nBUG14 | University of Huddersfield | 28th March 2025

Agenda

Time	Speaker	Presentation title
10:00	Registration + tea/coffee	Room JPSG/39 (to the right after entering the building)
10:30	Welcome	Room JPSG/18
10:35-11:15	Dr Marina Soares Da Silva, The Crick	Modern science from old bones: challenges and applications of ancient genomics
11:20-12:00	Dr George Foody, UK Biobank	Introduction to UK Biobank Research Analysis Platform
12:00-13:00	Lunch	Room JPSG/39 (by registration)
13:00-14:15	Clothilde Annabelle Francois, Rebecca Gullick-Shibata, Gagan Vishaya, University of York	New Ara-BOX-cis v2 trained on a single nucleus RNA-seq atlas across eight plant developmental stages
13:00-14:15	Surabhi Ranavat, University of York	Genomic signatures of inbreeding in a threatened African timber tree species, *Pericopsis elata* (*Fabaceae*)
13:00-14:15	Martin Carr, University of Huddersfield	Selection on both tRNA and mRNA facilitated efficient and accurate protein translation in ancestral eukaryotes
13:00-14:15	Rowan Green, University of Manchester	Environmental effects on mutagenesis in microbial populations
13:00-14:15	Aggie Turlo, University of Manchester	Practical challenges in breath metabolome analysis
14:15-14:25	Oxford Nanopore sponsor talk	
14:30-16:00	Tea/coffee break + poster session	Poster session is long as we have quite a few posters
16:00-17:15	Andrew Mason, University of York	Biologically-informed machine learning identifies a new clinically-actionable bladder cancer subgroup
16:00-17:15	Syed Murtuza Baker, University of Manchester	Challenges and approaches to spatial transcriptomics analysis

Time	Speaker	Presentation title
16:00-17:15	Mingkai Wang, University of Bradford	Analysis and Prediction of Changes in Protein Thermostability (ΔΔG) upon Point Mutations
16:00-17:15	Ophelia Forbes, University of Huddersfield	Computational analysis of human protein atlas antibody intensities in healthy and cancerous tissues
16:00-17:15	Dave Lunt, University of Hull	Bioinformatics at Hull: tales from the real world
16:00-17:15	Closing discussion + pub session	Warehouse, 200m from the meeting building :-)

Abstracts (in order of presentations)

Modern science from old bones: challenges and applications of ancient genomics

Marina Soares Da Silva, The Crick | marina.silva@crick.ac.uk

Ancient DNA (aDNA) refers to DNA molecules preserved in biological material (e.g. bone) and recovered postmortem, usually from museum collections or archaeological excavations. Biochemical processes triggered by environmental factors contribute to DNA degradation after death. As a result, aDNA is present in small amounts, is very fragmented (typically <100 bp) and displays a pattern of increased cytosine deamination. These characteristics can be used as a measure of authentication to rule out contamination, but they also introduce bioinformatic challenges and can confound downstream population genetic analysis. In this talk I will cover specific aspects of aDNA bioinformatic processing and showcase how, despite the technical challenges, aDNA research is uniquely positioned to provide insights into human past.

Introduction to UK Biobank Research Analysis Platform

George Foody, UK Biobank | George.Foody@ukbiobank.ac.uk

UK Biobank is a biomedical resource that enables health-related research in the public interest to be undertaken globally. With half a million volunteers and nearly 20 years of collected lifestyle and biological data linked to medical records, the UK Biobank has created a database with unique participant characterisation and follow up. This regularly augmented resource is already over 30 petabytes and includes health record, genomic, proteomic, environmental, and imaging data. To enable researchers to work with data at this scale, a cloud-based platform, the UK Biobank Research Analysis Platform (UKB-RAP) was developed. This platform continues to expand with increased tools and workflows available. The UKB-RAP allows researchers to control their compute and costs to suit their research needs.

