Samuel Rice – Bioinformatics Projects and Technical Portfolio

Master of Bioinformatics (MSc) at the University of Liverpool 2025

Bachelor of Biological Sciences (BSc) at the University of Liverpool 2024

Who Am I?

I pride myself on being a **hard worker**, combining my **passion** for **biology** with **strong programming skills** to solve complex problems in **genomics**, **proteomics**, and **metabolomics**.

Why Choose Me?

(I) Intuitive Problem-Solver (II) Analytical Thinker (III) Versatile Bioinformatician (IV) Efficient Communicator (V) Fast learner (VI) Team-Oriented and Independent

Portfolio Contents:

GITHUB Portfolio

UNIVERSITY OF LIVERPOOL

CBF COMPUTATIONAL BIOLOGY FACILITY

HiPy

NEXT GENERATION MIXED MARTIAL ARTS

Contact Details

Email: samuel.rice.bioinfo@gmail.com

LinkedIn: www.linkedin.com/in/samuel-rice-bbb979245

GitHub: https://github.com/samuel-rice-bioinfo

Technical Skills and Tools

Programming and Scripting

- R
- Python
- Linux /BASH
- MATLAB

Bioinformatics Tools and Databases

Sequence & Structural Analysis:
•BLAST, SAMtools, DIAMOND, InterPro, UniProt/PDB, Clustal Omega, MEGA-X, iTOL

Genomics & Transcriptomics Analysis
•DESeq2, Seurat, FreeBayes, BWA-MEM, SAMtools, InterPro, UniProt, HTSeq, DAVID, Cytoscape, GO enrichment

Proteomics & Phosphoproteomics:
•MaxQuant, Comet, Lorikeet, OrthoFinder, AlphaFold, ColabFold, PyMOL, InterProScan

Metabolomics & Pathway Analysis:
•MetaboAnalyst, KEGG, PathIntegrate, ipaPy2, BioCyc

Data Analysis & Visualisation:
•ggplot2, seaborn, matplotlib, Cytoscape, Pandas, NumPy, SciPy

Machine Learning & Modelling:
•Scikit-learn, COBRA Toolbox, PCA, SVM, LDA, Hierarchical Clustering, k-NN, Random Forests, Neural Networks, Cross Validation, ROC curves, Confusion Matrices, Overfitting control, Elbow method

Multi-Omics Research

Natures Crop Protection: A Multi-Omics Approach to Understanding Streptomyces Strain Interactions in Potato Scabies

Tools and Techniques:
•RNA-seq, **DESeq2**, Mass spectrometry (MS1), PQN + Glog, **KNN imputation**, Likelihood Ratio Testing (LRT), **DIAMOND, ipaPy2, UniProt, PathIntegrate, OrthoFinder, eggNOG-mapper**, Volcano plots, **KEGG enrichment**, single-sample pathway analysis (ssPA), ggplot2, KEGG maps, heatmaps, time-series expression plots, Gantt chart planning.

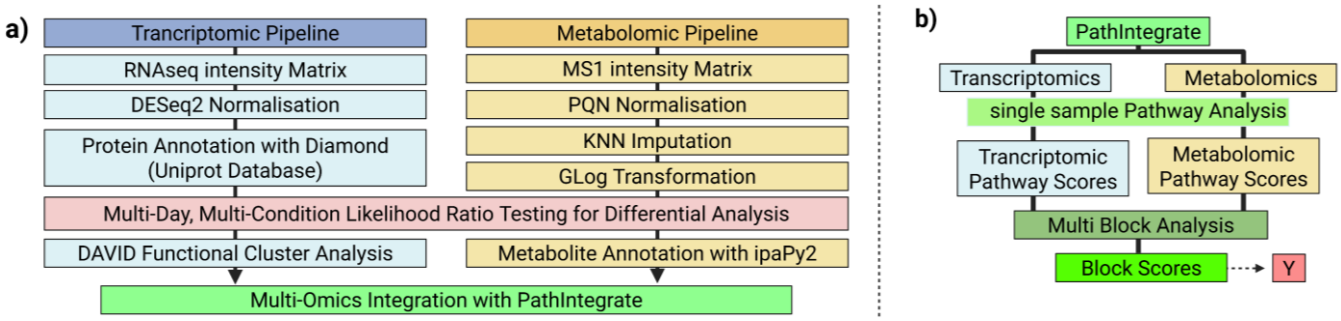


Figure 1. Multi-omics pipeline. a) Normalisation, annotation and differential analysis of transcriptomic and metabolomic data. b) Multi-omics integration using block scores to predict the response variable. Adapted from Wieder et al., (2024). Made with BioRender.

Summary:

Metabolomic MS1 and transcriptomic RNA-seq data were used in a multi-omics approach to identify inhibitory interactions between strains of *Streptomyces*. The project focused on **secondary metabolite production** and pathways that have potential for use in bio-inoculation strategies.

