BOOK OF ABSTRACTS
ISCB-AFRICA ASBCB
CONFERENCE ON BIOINFORMATICS

14–17 April 2025

Cape Town, South Africa

ASBCB

ISCB INTERNATIONAL SOCIETY FOR COMPUTATIONAL BIOLOGY

ISCB Student COUNCIL

ABI African Bioinformatics Institute

H3ABioNet

Aggregation in guts: on the link between neurodegeneration and bacterial functional amyloids

Authors: Alicja Wojciechowska (Wroclaw University of Science and Technology), Jakub Wojciechowski (Sano Centre for Computational Medicine) , Kinga Zielinska (Malopolska Centre of Biotechnology, Jagiellonian University), Johannes Soeding (Max Planck Institute for Multidisciplinary Sciences), Tomasz Kosciolek (Sano Centre for Computational Medicine) and Malgorzata Kotulska (Wroclaw University of Technology)

Functional amyloids, expressed mostly by microorganisms, play important physiological roles in the host organisms, e.g. bacterial biofilm stabilization. Using bioinformatics approach, we identified a significantly diverse set of putative bacterial functional amyloids in seven bacterial phyla of human gut proteome and analyzed their potential impact on health via the gut-brain axis. We showed that microbiomes of Parkinson's disease (PD) patients across all publicly available studies contain more bacterial functional amyloids than their healthy controls and cerebral meningitis (CM) patients. Although the overall greater abundance of gut bacterial functional amyloids was not shown for Alzheimer's disease (AD) patients, greater abundance of the amyloid curli protein family was observed similarly as for PD patients. The structural similarity between functional bacterial amyloids and pathological human amyloids could enable molecular mimicry, leading to altered aggregation rates of pathological amyloids or the activation of the same molecular pathways. For example, curli could interact with alpha-synuclein, as well as with Abeta, in vivo and in vitro, and inhibition of curli aggregation reduces neuronal death. Moreover, curli amyloid precursor is recognized by the human immune system in the same manner as Abeta-42 peptide, due to their structural features. Additionally, the amyloid proteins found in our study are 20-fold more likely to be extracellular than other bacterial proteins, which potentially enable them for molecular interactions with human proteins. Accordingly, the interactions between bacterial functional amyloids and human proteins could have a broader scale. The bacterial functional amyloids could be one of the factors contributing to pro-inflammatory characteristics of certain phyla. They can interact with immune response proteins involved in chemokine signaling and leukocyte migration. The bacterial functional amyloids, similar to human pathogenic amyloids, seem to have the potential to affect cell junctions and the abundance of bacterial functional amyloids in the gut may influence the gut permeability, similarly as human pathological amyloids affect blood-brain barrier. The negative effect of the protein aggregation phenomena on epithelial cell integrity has been already observed. Following our analysis, such direct interactions may affect proteins responsible for endo- and exocytosis, signaling and cellular transport. Our results demonstrate that bacterial functional amyloids are other important factors in the gut-brain axis.

Early Biomarkers of Long Covid (PASC)

Author: Malika Aid Boudries (Harvard School of Medicine, CVVR)

Long Covid, or Post-Acute Sequelae of COVID-19 (PASC), encompasses a range of chronic symptoms that persist after acute SARS-CoV-2 infection. The proposed mechanisms underlying Long Covid include factors such as persistent viral presence, reactivation of latent viruses, tissue damage, immune system dysregulation, and inflammatory responses. In a study involving 142 participants—spanning uninfected controls, acutely infected individuals, convalescent controls, and Long Covid patients—integrative bioinformatics and machine learning analyses were applied to immunologic, virologic, transcriptomic, and proteomic data. This approach revealed that Long Covid patients exhibited chronic immune activation, upregulated proinflammatory pathways (e.g., JAK-STAT and IL-6), and metabolic and T cell exhaustion signatures, differentiating them from convalescent patients six months post-infection. Advanced computational tools, particularly machine learning models, played a pivotal role in identifying these biological patterns, offering insights into biomarkers such as plasma IL-6R levels for Long Covid diagnosis. By leveraging these tools, we identified novel therapeutic targets to treat long Covid, such as JAK inhibitors, which are now under clinical investigation (NCT06597396). This integration of bioinformatics and machine learning not only accelerates discovery but also refines our understanding of complex disease mechanisms, underscoring their importance in shaping future medical research.

Computational algorithms to identify mechanism-centric biomarkers of treatment response in cancer

Author: Antonina Mitrofanova (Rutgers University, Rutgers Health, SHP)

We have developed a novel computational algorithm TR-2-PATH that reconstructs first-of-its kind mechanism-centric regulatory network, which connects molecular pathways to their upstream transcriptional regulatory programs, and prioritizes them as markers of therapeutic resistance in cancer. Such network offers a novel way to identify biomarkers that are mechanisms-centric, rather than based on individual genes or alterations - a new way to identify functional interactions and valuable therapeutic targets. As a proof of concept, we have applied TR-2-PATH to metastatic castration-resistant prostate cancer (mCRPC). Network mining step addressed a knowledge gap of multi-collinearity among upstream transcriptional regulators (TRs) and identified TR groups that collaborate to regulate downstream pathways. Interrogating this network with signatures of resistance to Enzalutamide, a second-generation androgen-deprivation drug commonly administered to mCRPC, identified a collaboration between NME2 TR program and MYC molecular pathways as a biomarker of primary resistance to Enzalutamide. In vitro and in vivo experimental validation confirmed cooperation of these mechanisms and demonstrated that their joined therapeutic targeting is not only effective to prevent resistance to Enzalutamide, but also re-sensitizes Enzalutamide resistant tumors in vivo, allowing Enzalutamide to work longer. We propose to use MYC and NME2 as markers to identify patients at risk of Enzalutamide resistance and as effective therapeutic targets for patients that failed Enzalutamide. Our novel algorithm is generalizable and could be applied to study a multitude of biologically and clinically important questions, including (but not limited to) therapeutic resistance, metastatic progression, tumor heterogeneity and plasticity across cancer types and in other diseases. TR-2-PATH was published in Nature Communications in 2024. We are now expanding this algorithm to include regulatory relationships with long non-coding RNAs.

Advancing gene function and disease mechanism insight through a containerized multi-omics integration tool

Authors: Kimberly Coetzer, Gian van der Spuy and Gerard Tromp
(Department of Biomedical Sciences, Biomedical Research Institute, Faculty of Medicine and Health Sciences)

Dissecting the biological mechanisms of disease will benefit from more than a single layer of "omics" data. Multi-omics is a discipline that has been developed to find methods and approaches to integrate various omics layers and provide greater insight into gene function and disease mechanisms. Integrating data from domains that operate at different time scales allows researchers to elucidate the complex relationship between genetic information (genotype) and observable characteristics (phenotype). The interactions and functions of biomolecules across multiple levels provide a more thorough understanding of the complexity of biological systems and diseases than a single-omics approach. While multi-omics data integration shows potential for improving our knowledge of complex biological systems, it faces challenges such as different data types, data variety, high complexity, and intricate interactions, all of which affect result interpretation. To address these issues, powerful computational methods must be employed. This necessitates robust frameworks that not only standardize data processing but also streamline the interpretation of multi-omics analyses, ensuring the biological relevance and utility of the findings. This project therefore aims to develop a containerized tool for multi-omics integration to uncover gene function and disease mechanisms in heritable disorders after identifying putative causative or contributing variants. The tool will be created and validated in Nextflow, a workflow management system that facilitates the integration of various data types. Python and R scripts will be employed in the development process to ensure robustness and accuracy. Several nf-core pipelines will be combined to prepare genomic, transcriptomic, and proteomic data for integration. These results will then be integrated using various methods for homology modelling, pathway and network analysis and functional annotation. The tool will undergo testing on a disease model using publicly available multi-omics data from Amyotrophic Lateral Sclerosis (ALS) patients, a widely prevalent neurodegenerative condition. ALS impacts the anterior horn motor neurons in the spinal cord and the pyramidal cells of the motor cortex, resulting in progressive degeneration and loss of motor function. Despite extensive investigation, the complete spectrum of genetic and molecular factors affecting its pathogenesis remains unclear. The utilization of ALS as a model seeks to deepen our understanding of neurodegenerative diseases, offering essential insights that could improve the knowledge and treatment of a wide array of genetic disorders, extending beyond neurodegenerative diseases alone. Moreover, outside this project's scope, the tool's potential use in various genetic disorders holds the promise of revealing new insights into their disease mechanisms, thus significantly contributing to the advancement of personalized treatment strategies for diverse genetic conditions.

A knowledge embedded deep learning framework for multi-parameter cytometry data

Elijah Willie, Ellis Patrick and Helen McGuire
(University of Sydney)

Multi-parameter cytometry technologies enable high-dimensional analysis of immune cell populations at single-cell resolution. Deep learning has emerged as a transformative tool for analyzing these datasets. Still, existing methods often struggle with transferability across datasets due to technical variability and batch effects, limiting their utility in clinical research. We present dioscRi, a transferable deep learning framework that integrates a Maximum Mean Discrepancy Variational Autoencoder (MMD-VAE) for normalization and denoising, enhancing cross-dataset compatibility. DioscRi incorporates both empirical and biologically derived cell type hierarchies to group cell type proportions and marker means, enabling a richer interpretation of immune cell features. These hierarchies capture the relationships between cell types, leveraging their inherent structure to improve predictive performance and interpretability. By combining these features with an overlapping group lasso model, dioscRi identifies associations between immune profiles and clinical outcomes, offering robust predictions while uncovering biologically meaningful insights. DioscRi validated known immune associations, such as those involving CD4+ regulatory T cells (Tregs) and plasmacytoid dendritic cells (pDCs) in coronary artery disease, while uncovering novel marker-level associations. Benchmarking across multiple datasets demonstrated dioscRi's ability to generalize and outperform existing methods, establishing it as a versatile and interpretable tool for cytometry data analysis.

Augmented kurtosis-based projection pursuit: a novel, advanced machine learning approach for multi-omics data analysis and integration

Tobias Karakach (Dalhousie University, Faculty of Medicine, Department of Pharmacology) , Fabian Bong (Dalhousie University, Faculty of Computer Science) , Nithya Ramakrishnan Institute of Bioinformatics and Applied Biotechnology (IBAB)), Karla Valenzuela (Dalhousie University, Faculty of Medicine, Department of Pharmacology) , Peter Wentzell (Dalhousie University, Faculty of Science, Department of Chemistry) and Jasmine Barra (Dalhousie University, Faculty of Medicine, Department of Microbiology and Immunology)

Due to the heterogeneity of multi-omics data, obtaining information from them remains a challenge. Whereas some solutions have been offered, most cannot overcome the large linear dynamic range associated with these data, while others require large biological effect sizes to produce meaningful models. Here, we (a) perform a comprehensive benchmarking of multi-omics data analysis tools, and (b) introduce kurtosis-based projection pursuit analysis, augmented with classification and regression trees (kPPA-CART) as a robust, easy-to-implement approach to model multi-omics data that are derived from next generation sequencing (NGS) and mass spectrometry (MS). Most of the available methods for unsupervised multi-omics integration suffer from an inability to model low-intensity (low count) features and instead focus on high variable (dominant) ones. While low-count features, such as genes involved in signaling, and non-coding RNAs (ncRNA) are associated with high analytical uncertainty, they exhibit significant biological impact upon perturbation. Methodologically, kPPA is an “unsupervised” data exploration approach that finds patterns in input data without a priori knowledge of class membership. The output of kPPA is projections of the original samples into “interesting” directions, which, when plotted against each other, show clustering of (dis)similar samples. We augment kPPA’s clustering with classification and regression trees (CART), which takes cluster identities derived from k-means classification as input to perform a quasi-supervised classification and decipher feature importance. Using ground truth data, we demonstrate that kPPA-CART exhibits superiority in inferring biological significance from low-intensity features. Moreover, when effect sizes (expected biological differences between conditions) are small, we show that kPPA-CART can recover important biological information better than available approaches. To provide biological context, we have re-analyzed prominent Breast Cancer (BC) data from The Cancer Genome Atlas (TCGA) and show that kPPA-CART identifies novel gene transcripts that provide a classification of BC into Basal, Her2, Luminal A, and Luminal B subclusters better than the original PAM50 panel. We validate these genes with an external set of data and show that the top kPPA-CART panel of genes is associated with poor overall survival for patients with BC for whom these genes are dysregulated. Finally, we provide an R package and an online implementation of kPPA-CART.