New Ara-BOX-cis v2 trained on a single nucleus RNA-seq atlas across eight plant developmental stages

Anabelle Francois, Rebecca Gullick-Shibata & Gagan Vishaya, University of York | xsp507@york.ac.uk, rgullick.shibata@gmail.com

Plants have experienced an expansion of transcription factor (TF) families, resulting in many TFs with similar DNA binding preferences. Previously, a bioinformatics tool called Ara-BOX-cis was developed to predict the downstream targets of bHLH and bZIP TFs in Arabidopsis. However, this tool was developed for seedlings only and was trained using bulk RNA-seq data. The University of York Bioinformatics MSc has developed Ara-BOX-cis v2, which is trained on a single nucleus RNA-seq atlas across eight developmental stages, showcasing how gene networks get rewired over development. Moreover, they have expanded the scope of the Ara-BOX-cis tool to incorporate WRKY TFs involved in agriculturally significant processes, like plant immune response and the reallocation of nutrients from leaves to reproductive structures. These results will be integrated into a Shiny App to enable wider access, showcasing how Bioinformatics MSc courses can be structured to deliver useful research outcomes that promote open research.

Genomic signatures of inbreeding in a threatened African timber tree species, *Pericopsis elata* (*Fabaceae*).

Surabhi Ranavat, University of York | surabhi.ranavat@york.ac.uk

Pericopsis elata is a large, light-demanding tree that is highly exploited in tropical Africa for its timber, and is listed as endangered by the IUCN. It is a peculiar tropical tree as it has a mixed-mating system, with a high selfing rate, and a lack of regeneration throughout its distribution range. Two highly differentiated gene pools (FST=0.53) were identified in Central Africa (Eastern and Western). Along the Western gene pool, we observe a steep westward decay of heterozygosity, and we hypothesize that this is a recent range expansion into Cameroon with founder effects, possibly facilitated by selfing. Whole genome resequencing data of individuals along the Western gene pool revealed longer runs of homozygosity and an increase in the inbreeding coefficient compared to the Eastern pool. These analyses, combined with phenotypic data, will help in identifying the effects of inbreeding and informing sustainable management strategies for this threatened timber tree.

Selection on both tRNA and mRNA facilitated efficient and accurate protein translation in ancestral eukaryotes.

Martin Carr, University of Huddersfield | m.carr@hud.ac.uk

tRNA molecules are highly modified in eukaryotic cells and modifications of the anticodon wobble position have the potential to alter the translational capacity of tRNA molecules. A major expansion of tRNA modification, through the deamination of adenosine to inosine at the anticodon wobble position, occurred during eukaryogenesis. Deaminated tRNA is shown to provide a selective advantage, at the levels of translational efficiency and accuracy, across extant eukaryotes which appears to be due to the abundance of the modified molecules within cells. Selection for efficient translation is shown to operate upon both mRNA and tRNA genes. The presented data indicate that the ancestral eukaryote possessed primarily GC-ending optimal codons. For eight amino acids the optimal codons were predominantly translated by deaminated tRNA. Extant species with strong genomic AT base composition have largely retained ancestral GC-ending codons, however novel optimal codons have convergently emerged in crown-group eukaryotes due to strong mutation pressure.

Environmental effects on mutagenesis in microbial populations.

Rowan Green, University of Manchester |rowan.green@manchester.ac.uk

By understanding how the environment affects the rate, and type, of mutations occurring in microbial populations we can hope to better predict and prevent the evolution of antimicrobial resistance. One example of such environmental dependency is the elevation of mutation rates at low population density in many microbial species including E. coli. The elevation of mutation rates at low densities is driven by an increase in the rate of A>G transition mutations. The effects of population density on mutation rates and types have previously only been explored using target locus specific methods. In contrast to this we examine genome wide patterns of environmentally dependent mutagenesis using deep whole genome sequencing. We identify newly evolved rare variants in 95 independent E. coli populations to uncovering patterns of mutagenesis associated with population density at a genome wide scale.

Practical challenges in breath metabolome analysis.

Aggie Turlo, University of Manchester | aggie.turlo@manchester.ac.uk

Exhaled volatile compounds are attractive source of non-invasive biomarkers in respiratory disease and beyond. One of techniques commonly used in breath biomarker discovery is gas chromatography-mass spectrometry. Analysis of untargeted mass spectrometry data, especially those obtained from longitudinal studies, presents challenges such as batch effects, instrumental drift and missing data. Bioinformatics approaches developed to address these challenges are not always available in analysis of breath metabolome data, due to e.g. unknown effects of sample storage and difficulty in creating pooled quality control samples. Moreover, breath samples are susceptible to environmental contamination, with the composition of inhaled air posing a major source of variation. This abstract discusses practical challenges encountered during analysis of breath metabolomics dataset obtained from large observational asthma diagnostic study, Rapid Access Diagnostics for Asthma (RADicA). This work highlights the need for development of bioinformatics workflows tailored to breath metabolomics data analysis.