My Roles:

- **Designed** and executed full **pipelines**.
- Performed appropriate **processing, normalisation, imputation,** and scaling techniques, followed by multi-condition statistical modelling (LRT).
- **Annotated metabolites (ipaPy2) and proteins (Diamond).**
- **Curated custom pathway database** and integrated data using *PathIntegrate*.
- Conducted biological interpretation.
- **Managed** project scope, timeline and deliverables over 12 months.
- Developed **presentations, posters and the final report**, written and presented to academic panels (Viva)

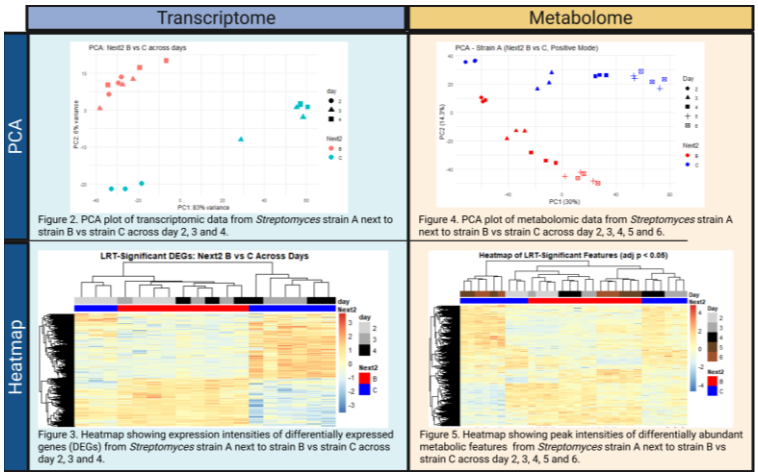


Figure 2. PCAs and heatmaps using transcriptomic and metabolomic data from strain A next to strain B vs strain C across days.

https://github.com/samuel-rice-bioinfo/Multi-Omics-Streptomyces

Machine Learning Applications in Bioinformatics

Supervised Learning – Breast Cancer Classification with K-Nearest Neighbours (k-NN)

Tools and Techniques:

- Python, **scikit-learn**, k-NN, Normalisation, Cross-Validation

Summary:

A k-NN model was trained to classify tumour types (benign vs. malignant). Normalisation techniques and cross-validation tuned the model, achieving high classification accuracy and identifying patterns in **30 numerical features from biopsy images**.

My Roles:

- Split datasets into training and testing sets.
- Compared **Logistic Regression (LR), Linear Discriminant Analysis (LDA)** and **Support Vector Machines (SVM)** methods.
- Applied Z-score normalisation and **optimised model parameters** (k-values).
- Created confusion matrices, elbow plots, accuracy vs k plots and ROC curves in Python.

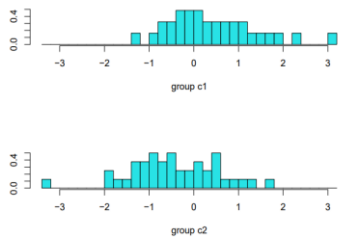


Figure 3. Base DLA structure.

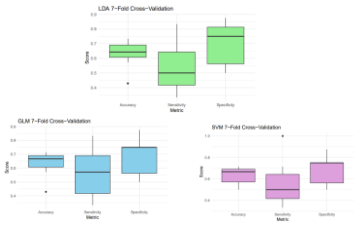


Figure 4. Generalised Linear Model (GLM), LDA and SVM 7-fold cross validation

Unsupervised Learning - Feature Clustering and Dimensionality Reduction in Tumour

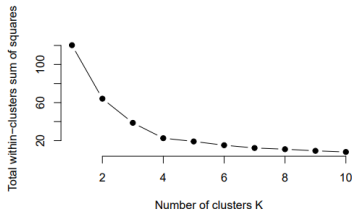


Figure 5. The elbow method score graph for cluster number optimisation.

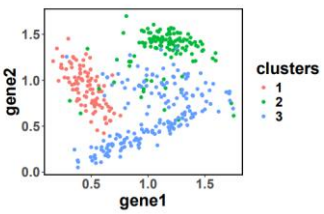


Figure 6. The elbow method score graph for cluster number optimisation.

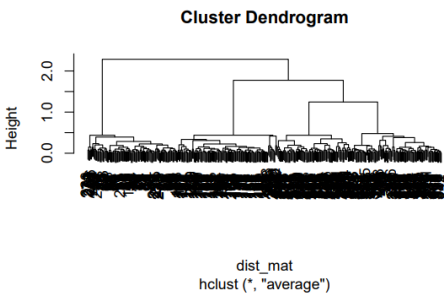


Figure 7. Clustered dendrogram using average linkage.

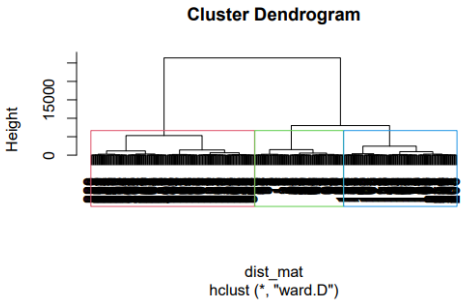


Figure 8. Clustered dendrogram using ward D linkage.

Tools and Techniques:

- PCA, Hierarchical Clustering, Agglomerative Clustering, Dendrograms

Summary:

- The project aimed to explore natural groups in the tumour dataset without feature labels.
- **PCA** was used to **reduce dimensionality**, and **agglomerative clustering** revealed hierarchical relationships across samples.

My Roles: Applied PCA, constructed dendrograms and compared clustering outputs against known tumour classes for biological validation.

https://github.com/samuel-rice-bioinfo/Tumour-Classification-Machine-Learning

Transcriptomics and Biomarker Discovery

PCA, Network Analysis and Biomarker Selection in Transcriptomics

Tools and Techniques:

- R, DESeq2, PCA, Cytoscape, DAVID, ggplot2

Summary:

This project involved the identification of gene expression biomarkers, differentiating cancerous and healthy samples using transcriptomic data.

My Roles:

- Implementing **study design**, differential expression analysis was performed with DESeq2, yielding candidate biomarkers.
- Principal Coordinate Analysis (PCA) highlighted biological variance and batch effects.
- **Enrichment analysis** and **network clustering** with DAVID and Cytoscape revealed functional gene modules associated with these biomarkers.