Graph Neural Network-Enhanced Multi-Task Learning for Menstrual Disorder Prediction Using Multi-Omics Data

Comfort Adesina (Federal University of Agriculture, Abeokuta)

It is already established that GNNs integrate multi-omics networks for functional genomics. However, the application of ML research on women's health is not discussed enough. The complex etiology of polycystic ovary syndrome (PCOS) presents challenges for early detection and personalized treatment. Current diagnostic methods lack the integration of biological and molecular data, limiting the ability to identify precise biomarkers and treatment pathways. Traditional machine learning models are limited in their ability to capture the complex, non-linear relationships present in such biological systems. Recently, graph neural networks (GNNs) have emerged as a powerful tool to model graph-structured data, including protein-protein interaction (PPI) networks, enabling the integration of both network topology and molecular data to improve prediction performance (Kipf & Welling, 2017). In this work, we develop a multi-task learning framework based on GNNs, designed to achieve two key objectives: a) to classify the presence of PCOS, and largely a multi classification system of menstrual disorders using PPI, and transcriptomic data (RNA Highseq), and b) to predict gene expression levels in a regression framework. To address this challenge, we employed graph convolutional layers to propagate node features through the PPI network here, enabling us to capture both local and global interactions (Szklarczyk et al., 2019). Node features included gene expression data derived from transcriptomic datasets, as well as centrality measures from the PPI network. The classification task was optimized using binary cross-entropy, while the regression task was optimized using mean squared error, leveraging multi-task learning to improve model generalization. Our approach significantly outperformed baseline methods that separately handle PPI and transcriptomics, highlighting the advantage of GNNs in multi-omics integration for predictive modeling (Edgar et al., 2002). Future work would include expanding the model to incorporate additional omics layers, such as proteomics and metabolomics, and adapting it to multi-class classification to differentiate between multiple menstrual disorders. This multi-omics, graph-based framework offers an advanced methodology for understanding the molecular drivers of PCOS, providing insights that could inform personalized medicine strategies.

Developing machine learning models for predicting cytochrome P450 ligand potency

Blessing Sitabule, Scott Hazelhurst and Houcemeddine Othman (University of the Witwatersrand)

Cytochrome P450 (CYP P450) enzymes are involved in over 90% of metabolic reactions that are known. To understand the interaction between CYP P450 enzymes and their ligands, a variety of experiments have been performed including in silico methods such as machine learning (ML) predictions. ML models can predict the potential interactions of enzymes with ligands of interest. In this study, we developed ML models based on several algorithms including the Support Vector Machine, Extreme Gradient Boosting (XGBoost), Random Forest, Neural-network and K-Nearest Neighbor using the Centre for High Performance Computing Lengau cluster. The models serve to predict the potency of ligands with respect to CYP P450 enzyme sequences. To develop the models, over 30 000 enzyme assay data was acquired from BindingDB. From the data, CYP P450 enzyme sequences were encoded using AAindex indices and the ligands were encoded using molecular descriptors, which were generated using RDkit and Mordred. In addition, the IC50 values of the enzyme assays were also encoded. Dimensionality reduction was performed on the enzyme and ligand features prior to training the models. The performance of the models was assessed using accuracy scores and the Area Under the Curve (AUC) of the Receiver of Operating Characteristic curves. The best performing algorithm was XGBoost with an accuracy of 0.8 and an AUC score of 0.9. This indicates that machine learning can be used to predict the potency of ligands with respect to enzymes of interest. Such models can potentially be employed for drug discovery, protein engineering and pharmacogenetics.

Development of a machine learning model for prediction of HIV virological failure in a resource-limited setting

Maria Magdalene Namaganda (Makerere University)

Background: Virological failure (VF) presents a significant challenge in HIV treatment, particularly in low and middle-income countries (LMICs) such as Uganda, where it threatens the long-term efficacy of antiretroviral therapy (ART) and increases the risk of drug resistance and adverse clinical outcomes. Early identification of treatment failure is crucial for guiding timely interventions that prevent the progression of resistance and optimise patient outcomes. Nevertheless, advanced methods such as genotypic antiretroviral resistance testing remain largely inaccessible in LMICs due to high costs, complexity and limited availability, further widening healthcare disparities. In parallel, vast amounts of routinely collected clinical and demographic data remain underutilised, despite their potential to improve predictive modeling for VF and inform personalised treatment strategies. Methods: This ongoing study involves the analysis of de-identified clinical data from a Ugandan cohort spanning from 2014 to 2024. The approach includes thorough data preprocessing, exploratory data analysis, and feature engineering to identify key predictors of VF. Supervised machine learning techniques, such as logistic regression, decision trees, and ensemble methods, will be applied to construct and validate the predictive models. The performance of these models will be evaluated using metrics such as accuracy, sensitivity, and specificity. Expected Outcome: The study aims to deliver a robust predictive tool capable of identifying patients at heightened risk of treatment failure before clinical deterioration occurs. The fitted machine learning models hold potential to revolutionise HIV care by providing clinicians with individualised risk profiles, thereby facilitating more informed treatment decisions. Overall aimed at reduction in the incidence of virological failure, minimise the emergence of drug resistance, and ultimately enhance the health outcomes for people living with HIV. Additionally, this research aligns with broader initiatives in precision medicine and data-driven healthcare in resource-constrained settings.

ESTROGEN RECEPTOR MEDIATED NEUROPROTECTION IN MODELS OF ALZHEIMER'S DISEASE

Heba Ali, Ivan Nalvarte, Mukesh Varshney and Per Nilsson
(Karolinska institutet)

Objectives Middle-aged women are 2-3 times more likely than men to develop Alzheimer's disease (AD) later in life. While women tend to live longer, factors like sex hormones, genetics, and environment heighten their AD risk. Understanding sex differences in AD has been complex. This project aims to uncover the molecular underpinnings of the female sex hormone estrogen in AD models and relate these findings to human disease. Methods We are using APPNLGF mice, which exhibit clear AD pathology by 6 months of age, to study the effects of biological sex and sex hormones on memory and AD pathology at 6 and 12 months. Surgical menopause is induced in young adult female mice and castration in male mice. Preliminary data indicate neuroprotective effects of estrogen receptor beta (ERβ) in APPNLGF mice. Consequently, we crossed ERβ-/- mice with APP-KI mice to study the effects of ERβ loss on AD pathology, utilizing single-cell RNA sequencing of hippocampal brain cell populations. Results Our ongoing analysis provides insights into ERβ's role in brain cell functions and identifies sex-dependent alterations in gene activity. Preliminary data show that ERβ has neuroprotective effects in APPNLGF mice. We will anchor our findings to human disease by studying gene and pathway regulations in human AD brains and analyzing GWAS data from women with or without AD diagnosis and with or without different menopausal hormonal treatments. Conclusions Collectively, these findings offer insights into mechanisms underlying sex-specific susceptibility to AD and identify regulatory proteins for potential treatments targeting sex-dependent AD pathology. Understanding the role of estrogen and ERβ in AD could lead to targeted therapies addressing the heightened risk in women.

Deepath: Deep learning for pathogenicity prediction from next generation sequencing

Salem A. El-Aarag (Bioinformatics Department, Genetic Engineering and Biotechnology Research Institute (GEBRI), University of Sadat City), Mohamed E Hasan (Bioinformatics Department, Genetic Engineering and Biotechnology Research Institute (GEBRI), University of Sadat City), Alaa E. Hemeida (Bioinformatics Department, Genetic Engineering and Biotechnology Research Institute (GEBRI), University of Sadat City), Mario Flores (Department of Biomedical Engineering, University of Texas at San Antonios) and Mahmoud Elhefnawi (Biomedical informatics and chemoinformatics group, Informatics and systems department, National Research Center)

In this presentation, I will review the recent works on pathogenicity prediction from NGS data and will highlight our current contribution to the field by developing DeePath, a Deep Learning model for pathogenicity prediction. Existing approaches for pathogenicity prediction can be categorized as protein content based or read based. Read-based approaches are faster and more applicable to metagenomic analysis. Machine learning methods, in contrast with taxonomic methods, can identify novel pathogens. Bartoszewicz et al. (2019) presented DeePaC and applied reverse-complement convolutional neural networks and LSTMs. Deneke et al. (2017) used a random forest approach for predicting pathogenicity potential from an Illumina read. Here, we tried to develop deep learning-based approach to accurately detect pathogenic phenotype from next-generation reads. We used a list of pathogenic and nonpathogenic bacteria retrieved from Integrated Microbial Genome and Mircrobiomes by Deneke et al. (2017). Non-pathogenic strains of well-known pathogenic species were discarded from further analysis. One strain per species is included. This resulted in a list of 446 species (342 pathogens and 67 non-pathogens). We simulated about 2.5 million paired-end Illumina reads per class using InSilicoSeq. Read length was set to 301 bases. We deleted shared reads between the two classes. One hot-encoding was used to represent DNA sequences. The final list is divided into 90% training, 5% validation, 5% test sets. We implemented convolutional neural networks (CNN) in our model achieving 0.919 accuracy, 0.977 AUC, 0.943 Recall, and 0.895 precision in test set.

GEMINI: A Breakthrough System for Robust Gene Regulatory Network Discovery, Enabling the Application of GRNs to Industrial Level Genetic Engineering

Ridhi Gutta (Academies of Loudoun)

In order to resolve crucial global issues, the widespread application of genetic engineering at an industrial level is key. Effective genetic engineering at an industrial scale hinges heavily on precise cellular control of the microorganism at hand. However, the majority of synthetically engineered strains fail at the industrial level due to disruptions in gene regulation. This stems from a lack of understanding and usage of gene regulatory networks (GRNs), which control cellular processes and metabolism. Research shows that effective manipulation of host GRNs and effective introduction of synthetic GRNs can improve product yield and functionality significantly. However, current GRN inference tools are extremely slow, inaccurate, and incompatible with industrial scale processes, because of which there are no complete expression based GRNs for any commonly used organism, limiting the application of GRNs as a practical tool in genetic engineering at the industrial level. This research proposes a novel computational system, GEMINI, to enable fast and efficient GRN inference for integration into industrial scale pipelines. GEMINI consists of two main parts. First, I create a novel information theoretic algorithm that replaces traditional sequential inference and calculation methods, ensuring compatibility with parallel processing. Second, I integrate a novel GNN architecture based on spectral convolution to bypass intensive eigenvalue computation and efficiently learn global and local regulatory structures. On the DREAM4 and DREAM5 in silico benchmarks, GEMINI outperforms all industry leaders in terms of AUROC and AUPRC, achieving a nearly 300% increase in AUPRC compared to the industry leading method, GENIE3. When applied on a real biological E. coli dataset, GEMINI not only recovered 98% of existing interactions, but discovered 468 novel candidate interactions, which were validated against literature. Thus, GEMINI was able to construct the most complete expression based GRN of E. coli to date, providing a novel biological blueprint for genetic engineers to use at the industrial level. GEMINI removes reliance on expensive computing equipment and enables fast and accurate GRN inference for the first time, opening doors to more efficient gene expression control and metabolic pathway manipulation for more effective application of genetic engineering at an industrial level.

Development and Validation of a Machine Learning Model for Identifying Novel HIV Integrase Inhibitors

Blessed Mukuhlani (University of Zimbabwe)

HIV integrase (IN) is a critical enzyme in the viral replication cycle, making it an essential target for antiretroviral therapy. Integrase inhibitors, including first-generation drugs like Raltegravir and Dolutegravir, have significantly improved HIV treatment. Recent advances focus on allosteric inhibitors and novel computational approaches for drug discovery. Machine learning has emerged as a powerful tool for predicting HIV integrase inhibitors, leveraging molecular descriptors and cheminformatics to enhance drug discovery efficiency. This study employs a Random Forest Classifier to predict the activity of HIV integrase inhibitors using bioactivity data from the ChEMBL database. A total of 7,520 compounds were retrieved and processed, with molecular descriptors—such as molecular weight, LogP, hydrogen bond donors (HBD), hydrogen bond acceptors (HBA), topological polar surface area (TPSA), and rotatable bonds—calculated via RDKit. Data preprocessing included standardization, conversion of IC50 to pIC50 values, and handling of missing data. The dataset was split (80/20) into training and testing sets. The model achieved an accuracy of 81.6%, an AUC-ROC of 0.886, precision of 0.792, recall of 0.790, and an F1-score of 0.791. These results indicate the model's reliability in distinguishing active from inactive inhibitors. Future work should focus on external validation using independent datasets and integrating deep learning techniques for improved prediction accuracy. This study highlights the potential of machine learning in computational drug discovery, offering a cost-effective approach for screening HIV integrase inhibitors. The findings contribute to ongoing efforts to develop novel antiretroviral agents, potentially improving treatment options for drug-resistant HIV.