Oxford Nanopore - Sponsor Talk followed by poster session

Biologically-informed machine learning identifies a new clinically-actionable bladder cancer subgroup.

Andrew Mason, University of York | andrew.mason@york.ac.uk

While personalised medicine is yet to be realised, unsupervised machine learning (ML) has been used to stratify cancer patient transcriptomes into “subgroups”. However, these have not translated to improved patient treatment. Rather than assuming ML will find clinically-relevant tumour biology, we forced it using experimentally-derived (drug response RNAseq) or physiologically-functional gene sets (TFs), and by incorporating expression data from healthy tissue. For proof-of-concept, we derived networks from healthy bladder urothelium (n=114) using PGCNA, allowing TFs to retain double the edges after pruning. Random prioritisation revealed 98 consistently highly-connected TFs we used to stratify The Cancer Genome Atlas bladder cancer cohort (n=408) identifying a novel, extremely-aggressive subgroup (4.9%) with cancer-protecting NRF2 activity. This group segregated canonical single-mode markers of NRF2 activity, revealing patients likely eligible for FDA-approved inhibitors used in lung cancer. Our study shows we can guide ML to identify subgroups not on final phenotype, but on clinically-targetable oncogenic processes.

Challenges and approaches to spatial transcriptomics analysis.

Syed Murtuza Baker, University of Manchester | syed.murtuzabaker@manchester.ac.uk

Spatial transcriptomics (ST) analysis faces several challenges, including identifying cell segmentation, matching cells to spots and difficulties in integrating multi-omics data. The cost of ST technologies further hinder widespread adoption in clinical settings. To overcome these challenges, we adopt different computational approaches and methods. We are also developing deep learning models that use ST technologies and make prognostic predictions of disease. In this talk I will explain different challenges associated to ST and the different computational approaches including machine learning models we are using to address these challenges.

Analysis and Prediction of Changes in Protein Thermostability (ΔΔG) upon Point Mutations.

Mingkai Wang, University of Bradford | mwang20@bradford.ac.uk

Thermostability is an important property of proteins and a critical factor for their wide application. Accurate prediction of ΔΔG enables the estimation of the impact of mutations on thermostability in advance. A range of ΔΔG prediction methods based on machine learning has now emerged. However, their prediction performance remains limited due to insufficiently informative training features and little effort has been made to integrate related feature calculation resources. Based on this, we integrated 12 computational resources to develop a pipeline capable of automatically calculating 1547 features. In addition, a feature-enriched ΔΔG dataset was created, including 15752 ΔΔG data. Furthermore, we performed feature selection and developed an accurate ΔΔG prediction model. It also outperformed several other representative prediction methods in comparisons with independent datasets.

Computational analysis of human protein atlas antibody intensities in healthy and cancerous tissues.

Ophelia Forbes, University of Huddersfield | U2170464@unimail.hud.ac.uk

Immunohistochemistry (IHC) is essential for studying protein expression in tissues using chromogenic stains, including DAB (3,3'-diaminobenzidine). While the gold standard remains manual assessment, there are computational methods for quantifying DAB-stained regions of the tissue, such as colour deconvolution (CD) in Fiji (open-source software for biological-image analysis). CD requires pre-defined colour matrices and manual adjustments, which limit reproducibility and efficiency. This study developed a novel computational pipeline scripted in Python to automate DAB quantification that works by isolating haematoxylin-stained regions using hue-defined pixel masking. The method is called hue-based quantification (HUE-Q).

Bioinformatics at Hull: tales from the real world.

Dave Lunt, University of Hull | dave.lunt@gmail.com

A talk about user experiences of running bioinformatics in a mixed academic community. Lessons learned, tips and tricks, screw-ups never to be repeated, and how we arrange bioinformatics projects now. I'll discuss reproducibility, electronic lab notebooks, software environments, and workflow managers. This isn't a science talk, rather a discussion of doing bioinformatics in the real and imperfect world.