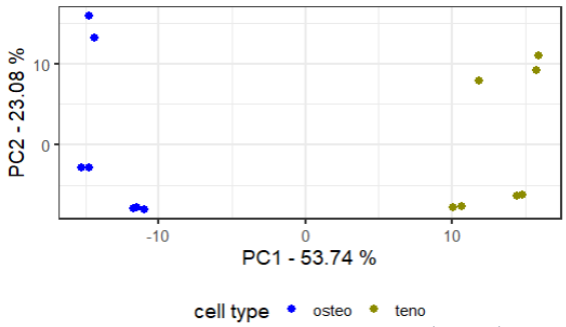


Figure 9. PCA plot of osteosarcoma (osteo) and tenosynovial (teno) giant cell tumor (TGCT).

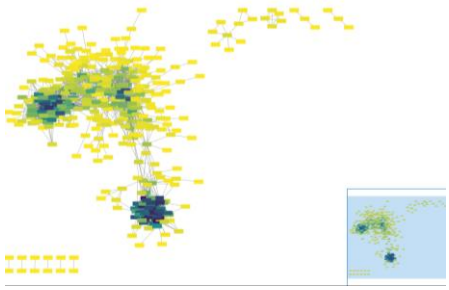


Figure 10. Cytoscape network clustering using a correlation network in osteo and teno cancer.

Uncovering Disease-Associated Mutations with SNP Variant Calling and Gene Expression in FKHR

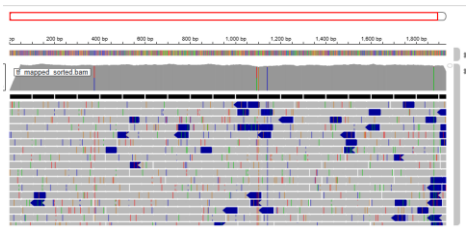


Figure 11. Integrative Genomics Viewer (IGV) visualising SNPs identified in the forkhead protein (FKHR) mRNA transcription factor coding sequence.

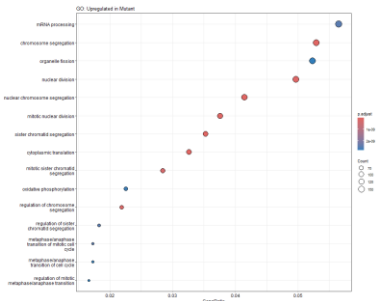


Figure 12. GO enrichment dot plot for upregulated genes in the mutant samples.

Tools and Techniques:

- **BWA-MEM, SAMtools, FreeBayes**, DESeq2, Interpro

Summary:

- Integrating genomics and transcriptomics, this study **identified a functional SNP** in the KFHR transcription factor.
- Differential expression analysis of RNA-seq data revealed over **7000 DE genes** across mutant and wild-type samples.
- **Gene ontology (GO) analysis** revealed mutation-associated autophagy and mitosis disruption

My Roles: **Novel Single Nucleotide Polymorphism (SNP) variant detection** with downstream gene expression analysis, interpreting biological consequences via GO enrichment.

https://github.com/samuel-rice-bioinfo/Biomarker-Discovery

Transcriptomics and Biomarker Discovery

Single-cell Transcriptomics in Glioblastoma

Tools and Techniques:

- R, **Seurat**, PCA, STRING database

Summary:

- This project used Seurat to analyse **single-cell RNA-seq data from glioblastoma** for the identification of distinct cell clusters.
- Significant markers and dimensionality reduction revealed five major clusters, with functional predictions explored using STRING to analyse gene associations.

My Roles:

- Identification of 480-725 markers per cluster, highlighting the top DE genes.
- Gene interactions were explored with **STRING**, creating **dimensionality reduction plots** and STRING interaction diagrams.

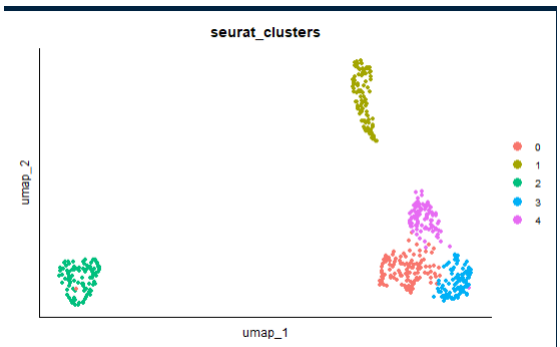


Figure 13. PCA plot of osteosarcoma (osteo) and tenosynovial (teno) giant cell tumor (TGCT).

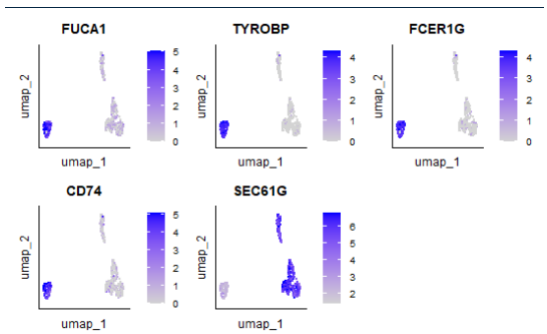


Figure 14. Cytoscape network clustering using a correlation network in osteo and teno cancer.

Spatial Transcriptomics of Cancer Subtypes

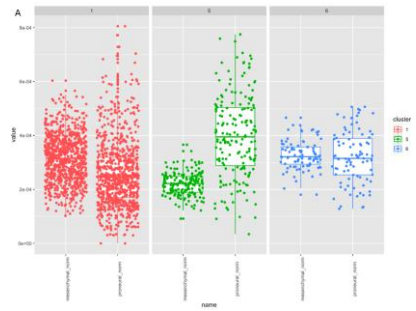


Figure 15. Spatial transcriptomics analysis comparing the subtype scores of 3 clusters between mesenchymal and proneural glioblastoma.