Democratising Bioinformatics in AMR Genome Analysis with AMRColab

Su Datt Lam (Department of Applied Physics, Faculty of Science and Technology, Universiti Kebangsaan Malaysia) , Sabarina Di Gregorio (Instituto de Investigaciones en Bacteriología y Virología Molecular, Universidad de Buenos Aires) , Mia Yang Ang (Department of Diagnostic & Allied Health Science, Faculty of Health & Life Sciences, Management & Science University), Emma Griffiths (Centre for Infectious Disease Genomics and One Health, Simon Fraser University), Tengku Zetty Maztura Tengku Jamaluddin (Department of Medical Microbiology, Faculty of Medicine and Health Sciences, Universiti Putra Malaysia), Sheila Nathan (Department of Biological Sciences & Biotechnology, Faculty of Science & Technology, Universiti Kebangsaan Malaysia) and Hui-Min Neoh (UKM Medical Molecular Biology Institute (UMBI); Universiti Kebangsaan Malaysia)

Antimicrobial resistance (AMR) is a growing global health crisis, projected to cause 39 million fatalities between 2025-2050. Timely and accurate surveillance of AMR pathogens is essential for tracking their resistance profiles and guiding interventions. However, many healthcare and public health professionals face barriers in bioinformatics expertise and computational resources, limiting their ability to analyse pathogen genomes effectively. To bridge this gap, with mentoring from Public Health Alliance for Genomic Epidemiology (PHA4GE), we developed AMRColab, an open-access bioinformatics suite hosted on Google Colaboratory (Lam et al. 2024). Designed for ease of use, AMRColab enables users with minimal bioinformatics experience to detect and visualise AMR determinants in pathogen genomes through an intuitive, 'plug-and-play' approach. Our platform integrates established tools such as AMRFinderPlus, ResFinder and hAMRonization, facilitating comparative analysis and transmission trend visualisation of AMR pathogens. A proof-of-concept study using methicillin-resistant Staphylococcus aureus (MRSA) strains demonstrated AMRColab's capability in identifying resistance determinants and enabling comparative genomic analyses between different laboratories. Additionally, two hands-on workshops were conducted, with participants expressing high confidence in using AMRColab and a strong interest in adopting it for their research. Recently, in an upgrade to the Colab, two genome assembly modules were introduced into AMRColab for better functionality: (1) a module supporting Illumina and IonTorrent reads, utilizing SPAdes for assembly and QUAST for quality assessment; and (2) a Nanopore-based module incorporating FastQC, FastP, NanoPlot, Flye for read processing and assembly, with Polypolish and BactInspector for polishing and quality assessment. To do this, Python codes to achieve both (1) and (2) were embedded into separate modules in the Colab. In keeping with the modular format of the Colab, these modules are standalone, and users can choose to only run the module(s) required for their research. These genome assembly modules are currently in beta and will be introduced and tested in future AMRColab workshops. With the addition of this genome assembly module, users will be able to upload sequences from either Illumina or Nanopore platforms, perform genome assembly, identify AMR determinants in these genomes, compare and visualize AMR genomes following only the prompts from the Colab. AMRColab's accessible design makes it valuable for medical laboratory technologists, clinicians, and public health researchers to perform genome analysis, regardless of their computational expertise. By lowering technical barriers, AMRColab contributes towards democratizing AMR surveillance and equipping healthcare professionals with essential genomic analysis tools. Reference: 1. Lam, Su Datt, Sabrina Di Gregorio, Mia Yang Ang, Emma Griffiths, Tengku Zetty Maztura Tengku Jamaluddin, Sheila Nathan, and Hui-min Neoh. "AMRColab–a user-friendly antimicrobial resistance detection and visualization tool." Microbial Genomics 10, no. 10 (2024): 001308.

Unravelling the Genetic Basis of Rare Diseases in Africa through Long-read Whole Genome Sequencing and Pangenomics: A Case Study of MPV17-associated Neuropathy

Gideon Akuamoah Wiafe (Neurogenomics Lab, Neuroscience Institute), Mohammed Farahat (Computational Biology Division, Faculty of Health Sciences), Jeanine Heckmann (Neurology Research Group, Neuroscience Institute), Melissa Nel (Neurogenomics Lab, Neuroscience Institute) and Elwazi African Refgraph Project (Computational Biology Division, Faculty of Health Sciences)
(University of Cape Town)

Short-read (sr) based next-generation sequencing (NGS) technologies have advanced our understanding of the genetic basis of rare diseases but they present limitations for resolving large structural variants and phasing alleles. While sr-based whole exome and genome sequencing identified putative compound heterozygous MPV17 variants, p.Arg125Trp and p.Gln36Ter, in four South African probands with inherited neuropathy, their trans configuration could not be confirmed to support a putative autosomal recessive disease entity enriched in an African neuromuscular disease cohort. Furthermore, targeted long-read sequencing uncovered both cis and trans configurations of these variants suggesting the possibility of multiple MPV17 gene copies. In this study, we explored the potential of long-read whole-genome sequencing (lr-WGS) and deconvolution of fully-phased haplotypes in the Human Pangenome Reference Consortium's (HPRC) pangenome graph to untangle the pathogenic role of these two MPV17 gene variants in a novel autosomal recessive neuropathy phenotype. We aligned 30X PacBio Revio lr-WGS data from 1 proband to the GRCh38 reference. The MPV17 gene locus was extracted from the HPRC graph, untangled and visualized to determine the MPV17 copy number state in each haplotype. The 30X lr-WGS data aligned to GRCh38 confirmed a trans configuration of the two MPV17 gene variants in 1 proband and did not reveal any additional pathogenic candidates. Furthermore, the MPV17 gene was single copy in all 94 haplotypes in the HPRC pangenome graph (including 48 of African ancestry). This study underscores the potential of lr-WGS combined with pangenome-based approaches to identify novel genetic causes of rare diseases in underrepresented populations. The ability of lr-WGS to determine variant phasing independently of trio data has significant implications for genetic research and clinical diagnostics in Africa, where recruitment for trio studies can be challenging.

Identification of operons in Clostridium difficile.

Hwenude Judicaelle Chance Gountin (South Africa National Bioinformatics Institute (SANBI) / University of Western Cape), Tracey Calvert-Joshua (South Africa National Bioinformatics Institute (SANBI)) and Alan Christoffels (South Africa National Bioinformatics Institute (SANBI))

Clostridium opportunistic pathogens are a genus of anaerobic, Gram-positive bacteria, found in soils and normal intestinal flora of animals and humans. The pathogenesis of C. difficile is worsened by the emergence of new hypervirulent strains and the development of antibiotic resistance. We aim to understand gene regulation in C. difficile by describing the operon organization in this pathogen. In pathogens, operons are a central feature of bacterial gene regulation. These operons represent a cluster of genes that are transcribed together to give a single messenger RNA (mRNA) molecule. The organization of the genes in operons can alter gene expression through specific regulatory mechanisms and create an opportunity to identify possible drug targets. Using a recently developed operon predictor, COSMO, we describe the operon network in C.difficile using a published dataset of RNAseq data for clostridium difficile 630 wildtye and its Vancomycin mutant strain. Understanding the regulation and expression of operons provides insights into how C. difficile develops and maintains resistance, which can inform strategies to combat antibiotic-resistant infections.

Tailoring PHA4GE's Wastewater Surveillance Training Through Lessons from LMICs

Tracey Calvert-Joshua (Public Health Alliance for Genomic Epidemiology (PHA4GE))

Workshops for participants from low- and middle-income countries LMICs offer a unique lens into the challenges and opportunities of delivering effective training in diverse contexts. Over time, we have identified key lessons that can make these workshops more impactful, practical, and sustainable for both participants and facilitators. 1) Understanding the varied experiences of attendees is critical. Many participants face logistical or contextual challenges, such as limited travel experience or unfamiliarity with local conditions, which can affect their ability to engage fully. 2) Managing expectations is essential. Participants may want to master advanced skills in a short time, but aligning these goals with what's achievable helps create a more focused and rewarding learning experience. 3) Technical issues are an inevitable part of any workshop, yet they also provide opportunities for creative problem-solving. 4) Incorporating hands-on projects has been particularly effective in reinforcing learning, as participants apply their skills to real-world scenarios that mirror the challenges they face in their own work environments. 5) The benefits of these workshops go beyond just the training. By creating support networks after the workshops, participants can keep learning and build a friendly community that keeps growing even after the sessions are over. These are just a few of the key insights which highlight how a well-designed workshop can go beyond teaching technical skills, to empower participants to collaborate effectively and confidently apply their skills to enhance data analysis practices within their institutions. As a result, what we've learned PHA4GE has shaped how we create our wastewater surveillance syllabus for both workshops and online training. By drawing on insights from both local and global experts, we're focusing on practical approaches that reduce theoretical content while keeping the training easy to follow and modular. The lessons are divided into small, timed and manageable sections, which allows for flexibility and caters to different skill levels with optional advanced modules. Link to training content (in progress): https://pha4ge.org/training/wes/

SeqWord Motif Mapper: Unlocking Bacterial Epigenetic Insights

Christophe Lefebvre, Rian Pierneef and Oleg Reva
(Centre for Bioinformatics and Computational Biology; University of Pretoria)

The SeqWord Motif Mapper (SWMM) is a newly developed tool designed to streamline the identification and visualization of complex patterns of epigenetic modifications in bacterial genomes using data obtained through single-molecule real-time (SMRT) sequencing technologies. Bacterial epigenetics, particularly through methylation, plays a crucial role in regulating processes such as gene expression, chromosome replication, symbiont-host interactions, and defense mechanisms against phages. However, there is a lack of computational tools for the detection, comparison, and visualization of patterns of epigenetically modified bases in bacterial genomes. SWMM addresses these challenges by providing a robust statistical framework and interactive visualization capabilities. Implemented in Python 3, the software utilizes input data from standard SMRT analysis pipelines, including GFF annotation files and reference genomes in GenBank format. The tool integrates advanced genomic analyses, such as motif distribution mapping and statistical assessment of the distribution of modified bases and motifs across coding, non-coding, and promoter regions; core and horizontally acquired regions; chromosomes and plasmids; leading and lagging replichores; and regions with alternative base composition. Its visualization outputs include circular and dot-plot representations, accompanied by statistical validation in both graphical and text formats. Applications of the tool have already yielded significant insights into epigenetic regulation mechanisms within various bacterial species, including motifs linked to antibiotic resistance and stress response [1-5]. SWMM can be deployed both locally and as a web application, making it accessible to users with varying levels of bioinformatics expertise. By offering a user-friendly interface and compatibility with multiple operating systems, it enables scalable and reproducible research. The program is freely available on GitHub (https://github.com/chrilef/BactEpiGenPro) and can also be accessed as a web application at http://begp.bi.up.ac.za. This tool represents a critical advancement in bacterial epigenetics, with promising implications for understanding bacterial adaptation, pathogenicity, and gene regulation in both clinical and environmental contexts.

