarchical clustering approach for multivariate data to a functional data context, thus introducing a novel method called functional iterative hierarchical clustering (funIHC) for clustering imbalanced functional data.

Discussion/Conclusion: Applying funIHC to gene expression data related to human influenza infection induced by the H3N2 virus, we identify five distinct and biologically meaningful patterns of gene expression. This method significantly enhances the clustering accuracy of time-course gene expression data compared to sixteen leading alternatives.

8: Exploring endocrine disrupting pathways using knowledge graph and network biology

Mariagiovanna Pais[1], Amin Arif[1], Ali Can[2], Jack Feltham[3], Richard Currie[3], Guillermo Lopez Campos[2], Gary Hardiman[1]

[1]Queen's University, Belfast, United Kingdom. [2]Queen's University, Belfast, United Kingdom. [3]Syngenta, Bracknell, United Kingdom

Background/Introduction: Endocrine disruptors (ED) pose significant risks to human health and the environment by interfering with hormonal systems. These substances, including chemicals such as pesticides and microplastics, can mimic or disrupt the body's hormones, leading to a range of adverse health outcomes. Given the complexity of these effects, this interdisciplinary project focuses on investigating the intricate pathways influenced by ED through the development of a computational model designed to analyse transcriptomics data from ED exposures.

Materials/Methods: By employing methodologies from knowledge graphs and network biology, this approach aims to facilitate the identification of critical genes and proteins involved in key pathways, such as Hormone Regulation, Oxidative Stress, MAPK Signalling, and MNT Pathways, within the Adverse Outcome Pathway (AOP) framework to understand how these disruptors exert their influence at the molecular level. A fundamental component of this project is the incorporation of Protein-Protein Interaction (PPI) mapping, which reveals protein interactions within the pathways impacted by ED. Additionally, literature mining is employed to help identify relevant gene and hub genes associated with these chemicals, enriching the overall analysis by integrating insights from existing scientific research.

Results: To support the findings, differential expression analysis is being exploited, alongside the implementation of machine learning techniques for pattern recognition, ensuring a comprehensive approach to data analysis.

Discussion/Conclusion: This multifaceted strategy is expected to uncover valuable insights into the pathways affected by endocrine disruptors, thereby contributing to a deeper understanding of their biological implications and the mechanisms through which they operate.

9: Deep Learning approach for detecting and segmenting Perineural Invasion in Colon, Prostate, and Pancreatic cancers

Ahmad Alkhan, Aedin Culhane, Meghana Kshirsagar, Conor Ryan

University of Limerick, limerick, Ireland

Background/Introduction: Perineural invasion (PNI) is a critical pathological feature associated with aggressive tumor behavior and poor prognosis, observed in cancers such as pancreatic, prostate, and colon cancer. This study aims to benchmark deep learning frameworks for the automated detection and segmentation of PNI in whole slide images (WSIs) from colon, pancreas, and prostate tumor tissues.

Materials/Methods: A dataset of 150 pathologist-annotated WSIs was used to evaluate the performance of four models: EfficientNet-b0, InceptionResNetV2, MobileViT_s, and MIT-b2. For preprocessing, 224 x 224 patches were extracted, and a UNet-based architecture was utilized for segmentation. Metrics such as Matthews Correlation Coefficient (MCC), precision, recall, and Area Under the Curve (AUC) were used for model evaluation, with specific attention to handling class imbalances.

Results: Among the models tested, MobileViT_s demonstrated superior performance, achieving the highest MCC scores of 0.36 for prostate, 0.35 for pancreas, and 0.28 for colon, reflecting its robustness in handling imbalanced datasets. Additionally, MobileViT_s achieved the highest AUC values: 0.82 for prostate, 0.78 for pancreas, and 0.74 for colon, outperforming other models in distinguishing PNI from non-PNI cases. It also maintained balanced performance in precision and recall across the datasets.

Discussion/Conclusion: The study successfully benchmarked multiple deep learning models for PNI detection, with MobileViT_s emerging as the most effective. Its performance across key metrics highlights its potential for clinical implementation, where it could improve diagnostic accuracy, incidence rate variability, and streamline workflows in oncology. Future work will explore the interaction between PNI and patient response markers, immune oncology, and tumor genomics.

23: Bioinformatics Refinement of CRISPR-Cas9 Knockout Screens Reveals Additional Genes Modulating Cellular Responses to CDC7 Inhibitors

Stefanus Bernard[1,2], Michael D. Rainey[2], Colm J. Ryan[1,3], Corrado Santocanale[1,2]

[1]SFI Centre for Research Training in Genomics Data Science, Galway, Ireland. [2]Centre for Chromosome Biology, School of Biological and Chemical Sciences, University of Galway, Galway, Ireland. [3]UCD Cancer Data Lab, Conway Institute of Biomolecular and Biomedical Research, School of Medicine, University College Dublin, Dublin, Ireland

Background/Introduction: CDC7 inhibitors (CDC7is) block the activity of CDC7, a kinase that promotes DNA replication and is overexpressed in cancers. Intriguingly the anti-proliferative activity of CDC7is varies across cell lines. To identify the genes determining sensitivity to CDC7is, we performed pooled CRISPR-KO screens in MCF10A cells using a library of ~150,000 sgRNAs targeting ~19,000 protein-coding genes with an average of 8 sgRNAs per gene. Bioinformatics analysis was conducted using a MAGeCK pipeline that integrates the results from multiple sgRNAs into gene level summaries. However, this analysis did not account for bias from sgRNAs that potentially target Cas9 to multiple genomic regions and falsely target another region (off-target), which could result in an inaccurate hit gene list and misleading biological conclusions. Therefore, we constructed a bioinformatics pipeline to refine the sgRNA library in our CRISPR screen.

Materials/Methods: The pipeline uses the sgRNA library file as input. The Hugo Gene Nomenclature Committee database of approved gene symbols was used to update the gene symbols associated with each sgRNA. All sgRNA sequences were aligned with a maximum of 2 mismatches into the human genome (GRCh38) to detect and filter out suspected off-target sgRNA. The sgRNA re-validation and annotation were performed using tools from a Bioconductor package.

Results: We used the pipeline to refine 150,076 sgRNAs from our library. We found at least 7% of the sgRNA's gene symbols were corrected. After refinement, the library was reduced to 134,371 annotated sgRNAs that uniquely targeted a single genomic locus.

Discussion/Conclusions: Reanalysis of the CRISPR screen using MAGeCK with refined sgRNA library allowed us to discover 90 additional hit genes that presumably regulate the cell responses to CDC7is. This pipeline should be validated using the public sgRNA libraries and screen datasets and can be developed to help all researchers performing CRISPR screen analysis.

26: Unravelling Resistance Mechanisms to Synthetic Lethal Therapies in Cancer Through Protein-Protein Interaction Networks

Metin Yazar[1,2,3], Colm Ryan[1,2,3]

[1]School of Computer Science, University College Dublin, Dublin, Ireland. [2]School of Medicine, University College Dublin, Dublin, Ireland. [3]Conway Institute of Biomolecular & Biomedical Research, University College Dublin, Dublin, Ireland

Background: Synthetic lethal interactions, wherein a mutation in one gene makes cells vulnerable to the disruption of another, can be exploited for the development of targeted cancer therapies. PARP inhibitors, which exploit these interactions, are now used in the treatment of specific breast, ovarian, and prostate cancer patients, with many other synthetic lethal therapies currently being evaluated in clinical trials. As with other targeted therapies, tumours have the potential to develop resistance to synthetic lethal treatments, thereby reducing their therapeutic efficacy. The mechanisms underlying the emergence of resistance to synthetic lethal therapies and whether the pathways to resistance might be somewhat predictable, remains unknown

Materials and Methods: Here, we tested the hypothesis that genes whose mutation confers drug resistance are enriched among the protein-protein interaction partners of synthetic lethal genes. We analysed 8 published genome-wide CRISPR-Cas9 drug resistance screens and determined the resistance genes for each synthetic lethal pair. Next, resistance and non-resistance genes were compared to the protein-protein interaction partners of each synthetic lethal gene pair obtained from STRING and BIOGRID physical databases

Results: Statistical analysis revealed that resistance genes were enriched among the protein-protein interaction partners of synthetic lethal pairs. These results suggest that mutations that confer resistance to synthetic lethal therapies are, to some extent, predictable

Discussion and Conclusion: As genome sequencing tumours that have developed resistance to targeted therapies often reveal hundreds of novel mutations, our results suggest that protein-protein interaction networks can be used to prioritise likely causal mutations. Furthermore, they may aid in the identification of combination treatments that delay the emergence of resistance or directly target resistance mechanisms.

31: A genetic network integrates regulation of the vegetative-reproductive phase transition in Arabidopsis thaliana

Pouya Motienoparvar[1], Ali Ebrahimi[2], Kaveh Kavousi[2], Mokhtar Jalali Javaran[3], Peter C. McKeown[1], Charles Spillane[1]

[1]University of Galway, Galway, Ireland. [2]University of Tehran, Tehran, Iran, Islamic Republic of. [3]Tarbiat Modares University, Tehran, Iran, Islamic Republic of

Background/Introduction: The transition from vegetative to reproductive phase (flowering) in *Arabidopsis thaliana* is regulated by a complex genetic network. Many genes involved in this process have been identified, but the integration of environmental and endogenous signals through this network remains poorly understood. This study focuses on identifying key genes within a genome-wide regulatory network using a systems biology approach, emphasizing the integration of flowering signals by central regulators.

Materials/Methods: Transcriptomic data for 22,810 genes from five *Arabidopsis thaliana* genotypes (Col-0, Ler, lfy-12, co-2, ft-2) under varying environmental conditions were analyzed. A gene co-expression network was constructed using the Weighted Gene Co-expression Network Analysis (WGCNA) and Hierarchical Complete Linkage Clustering (HCLC) methods. Control nodes were identified using Position Weight Matrices (PWMs) and control centrality metrics. Gene expression patterns were validated using quantitative real-time PCR (qRT-PCR).

Results: The analysis identified 77 core flowering genes and 31 controller genes within the regulatory network, with two key genes functioning as central regulators. Both genes exhibited oscillatory expression patterns during the vegetative-to-reproductive transition, implying a dual oscillation mechanism that integrates signals from various flowering pathways. A network control unit, the Transition Control Unit (TCU), was identified, comprising nine key genes, including the two central regulators.

Discussion/Conclusion: This study highlights the central roles of two key regulatory genes in controlling flowering time in *Arabidopsis*. Their oscillatory behavior suggests a "double oscillator" mechanism that integrates environmental and internal signals, facilitating robust control over the flowering transition. These insights could have practical applications in optimizing flowering time for crop improvement.

Keywords
Systems biology, chronobiology, flowering regulation, vegetative-to-reproductive transition, *Arabidopsis thaliana*, co-expression network, regulatory network, dual oscillation.

44: Factors in the Development of Extreme Ploidy States in Cancer

Elle Loughran, Ross P. Byrne, Russell L. McLaughlin, Máire Ní Leathlobhair, Aoife McLysaght

Trinity College Dublin, Dublin, Ireland

Background/Introduction: Aneuploidy is a hallmark of cancer. The vast majority of solid tumours carry at least one aneuploid chromosome, and in one third of cases ploidy has diverged severely from a diploid state towards polyploidy or low hypodiploidy. What distinguishes tumours that develop extreme ploidy aberrations from those that remain near-diploid?

Materials/Methods: Using data from the Cancer Genome Atlas and the Mitelman Database, we investigate specific factors that could promote or prevent the (a) origin (b) establishment or (c) expansion of a ploidy-aberrant clone, including germline and somatic mutations, environmental exposures and immunosurveillance.

Results: We find no common (GWAS) or rare germline variants linked with extreme ploidy in patient tumours, but report that somatic variants in TP53 and ARID1A are consistently implicated (positively and negatively, respectively) in both polyploidy and hypodiploidy. We identify mutational signatures significantly associated with extreme ploidy and search for associations with smoking, hypoxia and viral infection. Additionally, we will integrate CIBERSORT, neoantigen prediction and WGD timing to study the bidirectional relationship between tumour ploidy evolution and immune infiltration.

Discussion/Conclusion: Finally, we discuss the extent to which these extreme ploidies can be considered ‘natural kinds’ vs a continuum of aneuploid states, and consider the evidence for polyploidy and hypodiploidy as opposites vs as related manifestations of an underlying chromosomal instability phenotype.

50: Associating Mood Symptom Severity with Subcortical Brain Volumes in Bipolar Disorder and Major Depressive Disorder using an Item Response Theory Model

Emma Corley[1], John O'Connor[1], Brian Hallahan[1], Genevieve McPhilemy[1], Leila Nabulsi[2], Melody JY Kang[2], Elena Pozzi[3], Dick Veltman[4], Lianne Schmaal[3], Christopher R.K. Ching[2], Paul Thompson[2], Ole Andreasson[5], Colm McDonald[1], Dara M. Cannon[1]

[1]Clinical Neuroimaging Laboratory, Centre for Neuroimaging, Cognition, and Genomics (NICOG), University of Galway, Ireland, Galway, Ireland. [2]Imaging Genetics Center, Mark and Mary Stevens Institute for Neuroimaging and Informatics, University of Southern California, Los Angeles, California, California, USA. [3]Centre for Youth Mental Health, The University of Melbourne, Melbourne, Australia, Melbourne, Australia. [4]Department of Psychiatry, Amsterdam UMC, The Netherlands., Amsterdam, Netherlands. [5]Institute of Clinical Medicine, University of Oslo, Oslo, Norway, Oslo, Norway

Background: Mood rating scales are widely used to assess severity in Bipolar Disorder (BD) and Major Depressive Disorder (MDD), but the diversity of these scales complicates data aggregation and replication. This study optimally captures the common signature in mood symptom severity and its' relation to subcortical volume by examining item-level variance in a large global multisite dataset across disorders and rating scales.