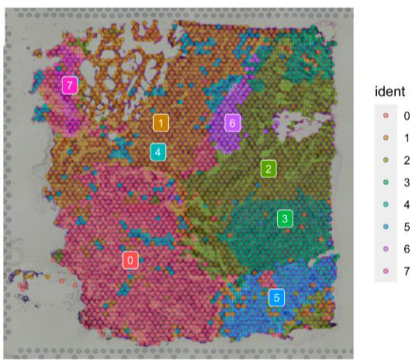


Figure 16. Map of each cluster's position on the tissue section.

Tools and Techniques:

- Spatial clustering, subtype scoring, literature integration.

Summary:

- Exploration of subtype expression in glioblastoma with spatial transcriptomics.
- Cluster five was identified as spatially distinct and biologically significant for differentiating proneural and mesenchymal subtypes.

My Roles:

- **Clustering** the **subtypes** with a comprehensive literature review to support the phenotypic differences, emphasising the **molecular stratification for therapeutic targeting.**
- Identified cancer subtypes using real biomedical data

Genomics and Phylogenetics

Calculating Polygenic Risk Scores for Blood Pressure using GWAS data

Tools and Techniques:

- **PLINK, LD clumping, PRS modelling**, awk

Summary:

- This project utilised genome-wide association study (GWAS) data to compute Polygenic Risk Scores (PRS) for blood pressure regulation.
- **PLINK** was used to **apply multiple p-value thresholds** and **linkage disequilibrium (LD) clumping** to assess the **genetic predisposition** to hypertension.

My Roles:

- Cleaning and processing **GWAS** datasets.
- Computing **multiple PRS models** with statistical thresholds.
- Analysing risk score distributions for interpretability and exploring PRS modelling for personalised **medicine applications**.

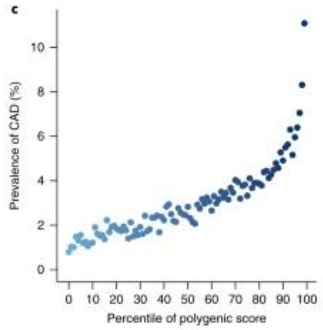


Figure 17. Prevalence of CAD amongst individuals with scores at various PRS percentiles

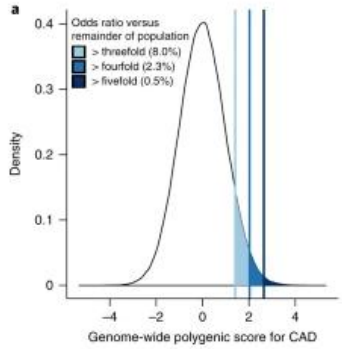


Figure 18. Odds ratio versus remainder of the population on coronary artery disease (CAD) risk.

Phylogenetic Analysis of Avian Influenza A (H5N1)

Tools and Techniques:

- ClustalW, MEGA-X, Sequence Alignment, Maximum Likelihood trees.

Summary: Exploring the evolutionary genomics of the HA gene sequence in H5N1 influenza viruses from various avian and mammalian hosts to reveal zoonotic transmission potential.

My Roles:

- Sequence alignment with **ClustalW with phylogenetic tree construction (MEGA-X).**
- **Maximum likelihood methods** were used with **bootstrapping**.
- Cross-species viral clusters were identified with potential public health impact.

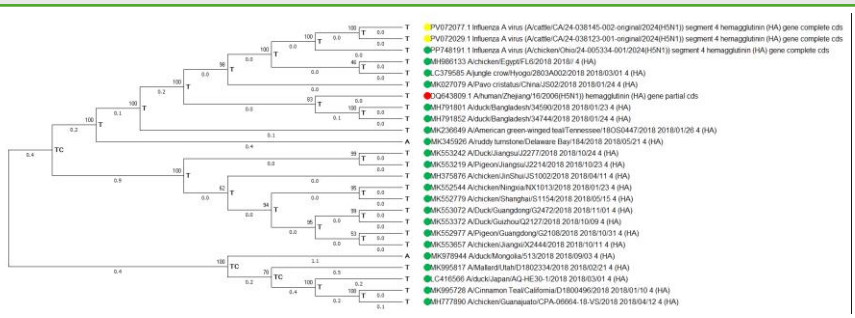


Figure 19. Maximum likelihood tree of HA gene in H5N1 influenza across various avian and mammalian hosts with bootstrap values and branch lengths.

Systems Biology and Metabolomic Modelling

Simulating Cellular Metabolism with Genome-Scale Metabolic Models (GSMMs)

Tools and Techniques:

- **COBRApy, Escher**, Flux Balance Analysis (FBA), Python

Summary:

This project constructed and analysed a genome-scale metabolic model (GSMM) of *E. coli* to simulate cellular metabolism under environmental and genetic perturbations.

My Roles:

- **Built** a **functional GSMM** using COBRApy in Python.
- Performed **FBA** to optimise biomass output, analysing the effects of gene knockouts on overall flux distribution.
- **Escher metabolic pathway maps** were used for interpretation and visualisation.