References: 1. Reva ON, La Cono V, Crisafi F, et al. Interplay of intracellular and trans-cellular DNA methylation in natural archaeal consortia. Environ Microbiol Rep. 2024;16(2):e13258. doi: 10.1111/1758-2229.13258. 2. Korotetskiy IS, Shilov SV, Kuznetsova T, et al. Analysis of Whole-Genome Sequences of Pathogenic Gram-Positive and Gram-Negative Isolates from the Same Hospital Environment to Investigate Common Evolutionary Trends Associated with Horizontal Gene Exchange, Mutations and DNA Methylation Patterning. Microorganisms. 2023;11(2):323. doi: 10.3390/microorganisms11020323.. 3. Korotetskiy IS, Jumagaziyeva AB, Shilov SV, et al. Transcriptomics and methylomics study on the effect of iodine-containing drug FS-1 on Escherichia coli ATCC BAA-196. Future Microbiol. 2021;16:1063-1085. doi: 10.2217/fmb-2020-0184. 4. Reva ON, Korotetskiy IS, Joubert M, et al. The Effect of Iodine-Containing Nano-Micelles, FS-1, on Antibiotic Resistance, Gene Expression and Epigenetic Modifications in the Genome of Multidrug Resistant MRSA Strain Staphylococcus aureus ATCC BAA-39. Front Microbiol. 2020;11:581660. doi: 10.3389/fmicb.2020.581660. 5. Reva ON, Swanevelder DZH, Mwita LA, et al. Genetic, Epigenetic and Phenotypic Diversity of Four Bacillus velezensis Strains Used for Plant Protection or as Probiotics. Front Microbiol. 2019;10:2610. doi: 10.3389/fmicb.2019.02610.

Bioinformatics Pipeline for Whole Genome Sequencing Data Generated To Investigate Genetic Variants Associated With COVID-19 Vaccination Status

Yentl Lamprecht (Stellenbosch University), Marlo Möller (Stellenbosch University; Genomics for Health in Africa, CoRE; NITheCS) and Desiree Petersen (Stellenbosch University; Genomics for Health in Africa, CoRE; NITheCS)

Background: Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) has had a devastating impact since 2019, resulting in over 7 million deaths worldwide, including more than 102 000 deaths in South Africa. Host genetic variants are associated with coronavirus disease 2019 (COVID-19) severity; however, due to the increased genetic diversity among African populations, it remains unclear whether previously reported associations from extensively studied European populations apply to Africans. This study compared variant interpretation tools and selected the most effective one to develop a bioinformatics pipeline identifying candidate host genetic variants associated with COVID-19 vaccination status. Results: Significant discrepancies were observed between the variant interpretation tools eVai and Franklin, reflecting a broader challenge with variant interpretation, as no universally accepted "gold standard" exists. Franklin was chosen for its cost-effectiveness, definitive classification, and closer correlation with our in-house variant interpretation software, Variant Interpretation and Prioritisation (VIPR). Whole genome sequencing (WGS) data from 48 individuals across four cohorts was analysed through the pipeline, identifying four genes of interest in the likely pathogenic moderate subcategory of Variants of Uncertain Significance: mucin 16, cell surface associated (MUC16), mucin 3A, cell surface associated (MUC3A), histone deacetylase 6 (HDAC6), and poly(A) binding protein cytoplasmic 3 (PABPC3). Conclusions: The pipeline demonstrated the potential for identifying genetic variants associated with severe COVID-19 risk, but the small sample size limited the generalisability of the results. A larger sample size would enable a more comprehensive investigation of genetic variants associated with COVID-19 vaccination status among African populations. Ultimately, the findings could guide targeted vaccination strategies, prioritising individuals based on their genetic risk profile.

Elucidation of the genomic basis of rare haematological disorders in South Africa.

Carli Loubser (Stellenbosch University) and Shahida Moosa (Stellenbosch University, Medical Genetics Tygerberg Hospital)

An estimated 4.2 million South Africans suffer from a rare disease, and many will wait years to receive a diagnosis. Individuals with rare haematological disorders can have clinical signs and symptoms which are non-specific, and clinical tests may be inconclusive, making it difficult to reach a definitive diagnosis. Furthermore, the exact molecular aetiology of a disorder may vary between patients. Thus, molecular testing provides a more reliable diagnosis, allowing for patient-specific prognoses and treatment recommendations to be made. Furthermore, informed family-planning, and cascade testing of at-risk family members can also take place. The aim of this project is to establish molecular diagnoses for patients with rare haematological disorders. In this study, we report on the diagnosis of patients enrolled in the "Undiagnosed Diseases Programme" using exome sequencing. Exome sequencing is able to detect approximately 85% of reported pathogenic variants. Following exome sequencing, variant calling was performed using a pipeline adapted from Genome Analysis ToolKit (GATK) best practices, and variant filtering and prioritisation was performed using an in-house pipeline. We present the preliminary results on the molecular findings which led to the diagnosis of three patients. For patient 1, a likely pathogenic variant was identified in ACTN1, leading to a diagnosis of platelet-type bleeding disorder 15 (BDPLT15). For patient 2, a sibling of patient 1, the same variant in ACTN1 was identified, as well as a pathogenic variant in GPR143. Thus, diagnoses of BDPLT15 as well as ocular albinism type I were made. For patient 3, a novel, likely pathogenic variant was identified in ENG, leading to the diagnosis of hereditary haemorrhagic telangiectasia type I (HHT1). Exome sequencing allows for the molecular diagnosis of patients with rare haematological disorders. From the preliminary results of this study, we identified a novel pathogenic variant in ENG causing HHT1 in one of the patients. The molecular diagnosis of these three patients allowed them to receive appropriate care, and cascade testing of at-risk family members will take place as a result of these diagnoses.

The African Pangenome Reference Graph Project

Mohammed Farahat (Computational Biology Division, IDM, University of Cape Town), Shaun Aron (Sydney Brenner Institute for Molecular Bioscience, University of the Witwatersrand) and Nicola Mulder (Computational Biology Division, IDM, University of Cape Town)

The conventional human reference genome, though essential for variant calling, lacks the genetic diversity needed to represent global populations, particularly African populations with high genetic variability. Use of a linear reference introduces both reference and allele bias, obscuring population-specific insights. To address this, we are constructing an African Pangenome Reference Graph, enabled by advancements in long-read sequencing and graph-based reference models. Our project leverages ~60X PacBio HiFi sequencing data from 27 individuals across Burkina Faso, Kenya, and South Africa, capturing a significant proportion of genetic diversity of Africa. This data allows us to build a pangenome graph that more accurately represents African genomes. Unlike traditional linear references, the pangenome graph integrates diverse sequence paths, improving variant calling for both single nucleotide and structural variants. The goal is to improve exploration of African-specific genetic variation and enhance variant discovery in related populations. To achieve this, we developed workflows for generating high-quality de novo assemblies and pangenome graphs using both reference-free (PanGenome Graph Builder, PGGB) and the reference-derived (Minigraph-Cactus) algorithms. A third workflow is under development to extract and analyze African variation within specific regions and call variants using the graph as a reference. Preliminary analysis based on a 30x coverage dataset has yielded high-quality assemblies with contig N50s between 31-49 Mb. While the recent dataset is 60x coverage. Supported by H3ABioNet and the eLwazi Consortium, the African pangenome graph provides a valuable resource advancing population-specific genomics. Initial applications include comparing variants called using the African graph versus linear and global references, investigating complex regions, and benchmarking graph-based variant calling. As part of our efforts to advance pangenome research and analysis, we hosted the Human Pangenome Bring Your Own Data (BYOD) Workshop in October 2024 in collaboration with the eLwazi Open Data Science Platform. During this hands-on workshop, participants explored methods for variant calling, graph-based analysis, and personalized genome graph creation, comparing results between linear reference and pangenome-based approaches. A second phase will generate assemblies from seven samples from the Democratic Republic of Congo, expanding the resource. This collaborative initiative is a crucial step toward a more inclusive genomic reference, enabling equitable genomic studies across African and global populations.

Developing an information system for integrating clinical and genomic infectious disease data in Tanzania

Melkiory Beti (Kilimanajaro Clinical Research Institute (KCRI)), Patrick Kimu (Kilimanajaro Clinical Research Institute (KCRI)), Boaz Wadugu (Kilimanajaro Clinical Research Institute (KCRI)), Willfred Senyoni (University of Dar es Salaam) and Tolbert Sonda (Kilimanjaro Clinical Research Institute (KCRI))

Background Infectious diseases continue to present significant public health issues in low- and middle-income countries like Tanzania, where the integration of clinical and genomic data is important for better disease diagnosis and surveillance. However, existing health information systems mostly operate in different sources limiting the ability to connect clinical data with genomic data for better patient diagnosis and infectious disease control. To tackle these challenges we developed an integrated information system that combines clinical data collected in a customized District Health Information System2 (DHIS2) with genomic data generated from Nanopore sequencing. The system aims to integrate these data, aiding clinicians and laboratory scientist in identifying multiple pathogens from a single patient sample and public health researchers in viewing infectious disease patterns. Methods Clinical data, including patient demographics and symptoms such as fever and diarrhoea, were collected from healthcare facilities using a customised DHIS2, an open-source software widely used for health data collection in Tanzania. R programming language scripts were used to securely fetch clinical data from DHIS2 using the DHIS2 API and integrate it with genomic data results that were produced from the analysis of the cgetools bioinformatics pipeline. This pipeline uses tools such as KmerFinder for pathogen identification, supporting the detection of the diverse pathogens from a single sample. R's shiny web framework was used to build an interactive web interface allowing the user to search for patient IDs on the system to view detailed clinical data alongside genomic data that displays the identified pathogens. Results The developed system successfully processed and integrated 21 datasets, connecting clinical information with genomic output results. The datasets included key clinical variables such as patient symptoms like fever and diarrhoea, gender, and the region of origin, while linked genomic data showed pathogens identified from patient samples. Proving an interactive web interface for users to search for patient IDs to view detailed clinical records alongside genomic data and also has features for interactive data visualisation capabilities, including bar graphs that show trends in pathogen occurrence according to Tanzania regions, enabling epidemiological monitoring and outbreak Discussion The developed system for integrating data shows several critical insights regarding the potential of clinical-genomic data integration in infectious disease control. The use of an open-source health information system like DHIS2 demonstrated the feasibility of leveraging existing digital health data collection software to enhance data integration in healthcare. Additionally, the application of the cgetools pipeline for pathogen detection proved effective in identifying multiple pathogens from a single sample. The integration process shows the significance in support real-time clinical decision-making. The interactive visualization tools provided valuable information on pathogen distribution patterns, emphasising their important role in outbreak detection and response. Conclusion By integrating clinical data from DHIS2 with genomic sequencing outputs, this system offers a powerful tool for infectious disease surveillance in Tanzania. It supports the identification of different pathogens, enabling timely diagnosis and supporting infectious disease control. The system’s flexible, scalable design makes it suitable for applications in infectious disease management across healthcare settings.

Whole genomic analysis uncovers high genetic diversity of rifampicin-resistant Mycobacterium tuberculosis strains in Botswana

Tuelo Mogashoa (Stellenbosch University), Johannes Loubser (Stellenbosch University), Ontlametse Choga (Botswana Harvard Health Partnership), Justice Tresor Ngom (Stellenbosch University), Wonderful Choga (Botswana Harvard Health Partnership), Mpaphi Mbulawa (Botswana National Tuberculosis Reference Laboratory), Tuduetso Molefi (Botswana National Tuberculosis Program), Topo Makhondo (Botswana National Tuberculosis Program), One Stephen (Botswana National Tuberculosis Reference Laboratory), Kedumetse Seru (Botswana Harvard Health Partnership), Patience Motshosi (Botswana Harvard Health Partnership), Boitumelo Zuze (Botswana Harvard Health Partnership), Joseph Makhema (Botswana Harvard Health Partnership), Rosemary Musonda (Botswana Harvard Health Partnership), Dimpho Otukile (Victus Global Botswana), Chawangwa Modongo (Victus Global Botswana), Botshelo Kgwaadira (Botswana-University of Maryland School of Medicine, Health Initiative (BUMMHI)), Keabetswe Fane (Botswana-University of Maryland School of Medicine, Health Initiative (BUMMHI)), Simani Gaseitsiwe (Botswana Harvard Health Partnership), Rob Warren (Stellenbosch University), Sikhulile Moyo (Botswana Harvard Health Partnership), Anzaan Dippenaar (University of Antwerp) and Elizabeth Streicher (Stellenbosch University)

Background: The emergence of drug-resistant Mycobacterium tuberculosis (M. tb) strains remains a threat to tuberculosis (TB) prevention and care. Understanding the drug resistance profiles of circulating strains is crucial for effective TB control. This study aimed to describe the genetic diversity of rifampicin-resistant M. tb strains circulating in Botswana using whole genome sequencing (WGS). Methods: This study included 202 stored M. tb isolates from people diagnosed with rifampicin-resistant TB (RR-TB) between January 2016 and June 2023. Genomic DNA was extracted using the cetyltrimethylammonium bromide (CTAB) method. Library preparation was performed using the Illumina DNA prep kit following the manufacturer's instructions. Sequencing was done on Illumina NextSeq2000. TBProfiler software was used to identify known M. tb lineages and drug resistance profiles. Statistical analyses were performed on STATA version 18. Results: WGS analysis revealed multidrug resistance (57.9%: 95% CI; 50.7–64.8), Pre-XDR (16.8%, 95% CI: 11.9–22.7), RR-TB (20.2%: 95% CI: 14.98–26.5), and HR-TB (0.5%, 95% CI; 0.01–2.7). We identified a high genetic diversity with three predominant lineages: lineage 4 (60.9%, 95% CI; 53.8–67.7), lineage 1 (22.8%: 95% CI; 17.2–29.2), and lineage 2 (13.9%, 95% CI: 9.4–19.4). The most frequently observed drug resistance mutations for rifampicin, isoniazid, ethambutol, streptomycin, pyrazinamide, and fluoroquinolones were rpoB S450L (28.6%), katG S315T (60.5%), embA_c.-29_-28delCT, embB Q497R (31.7%), rrs_n.517C>T (47.1%), pncA_c.375_389delCGATGAGGTCGATGT (36.0%) and gyrA A90V (79.4%), respectively. No bedaquiline and delamanid resistance-associated mutations were detected. Conclusions: This study highlights the high genetic diversity of M. tb strains, with a predominance of lineage 4 among people with RR-TB in Botswana. It provides valuable insights into the genetic diversity of rifampicin-resistant M. tb strains circulating in Botswana.