Method: The study included 9,625 individuals from 49 independent sites (5,355 controls, 2,804 BD, 894 MDD). Depression was assessed using the Hamilton Depression Rating Scale (HAM-D), Montgomery-Åsberg Depression Rating Scale (MADRS), or Beck Depression Inventory (BDI). A graded item response theory (IRT) model was used to calculate trait-level parameters for response categories and discrimination parameters for each item (mirt, R). T_1-weighted structural brain images were acquired by each site and subcortical volume determined by segmentation (Freesurfer) and harmonized using ComBat.

Results: Discriminative items were sadness, suicidal thoughts, and reductions in work and activities(α>0.3), while sleep and appetite were less informative. Total- and IRT-scores correlated highly(r>0.83). IRT scores captured greater variability overall (Δ > 8%) and showed significant associations with subcortical volumes. Expected volumetric differences were detected relative to controls and lower volumes between MDD and BD in reward-related structures (e.g., accumbens, putamen) and higher ventricular and thalamic volumes(*p*FDR<0.05). IRT-derived scores were significantly associated with subcortical regions (MDD *r*FDR range=-0.099; -0.249, BD *r*FDR range= -0.069; -0.143) with cross-disorder differences in the accumbens(MDD>BD), hippocampus (BD>MDD) and ventricles (MDD>BD).

Discussion: IRT scores provide a more precise measure of symptom severity and reveal distinct brain volume differences. Lower volumes in reward-related regions and larger thalamic and ventricular volumes in BD, provide further insight into the neurobiological distinctions between BD and MDD due to the improved precision of the discriminative estimation of severity.

76: Mapping Genetic Determinants of DNA Methylation Across Early Development

Anna Grossbach[1,2,3], Alexandre A Lussier[3,4,5], Erin C Dunn[6], Andrew J Simpkin[1,2]

[1]School of Mathematical and Statistical Sciences, University of Galway, Galway, Ireland. [2]The SFI Centre for Research Training in Genomics Data Science, Galway, Ireland. [3]Psychiatric and Neurodevelopmental Genetics Unit, Center for Genomic Medicine, Massachusetts General Hospital, Boston, USA. [4]Department of Psychiatry, Harvard Medical School, Boston, USA. [5]Stanley Center for Psychiatric Research, The Broad Institute of Harvard and MIT, Cambridge, USA. [6]Department of Sociology, Purdue University, West Lafayette, USA

Background: Epigenetic mechanisms, such as DNA methylation (DNAm), play a key role in genomic regulation, thereby shaping various phenotypic outcomes. DNAm patterns are influenced by both genetic architecture and environmental exposures. Understanding the genetic contribution is crucial for unravelling environmental interactions, particularly during early childhood - a period marked by rapid developmental changes and heightened susceptibility to environmental influences.

Central to the genetic-epigenetic interplay are methylation quantitative trait loci (mQTLs) - genetic loci influencing DNAm states, either locally (cis-mQTL) or across larger genomic distances (trans-mQTL). Previous large cohort studies have mostly identified stable cis-mQTL effects across different life stages. However, our knowledge of the dynamic nature of mQTLs during early development, particularly regarding trans-mQTLs, is still limited due to insufficient longitudinal cohort data and the analytical challenges posed by analyzing vast numbers of associations.

Methods: In this study, we investigate the dynamics of cis- and trans-mQTLs during early childhood using data from the Drakenstein Child Health Study (DCHS; n=1,143), a longitudinal South African cohort. DCHS provides allele information for over 6 million genetic loci as well as DNAm data across 900k sites from whole blood samples at ages 1, 3, and 5.

Results: Our results indicate a substantial overlap of shared cis-mQTLs across the three time points studied. However, we also observed that a considerable proportion of trans-mQTLs exhibit age-specific effects throughout early development.

Discussion The stable nature of cis-mQTLs aligns with previous research, but highly age-specific characteristics of trans-mQTLs represent a novel finding. These age-specific effects indicate that genetic influences on DNAm can shift as children develop. Understanding longitudinal contributions is crucial to illuminate the complex interplay between genetics and epigenetics in developmental trajectories and disease susceptibility, particularly in historically underrepresented populations. Our results can guide future studies seeking genetic causal anchors for DNAm in childhood observational research.

Posters

13: Whole genome sequencing of pre- and post- treatment biopsies in a rectal cancer patient using long-read and short-read technologies

Lauren McAuley, Sinead Toomey, Bryan Hennessy, Simon Furney

Royal College of Surgeons in Ireland, Dublin, Ireland

Background: Clinical interest in long-read sequencing (LRS) technologies has piqued in recent times as LRS technologies have improved significantly, with accuracies now approaching that of short-read sequencing (SRS) methods like Illumina. Long read DNA sequencing offers advantages over SRS technologies including improved ability to resolve complex structural variants and to detect native DNA methylation. LRS technologies like nanopore sequencing (Oxford Nanopore; ONT) are advantageous as ONT WGS can provide more information in a single assay compared to Illumina. In this pilot study, the abilities of ONT and Illumina sequencing technologies to detect somatic variants in the pre- and post-treatment genome of a locally advanced rectal tumour were compared.

Materials/Methods: WGS using ONT and Illumina sequencing technologies was performed on pre- and post-treatment tumour and adjacent normal tissue samples, as well as matched blood, of a patient with locally advanced rectal cancer. Somatic SNVs and indels, copy number variants (CNVs), structural variants (SVs) and DNA methylation profiles were detected using ClairS, HATCHet, nanomonsv, and modkit, respectively, in ONT-WGS data. For Illumina-WGS, somatic SNVs and indels, CNVs, and SVs were detected using Mutect, HATCHet, and Manta.

Results: Regression analysis demonstrated that the concordance of SNVs between ONT-WGS and Illumina-WGS was affected by variant allele frequency in normal tissues. Concordance of indels was poor between sequencing technologies. Similar copy number profiles were detected in ONT-WGS and Illumina-WGS data. Types of structural variants detected by the sequencing technologies differed (greater number of insertions in ONT-WGS, more breakends in Illumina-WGS).

Discussion/Conclusion: Small somatic variant calling in ONT-WGS data showed good agreement with Illumina data, however lower frequency variants (<5%) are more difficult to reliably detect. SNV and indel concordance may increase with improved ONT data quality. Further work remains to be done to compare SVs and investigate DNA methylation.

14: Multi-omic analysis implicates independent roles of KMT2C/D mutations in ER driven breast cancers

Emily Tinsley[1], Philip Bredin[1], Sinead Toomey[1], Bryan Hennessy[1,2], Simon Furney[1]

[1]RCSI, Dublin, Ireland. [2]Beaumont Hospital, Dublin, Ireland

Background/Introduction: Reprogramming of the epigenetic landscape of cells has been shown to contribute towards tumorigenesis in multiple solid cancers including breast. KMT2C/D are paralogous histone-lysine-N-methyltransferases responsible for depositing the epigenetic marker H3K4me1 at gene enhancer sites, a signal associated with open chromatin and gene expression. KMT2C/D are the most frequently mutated epigenetic related genes in breast cancer though the exact mechanisms of action in tumorigenesis is not well understood.

Materials/Methods: We utilised 44 breast cancer cell lines within DEPMAP to uncover KMT2C/D dependency profiles across breast cancer cell lines. Binding profiles of KMT2C/D in MCF-7 cells were determined using ChIPseq data from ENCODE and additional previously published sources. At the patient level, RNAseq data from 994 and 1866 breast cancer patients from TCGA and METABRIC respectively were used to carry out a differential expression analysis of KMT2C/D mutation with additional differential accessibility done on 75 TCGA samples with ATACseq.

Results: ER+ breast cancers show significantly increased dependency on WT KMT2D but not KMT2C and KMT2C mutations are ~4x more frequent in luminal breast cancers. In MCF-7 cells, KMT2C/D show differential binding locations, enriched motifs and overlaps with breast cancer related TFs though both show enrichment for enhancer regions associated with ER target genes. In patient samples, both KMT2C/D aberrant samples show significant down regulation of oestrogen response pathways.

Discussion/Conclusion: Whilst KMT2C/D both appear to regulate ER in luminal breast cancers, each gene acts via unique mechanisms which provides an oncogenic advantage of WT KMT2D but preferential loss of KMT2C.

17: Novel uncultured archaeal lineages illuminate mechanisms of microbial evolution and habitat adaptation

Paul Sheridan

University of Galway, Galway, Ireland

Background: Novel microbial lineages, discovered using culture-independent techniques, offer us the opportunity to answer central questions in microbial ecology and evolution, such as how microorganisms have adapted to specific ecological niches. Concurrently, gene tree-species tree reconciliation techniques have enabled us to examine the mechanisms of genome evolution across large evolutionary timescales and predict the gene content of ancient organisms.

Methods: We reconstructed genomes representing novel branches of the archaeal tree of life from soil metagenomic sequences and used them in combination with gene tree-species tree reconciliation techniques to address fundamental questions in archaeal genome evolution. We examined two phyla: Thaumarchaeota, whose history involved relatively few ecosystem transitions and Thermoplasmatota, whose history involved many ecosystem transitions.

Results: In Thaumarchaeota, two ancient periods of extensive lateral gene acquisition cooccurred with expansion into terrestrial environments. Subsequent duplication of these novel genes, including those for carbohydrate transport and coenzyme metabolism, drove genome expansion and likely facilitated niche specialisation in soils. Whereas Thermoplasmatota was punctuated by several periods of extensive lateral acquisition. Importantly, functional genes, such as those for aerobic respiration and acid tolerance, that appeared conserved across diverse lineages and different habitats were in fact laterally acquired multiple times from different donors and maintained through convergent evolution, rather than vertically inherited from the common ancestor.

Conclusions: Our results suggest a previously under-appreciated importance of gene duplication and convergent evolution in archaeal habitat adaptation and highlight the importance of culture-independent genome sequencing to our understanding of genome evolution across the tree of life.

18: Characterisation of diverse global ancestries within participants of the UK Biobank

Fiona Pantring[1,2,3], Gianpiero Cavalleri[1,2,3], Edmund Gilbert[1,2]

[1]School of Pharmacy and Biomedical Sciences, Royal College of Surgeons in Ireland, Dublin, Ireland. [2]The FutureNeuro Research Centre, Dublin, Ireland. [3]The SFI Centre for Research Training in Genomics Data Science, Galway, Ireland

Background/Introduction: The UK Biobank (UKB) is a large dataset containing in-depth phenotype and genotype data of 500,000 UK-based participants. To control for cryptic population genetic confounders, studies leveraging the UKB are typically restricted to a subset of the participants with homogenous European ancestry. By analysing the 78,573 UKB participants with non-UK-like ancestries using population genetic approaches, there is an opportunity to better understand the global genetic diversity in the UKB and enable their inclusion in disease association studies.

Materials/Methods: Here we characterise these diverse ancestries by identifying primary continental-like ancestry clusters and for the first time fine-scale communities. To determine continental ancestries, an individuals' ancestry proportions were estimated using the ADMIXTURE algorithm. The machine learning algorithm XGBoost was trained using ADMIXTURE results to assign each individual to one of eight continental-like ancestry clusters. These continental clusters were further divided by applying community detection to a network of Identity-By-Descent sharing.

Results: We find that the UKB is a repository of diverse ancestries primarily of European-, African-, and South Asian-like descent. Whilst capturing worldwide diversity, the 144 communities appear to reflect the immigration history of Great Britain and its Commonwealth in the 20th century and are likely less represented in other large global biobanks. Our communities also facilitate novel findings of community-specific genetic risk factors, such as one of the highest worldwide frequencies of idiopathic pulmonary fibrosis risk variant rs35705950 in individuals of Maltese-like ancestry.

Discussion/Conclusion: The work characterises the population history of diverse ancestry communities in the UK and provides a framework for future studies of health-related genetic variation specific to otherwise understudied genetic communities.

21: The South American Micro-Biome Archive (saMBA) enriches the microbiome field by evaluating neglected populations

Benjamin Valderrama[1,2], Paulina Calderón-Romero[3,4], Thomaz FS Bastiaanssen[1,2,5], Aonghus Lavelle[1,2], Gerard Clarke[1,6], John F Cryan[1,2]

[1]APC Microbiome Ireland, University College Cork, Cork, Ireland. [2]Department of Anatomy and Neuroscience, University College Cork, Cork, Ireland. [3]Center for Integrative Biology, Facultad de Ciencias, Universidad Mayor, Santiago, Chile. [4]Fondap Geroscience Centre for Brain Health and Metabolism, Santiago, Chile. [5]Department of Psychiatry, Amsterdam University Medical Centers Location Vrije Universiteit, Amsterdam, Netherlands. [6]Department of Psychiatry and Neurobehavioural Science, Cork, Ireland

Background/Introduction: Over 70% of the sequenced human microbiomes come from European and North American populations, skewing our understanding by focusing excessively on 15% of the global population. This knowledge is unlikely to generalize to other world-regions. Indeed, machine learning algorithms perform poorly when applied to under-represented populations, a blind spot with serious consequences in the medical field, and which can only be addressed by the analysis of microbiome data from currently neglected areas. Problematically, previous global gut microbiome compendiums that do exist have been compiled based on automated search strategies that have consistently failed to capture studies conducted in the South American continent.