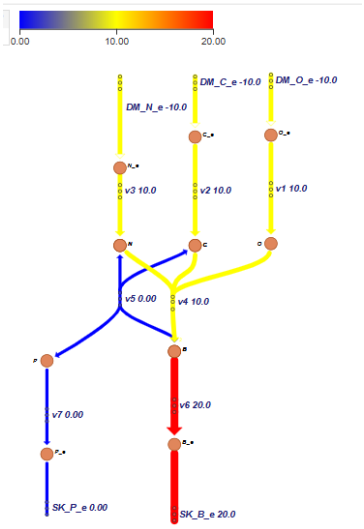


Figure 20. FBA of a metabolic model optimising v6 with the deletion of gene 5, visualised as an Escher map.

https://github.com/samuel-rice-bioinfo/Genome-Scale-Metabolic-Models

Metabolomic Profiling of Hypoxic Neuroblastoma

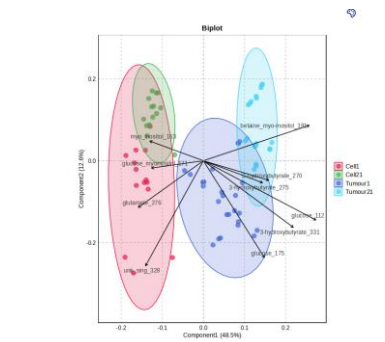


Figure 21. Biplot normal cells and cancer cells across hypoxic and normoxic conditions.

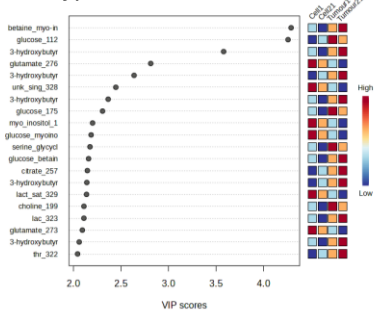


Figure 22. VIP plot of normal cells and cancer cells across hypoxic and normoxic conditions

https://github.com/samuel-rice-bioinfo/Normalisation-Methods-Metabolomics

Tools and Techniques:

- NMR Spectroscopy, **MetaboAnalyst**, Normalisation techniques, BioRender, **ANOVA, PCA, PLS-DA, VIP scores.**

Summary:

- Analysing the metabolic shifts in neuroblastoma cells under **hypoxic versus normoxic conditions** using NMR-based metabolomics.
- The chick embryo **CAM model** was utilised, with statistical analysis revealing hypoxia-induced alterations in key metabolites like choline and glutamate.

My Roles:

- Extracted and analysed metabolic profiles from tumour samples, applying and evaluating several advanced normalisation techniques.
- **Univariate** and **multivariate analyses** identified hypoxia-specific **biomarkers** relevant to cancer metabolism

Proteomics and PK/PD Modelling

Project 9: Phosphoproteomic Profiling of SARS-CoV-2 Infected Cells

Tools and Techniques:

- Mass Spectrometry (MS), Comet, Lorikeet, FDR Filtering, STRING Database, R, DAVID

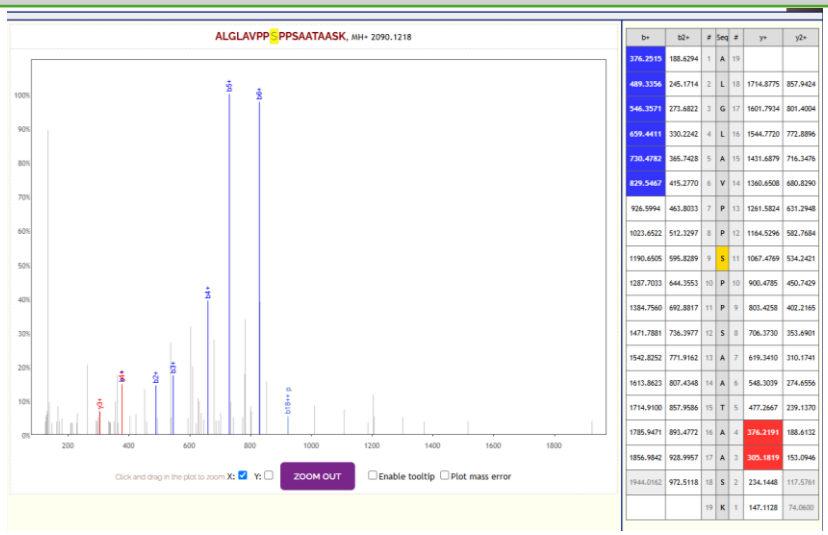


Figure 15. Lorikeet fragment ion spectra visualisation for the phosphorylation of ZN335_HUMAN (Q9H4Z2) protein peptide. The b+ and y+ ions are highlighted.

Summary:

- Investigating how SARS-CoV-2 infection alters cellular phosphorylation in human hosts.
- Using MS data processed with **Comet and visualised with Lorikeet**, high-confidence phophopeptides were identified across experimental conditions.
- Functional interpretation by **clustering phosphosites and STRING mapping** predicted vir impact on host signalling pathways.

National Institutes of Health
DAVID Bioinformatics

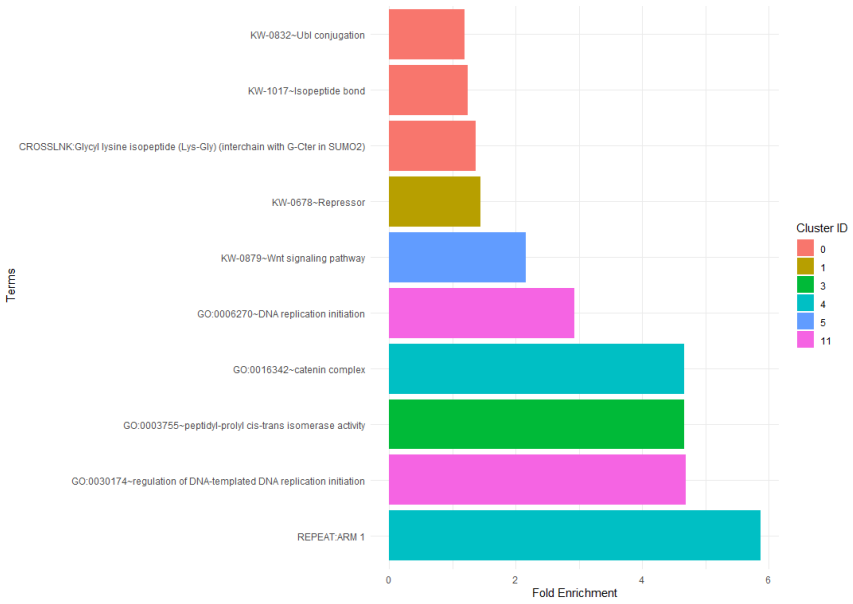


Figure 16. Bar plot of the top significantly enriched DAVID terms in SARS-CoV-2-infected cell lines.