Evaluating Innovative Data Science Methodologies for the Analysis of Disease Drivers and Outcomes: A Case Study for Kidney Disease in Khayelitsha, South Africa

Sohail Simon, Nicki Tiffin and Tsaone Tamuhla
(University of the Western Cape - SANBI)

Background: Non-communicable diseases (NCDs) are a growing global health concern, with cardiovascular diseases (CVD), diabetes, obesity, hypertension, and heart failure significantly contributing to the rising incidence of chronic kidney disease (CKD). In Sub-Saharan Africa (SSA), CKD prevalence varies widely, ranging from 6.4% to 13.9%, with certain high-risk groups, including older adults, women, people of African descent, and individuals with diabetes and hypertension, being disproportionately affected. Additionally, HIV infection and antiretroviral therapy (ART) further elevate CKD risk. Multivariate analysis of risk factors including infectious diseases, multimorbidity with other NCDs, medications taken, demographics and health facility access can assist with better understanding these drivers of CKD. Traditionally, clinical and epidemiological studies analyse data from bespoke research studies, but a wealth of information about risk factors and outcomes may also be derived from routine digital health data. Many SSA countries have relied on paper-based health records, with the transition to digital health information systems (HIS) limited by capacity constraints. In 2015, the Western Cape (WC) Province of South Africa, the Provincial Department of Health and Wellness (WCGHW) established the Provincial Health Data Centre (PHDC), a Health Information Exchange that integrates and links routine health data from multiple digital health platforms using a unique patient identifier, the Patient Master Index (PMI), in order to improve continuity of care for healthcare clients. Challenges: These routine health data provide a valuable resource for epidemiological research, offering longitudinal insights into population health, but are often collected at irregular intervals based on individuals' healthcare access. The longitudinal data are also large and highly granular, necessitating specialized analytical approaches. Multifactorial statistical approaches are routinely used for risk factor analyses, but due to the data type and structure, more complex approaches including latent factor analysis and machine learning methodologies may provide better insights. Study Objectives: This study aims to evaluate multifactorial statistical and machine learning methodologies for analysing large-scale longitudinal routine health datasets. This study aims to evaluate risk factors for rate of decline in kidney function in healthcare clients from the Khayelitsha subdistrict of the Western Cape Province, as a case study. We will appraise the use of multifactorial statistical analysis, latent variable modelling, and machine learning approaches to identify optimal methodologies for handling granular routine health data. Methods and Expected Outcomes: We present preliminary analyses of risk factors underlying decline in kidney function in this population, demonstrating the use of multifactorial analysis for this large, granular dataset. We describe appropriate feature reduction techniques and machine learning algorithms that will be compared for their effectiveness in evaluating the impact of a variety of risk factors on the decline in kidney function. By leveraging anonymised and perturbed PHDC data to evaluate the different methodologies, this study seeks to identify effective methodologies for risk factor analysis undertaken using large, granular routine health data for epidemiological research.

Developing a Bioinformatics Pipeline for Predicting Disease-Associated Variants in Human Leukocyte Antigen (HLA) Genotypes and Pharmacogenetic Profiles in South African Populations Using Data from the H3Africa Genotyping Chip

Tiffany Fredericks, Nicola Tiffin and Tsaone Tamuhla
(University of the Western Cape - SANBI)

Background: South African (SA) genomic data is increasingly available from projects like the Southern African Human Genome Programme and Human Heredity and Health in Africa (H3Africa) initiative. The Illumina H3Africa genotyping chip characterises around 2.5 million SNPs representing diverse African populations. This data which can enhance our understanding of complex diseases in Africa where genetic diversity is high, and inform personalized treatment strategies tailored to an individual's unique genetic makeup. The H3Africa Chip variants include the Human Leukocyte Antigen (HLA) gene regions and pharmacogenomic variants. These variants can inform susceptibility and diagnosis for autoimmune diseases, susceptibility to some infections, and drug efficacy in individuals. Whilst research-generated genotype data cannot be used to direct clinical interventions, it could alert clinicians to predict disease susceptibility, consider diagnostic possibilities, and to monitor potential adverse drug reactions (ADRs). Returning research data to healthcare providers could provide appropriate alerts to aid timely diagnoses, enable tailored drug doses to minimise severe ADRs, to improve patient outcomes, and reduce the burden on South Africa's already strained healthcare system and the economy. Problem Description: In Africa, genomic data from research participants are seldom returned to healthcare providers. Although research data cannot be used diagnostically, they can offer additional information to alert clinicians to possible phenotypes during their clinical management of patients, optimizing healthcare procedures. Currently, research data that could identify HLA type and pharmacogenomic variants in research participants are not being used in this way. Study Objectives: The Virtual Cohort for African MultiMorbidity (VCAMM) study integrates H3Africa Illumina genotype data and clinical phenotypes from routine health data for consenting participants to explore the genetics underlying disease in the Western Cape, SA. Using this data, this project aims to develop bioinformatics pipelines to analyse the genotype data and characterise (i) HLA genotypes of participants, linking any associated clinical phenotypes; and (ii) participants' pharmacogenetic variation profiles for essential medicines prescribed in the Western Cape. Where informed consent has been given by participants, these outputs will be returned to the Western Cape Department of Health and Wellness (WCGHW) to alert clinicians of the potential genetic factors which may influence patient care. Methods and Expected Results: We present methodology to develop a pipeline using genomic data processing tools, quality control steps, and comparisons with population-specific and global genetic data repositories to (1) identify HLA type in individuals linked to related disease phenotypes and (2) compile a pharmacovariant panel to assess drug responses and the risks of ADRs in relation to commonly prescribed medications. This genetic information will be used to generate individualized patient profiles, which will be returned to WCGHW to alert clinicians in the public health service to potential clinical phenotypes in each research participant. Impact of the Study: This study will test the feasibility of returning research genotype data from study participants to the WCGHW. By linking genetic data with clinical health information, clinicians may be encouraged to consider certain clinical phenotypes to improve the healthcare of study participants.

Harnessing Multi-Omics for Sustainable Agriculture: Evaluating Bio-Compost and Chemical Fertilization Impacts on Soil Health, Microbial Diversity, and Crop Quality

Dorra Rjaibi, Alia Benkahla and Oussama Souiai
(Laboratory of Bioinformatics, Biomathematics and Biostatistics)

Soil microbial diversity plays a crucial role in maintaining soil health, nutrient cycling, and overall ecosystem stability. However, modern agricultural practices,particularly the overuse of chemical fertilizers have led to soil degradation, and have been shown to negatively impact microbial communities, reducing biodiversity and altering soil functionality, diminished agricultural sustainability. This study leverages advanced multi-omics approaches, including metagenomics, transcriptomics, and metabolomics, to investigate the effects of bio-compost and chemical fertilizers on soil microbial communities, plant development , and fruit quality. By comparing soils amended with bio-compost (produced through aerobic and anaerobic methods) to chemically fertilized and untreated soils, we aim to assess their impact on microbial composition, metabolic activity, and nutrient retention. Our findings will contribute to the growing need for ecological transition in agriculture, promoting sustainable soil restoration through organic amendments. Integrative omics analyses will reveal how bio-compost influences microbial networks, enzymatic activities and plant-microbe interactions, leading to enhanced soil fertility and improved crop quality. Furthermore, long-term assessments of soil organic matter composition and microbial functionality will provide insights into resilience of soil ecosystems under different fertilization regimes. By reducing dependency on synthetic fertilizers and fostering microbiome-driven nutrient cycling, this study supports a more sustainable and climate resilient agricultural model.

Emergence and Evolution of Epizootic Hemorrhagic Disease Virus (EHDV) in the Mediterranean Region: Spatio-temporal Dynamics and Epidemiological Insights

Marwa Arbi (Laboratory of Bioinformatics, Biomathematics and Biostatistics), Emna Harigua-Souiai (Laboratory of Molecular Epidemiology and Experimental Pathology), Mariem Hanachi (Laboratory of Bioinformatics, Biomathematics and Biostatistics), Imen Larbi (Laboratory of Epidemiology and Veterinary Microbiology), Melek Chaouch (Laboratory of Bioinformatics, Biomathematics and Biostatistics), Dorra Rjaibi (Laboratory of Bioinformatics, Biomathematics and Biostatistics), Mohamed Fethi Diouani (Laboratory of Molecular Microbiology, Vaccinology and Biotechnological Development), Alia Benkahla (Laboratory of Bioinformatics, Biomathematics and Biostatistics) and Oussema Souiai (Laboratory of Bioinformatics, Biomathematics and Biostatistics)

Background: Epizootic Hemorrhagic Disease Virus (EHDV) is an arbovirus, transmitted to wild and domestic ruminants through Culicoides biting midges. Since 2006, high morbidity and mortality cases of EHDV have been reported among cattle and deer populations in several Mediterranean countries. The temporal and geographic origins of these incursions remained unclear. In this study, we aimed to investigate the evolutionary history of EHDV in the Mediterranean region and highlight the epidemiological features of viruses in relationship with genetic diversity and viral ecology. Methods: We extracted from Genbank the EHDV VP2 and VP5 segments isolated in the mediterranean region during the period 2006 to 2023 and blasted them to obtain a final dataset of 68 and 91 nucleotide sequences. Using these datasets, we conducted a Bayesian phylodynamic analysis, which inferred discrete models of 'geographic origin', 'Serotype' and 'Host' by employing the BEAST package. Results: RSPP and TMRCA analyses showed that the Mediterranean EHDV has as ancestral root the North America strains that circulated in the 17th century. Our study suggested that the first EHDV incursions in the Mediterranean region started in France and Tunisia during the 1800s. The latter countries were epicenters of EHDV in the region. Significant transition routes (BF>3) were detected revealing virus transmission between North African and European countries. Serotype model study revealed VP5 multiple inter-serotype events involving serotypes 1, 2, 6, 7 and 8 with high statistical support (BF>100). Significant virus transmission was detected for Cattle-deer and Culicoides-Cattle transition routes. Conclusion: The virus transmission was intense between North African and European countries of the Mediterranean region. EHDV spread in this region seems to be influenced mainly by vector/host distribution and abundance, ruminants' trade and prevailing winds.