Materials/Methods: An automated search of published studies and 16s sequencing data in four databases was followed by a comprehensive manual curation of the identified resources, which allowed us to build the **S**outh **A**merican **M**icro**B**iome **A**rchive (saMBA).

Results: The compendium includes 38 gut microbiome studies, 75% of which haven't previously been captured by automated global compendiums. This represents a 10-fold increase in the number of samples from that world-region, making saMBA the biggest South American compendium yet created. The analysis unveiled an extensive degree of uniqueness and richness of the gut microbiomes from South America when compared to other world regions. The conda environment and snakemake workflow developed for this analysis will be made available on GitHub, so researchers for other world-regions can create their own archives.

Discussion/Conclusion: saMBA better characterise the different states for the human gut microbiomes, thus expanding our understanding of a more globally representative healthy microbiome. Additionally, by sharing the resources used to build saMBA, our work facilitates the inclusion of other neglected populations to accelerate microbiome research globally.

24: Exploring MicroRNA Evolutionary Dynamics in Mammals

Sarahjane Power, Zixia Huang

University College Dublin, Dublin, Ireland

Background/Introduction: The evolutionary history of mammals involves continuous genetic innovation, leading to diverse adaptations. While extensive efforts have focused on understanding protein-coding gene evolution, genomic studies show significant conservation in these genes, which contrasts with the variation observed across mammals. This indicates that protein-coding genes alone cannot account for mammalian diversity, suggesting that non-coding gene regulation is likely crucial. MicroRNAs (miRNAs), short non-coding molecules that regulate gene expression post-transcriptionally, are implicated in organismal complexity. However, their role in mammalian evolution remains unclear due to limited characterisation within the clade, with existing miRNA annotations relying heavily on RNA-seq data from a few model species.

Materials/Methods: This study adopts a mapping-based approach to investigate miRNA evolution across mammals. By leveraging known miRNA sequences, orthologues will be identified in various mammalian genomes. The study will examine their evolutionary trajectories, assessing molecular changes like copy number variation and nucleotide substitutions across the mammal phylogeny.

Results: The research will uncover evolutionary patterns in miRNAs across mammals, identifying both conserved and divergent sequences. Results will illuminate how miRNAs contribute to species-specific adaptations and overall mammalian complexity through key molecular changes.

Discussion/Conclusion: This study offers a comprehensive investigation into miRNA evolution in mammals, expanding our knowledge beyond protein-coding genes. It will bridge the gap between genome and phenotype, highlighting the critical role of miRNAs in shaping mammalian diversity and providing a foundation for future studies on non-coding regulatory mechanisms and evolution.

25: Development of novel deep learning algorithm for neoantigen prediction incorporating patient response data and functional assay data

Noor Kherreh, Pilib Ó Broin

University of Galway, Galway, Ireland

Background: Neoantigens are mutated peptides, which arise from somatic mutations in cancer cells. Neoantigens hold immense promise for personalized cancer immunotherapy. With the advent of high-throughput sequencing technologies, numerous machine learning-based tools have been developed to predict neoantigens, aiding in the selection of optimal targets for immunotherapeutic interventions. The findings published by the TESLA (expand acronym) consortium however, have underscored the limitations and challenges associated with current computational neoantigen prediction tools. Here, we conduct a benchmark study.

Materials and Methods: We curated a comprehensive dataset integrating data from multiple published cancer studies, consisting of 1065 peptide-MHC complex, with 45 unique HLA types, and peptide lengths ranging from 8 to 14 amino acids. We used this dataset to benchmark the existing neoantigen prediction tools, measuring AUROC, F1-score etc.

Results: Among the benchmarked tools, BigMHC demonstrated the highest accuracy, achieving a score of 0.6 whereas DeepSeq had the highest precision score of 0.4. Performance across different peptide lengths showed that peptides of length nine were predicted with the greatest accuracy, likely due to the bias in MHC-I binding data favoring this length. When we grouped the data by HLA alleles, we found that although some tools performed consistently across all HLA types, tools like DeepNeo and DeepSeq showed significantly worse performance for HLA-C alleles compared to HLA-A and HLA-B.

Conclusion: In our benchmarking, existing neoantigen prediction tools exhibited significant limitations, particularly in terms of accuracy and precision. Future work will focus on the use of advanced NLP-based models, such as ProBert and ESM1b, which offer a promising approach to improving prediction accuracy by effectively capturing complex patterns in mutated protein sequences and their interactions with MHC molecules.

27: Genome-wide association analysis of social participation and occupational engagement in the UK Biobank

Evie Doherty[1,2], Aodán Laighneach[1,2], Mia Casburn[1,3], Fergus Quilligan[1,3], Gary Donohoe[1,4], Dara, M. Cannon[1,3], Derek, W. Morris[1,2]

[1]Centre for Neuroimaging, Cognition and Genomics (NICOG), University of Galway, Galway, Ireland. [2]School of Biological and Chemical Sciences, University of Galway, Galway, Ireland. [3]Clinical Neuroimaging Laboratory, School of Medicine, University of Galway, Galway, Ireland. [4]School of Psychology, University of Galway, Galway, Ireland

Introduction: Psychosis is a leading cause of disability worldwide. Although generally characterised by hallucinations and/or delusions, psychosocial disability (PD) in the form of impaired social participation (SP) and occupational function (OF), is also a key feature. While several environmental and cognitive factors have been identified as predictors of PD, the biological contribution to PD remains unclear. Here, we sought to identify genetic determinants of variability in SP and OF in the UKBiobank(UKB).

Methods: SP(*N*=404,403) was defined as a summed index of responses from frequency of friend and family visits and leisure/social activity questionnaires from UKB. OF was derived from individual employment status response(*N*=405,569). Mixed-linear-model genome-wide association (GWA) analysis was conducted on all phenotypes using fastGWA.

Results: GWA analysis of SP and OF revealed 17 and 1 independent loci respectively. Gene-based FUMA analysis of SP indicated 17 significant gene-phenotype associations at a Bonferroni correction threshold ($p < 2.62e-6$). The top genes identified for SP include CDH7(p=5.57e-12), GBE1(p=5.90e-11), and ZNF536(p=1.20e-10). Tissue expression analysis revealed brain tissues, including cerebellar, frontal lobular, and amygdalar were most specific to implicated genes in SP. Genetic correlations showed that lower SP was associated with increased risk for schizophrenia, socioeconomic deprivation, and loneliness.

Conclusion: Our findings propose that SP has a stronger genetic component than OF in a healthy population. Of the tissues implicated in this analysis, both the amygdala and frontal lobe have been linked to social behaviours in previous neuroimaging research. Overall, we outline genetic loci and tissue types with possible roles in PD and phenotypes with a shared genetic basis.

28: FOXP1 Dysregulation and its Association with Schizophrenia and Cognitive Function

Deema Ali, Derek Morris

Centre for Neuroimaging, Cognition and Genomics, School of Biological and Chemical Sciences, University of Galway, Galway, Ireland

Introduction: Rare mutations in FOXP1 (Forkhead-box protein P1), a transcription factor crucial for cortical neural development, cause FOXP1 syndrome, characterized by developmental delays, intellectual disability, with or without autistic features. Common SNPs in the gene are associated with schizophrenia (SCZ) and cognitive function. This study explores FOXP1's contribution in these conditions using RNA-seq data from FOXP1 knockout animal models, including neural stem cells from embryonic mice and cortical tissues from different postnatal stages (P0, P7, P47).

Methods: We performed pairwise comparisons and time-course expression analysis on the RNA-seq data from these stages. Linkage disequilibrium score regression was used to determine if differentially expressed genes (termed gene-sets) were enriched for heritability related to SCZ and cognitive function. Cell type enrichment and gene ontology analyses identified affected cell types, brain regions, and biological mechanisms impacted by FOXP1 knockout.

Results: Our findings show that FOXP1 gene-sets across all stages are enriched for SNP-based heritability related to educational attainment (EA) and/or intelligence. Most FOXP1 gene-sets are enriched for SCZ heritability, with the highest enrichment at the P7 stage. Gene-sets from P7 and P47 exhibit significant enrichment in excitatory glutamatergic neurons in the frontal and posterior cortex. FOXP1 influences neurogenesis and synaptic signaling, with P7 and P47 showing a particular association with ion transport and G protein-coupled receptor signaling. Time-course analysis identified 1,128 significant genes across stages, enriched for SCZ and EA heritability and involved in similar biological processes

Conclusion: FOXP1 disruption across stages dysregulates genes associated with SCZ risk and cognitive function, suggesting a diverse range of biological pathways involved in their aetiology.

29: Regional Structure-Function Coupling in Bipolar Disorder

Shir Dahan[1], Pilib Ó Broin[2], Dara M Cannon[1]

[1]School of Medicine, University of Galway, Galway, Ireland. [2]School of Mathematical & Statistical Sciences, University of Galway, Galway, Ireland

Introduction: Altered structural and functional connectivity is common in bipolar disorder (BD). Studies look at their coupling (SC-FC coupling) to provide insight into the structural substrate of neural activation. SC-FC coupling was found to vary between regions, whereby primary cortices show strong coupling and transmodal regions show weaker coupling. We aim to identify regional alterations in SC-FC coupling involving emotion-processing circuitry in BD.

Methods: The study included a BD and a control group from the UK Biobank. We quantified SC-FC coupling using a multilinear regression with the predicted variable as the vector of functional edges from one region to all others, and the predictors as vectors of structural network properties. We calculated the R^2 between predicted and observed functional vectors to estimate the regional SC-FC coupling. A gamma GLM was constructed for each node to identify SC-FC coupling differences between BD and controls.

Results: BD (n=163) and controls (n=326) were age (62±7) and sex (F=52%) matched. Average network coupling was highest in the visual (R^2=0.21) and dorsal attention (R^2=0.12) networks and lowest in the limbic network (mean R^2=0.07) and the subcortex (mean R^2=0.06). Significant group difference was found in the temporal pole (β=0.28, p=0.010), supramarginal gyrus (β=-0.26, p=0.001), precentral gyrus (β=-0.24, p=0.006), and frontal pole (β=-0.20, p=0.038).

Discussion: SC-FC coupling across regions mostly aligned with previous findings supporting the theory that coupling reflects a unimodal-transmodal anatomical organization hierarchy. Higher SC-FC coupling in BD compared to controls in the temporal pole of the limbic network extends previous demonstrations of altered connectivity within the limbic network as a trait feature of BD persisting into euthymia. Taken with lower coupling in the frontoparietal network (frontal pole), these novel findings extend our understanding of integrated structure-function underpinnings of cognitive and emotion processing in euthymic BD.

33: Switching switch: alternative XBP1 dimerization regulates the unfolded protein response in a calcifying sponge

Niño Posadas[1,2], Cecilia Conaco[1]

[1]Marine Science Institute, University of the Philippines, Diliman, Quezon City, Philippines. [2]Centre for Chromosome Biology, University of Galway, Galway, Ireland

Introduction: Marine sponges are predicted to be winners in the future ocean due to their exemplary adaptive capacity. However, while many sponge groups exhibit tolerance to a wide range of environmental insults, calcifying sponges may be more susceptible to thermo-acidic stress.

Methods: To describe the gene regulatory networks that govern the stress response of the calcareous sponge, *Leucetta chagosensis* (class Calcarea, order Clathrinida), individuals were subjected to warming and acidification conditions based on the climate models for 2100.

Results: Transcriptome analysis and gene co-expression network reconstruction revealed that the unfolded protein response (UPR) was activated under thermo-acidic stress. Among the upregulated genes were two lineage-specific homologs of X-box binding protein 1 (*XBP1*), a transcription factor that activates the UPR. Alternative dimerization between these *XBP1* gene products suggests a clathrinid-specific mechanism to reversibly sequester the transcription factor into an inactive form, enabling the rapid regulation of pathways linked to the UPR in clathrinid calcareous sponges.

Conclusion: Our findings support the idea that transcription factor duplication events may refine evolutionarily conserved molecular pathways and contribute to ecological success.

36: Using Genomics to aid in the Conservation of the Native Irish Honey Bee

Stephen Smith[1], Markus Neuditschko[2], Grace McCormack[1]

[1]University of Galway, Galway, Ireland. [2]Agroscope, Bern, Switzerland

Background/Introduction: *A. m. mellifera*, which belongs to the M-lineage of the European honey bee, is the sub-lineage that is native to the island of Ireland. Recent reports of an increase in the importation of foreign sub-lineages, namely the C-lineage from Southern Europe has led to conservation concerns regarding the conservation of the Irish honey bee population.

Materials methods: In this study analyses the population for evidence of C-lineage hybridisation using WGS pool-seq approach in a nationwide dataset (N=152), including the previously unsampled area of Inishowen in County Donegal. Hybrids were detected using f-statistics.

A European wide network model was created using K-nearest neighbours (N=442). The complementary sex determiner gene was analysed using nucelotide diversity.

A Cochran-Mantel-Haenszel test was used to discover regions that correlate with honey bee abdominal colouration.

Results:

18% of all sampled colonies showed significant signs of hybridisation with C-lineage.

The population model shows a distinct cluster of pure Irish honey bees.

The pure Irish population has a genetically diverse CSD gene.

Large region on chromosome 1 was significantly associated with abdominal colour.