My Roles:

- Filtered and annotated phosphoproteomic data, identifying statistically significant modifications associated with viral manipulation.
- Protein interaction and pathway visualisation to provide **insights into therapeutic targets.**

https://github.com/samuel-rice-bioinfo/Phosphoproteomic-Profiling-of-SARS-CoV-2-

PK/PD Modelling

Project 10: Pharmacokinetic and Tumour Growth Modelling

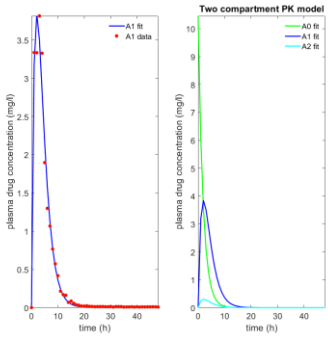


Figure 17. Two-compartment PK model fitting

Tools and Techniques:

- MATLAB, One-/Two-Compartment PK Models, Tumour Growth Simulation

Summary: Modelling drug distribution and tumour response using **pharmacokinetic (PK) and pharmacodynamic (PD) frameworks.** A tumour growth model was developed under drug treatment, evaluating the impact of dosing over time.

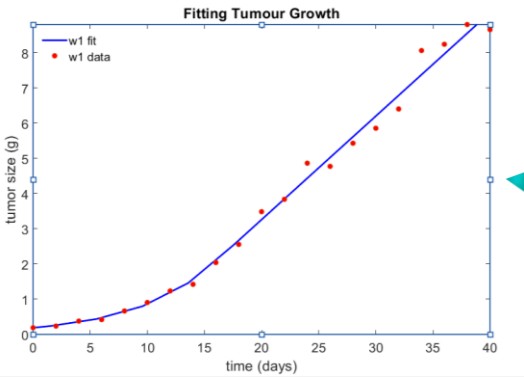
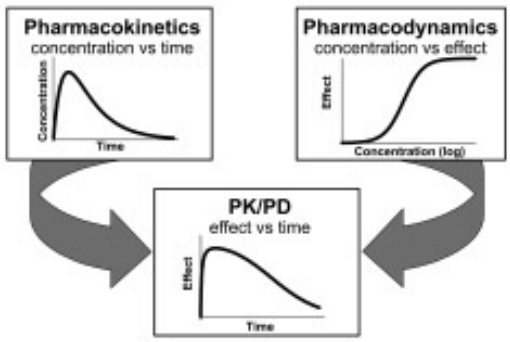


Figure 18. Tumour growth data (red dots) and the fitted growth model (blue line). The time is on the x-axis (days) and the tumour size is on the y-axis (g).

MATLAB®



My Roles:

- Built and analysed a one/two-compartment model in MATLAB, **simulating absorption, elimination, and inter-compartmental exchange.**
- **Utilised PK/PD models to compare behaviours** to support drug design strategy.

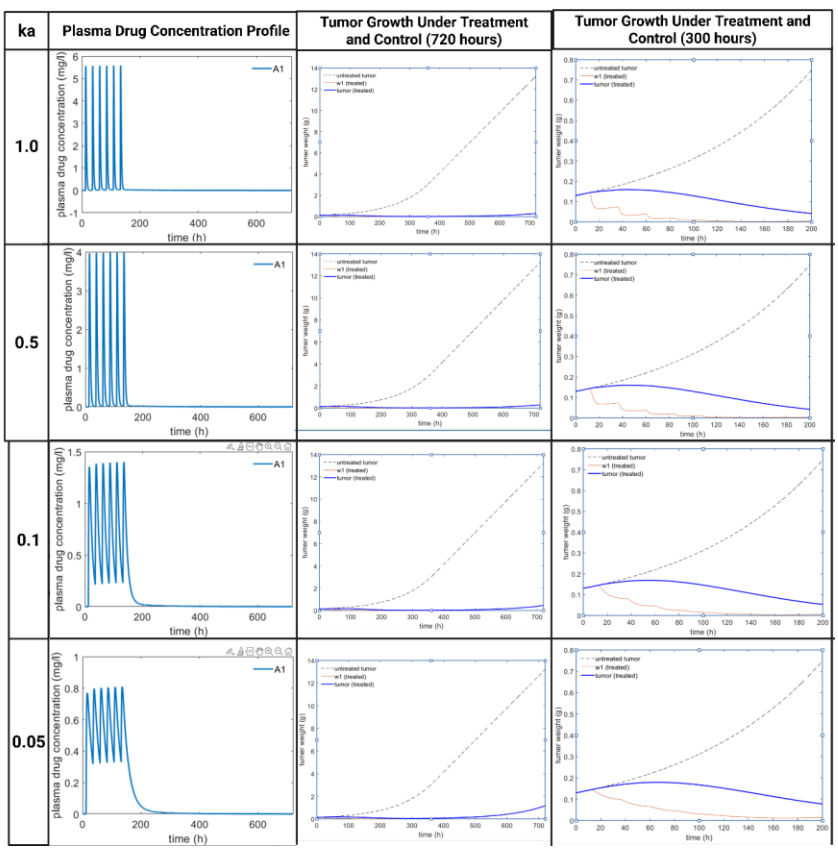


Figure 19. The effect of drug absorption rate (ka) on plasma drug concentration and tumour growth across 700 and 300 hours.