The Consortium of Genomics Students and Young Researchers in Africa (CoGSAYR Africa): a model for genomics capacity building in the young African population

Gladys Ibrahim, Ridwanullah Abdullateef and Enahoro Abhulimen
(The Consortium of Genomics Students and Young Researchers in Africa)

Africans, despite being the most genetically diverse, were said to represent about 3% of genomic data worldwide in 2019, with a drastic reduction to 1.1% in 2021. This emphasizes the critical need to build genomics and bioinformatics capacity in the African young population. The Consortium of Genomics Students and Young Researchers in Africa (CoGSAYR) was established with the aim of encouraging enthusiasm and involvement among young medical professionals in genomic research. CoGSAYR aims to create a collaborative platform that empowers the next generation of scientists to contribute significantly to genomic research in the African context. This is achieved through the establishment of Interest Groups where members collaborate on genomics-based research work in their focus areas. The organization was launched in September 2024, with a call for members which recorded 238 applications from 11 African countries. Countries from which the highest number of applications were received were Nigeria, Ghana, Kenya, and Zambia, respectively. The Infectious Disease Genomics Interest Group had the highest percentage of interested applicants (35.7%), closely followed by Global Health Genomics (30.7%), then Cancer Genomics (20.6%), Neurogenomics (8%), and then Epigenomics (5%). Of all the applicants, 55.9% have had some experience in genomics or related research areas, while 44.1% have never had any of such experience. This experience ranges from virtual capacity building workshops, lab internships, and free online courses. It is imperative that capacity building is prioritized in the young African population to bridge the gap in genomic data. Genomics institutions and leading genomics researchers have a role to play in engaging early career researchers and young Africans who have displayed interest in genomics. This can be achieved through collaborations with initiatives like CoGSAYR, to build a young and dependable workforce for the continuity of genomics-based projects.

Klebsiella pneumoniae Sequence Type 39: An emerging cause of hospital outbreaks in Malawi

Allan Zuza (Malawi Liverpool Wellcome Programme, London School of Hygiene & Tropical Medicine), Oliver Pearse (Liverpool School of Tropical Medicine, Malawi Liverpool Wellcome Programme), Patrick Musicha (Malawi Liverpool Wellcome Programme), Zoe Dyson, Kondwani Kawaza, Nicholas Feasey and Eva Heinz
Zoe Dyson (London School of Hygiene & Tropical Medicine, Wellcome Sanger Institute), Kondwani Kawaza (Malawi Liverpool Wellcome Programme), Nicholas Feasey (University of St Andrews, Malawi Liverpool Wellcome Programme) and Eva Heinz (University of Strathclyde)

Introduction Klebsiella pneumoniae (KPN) is a significant contributor to multidrug-resistant healthcare-associated infections, particularly prevalent in neonatal bloodstream infections at Queen Elizabeth Central Hospital (QECH), Blantyre Malawi. In a large-scale genome sequencing analysis of KPN from invasive disease at QECH between 1998 and 2020, ST39 emerged as the second most prevalent lineage. Majority (n = 100, 74.07%) of all ST39 KPN isolates were isolated in 2017, indicating a potential outbreak. This study aims to elucidate the genomic changes that might have provided this lineage with selective advantage over other KPN to cause this lineage's expansion in 2017. Methods We performed a genomic analysis of 135 ST39 KPN genomes from QECH. Antimicrobial resistance (AMR) and heavy metal resistance genes were identified using AMRFinderPlus, plasmid replicons were detected using Abricate and the PlasmidFinder database. Virulence genes and capsular typing was performed using Kleborate. Whole-genome single-nucleotide polymorphisms were identified using Snippy, recombinant regions were masked using Gubbins and a phylogenetic tree constructed using IQ-TREE. Lineage assignment was performed using rPinecone and data was visualized in R using the ggtree and ggplot2 packages. Five genomes from this collection had long reads sequenced to use as references for phylogenetic analysis and genome interrogation. Results All the 135 genomes were multi-drug resistant and only one isolate encoded a gene conferring resistance to carbapenems. The majority of genomes (n = 105, 77.8%) formed a single phylogenetic clade with less than 5 single nucleotide polymorphisms between genomes within the clade. This clade had most genomes from isolates collected in 2017 (n = 98, 93%) and from a single neonatal ward (86, 81.9%) suggesting continued within ward transmission of isolates from this clade in 2017. Isolates in this clade had three large-genomic regions which had coding sequences absent in the rest of the collection. Additionally, the catecholate siderophore esterase genes IroE and iroD were identified in genomes of the outbreak clade suggesting enhanced salmochelin and enterobactin degradation in these isolates. Discussion and conclusion This study explores the genomic epidemiology of ST39 KPN, a cause of hospital-acquired infections in Malawi. The study identified variable regions within the ST39 genome which facilitated the persistence of this lineage in the hospital in 2017. Such findings reveal distinct epidemiological factors leading to success of certain lineages in the local setting. Consequently, it emphasizes the need to analyze locally collected isolates to inform the design of infection prevention and control strategies suited local context and pathogens.

RNA-seq and gut microbiota analyses of changes in the Anopheles arabiensis transcriptome and gut microbiota associated with the endosymbiotic Microsporidia MB

Jacqueline Wahura (UCT ; icipe), Nicola Mulder (UCT), Joseph Gichuhi (icipe), Cynthia Nyambura (icipe), Irene Muiruri (icipe), Edward Edmond Makhulu (icipe), Mwatum Maloba Alakonya (icipe), Lilian Mbaisi (icipe), Daniel Masiga (icipe) and Jeremy Herren (icipe)

Insect endosymbionts play a key role in modulating vector competence in mosquitoes. One key endosymbiont shown to possess a Plasmodium transmission blocking phenotype is the recently discovered Microsporidia MB in Anopheles arabiensis mosquitoes. However, its adoption for the development of a malaria transmission-blocking strategy could be impacted by how it influences the physiology of its host. We assessed the possible effect of natural mosquito infection with the endosymbiont Microsporidia MB on the gene expression profiles and the gut microbiota composition and diversity in non-blood-fed and blood-fed (24, 48 and 72 hours post blood meal) filial 1 generation An. arabiensis mosquitoes. We sequenced the transcripts from the fat body and guts using the oxford nanopore technology (ONT) MinION device and profiled the gut microbiota by sequencing the V3-V4 region of the bacterial 16S rRNA gene in midgut samples. Differences in microbiome diversity were recorded between Microsporidia MB positive and negative mosquitoes at all sampling points. We observed that mosquito infection with Microsporidia MB upregulated the juvenile hormone biosynthesis pathway in the non-blood-fed mosquitoes which could be associated with the distinct gut microbiota composition in the positive mosquitoes. In addition, blood feeding in Microsporidia MB positive mosquitoes was associated with an activated immune system 24 hours post a blood meal where the lipopolysaccharide tumor necrosis factor was upregulated. Interestingly, an activated immune system 24 hours post blood meal was associated with a microbiota shift to favor the proliferation of microbes with anti-Plasmodium properties including Pseudomonas and Serratia. The immune system in Microsporidia MB positive mosquitoes persisted even 48 hours post a blood meal where factors such as the peptidoglycan recognition SC2-like and lysozyme c-1 were activated. Notably, an upregulated immune system at this time point was associated with downregulated metabolism systemically and the re-instatement of Serratia, Pseudomonas and other key microbes to relative abundances closely like those recorded in non-blood-fed mosquitoes. At the 72-hour time point, genes associated with immunity including cecropins and defensins were mostly downregulated while metabolic processes were upregulated. Our results provide insights into the effect of Microsporidia MB infection on the An. arabiensis gene expression and gut microbiota profiles and will be fundamental in understanding the mechanistic basis of Microsporidia MB-malaria transmission blocking phenotype.

Integrating Multiple Learning Strategies into the PHA4GE Wastewater Surveillance Bioinformatics Course

Keaghan Brown, Farzaana Diedericks and Tracey Calvert-Joshua
(Public Health Alliance for Genomic Epidemiology)

One of the main objectives of the Public Health Alliance of Genomic Epidemiology (PHA4GE) is to equip those working in the public health sector with the necessary skills and knowledge to efficiently and effectively respond to disease outbreaks. A key component of our courses is the inclusion of domain experts from diverse communities, whose insights and experiences provide invaluable perspectives that enrich the learning process and ensure that what is being integrated is relevant for the target group. For our wastewater surveillance bioinformatics course, participants are presented with an outbreak scenario which imitates the process of forming a response in the case of a real-world event, with the overarching aim of using the skills and knowledge acquired to build a wastewater surveillance health dashboard. The foundational concepts such introduction to wastewater surveillance epidemiology, and waste water surveillance bioinformatics use a pedagogical approach (Bansal et al., 2020), where lessons are broken down into short, modular presentations. A combined direct learning model and cooperative learning model (Guzmán and Payá, 2020) allow participants to rapidly engage with the material to grasp concepts such as public health and community impact and to develop fundamental analytical skills. Andragogical strategies (Bansal et al., 2020) such as outbreak case studies and discussions dispersed between modules, further enhance the peer learning (Tullis and Goldstone, 2020) process by promoting teamwork and idea sharing through the application of acquired knowledge. The main learning approach involves following an avatar who needs to solve day-to-day bioinformatics problems. This incorporates an interactive and playful learning experience, where tasks become progressively more challenging, thus incorporating both narrative and game-based learning approaches (Breien and Wasson, 2021). For the final project, participants engage in a heutagogical exercise, where they are required to build a dashboard in pairs, incorporating all their accumulated skills and knowledge. Participants have to decide which of the principles, techniques and tools they have learned throughout the course are relevant to the assignment. Heutagogy is also a key component of online curriculum, empowering participants to take increasing responsibility for their own learning (Mwinkaar and Lonibe, 2024). This not only develops technical proficiency but also facilitates self-directed learning and problem-solving. Additionally, this reinforces the effective application of bioinformatics solutions in wastewater surveillance. These are just some of the learning approaches which have consolidated diverse teaching strategies and models, facilitating the inclusivity of people with different learning styles, varying attention spans, and time constraints. The course's modular design allows for focused exploration of topics relevant to public health practitioners and academics. The training program is designed so that participants are expected to systematically and progressively develop expertise in public health and computational techniques.

Incorporating RNA sequencing bioinformatics pipeline into the Undiagnosed Disease Program

Jana Van der Westhuizen (Stellenbosch University) and Shahida Moosa (Stellenbosch University, Medical Genetics Tygerberg Hospital)

Approximately 3.5-6% of the human population has a rare disease, globally, affecting 4.2 million South Africans. There are approximately 7000 rare diseases, with an estimated 80% of these thought to have a genetic basis. Establishing an accurate molecular diagnosis for a Mendelian disease unlocks valuable information to the patient, empowering them with guidance on management and therapy, informing family planning decisions, and enabling personalised treatment options. RNA sequencing allows for the identification of alternative splicing, aberrant expression levels and mono-allelic expression of variants. This reveals the mechanisms underlying variant pathogenicity at the transcriptomic level and can help reclassify variants of uncertain significance (VUS) as likely pathogenic. It also enables the discovery of pathogenic variants that may be missed by exome sequencing. This study will incorporate a RNA sequencing bioinformatics pipeline to elucidate unsolved cases in the Undiagnosed Disease Program, based at Tygerberg Hospital. A candidate VUS in FREM2 was identified with previous exome analysis, in an unresolved affected child, suspecting Fraser syndrome. RNA sequencing of both the affected child and unaffected mother will potentially reclassify the VUS as likely pathogenic. A second case clinically indicated Sotos syndrome, but no candidate variants were found in the exome. Prior genome analysis identified a deep intronic variant in NSD1, warranting the use of RNA sequencing to identify probable novel splice variants. An in house RNA-sequencing bioinformatics pipeline will therefore be used to identify probable pathogenic splicing abnormalities. This project will contribute evidence to reclassify a suspected pathogenic VUS found by genetic testing and identify possible variants missed by prior clinical genetic testing, ending the diagnostic odyssey for these patients.

cOmicsArt – a customizable Omics Analysis and reporting tool

Lea Seep (University of Bonn), Paul Jonas Jost (University of Bonn), Clivia Lisowski (University of Bonn), Hao Huang (University of Bonn), Stephan Grein (University of Bonn), Hildigunnur Hermannsdottir (Technical University of Munich), Katharina Kuellmer (Technical University of Munich), Tobias Fromme (Technical University of Munich), Martin Klingenspor (Technical University of Munich), Elvira Mass (University of Bonn), Christian Kurts (University of Bonn) and Jan Hasenauer (University of Bonn)