Discussion/Conclusion: This is a significant increase in hybridisation levels compared with previous large scale studies of this population (5%). This result proves that the reports of increased importations of foreign honey bees into Ireland are true.

The Irish honey bee is still extremely pure in comparison to other M-lineage populations and is worth conserving, hybridisation threatens this distinctness.

The GO analysis found the gene *Ebony* on chromosome 1 to be involved in cuticle pigmentation in insects.

38: Insights into the hybrid origins and evolutionary history of the Irish shamrock, *Trifolium dubium*, through annotation and comparative genomics

Katie Herron[1], Ann Mc Cartney[2], Graham Hughes[1]

[1]School of Biological and Environmental Sciences, University College Dublin, Dublin, Ireland.
[2]University of California, Santa Cruz, Santa Cruz, USA

Background/Introduction: *Trifolium dubium*, or the lesser trefoil, is a small yellow-flowered clover native to Western Eurasia but found in temperate climates globally as an introduced species. In Ireland, the species is of particular cultural significance, widely regarded as being the 'true shamrock'. *T. dubium* is an allotetraploid, arising from the suspected hybridisation of two diploid clovers, *T. campestre* and *T. micranthum*. Recently, as part of the European Reference Genome Atlas (ERGA) project, the first high-quality reference genome for *T. dubium* has been assembled, enabling a comparative genomic study across the *Trifolium* genus to better understand the species' hybrid origins.

Materials/Methods: The *T. dubium* genome has been annotated to identify coding and non-coding elements, with a focus on distinguishing and comparing the species' two subgenomes. Comparative analysis of repeat elements, particularly long terminal repeat retrotransposons (LTR-RTs), and gene families across the *Trifolium* genus is ongoing using publicly-available nuclear and organellar genomic data.

Results: Sequence-based and annotation-based methods reveal distinct differences between subgenomes, showing patterns of differential activity of repetitive elements, indicating subgenome-specific evolutionary trajectories. Broader comparative analysis places the species in evolutionary context within the *Trifolium* genus.

Discussion/Conclusion: Subgenome partitioning and analysis of repetitive element dynamics enhances our understanding of *T. dubium'*s evolutionary history, particularly the role of LTR-RTs in shaping genome structure. Future work will include extending into a population genomics study to aid conservation and agricultural applications.

40: Developing causal gene signatures to identify novel therapeutic vulnerabilities in cancer

Alanah McIntosh[1,2], Luis Iglesias-Martinez[1], Jonathan Bond[1,3], Walter Kolch[1]

[1]University College Dublin, Dublin, Ireland. [2]SFI Centre for Research Training in Genomics Data Science, Galway, Ireland. [3]Childrens Health Ireland, Dublin, Ireland

Background: Gene signatures are distinct expression patterns linked to a phenotype or outcome. While oncogenic mutations often activate unique gene signatures, similar signatures can arise lacking these mutations, pinpointing similar therapeutic vulnerabilities. Childhood cancers have lower mutational burdens, limiting targeted therapy use. We aim to discover gene signatures which correlate to genetic alterations and use them to repurpose targeted therapies for childhood cancer.

Materials/Methods: Signatures are traditionally derived using correlative methods, which can be confounded by the mutational patterns present in cancer. I aim to benchmark acausal method, 'The Deconfounder,' a causal inference algorithm, on simulated cancer data and compare the performance with other more classical methods of deriving signatures. We applied Probabilistic Principal Component Analysis to genetic alterations, deriving latent variables, which serve as surrogates for unidentified confounders. Cross-validated LASSO regression established relationships between genetic alterations, latent variables, and gene expression, yielding signatures where each alteration is associated with a list of genes whose expression is causally affected by the alteration. We compared our results with results from a standard Differential Expression (DE) analysis (DESEQ2). Signatures were based on genes which were differentially expressed, based on an adjusted p-value <0.05.

Results/Conclusions: This analysis will help to identify the most effective method for deriving gene signatures. The most effective method will be applied to publicly available data, with the signatures derived, analysed and validated. Our study's objective is to develop causal signatures and assess the clinical implications in childhood cancer treatment. Our preliminary results highlight the potential of causal methods in studying genetic alteration effects on gene expression.

41: Preliminary Single-Variant Survival Analysis Identifies Novel Genetic Loci Associated with Idiopathic Pulmonary Fibrosis Progression

Dominic Sayers[1], Josyf Mychaleckyj[2], Shwu-Fan Ma[2], Jose Miguel Lorenzo-Salazar[3], Iain Stewart[4], Toby Maher[4,5,6], Phil Molyneaux[4,5], Gisli Jenkins[4], Louise Wain[1,7], Carlos Flores[8,9,3], Imre Noth[2], Katherine Fawcett[1], Richard Allen[1]

[1]Department of Health Sciences, University of Leicester, Leicester, United Kingdom. [2]Division of Pulmonary & Critical Care Medicine, University of Virginia, Charlottesville, USA. [3]Genomics Division, Instituto Tecnológico y de Energías Renovables, Santa Cruz de Tenerife, Spain. [4]National Heart and Lung Institute, Imperial College London, London, United Kingdom. [5]Royal Brompton and Harefield Hospitals, London, United Kingdom. [6]Division of Pulmonary and Critical Care Medicine, University of Southern California, Los Angeles, USA. [7]National Institute for Health Research, Leicester Respiratory Biomedical Research Centre, Glenfield Hospital, Leicester, United Kingdom. [8]Research Unit, Hospital Universitario Nuestra Señora de Candelaria, Santa Cruz de Tenerife, Spain. [9]CIBER de Enfermedades Respiratorias, Instituto de Salud Carlos III, Madrid, Spain

Introduction: Idiopathic pulmonary fibrosis (IPF) is a lung disease characterised by tissue thickening and scarring, leading to shortness of breath and fatigue. Life expectancy is 3–5 years post-diagnosis, and current treatments (nintedanib and pirfenidone) slow progression but do not cure IPF. Lung transplants can restore function but have risks. Common genetic variants have been shown to affect IPF progression, however, the effect of rare variants on IPF progression is unclear.

Methods: A single-variant survival analysis was conducted on whole-genome sequencing data from PROFILE. This software uses the Cox proportional hazard function to calculate the p-values for each genetic variant in the cohort (significance at $p=5x10^{-8}$). The base model included the following covariates from the phenotype file (along with the snp effect): sex, age and the top 5 principal components All variants were included regardless of minor allele frequency meaning both common and rare variants were included.

Results: Three novel common variants (minor allele frequency >5%) were statistically ($p<5x10^{-8}$) associated with IPF progression; one in chromosome 21 in the *BAGE2* pseudogene, one in chromosome 19 that lies in an intergenic region and one in chromosome 9 that appears to be near the centromere.

Discussion: This study demonstrates that our analysis of whole-genome sequencing data identifies novel and known genetic loci involved in IPF survival. These results will be combined with those from an independent study in a meta-analysis to ensure replication. In future work, we will extend these analyses to rare variants including collapsing gene-based approaches to improve statistical power.

42: UNRAVELLING GENETIC FACTORS IN RELATION TO THE AGING BRAIN: PRELIMINARY GENOMIC ANALYSIS WITH MACHINE LEARNING APPROACHES FROM THE TUDA STUDY

Shane Gordon[1], Leane Hoey[1], Helene McNulty[1], Faith Pangilinan[2], Lawrence Brody[2], Catherine Hughes[1]

[1]Nutrition Innovation Centre for Food and Health (NICHE), Ulster University, Coleraine, Derry, United Kingdom. [2]Genetics and Environment Interaction Section, National Human Genome Research Institute, National Institutes of Health, Maryland, USA

Background: As global rates of dementia rise, understanding the drivers of cognitive decline is crucial for developing preventative strategies. Nutrition, particularly B-vitamin status, and genetic factors such as those identified through genome-wide association studies (GWAS), are emerging as important determinants of cognitive health in ageing. Machine learning (ML) can further analyse interactions between genetics, lifestyle, and dietary factors. The aim was to identify genetic markers associated with cognitive dysfunction in older adults from the TUDA cohort using GWAS and ML techniques.

Methods: The TUDA cohort of 5186 Irish adults (aged 60–102 years) provided comprehensive data on nutrition, health, and lifestyle, alongside genome-wide genotyping. GWAS identified SNPs associated with cognitive dysfunction, measured by the Repeatable Battery for the Assessment of Neuropsychological Status (RBANS), with dysfunction defined as a score <70. ML models were applied to identify predictors, using decision tree, random forest, and Naïve Bayes classifiers.

Results: GWAS identified the rs429358 SNP in the APOE gene ($P = 5.6 \times 10^{-9}$) as significantly associated with cognitive dysfunction. Other SNPs, such as rs3771791 (HK2 gene), did not reach genome-wide significance. ML models identified frailty, family-reported memory concerns, and years of education as key predictors. Random forest also highlighted age, glomerular filtration rate (GFR), pyridoxal phosphate (PLP), and total plasma homocysteine (tHcy) as important.

Conclusion: Combining GWAS and ML revealed genetic and metabolic markers, notably APOE and PLP, that could inform strategies for predicting and preventing cognitive decline. Further research is needed to validate these findings.

43: Predicting biochemical recurrence in prostate cancer using a prognostic model derived from a novel ceRNA regulatory network

Barry Digby[1], Pilib Ó Broin[1], Stephen Finn[2]

[1]University of Galway, Galway, Ireland. [2]Trinity College, Dublin, Ireland

Biochemical recurrence (BCR) occurs in one-third of prostate cancer patients treated with local therapy and inevitably develops in all patients treated with enzalutamide. Circular RNAs (circRNAs) are covalently closed transcripts involved in cancer pathways and disease mechanisms, harbouring microRNA (miRNA) response element sites within their sequence suggesting a regulatory role within the competing endogenous RNA (ceRNA) network. In this study, we combine prostate adenocarcinoma (PRAD) and enzalutamide-resistant circRNA, miRNA and messenger RNA (mRNA) expression datasets to derive a ceRNA network for the prediction of BCR in prostate cancer. A prognostic model was constructed using mRNAs from the ceRNA network via Cox regression and stepwise regression, producing a prognostic index that was used to effectively stratify patients into high-risk and low-risk groups with significantly different BCR outcomes. High-risk patients exhibited unfavorable prognosis and elevated immune infiltration compared to low-risk groups. Moreover, the prognostic model was externally validated in six external validation datasets. Finally, this study presents a clinical nomogram that combines clinical features and pathological factors in conjunction with the prognostic index, demonstrating enhanced accuracy compared to the prognostic index in standalone use.

45: Predicting auditory verbal hallucinations using sociodemographic, environmental, psychological and biological data in the general population and psychosis samples from the UK Biobank using machine learning

Dijana Ostojic, Gary Donohoe, Derek W Morris

School of Biological and Chemical Sciences and School of Psychology, Centre for Neuroimaging, Cognition and Genomics (NICOG), University of Galway, Galway, Ireland

Introduction: Auditory verbal hallucinations (AVH) in the form of hearing voices are defined as core symptoms of psychosis. AVHs are also experienced by members of the general population who have not been diagnosed with psychosis. The prevalence of AVH ranges between 5% and 28% in the general population.

Methods: The current study aims to identify the most important predictors of AVH from 41 sociodemographic (23%), environmental (24%), biological (23%) and psychological features (30%) in two samples from the UK Biobank; a sample of individuals diagnosed with psychosis (*n* =308 voice-hearers and *n* = 637 non-voice-hearers) and a general population sample with no diagnosis of psychosis (*n* = 1,786 voice-hearers and *n* = 126,709 non-voice-hearers). A machine learning (ML) data-driven approach was applied using the XGBoost classification algorithm to develop a predictive model for predicting AVH, which enabled us to compare the importance of the features using different metrics.

Results: Passive suicidal ideation was the most important predictor of AVH in the general population, while other important predictors included the presence of mental distress, past trauma experience, the severity of post-traumatic stress disorder, the severity of depressive symptoms, and seeking professional help. Engagement in self-harm and severity of depressive symptoms were the most important predictors within the psychosis sample followed by the severity of depressive symptoms, presence of mental distress, past traumatic experiences, post-traumatic stress disorder symptoms, and polygenic risk score for schizophrenia.

Discussion: This is the first ML-based analysis of predictors of AVHs, revealing that the most important features are largely consistent across both psychosis and general population samples.

46: Fragmentomic characterisation of ovarian cancer samples

Devesh Haseja[1], Seyed Aghil Hooshmand[1], Charlotte McBrien[2], Shannon Beattie[2], Charity Hall[2], Micheal Ryan[3], Alexander McIntyre[3], Paul Mullan[3], Pilib Ó Broin[1]

[1]School of Mathematical & Statistical Sciences, University of Galway., Galway, Ireland. [2]GenoME Diagnostics, Queen's University Belfast., Belfast, United Kingdom. [3]Patrick G. Johnston Centre for Cancer Research, Queen's University Belfast., Belfast, United Kingdom

Background: Liquid biopsies are non-invasive tests that detect genetic material shed from tumors into bodily fluids, such as blood, urine, or tears. This approach allows for real-time or longitudinal monitoring of cancer dynamics, enabling early detection, tracking of disease progression, and assessment of treatment response. Fragment analysis of cell-free DNA (cfDNA) in the biological fluid of interest allows the detection and classification of the circulating tumor DNA (ctDNA), fragments released from cancer cells following apoptosis or necrosis. The characteristics of these fragments, for example, their size and distribution, as well as their molecular profiling (mutations, methylation status etc.) may indicate the type of cancer or its stage, as well as providing a measure of tumor heterogeneity, a crucial step for providing personalized treatment strategies.