Genomics and Antimicrobial Resistance

Comparative genomic and structural analysis of antimicrobial resistance (AMR) genes across *Escherichia coli/Shigella* strains.

Tools and Techniques:

- Linux/BASH, Panaroo, RGI, Diamond-BLASTx, VFDB, InterPro, Clustal Omega, MEGA11, ColabFold, PyMOL

Summary:

- The project conducted a comprehensive **comparative genomic and structural analysis of AMR genes** across E. coli/Shigella clinical isolates.
- Pangenome analysis and **Resistance Gene Identifier (RGI)** identified conserved, strain-specific resistance elements.
- Efflux pump regulator AcrR was among the most common across strains; therefore, the **paralogs were modelled** for structural analysis of the mutations.

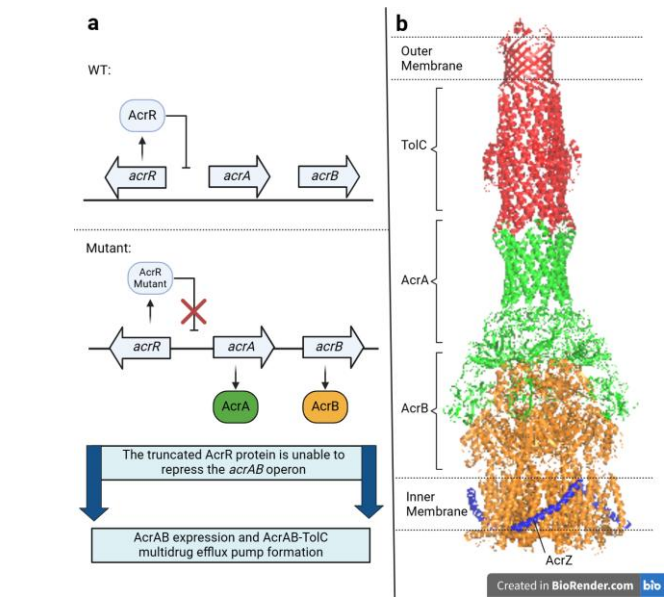
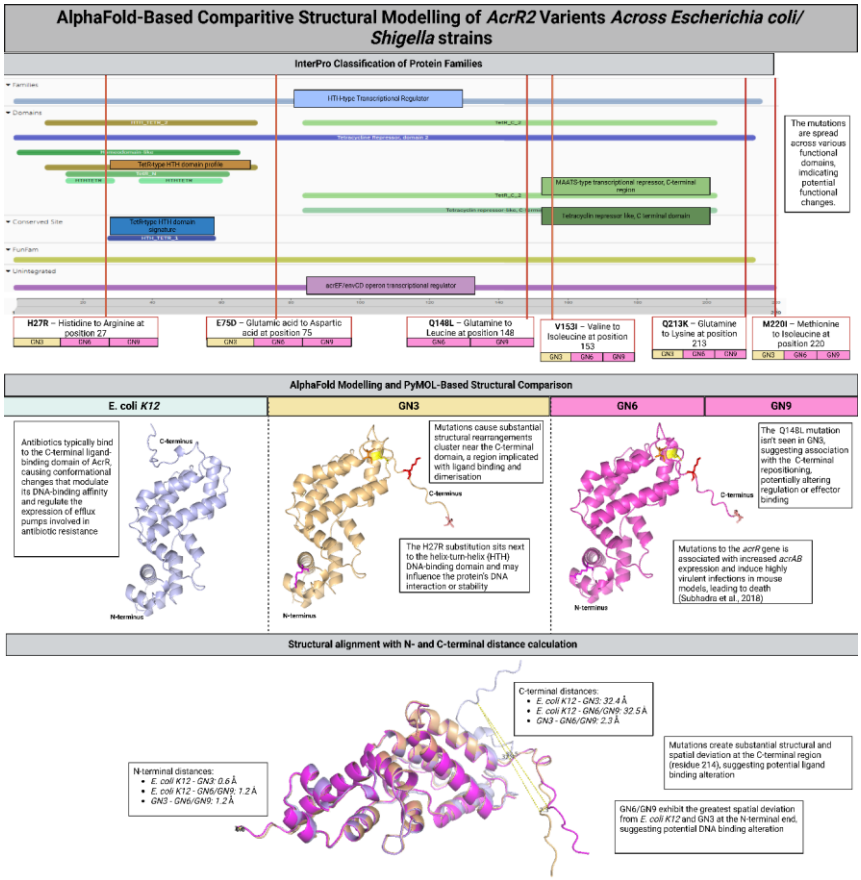


Figure 19. AcrAB-TolC multidrug efflux pump regulation and structure.



My Roles:

- Study design and evaluation.
- **Pangenome analysis**, identifying core and unique AMR gene distributions.
- Detected major resistance mechanisms with target modification and antibiotic inactivation.
- **Analysed AcrR paralog evolution and 3D structural divergence** with ColabFold and PyMOL.

https://github.com/samuel-rice-bioinfo/AMR-Analysis-E.coli-Shigella

Phage Biology and Non-Coding RNA Dynamics

The role of *Wolbachia* phage WO and Mobile Genetic Elements (MGEs) on the Horizontal Gene Transfer (HGT) of non-coding genetic elements

Tools and Techniques:

•NCBI BLASTn, BLASTp, tBLASTn, Rfam, RNAcentralGenomic, FASTA conversion, accession ID extraction.