The availability of bulk-omic data is steadily increasing, necessitating collaborative efforts between experimental and computational researchers. While software tools with graphical user interfaces (GUIs) enable rapid and interactive data assessment, they are limited to pre-implemented methods, often requiring transitions to custom code for further adjustments. However, most available tools lack GUI-independent reproducibility such as direct integration with R, resulting in very limited support for transition. We introduce the customizable Omics Analysis and reporting tool – cOmicsArt. cOmicsArt aims to enhance collaboration through integration of GUI-based analysis with R. The GUI allows researchers to perform user-friendly exploratory and statistical analyses with interactive visualizations and automatic documentation. Downloadable R scripts and results ensure reproducibility and seamless integration with R, supporting both novice and experienced programmers by enabling easy customizations and serving as a foundation for more advanced analyses. This versatility also allows for usage in educational settings guiding students from GUI-based analysis to R Code. Availability: cOmicsArt is freely available at https://shiny.iaas.uni-bonn.de/cOmicsArt/ User documentation is available at https://icb-dcm.github.io/cOmicsArt/ Source code is available on GitHub https://github.com/ICB-DCM/cOmicsArt A docker image can be retrieved from https://hub.docker.com/r/pauljonasjost/comicsart/tags A snapshot upon publication can be found on Zenodo: https://zenodo.org/records/13740904 A screen recording of cOmicsArt is available at: https://www.youtube.com/watch?v=pTGjtIYQOak p

PHYLOGENETIC RECONSTRUCTION OF SOME BACTERIAL SPECIES FROM THE ORAL FLORA OF DOGS AND THE ANTIBIOFILM POTENTIALS OF EMULSIONS OF BASIL AND GINGER ESSENTIAL OILS ON THE BACTERIA ISOLATES

Mukhtar Adeiza Suleiman (Department of Biochemistry, Ahmadu Bello University Zaria), Diana Obistioiu (Faculty of Agriculture, University of Life Sciences "King Michael I" from Timisoara), Mohammed Auwal (Department of Biochemistry, Ahmadu Bello University Zaria), Ibrahim, Anca Hulea (Faculty of Agriculture, University of Life Sciences "King Mihai I" from Timisoara), Mohammed Nasir Shuaibu (Department of Biochemistry, Ahmadu Bello University Zaria), Iuliana Popescu (Faculty of Agriculture, University of Life Sciences "King Michael I" from Timisoara) and Jacob Kwada Paghi Kwaga (Department of Veterinary Public Health and Preventive Medicine, Ahmadu Bello University Zaria)

The continuous occurrence of antimicrobial resistance has advanced microbial infections leading to public health alert. Biofilm-associated virulence is an important mechanism microorganisms utilize to escape the desired effects of antibiotics and other therapeutic regimen. The current study aimed to determine the antibiofilm effects of two essential oil-based emulsions prepared to inhibit and eradicate both preformed and formed biofilms of Staphylococcus, Streptococcus and Pasteurella species isolated from the oral cavity of dogs. The species delineation of the bacterial phylogenetic reconstruction will involve the 16S ribosomal ribonucleic acid (rRNA) gene. The emulsions of basil and ginger essential oils were prepared to target the biofilms of the three bacteria species. The efficacy of the emulsions was elucidated through molecular docking analysis to demonstrate the activity of compounds in the oils as identified by Gas Chromatography-Mass Spectrometry (GC-MS). Results in the phylogenetic analysis revealed distinct cluster for Staphylococcus species, Streptococcus species and Pasteurella species with a strong coverage of their respective clades. Antibiofilm analysis showed lower concentrations of emulsions of basil and ginger essential oils were able to inhibit (70% maximum, 39% minimum) and eradicate (67% maximum, 4% minimum) biofilms of Staphylococcus, Streptococcus and Pasteurella species, respectively. Furthermore, the forty-six (46) compounds identified in basil and ginger essential oils were docked against the crystal structure of four proteins involved in bacterial biofilm formation (sortase A, sortase B, master biofilm regulator, biofilm-associated protein). Interestingly, docking outcome showed six compounds, 5-Hepten-2-one, 6-methyl-, p-Anisic aldehyde, trifluoroacetic acid, cyclohexyl ester, linalool acetate, alpha-citral, geranyl acetate had high binding affinity to the four protein targets which are stabilized by the free binding energy, strong complex formation, hydrogen bonds and hydrophobic interactions with key amino acid residues of the proteins. This study highlights the potential use of basil and ginger essential oil emulsions towards reducing biofilm production, thus serving as alternative natural antimicrobial agents for these species of bacteria.

Making models work: Matching Human in vitro models to deliver precision medicine

Pourya Naderi , Sang Su Kwak, Mehdi Jorfi, Weiming Xia, Rudolph Tanzi, Doo Kim and Winston Hide (Harvard Medical School)

Objectives Successful Alzheimer's disease (AD) interventions in preclinical models often fail in human trials. While preclinical models offer insights into AD mechanisms, there is no systematic approach to verify whether preclinical target mechanisms retain therapeutic relevance in humans. Bridging this preclinical-to-clinical translational gap accelerates therapeutic development by precisely addressing whether failures are due to testing ineffective drugs, targeting the wrong mechanism, or relying on unrepresentative models. Methods We have developed a novel bioinformatics platform, named Integrative Pathway Activity Analysis (IPAA), that maps pathway activity from omics data. IPAA precisely captures the degree to which disease functions in models match those in human brains and prioritizes targetable pathways in the most representative models. We assessed the mechanistic similarities between the transcriptomes of three AD brain regions and multiple 2D/3D human AD cellular models to define targetable functions. We performed phosphoproteomics analysis and compared pathway activity changes with transcriptomic findings. Top pathways were pharmacologically evaluated for their impact on AD pathology in 3D models. Results IPAA found high correlation of pathway dysregulation between brain regions (r=0.84, temporal cortex and parahippocampal gyrus), suggesting IPAA's ability to detect conserved AD functions. IPAA found 83 dysregulated transcriptomic pathways shared between AD brains and a 3D model with a high Amyloid-beta (Aβ) 42/40 ratio. Shared dysregulated pathways included p38 MAPK, YAP1/TAZ, E-cadherin, CDC20, and APC/C, which were confirmed at the protein level. Elevated active p38 MAPK was observed in the 3D models, human AD brains, and 5XFAD mice, localized to presynaptic dystrophic neurites. Phosphoproteomic analysis confirmed an increase in p38 MAPK substrate phosphorylation driven by Aβ42 accumulation. Targeting p38 MAPK with a clinical p38α/β MAPK inhibitor (Losmapimod)– which has not been tested for AD– significantly reduced Aβ-induced tau, Aβ accumulation, neuronal loss, and microglial activation in 3D models and human microglia. We further found that MAPK-activated protein kinase 2 (MK2) plays crucial roles in mediating Aβ-induced tau pathology. Conclusions IPAA enables rapid preclinical assessment of target pathways with confidence for impact on AD pathology prior to clinical trials. Our findings highlight the critical role of protein kinase networks, particularly the p38 MAPK-MK2 axis, in driving AD pathology in humans.

Linkage of ApoE gene polymorphisms with the risk of ACVD Intype 2 diabetes patients at Gauteng province in South Africa

Siphesihle Mkhwanazi, Mashudu Nemukula, Tumelo Maphetho and Stanley Sechene Gololo
(Sefako Makgatho Health Sciences University)

Background Lipoprotein-related metabolisms have been associated with the damage of cardiovascular system in diabetic patients, particularly in type 2 diabetic mellitus (T2DM) patients. ApoE gene play a significant role in the metabolism of lipoproteins Higher levels of total cholesterol (TC), triglycerides (TG), low density lipoprotein cholesterol (LDL-C), non-high-density lipoprotein (non-HDL-C), ApoB-100, and very low-density lipoprotein (VLDL-C) were observed in the T2DM+ACVD group compared to both T2DM-ACVD and control groups. The distribution of ApoE3 allele showed no significant differences across all the study group Conclusion The study confirms that ApoE gene polymorphisms influence ACVD risk in T2DM patients, displaying multifaceted interactions with metabolic risk factors. Notably, age and diabetes duration (>10 years) were found to be significant risk factors for ACVD among T2DM, highlighting prolonged hyperglycaemia's impact. Surprisingly, the E2/E3 genotype and ApoE2 allele, were associated with increased risk, while the E4/E4 genotype, usually a risk factor, appeared to be protective.

A novel normalisation approach for RNA-Seq data to detect cancer progression and prognostic subtypes.

Hocine Bendou and Michelle Livesey
(University of Cape Town)

Cancer heterogeneity, characterised by omics diversity, plays a critical role in tumour progression, therapeutic response and clinical outcome. Among its dimensions, intertumour heterogeneity, the variability between tumours across patients or within the same individual, poses significant challenges for accurate prognosis and treatment selection. To address this complexity, a novel RNA-Seq normalisation method developed by my research group uses early-stage cancer samples as a reference to adjust gene expression levels in advanced-stage tumours. By calculating expression ratios between different stages, this method enhances the detection of progression-related gene expression changes, refines patient stratification and improves the identification of key molecular drivers of cancer progression. To assess the effectiveness of this approach, the normalisation method was applied to transcriptomic data from renal clear cell carcinoma (RCC). Hierarchical clustering of the normalised gene expression profiles identified three distinct patient subgroups, each with unique molecular characteristics. Notably, these subgroups showed significant differences in survival outcomes, suggesting that the normalised expression data can reveal novel cancer subtypes with different prognoses. These findings highlight the importance of accounting for tumour progression in transcriptomic analyses to improve precision oncology efforts. Integrating stage-appropriate normalisation into RNA-Seq analysis can refine cancer classification, uncover progression-related biomarkers and improve patient stratification for targeted therapies. Applying this method to other cancer types may further improve our understanding of tumour evolution and heterogeneity.

textToKnowledgeGraph: Generation of Molecular Interaction Knowledge Graphs Using Large Language Models for Exploration in Cytoscape

Favour James (Department of Electronic and Electrical Engineering, Obafemi Awolowo University), Christopher Churas (Department of Medicine, University of California San Diego), Trey Ideker (Department of Medicine, University of California San Diego), Dexter Pratt (Department of Medicine, University of California San Diego) and Augustin Luna (National Library of Medicine and National Cancer Institute)

Motivation Knowledge graphs (KGs) are powerful tools for structuring and analyzing biological information due to their ability to intuitively represent data and improve query performance across heterogeneous datasets. However, constructing KGs from unstructured scientific literature remains challenging due to the high cost and expertise required for manual curation. Prior works have explored text-mining techniques to automate this process but have limitations that impact their ability to capture complex biological interactions fully. Traditional text-mining methods struggle with understanding context across sentences. Additionally, these methods lack expert-level background knowledge, making it difficult to infer relationships that require awareness of biological concepts indirectly described in the text. Large Language Models (LLMs) present an opportunity to overcome these challenges. LLMs are trained on large amounts of diverse biological literature, equipping them with contextual knowledge that enables more accurate extraction. Additionally, LLMs can process the entirety of an article's text, capturing relationships across several sections rather than analyzing sentences in isolation; this allows for more precise extraction. Results To address these challenges, we present textToKnowledgeGraph (https://pypi.org/project/texttoknowledgegraph), an artificial intelligence (AI) tool using LLMs to extract interactions from individual publications directly in Biological Expression Language (BEL). BEL was chosen for its compact and detailed representation of biological relationships, allowing for structured and computationally accessible encoding. The tool provides two usage modes: 1) a Python package usable through the command line or within other projects, or 2) an interactive application within Cytoscape Web to simplify extraction and online exploration. In the text processing pipeline, we leverage LangChain with GPT-4o for information extraction using a predefined schema implemented with Pydantic to ensure structured outputs for BEL generation. The extracted BEL statements are outputted in CX2 format, enabling visualization and exploration within the Cytoscape ecosystem. Additionally, the ndex2 package is used for CX2 conversion and to support optional storage and sharing of extracted networks on NDEx. In this initial version of textToKnowledgeGraph, we only support the extraction of interactions into BEL. Future updates will enable greater customization, making it more adaptable for broader applications. To evaluate the accuracy of extracted interactions, we applied textToKnowledgeGraph to various published articles. The extracted interactions were manually reviewed by BEL experts, ensuring the biological accuracy and completeness of captured relationships. Finally, we present a use case example in which a topic-specific BEL knowledge graph provides relevant information to augment queries to an LLM using a technique known as Graph Retrieval Augmented Generation (Graph RAG).