Methods: This study includes plasma samples from 7 non-cancer controls and ascites samples from 7 ovarian cancer patients. Library preparation and sequencing was carried out using PacBio's Onzo sequencing protocol and computational analysis was performed using the FinaleDB, cfDNApipe, and OpenGene ctDNA workflows.

Results: Results are grouped into four main categories: fragment profile (fragment length), fragmentation patterns (window protection score, WPS), end-motif analysis, and variant analysis. A significant distinction in fragment length distribution was observed between the cancer and the normal plasma samples as well as significant differences in their end motif k-mer usage. A number of SNPs were also identified for further follow-up.

Discussion: Fragmentomic analysis can determine the presence of cancerous cells without the need for invasive biopsies, and may prove especially useful for early detection and continuous monitoring of cancer. Future work will focus on additional multi-omic analysis including the integration of window protection score and methylation profiles.

47: Evolution of Tissue-Specific Expression Following Salmonid Whole Genome Duplication

Yuanshuo Li, Aoife McLysaght

Trinity College Dublin, Dublin, Ireland

Background/Introduction: Whole genome duplication (WGD) events are extreme occurrences in eukaryotic genome evolution that provide abundant genomic material and facilitate the potential for widespread functional divergence of genes. While most duplicated genes are lost due to functional redundancy, those retained after WGD—known as "ohnologs"—exhibit distinct characteristics in their expression evolution, diverging from singleton genes due to gene dosage balance. However, the evolution of tissue-specific expression in ohnologs remains poorly understood.

Materials/Methods: In this study, we investigate the relatively recent salmonid WGD event, which occurred 80–100 million years ago, using transcriptomic data to construct a meta-expression dataset encompassing 10 tissues from 8 species. We applied the Expression Variance and Evolution (EVE) model to describe the phylogenetic evolution of expression levels both between and within species.

Results: We found ohnologs and singletons exhibit distinct functional differentiation: ohnologs are more involved in the regulation of transcription and RNA metabolism, while singletons are associated with DNA repair and epigenetic regulation.

Discussion/Conclusion: Compared to singletons, ohnologs demonstrate a greater change in tissue specificity after the WGD event, with their expression levels evolving primarily asymmetrically. This highlights the significant roles of tissue-regulated expression and gene dosage constraints in the adaptive evolution of gene expression following WGD.

49: Predicting social participation using brain MRI, environmental and genetic measures in the UK Biobank using Extreme Gradient Boosting (XGBoost)

Aodán Laighneach[1,2,3], Dijana Ostojic[1,2], Mia Casburn[1,3], Fergus Quilligan[1,3], Evie Doherty[1,2], Pilib Ó Broin[1,4], Gary Donohoe[1,5], Dara M. Cannon*[1,3,6], Derek W. Morris*[1,2,6]

[1]Centre for Neuroimaging, Cognition and Genomics (NICOG), University of Galway, Galway, Ireland. [2]School of Biological and Chemical Sciences, University of Galway, Galway, Ireland. [3]Clinical Neuroimaging Laboratory, School of Medicine, University of Galway, Galway, Ireland. [4]School of Mathematical and Statistical Sciences, University of Galway, Galway, Ireland. [5]School of Psychology, University of Galway, Galway, Ireland. [6]*Authors jointly directed this work, Galway, Ireland

Background/Introduction: Experience of psychosis is often associated with poor social outcomes, even with successful antipsychotic treatment. Understanding which factors influence an individual's social behaviour is an important step in understanding this problem. Here, we aim to identify to what degree social participation (SP) can be predicted in healthy individuals and by which specific brain MRI metrics, and environmental and genetic factors.

Materials/Methods: SP (range:0-10) was defined based on a metric comprised of frequency of friend/family visits (UKB:p1031) and leisure/social activities (UKB:p6160). Individuals were classed as low SP (n=8,288) or high SP (n=5,681) using cutoffs related to the top and bottom quartile of the SP distribution. Measures of brain MRI volume (n=451), environmental exposures (n=32) and genetic polygenic risk scores (PRS) (n=36) with <10% missingness, as well as age, sex and educational attainment were used to predict SP in healthy individuals in the UK Biobank. Data were split into 75:25 training:test cohorts.

Results: Balanced prediction accuracy for classifying high/low SP in the test set was 61.4%. Receiver operating characteristic AUC was 69.4%. Precision and recall were 66.7% and 84.5% respectively. Age had the highest variable importance, followed by friendships satisfaction, length of time at current address, sex and household income. Among the top 20 predictors, 8 (40%) were environmental, 8 (40%) were brain imaging measures, and 1 (5%) was genetic.

Discussion/Conclusion: This work suggests that SP is a predictable measure and that brain MRI features and environmental exposures are more important factors than genetic factors. Although literature details the link between brain MRI features and behaviour, the role and causality of environmental factors appears less clear. Further work including robust feature selection and validation in affected individuals is required to fully understand the environmental and biological factors that influence social behaviour in psychosis.

51: Mapping the Evolution of Translational Regulation Using Ribosome Decision Graphs

Elizaveta Siling[1,2], Jack Tierney[1,2], Pavel Baranov[1]

[1]School of Biochemistry and Cell Biology, University College Cork, Cork, Ireland. [2]SFI Centre for Research Training in Genomics Data Science, University College Cork, Cork, Ireland

Background/Introduction: Eukaryotic RNA translation is less straightforward than traditionally shown. It usually involves multiple translation initiation sites leading to diverse protein products from a single mRNA. The process of translation initiation and its impact on protein synthesis remains poorly understood, especially in the context of genetic variability. Ribosome Decision Graphs (RDGs) provide a novel framework for representing translation initiation as a network of probabilistic decisions made by ribosomal complexes as they traverse the RNA.

Materials/Methods: We are creating computational approaches to examine the topologies of RDG across various evolutionary levels. Our methodology will compare RDG structures between homologous genes across species and within human population variants. We intend to implement clustering algorithms based on RDG structural similarities and explore their correlation with traditional phylogenetic patterns. Ribosome profiling data will be used to validate predicted translation initiation sites**.**

Results: The initial phase of our work involves developing reliable techniques for comparing RDGs and creating metrics to measure the structural similarities between graphs. Our primary objective is to identify patterns of RDG conservation and variability that may suggest functional limitations on translation regulation.

Discussion/Conclusion: This study aims to demonstrate how RDGs can provide a new framework for understanding the evolution of translation regulation. By analyzing RDG conservation patterns, we hope to gain insights into the selective pressures that act on alternative translation initiation, and develop new approaches for interpreting genetic variations in protein-coding regions. These findings may ultimately enhance our understanding of how sequence variations impact translation regulation, both in evolutionary and disease contexts.

52: Oral Pathobiont Colonisation in Patients with Treatment Resistant Schizophrenia Prescribed the Antipsychotic Clozapine

Francesca McDonagh[1], Elaine K. Murray[2], Brian Hallahan[3], Georgios Miliotis[1,4]

[1]Antimicrobial Resistance and Microbial Ecology Group, School of Medicine, University of Galway, Galway, Ireland. [2]Personalised Medicine Centre, School of Medicine, C-TRIC, Altnagelvin Hospital, Ulster University, Derry, United Kingdom. [3]Discipline of Psychiatry, School of Medicine, University of Galway, Galway, Ireland. [4]Centre for One Health, Ryan Institute, University of Galway, Galway, Ireland

Introduction: Antipsychotic drugs are increasingly recognised to be associated with a heightened risk of infection. Clozapine, a second generation antipsychotic used for treatment-resistant schizophrenia (TRS), has been specifically highlighted for its association with an increased risk of pulmonary infection. Additionally, antipsychotics have been reported to contribute to microbiome dysbiosis, creating opportunities for pathobiont colonisation. The aim of this study is to characterise the rate of transient oral pathobiont colonisation associated with the use of the antipsychotic clozapine.

Materials/Methods: Salivary samples were collected from individuals diagnosed with TRS and prescribed clozapine (n=19) and screened on selective agars. A questionnaire on the study participants oral health and general medical history was conducted. Isolates positively identified as *Enterobacteriaceae* or *Moraxellaceae* underwent analysis by molecular and bioinformatic methods including identification of genes of interest such as antimicrobial resistance genes (ARGs) and virulence factors (VFs).

Results: 15 (~79%) participants were colonised with pathobionts. Of those, 5 (~33%) were identified to be multi-colonised. The colonisation rate of the control cohort was identified to be approximately 10%. 16 species of *Enterobacteriaceae* and 3 species of *Moraxellaceae* were detected. All genomes exhibited a virulent genotype with a mean 77 VFs per genome. Virulence factor genes identified included *aslA* and *cnf1* which contribute to penetration of the blood brain barrier and entry to the central nervous system.

Conclusion: Use of the antipsychotic clozapine is associated with oral colonisation by transient pathobionts that may contribute towards the increased risk of respiratory infection observed this patient cohort.

53: Investigation of Alternative Splicing in Patient-Derived Cancer-Associated Fibroblasts

Nupur Dubey[1], Kevin Ryan[1,2], Domhnall O'Connor[3], Barry Digby[1,2], Laura Barkley[3], Pilib Ó Broin[1]

[1]School of Mathematical & Statistical Sciences, University of Galway, Galway, Ireland. [2]The SFI Centre for Research Training in Genomics Data Science, Galway, Ireland. [3]Lambe Institute for Translational Research, University of Galway, Galway, Ireland

Background: Understanding the contributions of individual cells in the tumour microenvironment (TME) can shed light on the processes of angiogenesis and tumour growth. A key constituent of the TME are cancer-associated fibroblasts (CAFs), which have shown evidence of promoting tumour growth, angiogenesis, and metastasis, by releasing signals that are pro-oncogenic and by remodelling of extracellular matrix. Evidence of discrepancies in alternative splicing events in proteins associated with key cellular processes, demonstrates contribution to initiation and advancement of cancer, and therapeutic resistance. In this study, we investigated the role of alternative splicing in CAFs and whether or not specific isoforms contributing to pathogenesis in breast and lung cancer could be identified.

Methods: RNA-sequencing was carried out on CAF and tumour-adjacent normal (TAN) fibroblasts from n=12 breast cancer patients and n=10 lung cancer patients. Analysis of differential exon usage (DEU) was performed using DEXSeq and rMATS-turbo.

Results: A total of 44 genes in breast cancer samples and 12 genes in lung cancer samples were identified as demonstrating DEU. Many of the identified genes play a role in cancer-relevant biological pathways including: cell cycle, focal adhesion, endocytosis, and immune response to infection.

Discussion: Among the identified genes, several (e.g., *PDGFRA*, *VEGFA* and *MOB1B*) play an important role in cancer and are linked to CAF biology, highlighting their potential as targets for cancer treatments designed to disrupt the cancer-promoting role of certain CAFs in the TME. Future work will focus on validation of these isoforms and further analysis of their potential role in other solid tumours.

55: Genomic analysis of clinical *Acinetobacter radioresistens* strains associated with the hospital environment, human colonisation and infection

Aneta Kovarova[1], Anna Tumeo[1], Francesca McDonagh[1], Andy O'Connor[1], Christina Clarke[2], Brigid Hooban[3], Dearbháile Morris[1,4], Kasthuri Venkateswaran[5], Georgios Miliotis[1,4]

[1]Antimicrobial Resistance and Microbial Ecology Group, School of Medicine, University of Galway, Galway, Ireland. [2]University Hospital Galway, Ireland, Galway, Ireland. [3]Atlantic Technological University, Galway, Ireland. [4]Centre for One Health, Ryan Institute, University of Galway, Galway, Ireland. [5]NASA Jet Propulsion Laboratory, California Institute of Technology, Pasadena, California, USA

Introduction: The four *A. radioresistens* clinical isolates were obtained from Irish hospitals via the National CPE Reference Laboratory at University Hospital Galway. *A. radioresistens* is well-adapted to hospital environments and human skin due to its exceptional resistance to desiccation and radiation. It holds clinical importance for its potential to cause hospital-acquired or opportunistic infections and its role as a silent reservoir for antimicrobial resistance, especially carbapenem resistance.

Methods: Four clinical genomes were sequenced using second generation sequencing (PE300), assembled with Shovill, and quality-assessed using Quast and CheckM. Annotation was done with PROKKA. Pan-genome all-vs-all analysis was conducted by calculating ANI relative to the reference genome, type strain and all *A. radioresistens* genomes available on NCBI (n=62), with taxonomic placement confirmed using GTDB-Tk. CARD, VFDB, and BacMet databases were used to identify resistance genes, virulence-factors (VFs), and biocide/metal resistance genes. MIC testing was performed for 14 antimicrobials.

Results: GTDB-Tk confirmed the isolates as *A. radioresistens*, forming a distinct clade with *A. lwoffii* as the closest relative. ANI analysis showed >97.90% similarity among clinical isolates and available *A. radioresistens* genomes. Resistome analysis revealed 46 ARGs across the species potentially conferring resistance to carbapenems ($bla_{\text{OXA-23-like}}$, $bla_{\text{NDM-1}}$, *PER-2*) and colistin (*LpsB*), which are considered second-line and last-resort treatments, respectively. Notably, $bla_{\text{OXA-23}}$ and $bla_{\text{OXA-23-like}}$, were detected in 100% (n=67) of the genomes. In addition, the *adeI/J/K* genes, coding for the three subunits of the *AdeIJK* multidrug efflux-pump typically found in *A. baumannii*, were detected in 100% (n=67) of the genomes. VFs screening revealed 66 VFs, with 41 shared across 98.51% of genomes.