Posters (titles and abstracts where available)

[1] Dora Marčec, University of Manchester | dora.marcec@postgrad.manchester.ac.uk

Impact of ageing on the tissue microenvironment and identification of ageing biomarkers (placeholder title).

The process of ageing induces physiological changes often leading to complex chronic diseases. Whilst a large body of literature describes age-induced changes at a tissue level, little is known about the impact of ageing on the tissue microenvironment, defined by its extracellular matrix. This project investigated matrisome and mitochondrial proteome expression changes with age across 44 human tissue types, aiming to develop a workflow to identify existing and novel ageing biomarkers. Data mining was used to obtain immunohistochemistry images and corresponding patient metadata from the Human Protein Atlas (HPA) database. The patient metadata from the HPA database has never been explored in the context of ageing, making this project the first exploration of the HPA data in an ageing study. The purpose was to identify a subset of mitochondrial and extracellular matrix proteins whose expression levels significantly change during ageing.

[2] Adrienne Unsworth, University of Newcastle | a.unsworth2@ncl.ac.uk

To reconstruct retinal cell lineages using endogenous barcodes or other markers of cell age, such as somatic variants, to take advantage of existing single cell data. (placeholder title).

[3] Ana R. Martinez, University of Leeds | Contact: a.i.martinezrodriguez@leeds.ac.uk

Investigating the somatic mutational landscape of UBA1 within large genomic repositories and diagnostic screening cohorts based in the UK.

[4] Hugo Aguado Robert, University of York | ha1238@york.ac.uk

Differential Transcript Isoform Usage In Bladder and Upper Tract Urothelium.

[5] Pasky Miranda, University of York | pasky.miranda@york.ac.uk

The UKRI Digital Research Skills Catalyst project: to provide a central website giving streamlined access to DaSH data analysis courses. (placeholder title).

The UKRI Digital Research Skills Catalyst project aims to provide a one-stop shop for UKRI-funded training and upskilling resources for data analysis in biology and health sciences. In 2021 DaSH projects were funded, including:

- Cloud-SPAN (omics analyses using HPC)
- ELIXIR-UK (FAIR practice and data stewardship)
- IDEAS (Imaging analysis in health and biosciences)
- Learn to Discover (Python programming and machine learning)
- DATA CAMPP (Automated data capture, analysis, and management)
- Ed-DaSH (Statistics and data science)
- Innovation Scholars (Big data skills)

The Digital research Skills Catalyst will provide a central website giving streamlined access to all the courses provided by the above DaSH projects, in addition to a brand new booking system to book a timeslot for one-to-one help with one of the DaSH specialists, empowering researchers with access to cutting edge data skills and expertise.

[6] Lewis Ward, University of Huddersfield | U2283985@unimail.hud.ac.uk

Maudr: an R package for analysing data and generating model answers for enzyme kinetics undergraduate laboratory practicals (placeholder title).

maudr is an R package that generates and analyses enzyme kinetics data commonly employed in undergraduate laboratory practicals. Initially developed as a series of scripts during the COVID-19 pandemic to support laboratory work when students could not come to campus, maudr is now used internally as a pedagogical tool for ad-hoc exercises, formative assignments and to support data transformation techniques in R programming classes. Maudr generates unique datasets of enzyme activity with or without inhibitors for each students to analyse. By default, parameters for alcohol dehydrogenase (ADH) is provided, but : if Km, Kcat, and extinction coefficient are known, maudr can generate other enzyme activity data. Maudr then takes the generated datasets, analyses them and produces Michaelis-Menten and Lineweaver-Burke plots to support marking of students assignments. The package is currently hosted on GitHub.

[7] Matthew Merkin, University of Liverpool | hlmmerki@liverpool.ac.uk

Effect of anthropogenic factors on genomic patterns of selection and demographic changes in lepidopterans over the last century (placeholder).

Whole-genome sequencing data has been collected for 20 species of British lepidopterans (butterflies and moths) from both Natural History Museum collections (pre-1930s) and recent sampling (late 2010s). Here, I outline my plan (and share some preliminary results) to use population-based statistical methods to infer genomic patterns of selection and demographic changes that have arisen over the last 80-100 years, and whether they have been driven by anthropogenic factors, such as agricultural intensification and recent climatic changes. As such, this analysis should provide insights into the adaptations of particular species, convergent traits that have arisen across multiple species and the major threats that lepidopterans have faced in the last century. In addition, I will discuss the challenges of working with DNA from historical collections.