•Genome annotation referencing, NCBI Conserved Domain Database (CDD), protein domain prediction, Clustal Omega, iTOL (Interactive Tree of Life)

•Pseudogene identification, Tree annotation, comparative gene structure mapping, sequence logos

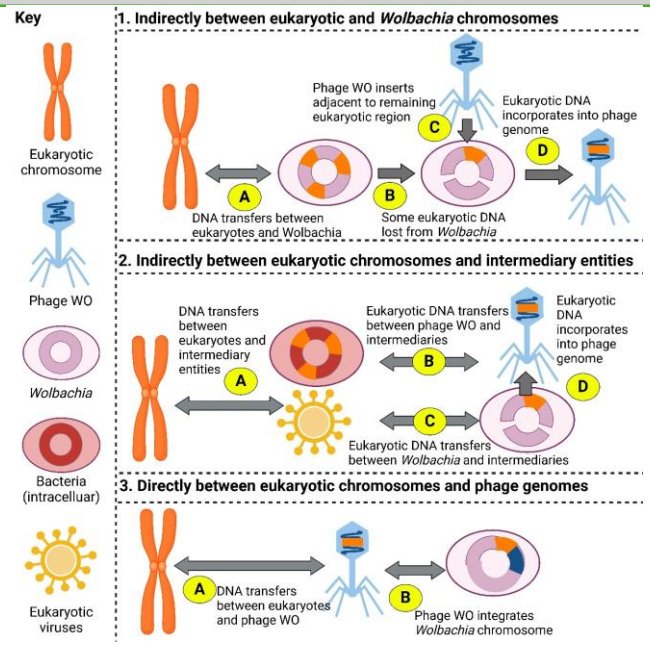


Figure X. *Modelling horizontal gene tranfer (HGT) between eukaryotes and bacteriophages (phage WO).*

Summary:

- The project explored the role of *Wolbachia* phage WO in the **HGT of non-coding RNA** elements across eukaryotic hosts.
- Conserved TSS-derived sequences were used, mapping their conservation and potential functional domains.
- There was a focus on **cytoplasmic incompatibility (CI)** and **mosquito vector control to reduce arboviral spread**.

My Roles:

- Processed experimental RNA data for genomic search with annotation.
- Conducted sequence alignment and conservation analysis via BLAST.
- **Built phylogenetic trees** and explored adjacent gene context
- **Identified conserved domains** and RNA families linked to phage-derived elements
- Interpreted data for evolutionary relevance and gene transfer potential.
- **Comprehensive literature search** to support findings about non-coding RNA

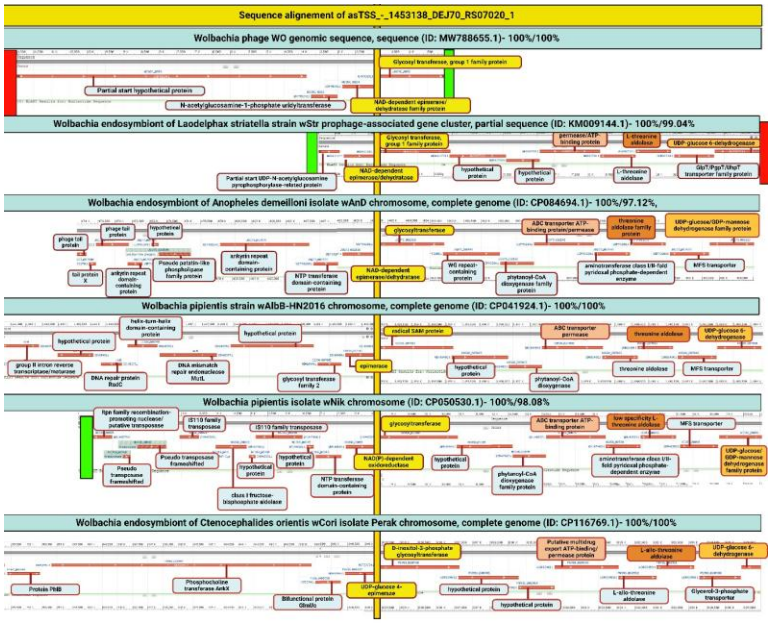


Figure 4. Sequence Alignment of a NAD-dependent epimerase/dehydratase family protein and Glycosyl transferase group 1 family protein-associated Transcription Start Site-Derived Sequence (TSS-DS) across Wolbachia strains and phage WO genomes

Personal Skills and Development

Leadership and Teamwork

Captain of my club's **rugby team for eight years** (age 10-18), developing early leadership skills. I learnt to motivate a team, manage setbacks and intragroup conflict, and lead by example.

Volunteering, Passion, and Independence

At the age of 18, I spent time **volunteering** at **animal rescue centres in Peru and Costa Rica**. This work developed my compassion, responsibility and ability to adapt in new environments. I led teams whilst facing language barriers and the threat of dangerous animals around people I was in care

Discipline, Responsibility, and Community: Brazilian Jiu Jitsu

Training **BJJ** at the world-renowned **Next Generation Liverpool** has instilled **focus, discipline and humility**. Competing and learning alongside others has sharpened my ability to stay strategic and calm in high-pressure situations.

Public Speaking, Scientific Communication, and Creativity

Interviewing professional boxer Johny Fisher live at a major event, privately and **in front of over 1000 people**, built my communication confidence and stage presence, which I now bring to academic presentations and discussions.

Presenting a vast array of projects throughout my academic history, including my MSc research project, where I **communicated complex scientific information clearly** and interpretably **under strict time constraints**. Designing the poster and numerous other figures using BioRender required strong creative refinement to make complex information digestible.