Conclusion: Comprehensive genomic analysis shows *A. radioresistens* carries either $bla_{\text{OXA-23}}$ or $bla_{\text{OXA-23-like}}$ genes and MDR efflux pumps, emphasising its role as a potential carbapenem resistance reservoir and the risk it poses in healthcare settings.

56: R and Bioconductor-based Workflow for Downstream Analysis of LFQ Meat Proteomics Data: Enhancing Reproducibility in Meat Proteomics Research

Sudipta Hazra[1,2], Claudia Terlouw[3], Mohammed Gagaoua[4], Joe P. Kerry[2], Brigitte Picard[3], Ruth Hamill[1]

[1]Department of Food Quality and Sensory Science, Teagasc Food Research Centre, Dublin, Ireland. [2]School of Food and Nutritional Sciences, College of Science, Engineering and Food Science, University College Cork, Cork, Ireland. [3]Université Clermont Auvergne, INRAE, VetAgro Sup, UMR Herbivores, Saint-Genès-Champanelle, France. [4]PEGASE, INRAE, Institut Agro, Saint-Gilles, France

Background: Proteomics have been extensively used in meat science to understand the biochemistry of meat quality. Identified and quantified proteomics data obtained through processing with analytical software tools, such as Mascot and Progenesis QI, among others, serve as a common starting point for meat scientists to analyze shotgun proteomic data. However, a standardized workflow for the downstream statistical analysis and visualization of these factorial design datasets is lacking.

Methods: The workflow was developed using several R and Bioconductor packages including Tidyverse, UniprotR, QFeature, limma, EFS, ggplot2, pheatmap, ggvenn, and UpSetR, along with custom functions. The methodology is demonstrated using an experimental shotgun dataset preprocessed via Mascot and Progenesis QI. This dataset was derived from postmortem muscle tissues from a 2×4 factorial study of cattle reared under two slaughter conditions and four feeding regimes.

Results: This workflow produced a result table with the gene name, UniProt accession, log2 fold change, p-value, and the corresponding adjusted p-value from the LFQ data. Publication-quality Venn diagrams, UpSet plots, volcano plots, and heat maps were generated. Dimensional reduction techniques such as PCA, t-SNE, and UMAP were integrated into the workflow. In addition, eight distinct feature selection techniques based on ensemble learning were applied using the EFS package, and the results were compared with differentially abundant proteins and visualized using a volcano plot. To enhance the readability, we implemented a strategy to present the correlation of large datasets more effectively.

Discussion: This R and Bioconductor-based workflow presents a consolidated resource of selected tools that facilitates downstream processing, statistical evaluation, and comprehensive visualization of shotgun proteomic data, thereby potentially enabling meat science researchers to address specific biological questions.

57: Investigating the Relationship Between Bacteriophage Presence and Antimicrobial Resistance Phenotypes in *Escherichia coli* Against Amoxicillin

Corey Woods, Chris Creevey, Lucy Dillon, Emmet Campbell

Queen's University Belfast, Belfast, United Kingdom

Background: Antimicrobial resistance (AMR) poses a severe global health challenge. Defined as the ability of a microorganism to adapt and grow in the presence of previously effective medications, AMR leads to ineffective treatments and persistent infections, largely driven by the overuse and misuse of antibiotics. This project investigates the role bacteriophages play in influencing AMR, hypothesizing that bacteriophages directly impact AMR phenotypes in *Escherichia coli* against the antibiotic Amoxicillin.

Materials and Methods: Bioinformatic and machine learning techniques were applied to a dataset of 272 bacterial genomes to identify resistance genes and prophage communities potentially linked to AMR. Using the J48 Decision Tree algorithm within Weka, the study predicted AMR outcomes across three datasets—bacterial genomes, prophage communities, and a combined dataset of both—with true positive rates of 0.934, 0.886, and 0.926, respectively.

Results: The analysis highlighted 10 significant AMGs and 17 prophage communities. Key findings included the beta-lactamase gene *bla-TEM1B* and the aminoglycoside gene *aac(3)-ila*, both conferring resistance to Amoxicillin and illustrating the complexity and cross-resistance potential of AMR mechanisms. A combined dataset analysis also revealed a resistance route involving both an AMG (*drfA5*) and a prophage community (Community 58), demonstrating an interactive relationship between AMGs and prophages in conferring resistance phenotypes.

Discussion: The prophage communities were labeled in a manner that limited direct association with known prophage sequences, restricting our ability to determine their specific contributions to AMR. However, further analysis using StarAMR confirmed that some prophage sequences harbor AMGs, including *blaTEM-1B* and *blaCTX-M-15*, reinforcing the hypothesis that bacteriophages play a significant role in AMR phenotypes by carrying AMGs themselves. Future research should focus on addressing bacteriophage-mediated AMR, potentially through therapeutics targeting phage infections.

58: Pangenomics and machine learning reveal mutually exclusive routes to resistance

Lucy Dillon, Christopher J. Creevey

Queen's University Belfast, Belfast, United Kingdom

Background: Prokaryotic pangenome variation is not random but shows highly structured associations in gene content likely driven by natural selection. How these pangenomic selective patterns relate to an organism's traits and phenotypes is still unclear, but it follows that the essentiality of any gene to a particular phenotype may also be subject to the wider genetic environment. Supporting this, we previously showed that predicting antimicrobial resistance (AMR) phenotype requires whole-genome context rather than individual AMR genes alone.

Methods: We aimed to further understand the routes to resistance within and between species by using a combination of pangenomics and machine learning, using *Pseudomonas aeruginosa* and *Escherichia coli* due to their clinical importance. The machine learning methods identified genes statistically linked to particular AMR phenotypes and then pairs of these key genes were searched in networks evaluating gene presence within the pangenome (if genes are associated/disassociated).

Results: We found that AMR genes were present within the core genome in both species. These genes often conferred resistance to multiple drug classes, suggesting multidrug resistance genes are core. We revealed genes linked to multidrug phenotypes that were disassociated in *E. coli*, suggesting there are mutually exclusive routes to MDR phenotypes. We then compared *P. aeruginosa*'s pangenome network to *E. coli*'s, in which we showed that the gene pairs had different associations in the different networks.

Conclusion: Our results highlight the importance of understanding genomic context of AMR mechanisms to explore how AMR arises in different strains and this may offer potential solutions to tackle AMR in the future.

59: Genomic and phylogenetic analysis of hypervirulent *Klebsiella pneumoniae* ST23 in Ireland

Mark Maguire[1,2,3], Niall DeLappe[4], Christina Clarke[4], Wendy Brennan[4], Alma Touhy[4], Genevieve Devane[4], Martin Cormican[4], Ulrike Carlino-MacDonald[5,6], Alan Hutson[7], Dearbháile Morris[1,2], Simone Coughlan[3], Georgios Miliotis[1,2], Thomas Russo[5,6,8,9], Liam Burke[1,2]

[1]Antimicrobial Resistance and Microbial Ecology Group, School of Medicine, University of Galway, Galway, Ireland. [2]Centre for One Health, Ryan Institute, University of Galway, Galway, Ireland. [3]SFI Centre for Research Training in Genomics Data Science, Galway, Ireland. [4]Galway Reference Laboratory Service, University Hospital Galway, Galway, Ireland. [5]Veterans Administration Western New York Healthcare System, Buffalo, USA. [6]Department of Medicine, University at Buffalo-State University of New York, Buffalo, USA. [7]Department of Biostatistics and Bioinformatics, Roswell Park Comprehensive Cancer Center, Buffalo, USA. [8]Department of Microbiology and Immunolog, University at Buffalo-State University of New York, Buffalo, USA. [9]The Witebsky Center for Microbial Pathogenesis; University at Buffalo, State University of New York, Buffalo, New York, Buffalo, USA

Introduction: Hypervirulent *Klebsiella pneumoniae* (hvKp) has emerged as a pathogen of global concern associated with invasive community acquired infections. The combination of hypervirulence and carbapenem resistance can result in severe and difficult to treat infections. This retrospective study aimed to investigate the spread of hypervirulent (hv) *Klebsiella pneumoniae* (hvKp) ST23 in Ireland and the convergence of hypervirulent and antimicrobial resistance (AMR) genotypes.

Methods: Short-read sequences (PE300) for 90 *Klebsiella pneumoniae* ST23 isolates were generated by the Galway Reference Laboratory Services (GRLS). Isolates were from screening swabs (n = 59), invasive infections (n = 18), non-invasive sites (n = 12), and the hospital environment (n = 1). The virulence and resistance content were assessed genomically using Kleborate (v2.2.0), ABRicate (v1.0.1) and Platon (v 1.6). The in-vivo virulence of the isolates was assessed using a murine model.

Results: All isolates were genotypically hv with 88/90 isolates having a maximal Kleborate virulence score of 5 including carriage of key genes. Eighty-two percent of isolates (74/90) carried a carbapenemase gene ($bla_{\text{OXA-48}}$/$bla_{\text{OXA-181}}$/$bla_{\text{NDM-1}}$) and 42% carried resistance genes to 3 or more antimicrobial classes. Core genomic delineation revealed the isolates to be clonal with similar resistance and virulence profiles. Two distinct clusters of Irish isolates were detected consisting of 82/90 of the isolates. Isolates associated with carriage and infection demonstrated similar *in vivo* virulence.

Conclusion: An established clone of hvKp ST23 is circulating within Ireland and causing both colonisation and infection of patients. The lack of reliable screening methods for hvKp makes its detection and control in the healthcare setting challenging.

60: Diverse morphology and proteomic phenotypes of calcification in bioprosthetic structural valve degeneration

Rachel Cahalane[1], Cassandra Clift[2], Mark Blaser[2], Mandy Turner[2], Taku Kasai[2], Alesandra Campedelli[2], Amber Hendrickx[3], Filip Rega[3], Marie Billaud[2], Jochen Muehlschlegel[4], Masanori Aikawa[2], Laoise McNamara[1], Bart Meuris[3], Sasha Singh[2], Elena Aikawa[2]

[1]University of Galway, Galway, Ireland. [2]Brigham and Women's Hospital, Boston, USA. [3]KU Leuven, Leuven, Belgium. [4]Johns Hopkins University School of Medicine, Baltimore, USA

Background/Introduction: Fibrocalcific remodelling is an end-stage feature of bioprosthetic (BP) valve degeneration and calcific aortic valve (AV) disease. However, the processes governing BP calcification are understudied. Here we conduct a histopathological assessment of BP degeneration and build a proteomic comparison map of human BP degeneration versus AV disease.

Materials/Methods: Macroscopic segmentation was performed on entire explanted degenerated bovine pericardial BP leaflets (n=48) and diseased AV valves (n=19) and validated with histology to classify their state (BP: non-degenerated/thrombotic/neotissue/calcified, AV: non-diseased/fibrotic/calcified). Segment-specific mass spectrometry-based proteomics was performed on BP/AV tissues. Laser capture microdissection enabled spatially resolved proteomics of BP calcification within discrete bioprosthetic/thrombotic/neotissue matrices versus explanted non-calcified regions.

Results: Principal component analysis revealed BP and AV proteomes (2,005 and 2,012 proteins) clustered according to their degenerated and diseased segments; BP thrombotic and BP calcified sample clusters overlapped. Correlations of segment proteome-wide abundances revealed the highest intra- and inter-tissue similarity between non-degenerated BP and calcified BP (r_p=0.87) and BP neotissue and calcified AV (r_p=0.69), respectively. Histopathological observations supported these proteomic findings, confirming the presence of calcification within discrete BP matrices: bioprosthetic (22/59), thrombus (22/59), and neotissue (15/59). 252 proteins were enriched in either bioprosthetic, thrombotic, or neotissue matrix calcification compared to explanted non-calcified regions. Only 3% of differentially enriched proteins overlapped, suggesting different mechanisms.

Discussion/Conclusion: This is the first comparative proteomic study of segmented degenerated BP and diseased AV tissue. We identified 3 subtypes of BP calcification within bioprosthetic, thrombotic, or neotissue matrices and revealed different proteins associated with their calcification.

63: A comprehensive and user-friendly database of vertebrate ohnologs

Lukasz Niezabitowski, Aoife McLysaght

Trinity College Dublin, Dublin, Ireland

Background/Introduction: All vertebrate genomes are impacted by multiple rounds of ancestral polyploidisation that have shaped their structure, organisation, and function. Genes arising from these events, termed ohnologs, amount to 20-35% of the human genome and have played a major role in the evolution of vertebrate complexity, development, and susceptibility to genetic disease. While this class of genes is of broad interest, high-quality ohnolog datasets are hard to find and are often generated as needed using ad hoc methods, leading to inconsistencies across studies.

Materials/Methods: Here we create a high-confidence ohnolog dataset with a focus on the 1R and 2R whole genome duplication events shared by all jawed vertebrates. By integrating phylogenetic and synteny data, we are able to distinguish between ohnologs with differing retention patterns.

Results: Ohnologs retained after 1R are enriched for cytoskeleton components, while those retained after 2R are highly involved in signalling pathways. Furthermore, only duplicates retained after both 1R and 2R are enriched in immune system functions. In order to make our dataset easily accessible, we develop tooling for visualising and analysing our data using a modern web interface.