[8] Marvellous Oyebanjo, University of Bradford | marvel.oye.bioinformatics@gmail.com

Gene analysis of mallard and Muscovy duck prolactin (dPRL) (placeholder title)

Regions of mallard and muscovy duck prolactin (dPRL) (exon 5 and parts of intron 4) were sequenced. 32 SNPs were identified in Mallard and Muscovy ducks; notably, SNPs 526A>G and 684T>C were found recurrently in Muscovy alone, 584T>A occurred in Mallard alone, and SNP 581G>A was detected in both breeds. INDELs were found in intron 4 and 3' flanking regions. A highly conserved motif (VHSGDIGNEVYSQWEGLPSLQLADED/VHSGDIGNEIYSQWEGLPSLQLADKD) was observed in both breeds – and this motif is linked with the somatotropin/prolactin hormone (SPH) superfamily. Additionally, a Leucine-rich functional motif (CLRrDShKIDnYlkVlkC) was found in Muscovy dPRL alone. The evolutionary conservation profile of dPRL in both breeds revealed a moderate degree of conservation, involving functional and structural residues. The PPI network of dPRL gene involved: GH1 (Growth Hormone 1), GHR (Growth Hormone Receptor), PRLR (Prolactin Receptor), FSHB (Follicle Stimulating Hormone Subunit Beta), INS (Insulin), POMC (Proopiomelanocortin), TSHB (Thyroid Stimulating Hormone Subunit Beta), JAK2 (Tyrosine-protein kinase), and LEPR (Leptin Receptor).

[9] Claudio Silva De Freitas, University of Hull | C.F.SILVA-DE-FREITAS-2018@hull.ac.uk

Are there ecological consequences of a rhodopsin polymorphism in cichlids?

Visual sensory systems of animals are vital for perceiving their environment and are constantly under strong selection pressures. Selection for vision is greater in aquatic environments compared to terrestrial as light intensity and spectrum greatly change along a depth gradient because of the properties of water. Turbidity also adds greater selection pressure on the sensory system for aquatic animals. Rhodopsin (RH1) is responsible for gathering achromatic visual information vision as well as playing a pivotal role in recognising light spectrum around the λmax-value of 500nm (blue-green part of the spectrum). Rhodopsin (RH1) is responsible for gathering achromatic visual information vision as well as playing a pivotal role in recognising light spectrum around the λmax-value of 500nm (blue-green part of the spectrum). In this study we examine the role of a single base pair mutations (A292S) in RH1 that leads to a shift in light spectrum sensitivity and affecting the ability of food identification. We suspect the single base pair mutation leads to a shift of light sensitivity towards the bluer region of the light spectrum of about 15nm. We tested this idea on Astatotilapia calliptera who are "the general" cichlid of Lake Malawi and are known to have distinguished populations with various genotypic and phenotypic traits because of their respective environments they occupy. We found A. calliptera near the chisumulu site (open water), who obtained the mutation in Rh1, were better suited in identifying food in the bluer region of the light spectrum compared to other individuals and populations who did not. As such we believe that A. calliptera individuals with this mutation are better equipped to survive in harsher light environments as they are likely to identify food easier and therefore giving them an ecological advantage compared to individuals without.

[10] Kiran Gok Lune, University of Sheffield | kgllee1@sheffield.ac.uk

Parentage assignment from low-coverage whole genome data in Seychelles warblers (placeholder title).

Genomic data may soon be an efficient way to gather detailed data from long-term sampled populations. We have ~2000 imputed low-coverage whole-genome sequences of Seychelles warblers from a closed, island population. This data can accurately assign sex and parentage to sampled individuals. Sex and parentage assignment is labour intensive, each requiring PCR amplification of markers followed by gel electrophoresis. Sex assignment in birds uses a gene, sexually dimorphic in length. Parentage assignment uses hypervariable microsatellite markers and in Seychelles warblers, has also relied on detailed observations from fieldwork by tracking behaviours of birds. Both these assignments can now be done with low-coverage whole-genome sequencing, which is more efficient, in the lab and field, and it comes free with the range of downstream population genetic analyses that can be done on the genomic data.