Discussion/Conclusion: Our dataset is the first to distinguish between 1R/2R ohnologs with differing retention patterns, allowing future research to determine the relative impact of these ancient polyploidisation events on vertebrate genomes. We believe this data will be of particular interest to studies involved in human development, disease, and immunity.

64: Chromosomes, the more the merrier? Evidence for increased stress resistance due to polyploidy from synthetic autotetraploid *Caenorhabditis elegans*

Chauve Laetitia, Clément Verdier, Bazzani Emma, Liam Butler, Aoibhin McGarry, Martha Atimise, Aoife McLysaght

Smurfit Institute of Genetics, Trinity College Dublin, Dublin, Ireland

Introduction: Whole genome duplication (WGD) is a well-studied yet enigmatic phenomenon. Circumstantial evidence of many WGD events coinciding with periods of mass extinction is consistent with the hypothesis that polyploidy is advantageous under stress conditions. While support for this comes from both theoretical work and field studies, direct evidence is scarce, especially in animals.

Materials/Methods: This study investigated tetraploid *C. elegans* derived from diploid strains to examine ploidy effects on physiology and stress resistance. Body size, fertility, lifespan, and stress responses to temperature and pathogens were assessed. RNAi treatments targeting specific genes were performed and transcriptional responses in diploids and tetraploids were analysed using qRT-PCR.

Results: Although tetraploid *C. elegans* show reduced fitness, fertility, and lifespan under normal conditions, upon severe cold, gravid adult tetraploids strikingly escape cold-induced death and therefore produce more progeny than cold-shocked diploids. Moreover, as shown previously, cold stress activates the *zip-10* transcription factor pathway responsible for cold-induced death. Interestingly, this pathway is differently activated in diploids and tetraploids.

Discussion: The *zip-10* pathway activation alone does not explain the striking phenotype observed in cold-shocked tetraploids. A genome-wide RNA sequencing analysis could reveal more transcriptional differences between tetraploid and diploid individuals. Nonetheless, for the first time, our study reveals that tetraploidy in animals can be seen as a potential adaptive strategy against stress, with tetraploid adult *C. elegans* surviving cold stress and therefore producing an increased number of progeny of equal fitness compared to normal diploid individuals.

65: Identification and Analysis of Transcriptomic Changes of MSCs Cells from People with Type 2 Diabetes Mellitus

Jingyan Wang[1], Katarzyna Goljanek-Whysall[2], Pilib Ó Broin[3], Cynthia M Coleman[1]

[1]Regenerative Medicine Institute, School of Medicine, College of Medicine, Nursing and Health Sciences, University of Galway, Galway, Ireland. [2]School of Medicine, College of Medicine, Nursing and Health Sciences, University of Galway, Galway, Ireland. [3]School of Mathematical & Statistical Sciences, College of Science and Engineering, University of Galway, Galway, Ireland

Background: Type 2 diabetes mellitus (T2DM) is a chronic disease, characterized by elevated blood glucose, which can lead to complications such as osteopathy. T2DM-induced osteopathy is marked by reduced cortical bone mass and increased mineral density. Human bone marrow mesenchymal stromal cells (hBM-MSCs) are progenitors of osteoblasts and osteocytes, essential in maintaining bone homeostasis and repair.In a T2DM environment, the number and osteogenic potential of hBM-MSCs are at risk of reduction. It was therefore hypothesized that transcriptional activity and the proportion of osteoprogenitors, are altered in the T2DM group, with age and sex identified as key factors that further exacerbate T2DM-related osteopathy.

Methods: hBM-MSCs from people living with and without T2DM were profiled by bulk RNA-seq. The dataset(n=3) was augmented with publicly database, yielding 65 samples (T2DM=23, Non T2DM=42). Differentially expressed genes (DEGs) were identified using DESeq2 in T2DM group, considering sex and age subgroups (20-40, 41-60 and >60 years). BayesPrism deconvolution was conducted, using scRNA data from Yuchen et al. (2023, PMID: 37474525) as the reference. DEG analysis was performed on each hBM-MSCs subpopulation from the deconvoluted results.

Results: 18 DEGs directly or indirectly related to skeletal system development were identified in hBM-MSCs from T2DM group based on the bulk RNA data. Using marker genes, 9 MSC subpopulations were identified in the scRNA data from Yuchen et al, ranging from undifferentiated MSC to more restricted in biologic potential cells. The osteogenesis-related gene ALPL showed downregulated in T2DM samples across hBM-MSCs subpopulations in deconvoluted data, although changes were not statistically significant.

Discussion: DEGs comparison within bulk RNA-seq data and across hBM-MSC subpopulations in deconvoluted data, highlighted differences between T2DM and non-diabetic groups. A consistent subpopulation pattern in deconvoluted results, aligned with previous studies, was observed. While hBM-MSCs retain multilineage differentiation capacity, heterogeneous subpopulation with diverse functions and characteristics remains.

66: Delayed rediploidisation following whole genome duplication at the base of the vertebrate lineage

Lukasz Niezabitowski[1], Róisín Long[1], Anthony Redmond[2], Aoife McLysaght[1]

[1]Trinity College Dublin, Dublin, Ireland. [2]University College Dublin, Dublin, Ireland

Background/Introduction: Whole genome duplication (WGD) events have occurred multiple times in the evolution of lineages across the tree of life. Two WGDs, 1R and 2R, occured at the base of the vertebrate tree, with 1R occuring before the divergence of gnathostomes and cyclostomes, and 2R being gnathostome-specific. A hexaploidisation also occurred in cyclostomes. WGD events provide raw material for evolution and such events may have allowed for lineage survival during mass extinctions.

Following WGD, chromosomes are inherited tetrasomically. Over time, alleles transition to diploid in a process called rediploidisation. Only after this, functional and sequence divergence of WGD duplicates (ohnologs) can occur.

Rediploidisation has been shown in some lineages to occur asynchronously over millions of years, meaning subsequent evolutionary processes are delayed. Additionally, rediploidisation can occur after speciation.

We investigate evidence for lineage-specific rediploidisation following 1R, to shed light on the evolutionary history of the vast diversity of vertebrate life seen today.

Materials/Methods: Past analysis of rediploidisation following 1R was complicated due to the lack of available hagfish genomes. Our analysis investigates evidence for rediploidisation after cyclostome-gnathostome divergence. We improve upon previous studies using recently released hagfish genomes, high-confidence ohnolog datasets, better phylogenetic models accounting for composition bias in cyclostomes, and in-depth analysis of hox clusters.

Results: Initial hox gene-tree analysis shows evidence for lineage-specific rediploidisation following 1R, and extensive delayed rediploidisation following the cyclostome hexaploidisation event. We intend to investigate rediploidisation across other regions of vertebrate genomes.

Discussion/Conclusion: Lineage-specific rediploidisation allows for the evolution of lineage-specific functions following shared WGD, which could have interesting implications for gnathostome genomes, including human genomes. Our analysis will shed light onto previously unreported aspects of rediploidisation following 1R.

68: Comparative genomics analysis and complete resistome characterisation of A. johnsonii strains isolated during and post NASA Phoenix spacecraft assembly and metagenomics-based analysis of the spatiotemporal persistence of A. johnsonii in NASA Jet Propulsion Laboratory and Kennedy Space Center clean rooms

Anna Tumeo[1], Georgios Miliotis[1], Andy O'Connor[1], Christina Clarke[2], Francesca McDonagh[1], Aneta Kovarova[1], Varsha Vijaikumarx[3], Pratyay Sengupta[3], Brigid Hooban[4], Karthik Raman[3], Dearbhaile Morris[1], Alexandre Rosado[5], Katshuri Venkateswaran[6]

[1]University of Galway, Galway, Ireland. [2]Galway University Hospital, Galway, Ireland. [3]Indian Institute of Technology Madras, Chennai, India. [4]Atlantic Technological University, Galway, Ireland. [5]KAUST, Thuwal, Saudi Arabia. [6]JPL-California Institute of Technology, Pasadena, USA

Introduction: *Acinetobacter johnsonii* is a Gram-negative Gammaproteobacterium typically associated with environmental strains. Human infection case reports are rare; however, evidence highlights *A. johnsonii* as reservoir of antimicrobial resistance (AMR). To date, its prevalence in manufacturing clean rooms and clinical settings remains largely unknown.

Materials/Methods: The study includes 16 strains isolated during and after the assembly of the Phoenix spacecraft at NASA spacecraft assembly facilities (SAF). Isolates were taxonomically classified using a combination of in-vitro (MALDI-ToF) and in-silico typing methods (16S rRNA, ANI, dDDH, gyrB). A pan-genome analysis of all *A. johnsonii* genomes available on GenBank (n=89) was conducted to contextualise the findings within a broader genomic framework. An all-vs-all ANI was performed alongside core-genome phylogeny, and the species' total resistome, virulome, and plasmidome were characterized. Genome-to-metagenome mapping to SAF metagenomes was also conducted.

Results:

- All *A. johnsonii* genomes encoded a blaOXA gene in their chromosome. The SAF *A. johnsonii* clade shared *blaOXA-652*, which was not found in other *Acinetobacter spp.* genomes except for seven additional *A. johnsonii* genomes.
- MICs of a representative SAF isolate exhibited a multidrug resistance phenotype.
- The adeA/B/C and adeI/J/K efflux pumps were also identified in the SAF genome cluster, suggesting potential adaptability to cleaning-intensive, resource-poor environments.
- Genomic signatures of *A. johnsonii* 2P07AA persisted in SAF up to 10 years post initial isolation.

Discussion/Conclusion: SAF-originating *A. johnsonii* isolates displayed multidrug resistance phenotypes attributable to the presence of *blaOXA-652*. They showed genomic persistence in SAF and divergence from the main species lineage. The genomic presence of two ade efflux pump systems indicate their ability to survive in nutrient-poor environments with intense cleaning. Overall,

this study highlights the potential for non-spore-forming Gram-negative MDR opportunistic pathogens to persist in SAF settings, with implications for planetary protection and biotechnological clean rooms.

69: Genome-wide identification of super-enhancers and enhancers-driven genes in chronic kidney disease using single-cell multiomics

Hesborn Obura, Brendan Loftus

University College Dublin, Dublin, Ireland

Background: Fibrosis represents the uncontrolled extracellular matrix (ECM) deposition produced by myofibroblasts. While genetic fate-tracing and single-cell technologies have helped uncover fibroblast heterogeneity and fibroblast-to-myofibroblast transition. Illuminating the genetic programs directing kidney development provides insights into the mechanisms of renal fibrosis and strategies for future renal therapy. Super-enhancers (SE) play a key role in cell lineage development and can be used to explain cell-type-specific expression patterns. Fibrosis is the hallmark of chronic kidney disease (CKD) and the dysregulation of transforming growth factor beta (TGFβ1) is implicated in renal fibrosis. CKD is characterized by excessive proliferation and activation of fibroblasts and their subsequent differentiation into myofibroblasts. We employed ScRNA-seq and scATAC-seq to identify gene expression programs and cis-regulatory landscapes in a cell-type-specific manner

Methods: The ScRNA-seq and scATAC-seq data of TGFβ1 treated IPSCs-kidney-derived organoids and control were analyzed using the Seurat and Signac pipelines, respectively. Labels were transferred from the annotated clusters in the scRNAseq dataset to the predicted clusters in the scATAC-seq dataset. Trajectory analysis was performed for lineage development inference. Rose algorithm was applied for SE analysis. Using the default settings, HOMER was used for motif analysis for SMAD3 motif-enriched SEs.

Results: The integrated analysis of ScRNA-seq TGFβ1 treated IPSCs-kidney-derived organoids and control showed 18 clusters with a new cluster of myofibroblast induced by TGFβ1, which is characterized by the expression of *POSTN, ACTA2, FN1,* and *TAGLN*. Using Seurat and Signac, we identified 18 scRNA-seq and 15 sATAC-seq clusters, annotated using known cell-type-specific marker genes to arrive at 12 distinct cell populations via label transfer. Trajectory inference showed myofibroblast originated from the fibroblast cluster. Differential analysis of fibroblast-myofibroblast transition showed 1743 accessible genomic loci currently under annotation for potential enhancers and super-enhancers using the Rose algorithm and Vista enhancer database.

70: Building a Placenta: Roles for Co-opted Retroviral Envelopes of Livebearing Fish

Amy Ó Brolcháin, Máire Ní Leathlobhair

Moyne Institute of Preventive Medicine, Trinity College Dublin, Dublin, Ireland

Background: Livebearing fish (Family Poeciliidae) have independently evolved a placenta several times over the past ~2 million years, making them ideal models for investigating factors contributing to placental evolution. Endogenous retroviruses (ERVs) have long been known to play a role in placentation, primarily through the activity of fusogenic syncytins, which are co-opted retroviral envelopes. Syncytins are also known to have immunosuppressive properties that may be necessary for tolerance of the feto-placental unit. It remains to be confirmed, however, whether or not these elements are necessary for placental evolution.

Methods: All publically available poeciliid assemblies were searched using a custom-built Python pipeline designed for identification of retroviral envelopes. This pipeline takes genomic assemblies as input and screens them using similarity search methods to identify candidate syncytins. Putative orthologues of these candidates were then identified by taking reciprocal best BLAST hits. Insertion dates of identified putatively functional envelopes were estimated from long terminal repeats.

Results: We identified several candidates for functionally active retroviral envelopes across 28 poeciliid species. Based on analysis of ERV integration time, orthologue groups, and structural motifs, we suggest candidate syncytins and other functional ERVs which may play a role in placentation or other biological processes.