[11] Daphne Ezer, University of York | daphne.ezer@york.ac.uk

New Ara-BOX-cis v2 trained on a single nucleus RNA-seq atlas across eight plant developmental stages (placeholder title).

Plants have experienced an expansion of transcription factor (TF) families, resulting in many TFs with similar DNA binding preferences. Previously, a bioinformatics tool called Ara-BOX-cis was developed to predict the downstream targets of bHLH and bZIP TFs in Arabidopsis. However, this tool was developed for seedlings only and was trained using bulk RNA-seq data. The University of York Bioinformatics MSc has developed Ara-BOX-cis v2, which is trained on a single nucleus RNA-seq atlas across eight developmental stages, showcasing how gene networks get rewired over development. Moreover, they have expanded the scope of the Ara-BOX-cis tool to incorporate WRKY TFs involved in agriculturally significant processes, like plant immune response and the reallocation of nutrients from leaves to reproductive structures. These results will be integrated into a Shiny App to enable wider access, showcasing how Bioinformatics MSc courses can be structured to deliver useful research outcomes that promote open research.

[12] Alex Marks, University of York | alexandermgmarks@gmail.com

Use of machine learning to assign functional guilds to microbial species found in anaerobic digestion reactors (placeholder title).

Anaerobic digestion (AD) is essential for sewage treatment and agricultural waste management, producing carbon-neutral biogas. The process has four stages: hydrolysis, acidogenesis, acetogenesis, and methanogenesis, which are facilitated by a diverse microbial community. However, assigning functional roles to microbes remains challenging. This exploratory study aimed to use machine learning to assign functional guilds to species found in AD reactors. We assessed microbial functional diversity using KEGG module completeness data from 206 whole-genome sequences. Genomes were annotated and clustered using k-means, identifying 12 functional groups visualised with UMAP. Clusters did not align with taxonomic classifications, suggesting functional rather than taxonomic distinctions. The methanogen cluster was best defined and only contained archaea. However, archaea did appear in other clusters. Our findings pave the way for machine learning to determine biologically relevant functional guilds that could enhance understanding and optimise AD reactor performance and stability.

[13] Hayley Chesman, University of Huddersfield | U2167501@unimail.hud.ac.uk

A deep dive into codon usage and gene optimisation in killer whales and other *Cetacean* species.

Synonymous codons have an unequal distribution within the genome, creating patterns of biased codon usage. A previous study on single-celled relatives of the metazoans presented strong evidence for natural selection being the major force driving codon bias for translational efficiency and accuracy. This study investigates codon usage in killer whales in comparison to four other members of the Cetacea infraorder, revealing new information on the extent of codon bias in aquatic mammals, the direction of bias against a neutral mutational model and the role of modified tRNA in translation. These species exhibit similar patterns of codon usage in highly biased genes, that are enriched with codons complimentary to the most abundant tRNA genes. Optimal codons appear to be conserved among these species and their metazoan ancestors. Furthermore, there is some preference shown for deaminated modified tRNA molecules from adenosine to inosine.

[14] Aneeza Barkat, University of Huddersfield | u1958957@unimail.hud.ac.uk

Gene optimisation in brain tumours.

[15] Ahmed Boubaker, University of Huddersfield | U2265718@unimail.hud.ac.uk

Investigation of Optimal Codons in Deuterostomes.

Codon usage bias in deuterostomes is influenced by mutational forces, genetic drift, and translational selection. We analysed optimal codons across 70 species, identifying ~1,890 translationally optimal codons, all GC-ending. Nc-plots showed most genes align with neutral expectations, though subsets exhibited stronger GC bias, suggesting translational selection. The TAPSLIVR amino acids consistently had C-ending optimal codons, aligning with the role of A34 deaminated tRNAs. These findings highlight conserved codon preferences and the interplay of genomic composition and tRNA availability in deuterostome evolution.

[16] Holly Anne-Albans, University of Huddersfield | u2167501@unimail.hud.ac.uk

[17] Nunnapat Jantorn, University of York | thc546@york.ac.uk

[18] Khaled Jum'ah, University of Bradford | K.jumah@bradford.ac.uk

[19] Gagan Shetteppanavar Vishaya, University of York | ltz508@york.ac.uk