Conclusions: Conservation of potentially functional retroviral envelopes suggests a role in the biology of poeciliid fishes, possibly related to the placenta, as observed in mammals. Given that retroviruses are currently understudied in non-mammalian species, this research aims to address this gap. Further functional characterisation of identified envelopes using transcriptomic data, immunohistochemistry, and RT-qPCR on placental tissues will help to confirm whether identified envelopes are active in placental tissues.

72: nf-hlamajority: a Nextflow pipeline for consensus MHC class I typing and its application to neoantigen prediction in breast cancer stromal cells

Kevin Ryan[1,2], Domhnall O'Connor[3], Barry Digby[1,2], Laura Barkley[3], Pilib Ó Broin[1]

[1]School of Mathematical & Statistical Sciences, University of Galway, Galway, Ireland. [2]SFI Centre for Research Training, Galway, Ireland. [3]Lambe Institute for Translational Research, University of Galway, Galway, Ireland

Introduction: Cancer-associated fibroblasts (CAFs) are a heterogeneous cell type found in the tumour microenvironment. CAFs can support tumour growth and metastasis and can contribute to therapeutic resistance, making them a potential therapeutic target. Here, we aim to identify neoantigens resulting from somatic mutations in CAFs. HLA genotyping, a critical step for neoantigen prediction, can be performed using DNA sequencing data, with various tools available. Claeys et al. (PMID:37161318) found that a majority voting approach improved HLA typing performance, however, no end-to-end pipeline currently exists to apply this approach, making it difficult for non-informaticians to implement. Our objectives were to 1) develop a Nextflow pipeline implementing majority voting for MHC class I typing from DNA sequencing, and 2) use HLA calls from this pipeline to identify neoantigens in CAFs.

Materials/Methods: CAFs and corresponding tumour-associated normal fibroblasts (TANs) were cultured from the tissue of 12 breast cancer patients (11 Luminal A and one triple-negative). Bulk RNA-sequencing (n=12) and whole-exome sequencing (WES, n=11) were carried out on CAFs and TANs. Our pipeline, *nf-hlamajority*, performed HLA typing for each patient, and 12 NCI-60 Human Tumor Cell lines with matched PCR-based genotyping, using Optitype, Polysolver, HLA-LA, and Kourami. For each HLA gene, *nf-hlamajority* assigned the genotype called by the highest number of tools. These HLA genotypes were input to Landscape of Effective Neoantigens Software (LENS) to identify CAF-specific neoantigens (PMID:37184881).

Results and Discussion: Results from the NCI-60 dataset showed 97% accuracy, with 68 out of 70 calls matching PCR-based genotyping. LENS identified potential neoantigens resulting from missense mutations, including genes involved in lipid metabolic pathways. CAFs contribute to lipid metabolism, impacting cancer progression and tumour immunogenicity. Future efforts will focus on neoantigen validation using T-cell immunogenicity assays, offering insights into the potential benefit of targeting CAF neoantigens to enhance cancer therapies.

75: IDENTIFICATION OF PUTATIVE INHIBITORS OF BACE-1 USING MOLECULAR DOCKING AND CHEMINFORMATICS TECHNIQUES

Nurudeen Owolabi

Ladoke Akintola University of Technology, Ogbomoso, Nigeria. Foresight Institute of Research and Translation, Ibadan, Nigeria

Background/Introduction: Alzheimer's disease (AD) is a chronic and irreversible neurodegenerative disease and is the most common cause of dementia, accounting for 60-80% of all cases. AD is characterized by cognitive decline and memory loss, consequently leading to factors impairing daily activities: depression, disorientation, apathy, impaired communication, poor judgement, and behavioural problems, and eventually leads to death. This study aims to identify potent inhibitors against beta-site amyloid precursor protein cleaving enzyme (BACE) 1.

Material/Methods: We used computational and cheminformatics techniques to identify the best promising (lead) compounds among 7,171 BACE-1 inhibitors obtained from the ENAMINE database. This includes molecular docking techniques, which determine compounds with the best binding affinity and interaction with catalytic residues and compounds' drug-likeness analyses based on Egan's, Verber's, and Lipinksi's rules.

Results: We observed that four compounds (2752, 4074, 6613, and 5085) demonstrated promising potentials in terms of binding affinities (-9.35, -8.76, -8.66, and -8.66, respectively), hydrogen bonding with catalytic residues (5, 2, 4, and 3), and drug-likeness analysis (all demonstrated excellent profiles).

Discussion/Conclusion: The observed features demonstrated by these lead compounds are widely understood to be a marker of potent inhibitory abilities.

Conclusively, our lead compounds demonstrated excellent potential as a viable drug candidate against BACE-1 if developed into drug molecules. However, further studies, such as molecular dynamics simulation, toxicity, preclinical, and clinical tests, are required to evaluate these compounds further.

78: Combining individual and wastewater whole genome sequencing improves SARS-CoV-2 surveillance

Evan P. Troendle[1], Andrew J. Lee[2], Marina I. Reyne[2], Danielle M. Allen[2], Stephen J. Bridgett[1], Clara H. Radulescu[1], Michael Glenn[1], John-Paul Wilkins[2], Francesco Rubino[2], Behnam Firoozi Nejad[3], Cormac McSparron[3], Marc Niebel[4], Derek J. Fairley[4], Christopher J. Creevey[2,5], Jennifer M. McKinley[3], Timofey Skvortsov[6], Deirdre F. Gilpin[6], John W. McGrath[2,5], Connor G. G. Bamford[2,5], David A. Simpson[1]

[1]Wellcome-Wolfson Institute for Experimental Medicine, School of Medicine, Dentistry and Biomedical Sciences, Queen's University Belfast (QUB), Belfast, United Kingdom. [2]School of Biological Sciences, QUB, Belfast, United Kingdom. [3]Geography, School of Natural and Built Environment, QUB, Belfast, United Kingdom. [4]Regional Virology Laboratory, Belfast Health and Social Care Trust, Royal Victoria Hospital, Belfast, United Kingdom. [5]Institute for Global Food Security, QUB, Belfast, United Kingdom. [6]School of Pharmacy, QUB, Belfast, United Kingdom

Background: Robust methods to track pathogens support public health surveillance. Both wastewater (WW) and individual whole genome sequencing (WGS) are used to assess viral variant diversity and spread. However, their relative performance and the information provided by each approach have not been sufficiently quantified. Therefore, we conducted a comparative evaluation using extensive individual and wastewater longitudinal SARS-CoV-2 WGS datasets in Northern Ireland (NI).

Methods: WGS of SARS-CoV-2 was performed on >4k WW samples and >23k individuals across NI from 14th November 2021 to 11th March 2023. SARS-CoV-2 RNA was amplified using the ARTIC nCov-2019 protocol and sequenced on an Illumina MiSeq. Wastewater data were analysed using Freyja to determine variant compositions, which were compared to individual data through time series and correlation analyses. Inter-programme agreements were evaluated by mean absolute error (MAE) calculations. WW treatment plant (WWTP) performances were ranked by mean MAE. Volatile periods were identified using numerical derivative analyses. Geospatial spreading patterns were determined by horizontal curve shifting.

Results: Strong concordance was observed between wastewater and individual variant compositions and distributions, influenced by sequencing rate and variant diversity. Overall variant compositions derived from individual sequences and each WWTP were regionally clustered rather than dominated by local population size. Both individual and WW sequencing detected common nucleotide substitutions across many variants and complementary additional substitutions. Conserved spreading patterns were identified using both approaches.

Conclusions: Both individual and wastewater WGS effectively monitor SARS-CoV-2 variant dynamics. Combining these approaches enhances confidence in predicting the composition and spread of major variants, particularly with higher sequencing rates. Each method detects unique mutations, and their integration improves overall genome surveillance.

79: Full-length 16S rRNA Amplicon Profiling of Clinical Microbial Keratitis Samples

Michael Glenn

Queen's University, Belfast, United Kingdom

Background/Introduction: Microbial keratitis is a serious infection of the cornea, and identifying the specific microorganism responsible for infection is often challenging but is essential for guiding effective treatment. Culturing is considered the current gold-standard technique for pathogen identification in microbial keratitis. However, many clinical cases fail to yield microbial culture growth, and even when growth is achieved, the isolated organism(s) may not necessarily be responsible for infection. Previous studies have indicated the value of 16S sequencing but further validation with a larger clinical cohort, combined with a full-length sequencing approach, would enhance our understanding of its efficacy and could facilitate broader clinical adoption.

Materials/Methods: 61 clinical keratitis specimens were collected from patients at The Royal Liverpool University Hospital using corneal impression membranes (CIMs). Each specimen was processed for both standard bacteriology culturing and 16S rRNA sequencing. DNA extraction, full-length 16S rRNA amplification, nanopore sequencing, and subsequent computational analysis were conducted at Queen's University, Belfast.

Results: Of the 61 CIM samples, 37 (61%) showed positive microbial culture growth, while 24 (39%) exhibited no growth. A bacterial taxonomic profile was successfully generated for all samples using 16S rRNA sequencing. There was a 46% concordance between culture results and 16S profiling, with the isolated bacteria from culture appearing among the top two taxa with the highest relative abundances in the 16S profiles. In several culture-negative cases, the taxonomic profiles revealed high relative abundances of established keratitis pathogens *Moraxella nonliquefaciens* and *Corynebacterium mastitidis*, suggesting these bacteria as likely causative agents that were overlooked by standard bacteriology culture.

Discussion/Conclusion: Full-length 16S sequencing demonstrates significant potential as a tool for pathogen detection in clinical practice, offering improved identification of causative bacteria in keratitis compared to traditional culture methods.

80: HIVE (Heatmap for Interactive Viewing of gene Expression data) Browser

Stephen Bridgett, Evan P Troendle, Clara Radulescu, Sara El Jadid, Michael Glenn, Guillermo Lopez-Campos, David Simpson

Queens University, Belfast, United Kingdom

Introduction: Genomics technologies are continually advancing, producing datasets of increasing size, number, and complexity - for example: a single cell RNA-sequencing experiment can generate multiple lists of genes that are differentially expressed within multiple cell-types, from samples at different conditions or timepoints.

Materials/Methods: We therefore developed this 'Heatmap for Interactive Viewing of Expression data (HIVE)' web-browser-based application to enable users to easily load and analyse their data-files. They can quickly generate a heatmap of expression changes; filter genes by fold-change and p-value; then submit selected genes to external tools for network and gene set enrichment analyses; and download a heatmap png/jpg image. HIVE's single webpage tabular format is familiar for those accustomed to spread-sheets.

The browser is developed as a single-page application, using HTML, Javascript and CCS. The String DB API javascript file as its only dependency (which is automatically used if the computer/tablet has an internet connection).

Results: Tabular text files containing differential expression data (eg scRNA-Seq, proteomics etc.) can be loaded directly into the HIVE browser. The data can be multiple files (one file per sample/condition eg: from DESeq), or all in one file. The columns should be separated by commas (csv) or tabs (tsv). HIVE will try to identify the gene-id, fold-change, (and optional p-value and percentage of cells) columns from the column headings. The user can confirm the automated selection or choose their own custom columns.

The files are loaded directly into your local webbrowser. Files are not uploaded to a web-server, thereby keeping data private. HIVE can also load data from a dropbox or github web-url, making sharing with collaborators easy.

Conclusion: HIVE is available free on GitHub-Pages: https://qub-simpson-lab.github.io/HIVE_browser/ (or: bit.ly/hivebrow), or can be downloaded from Github repository: https://github.com/QUB-Simpson-lab/HIVE_browser along with sample expression data in standard csv (comma-separated-value) or tsv (tab-separated-value) formats.

81: A20: The Link Between Weight Loss and Inflammation Control in Obesity

John Scanlan[1], David Finlay[1], Lydia Lynch[2]

[1]Trinity College Dublin, Dublin, Ireland. [2]Princeton, Princeton, USA

Background/Introduction: Over 10% of the global population and a staggering 60% of Irish adults are classified as overweight or obese. Obesity is not just a disorder of excess weight; it's a complex condition intricately linked to an array of comorbidities including cancer, cardiovascular diseases, and diabetes. Central to obesity's pathophysiology is the immune system, particularly the chronic low-grade inflammation observed in obese individuals. This inflammation is both a result and a driver of the metabolic dysregulations associated with obesity. Our research aims to elucidate the cellular alterations in the immune systems of obese individuals, focusing on how weight loss impacts these changes. Specifically, we investigate the role of TNFAIP3 (A20), a well-documented negative regulator of the NF-κB pathway, which is known to be hyperactivated in obesity.

Materials/Methods: Utilizing single-cell RNA-sequencing datasets, we analysed immune cells from lean, obese, and weight loss subjects across human and mouse models. Our approach utilised both adipose and blood tissues to provide a comprehensive overview of immune cell responses under these conditions.

Results: Meta-analysis reveals a consistent downregulation of TNFAIP3 in obese subjects, a pattern which is reversed to match levels seen in lean individuals upon weight loss.

Discussion: The downregulation of TNFAIP3, observed across all lymphocyte populations, suggests a systemic mild activation of the NF-κB pathway in obesity that can be reversed. The restoration of TNFAIP3 expression through weight loss highlights a mechanism through which obesity-induced inflammation can be mitigated. By deciphering how weight loss restores TNFAIP3 function and suppresses NF-κB-mediated inflammation, our study paves the way for developing targeted interventions that mimic these effects.