Leveraging Data Balancing and Chemical Encoding Strategies for Robust AI-Based Drug Discovery Pipeline

Ons Masmoudi (Laboratory of Molecular Epidemiology and Experimental Pathology, Institut Pasteur de Tunis), Afef Abdelkrim (Research Laboratory Smart Electricity & ICT, National Engineering School of Carthage, University of Carthage) and Emna Harigua-Souiai ((Laboratory of Molecular Epidemiology and Experimental Pathology, Institut Pasteur de Tunis)

Artificial intelligence (AI) has emerged as a revolutionary approach in the field of drug discovery, with the increased availability of large datasets for training AI models to predict the properties and potential biological activities of chemical compounds. The AI-driven framework essentially consists of three main components: the dataset, the combination encoding system-model, and the prediction task. The present work introduces an AI-based Ligand-Based Drug Design approach focused on optimizing the different components of such a pipeline to provide robust predictive tools of chemical compound activities against various diseases. In this study, we investigated the impact of class imbalance on the performance of various classifiers in predicting the biological activity of chemical compounds. We trained two machine learning models, four graph-based models, and two pre-trained models on highly imbalanced bioassay datasets. To address the class imbalance, we first employed two oversampling methods namely Random Oversampling (ROS) and SMOTE and two undersampling methods namely Random Undersampling (RUS) and NearMiss. Additionally, we proposed a novel strategy called K-Ratio Undersampling. Through this approach, based on RUS, we created three specific ratios (1:50, 1:25, and 1:10) for each dataset. The impact of these ratios on model performances was evaluated using F1-scores. To ensure the robustness of our models, we conducted an external validation on unseen data. As a last step, we performed an analysis of each dataset content to better understand the factors behind the models' misclassifications. Across all simulations, the comparison of the classical resampling techniques revealed that RUS outperforms ROS across various evaluation metrics, supporting our hypothesis that reducing majority class instances through undersampling improves model performance. Through the investigation of the impact of the various imbalance ratios on the ML and DL models, we demonstrated that moderate imbalance ratios of (1:25 - 1:10) significantly enhanced the models performances, achieving higher F1-scores compared to previous results. Among the evaluated models, the top-performing models for each dataset were optimized through hyperparameter tuning. The external validation step confirmed that the 10-RUS configuration yielded the best configuration in achieving a good balance between true positive and false positive rates. Although no particular model showed optimal performances on all datasets. Through the previous results, the HIV dataset was particularly challenging. The analysis of the similarity between active and inactive compounds through a chemical space network showed that high similarity between both classes reduced predictive accuracy. Our findings highlighted the importance of optimizing both the chemical data content and the class imbalance to improve the model performances in predicting the biological activity of chemical compounds.

Surveillance Capacity Building through Pathogen Genomics and Bioinformatics Training Across Africa

Siddiqah George, Kirsty Lee Garson, Tony Yiqun Li, Perceval Maturure and Nicola Mulder
(NGS Academy for the Africa CDC)

Abstract: The recent emergence and re-emergence of infectious diseases in Africa highlight the critical need for robust pathogen genomic surveillance systems across the continent. Effective surveillance depends on comprehensive training and capacity development in pathogen genomics and bioinformatics, as rapid public health responses to disease outbreaks rely on continuously enhancing these essential skills. To ensure quality and consistency in training, the development and implementation of a standardised curriculum are crucial; enabling uniform skill-building and knowledge dissemination across diverse regions. Over the past four years, we have delivered hybrid training in pathogen genomic surveillance and bioinformatics to over 290 participants from 36 African countries. These initiatives, tailored to diverse personas in national public health institutions, leveraged trainers and facilitators from across the continent to address varying competency levels. We have also developed and implemented resources to support our training initiatives, including a user-friendly helpdesk ticketing system, a robust trainer database, and intuitive websites hosting training materials. These tools work jointly to ensure that training and related resources are widely accessible, while also providing participants with support and engagement opportunities long after receiving training. To ensure consistency in the training of public health staff in Africa, a standardised pathogen genomics surveillance training curriculum has been developed. The curriculum is designed to serve as a comprehensive resource for trainers, encompassing content that ranges from foundational courses in generic, wet-lab, and bioinformatics topics to advanced pathogen-specific courses that include tailored genomic surveillance workflows. The next step is implementing this curriculum in future training initiatives across African public health institutes. Additionally, we are exploring the integration of AI in pathogen genomics curriculum development and training. Our training efforts have highlighted the need for ongoing training and capacity building in pathogen genomic surveillance in Africa. A standardised curriculum can be used in addressing this need and facilitate consistent skills development and collaboration across the continent's public health institutes. Implementing this curriculum and exploring AI-driven training and decision-making will enhance preparedness for future disease outbreaks and public health responses.

Genetic Diversity and Spatiotemporal Distribution of SARS-CoV-2 Variants in Guinea: A Meta-Analysis of Sequence Data (2020–2023)

Thibaut Armel Chérif Gnimadi (Centre de Recherche et de Formation en Infectiologie de Guinée (CERFIG)), Kadio Jean-Jacques Olivier Kadio (Centre de Recherche et de Formation en Infectiologie de Guinée (CERFIG)), Mano Joseph Mathew (EFREI Research Lab, Panthéon Assas University), Castro Gbêmêmali Hounmenou (Centre de recherche et de formation en infectiologie de Guinée (CERFIG)), Nicolas Fernandez-Nuñez (Institut de Recherche pour le Développement (IRD), INSERM, TransVIHMI, University of Montpellier), Nicole Vidal (Institut de Recherche pour le Développement (IRD), INSERM, TransVIHMI, University of Montpellier), Ahidjo Ayouba (Institut de Recherche pour le Développement (IRD), INSERM, TransVIHMI, University of Montpellier), Martine Peeters (Institut de Recherche pour le Développement (IRD), INSERM, TransVIHMI, University of Montpellier), Abdoulaye Touré (Centre de Recherche et de Formation en Infectiologie de Guinée (CERFIG)) and Alpha Kabinet Keita (Centre de Recherche et de Formation en Infectiologie de Guinée (CERFIG))

In Guinea, genomic surveillance has been established to generate sequences of and to identify locally circulating SARS-CoV-2 variants. This study aims to describe the distributions, genetic diversity, and origins of SARS-CoV-2 lineages circulating in Guinea during the COVID-19 pandemic. A migration analysis was performed by selecting all sequences generated in Guinea for variants of concern and interest. From March 2020 to December 2023, 1038 sequences were generated in Guinea and submitted to the Global Initiative on Sharing All Influenza Data (GISAID) database. Of these, 73.1% corresponded to SARS-CoV-2 variants of concern, which were further grouped into Omicron (69.4%), Delta (21.9%), Alpha (6.6%), and Eta (2.1%). Other variants accounted for 26.9% of the total. Among the total variants analyzed, 75 importations into Guinea from various countries worldwide were identified. Most of the importations (40%) originated from African countries, followed in significance by those from European countries (25.3%) and Asia (18.6%). A significant migratory flow was observed within Guinea. The genomic surveillance reported in this study revealed the diversity of SARS-CoV-2 variants circulating in Guinea, emphasizing the importance of large-scale sequencing analyses in understanding the dynamics of the pandemic.

Emergence of novel mosaic G9P[6] rotaviruses through multiple intragenogroup reassortment events post vaccine introduction in Blantyre Malawi

Chimwemwe Mhango (Malawi-Liverpool-Wellcome Programme), End Chinyama (Malawi-Liverpool-Wellcome Programme), Ernest Matambo (Institute of Infection, Veterinary and Ecological Sciences, University of Liverpool), Landilani Gauti (Malawi University of Science and Technology), Flywell Kawonga (Institute of Infection, Veterinary and Ecological Sciences, University of Liverpool), Benjamin Kumwenda (School of Life Sciences and Allied Health Professions, Kamuzu University of Health Sciences), Arox Kamng'Ona (School of Life Sciences and Allied Health Professions, Kamuzu University of Health Sciences), Celeste Donato (The Peter Doherty Institute for Infection and Immunity) and Khuzwayo Jere (Institute of Infection, Veterinary and Ecological Sciences, University of Liverpool)

Background: Rotavirus remains a leading cause of severe gastroenteritis in children under five, particularly in low- and middle-income countries (LMICs). In Malawi, G9P[6] strains re-emerged in 2017, five years after the introduction of Rotarix rotavirus vaccine, necessitating an in-depth investigation of their genetic diversity, evolutionary origins, and public health implications. Methods: Using whole-genome sequencing (WGS), we analysed and assigned complete genotype constellations and employed phylogeographic and phylogenetic network analyses to trace the evolutionary pathways of G9P[6] strains (n=11) between 2017 to 2022. Findings: The re-emergent G9P[6] strains were characterised by a DS-1-like G9-P[6]-I2-R2-C2-M2-A2-N2-T2-E2-H2 genotype constellation. Phylogeographic analysis of the VP7 gene revealed monophyletic clustering with contemporary G9P[6] strains from Mozambique. Phylogenetic network analysis demonstrated high genetic similarity of the inner capsid and non-structural genes of G9P[6] strains to previously circulating Malawian G2P[4], G2P[6], G3P[4], and G3P[6] strains. Time-resolved phylogenies dated the most recent common ancestor for the inner capsid and non-structural genes between 2009–2015. Evolutionary analysis suggested lineage spillover events associated with the VP6 segment Conclusion: This study, for the first time in Malawi, elucidates the role of reassortment and zoonotic transmission in the re-emergence of G9P[6] strains. These findings highlight the evolutionary dynamics of rotaviruses and the need for continuous genomic surveillance. Considering the limited heterotypic protection provided by the Rotarix (G1P[8] strain) vaccine, tailored vaccination strategies and ongoing vaccine effectiveness studies are critical to addressing the emergence of novel rotavirus strains and improving vaccine performance in LMICs

Dual Effect of the Tumor Microenvironment on Newly Identified Subtypes of Colon Cancer

Christianah Kehinde (Computational Biology Division), Bendou Hocine (Computational Biology Division) and Michelle Livesey (Department of Pathology) (Faculty of Health Sciences, University of Cape Town)

Colon cancer is a highly heterogeneous disease, marked by substantial intra-tumor and inter-tumor variability. Investigating transcriptomic profiles can offer deeper insight into this heterogeneity. However, most genome-transcriptome studies on colon cancer have primarily focused on examining primary tumors and matched normal tissues, often neglecting the multi-stage progression from early to late stages of the disease. The research aimed to establish unique molecular colon subtypes based on the progression in transcriptomic profiles. Additionally, to investigate the implicating factors, such as mutation and tumor microenvironment (TME), affecting the progression of colon cancer and the implications in diagnosis and therapy. RNA-sequencing data from 39 colon cancer patients were obtained from the UCSC Xena database. An in-house, novel normalization method was applied to capture the heterogeneity in the early- to late-stage colon cancer development. Hierarchical clustering revealed colon subtypes with varying progression, and differentially expressed genes (DEGs) between these subtypes were identified with Limma. The DEGs were then subjected to Recursive Feature Elimination (RFE) and mutational analysis to reveal driver genes. The tumor microenvironment was assessed using the EPIC tool, and biological pathways were analyzed using the clusterProfiler R packages. Finally, an independent GSE17538 dataset was used to validate the study. Two novel colon subtypes were identified, comprising 23 and 16 samples, designated as L (Large) and S (Small), respectively. Each subtype displayed heterogeneous transcriptomic profiles. A total of 1,855 DEGs were identified, interestingly, all were downregulated in the S subtype. Two significant enrichment pathways and varied mutations in cancer driver genes were identified in both subtypes. The concurrent downregulation of oncogenes and tumor suppressor genes (TSGs) were discovered, with a link to the dual functionality of CD4 and CD8 T-cells in the TME. This study supports the existence of a complex relationship between TME and gene expression in explaining the molecular heterogeneity of novel colon cancer subtypes. In the S subtype, TME played a dual anti-tumorigenic and pro-tumorigenic role, ensuring a balanced progression that could not be identified without the normalization method. The findings provide insights into colon cancer oncogenesis, with implications for improved prognosis and development of targeted therapies.

Predicting SARS-CoV-2 variant-specific infection risk using a joint survival model

Clemens Peiter (Life and Medical Sciences (LIMES) Institute and Bonn Center for Mathematical Life Sciences, University of Bonn), Simon Merkt (Life and Medical Sciences (LIMES) Institute and Bonn Center for Mathematical Life Sciences, University of Bonn), Solomon Ali (Saint Paul's Hospital Millennium Medical College), Esayas Kebede Gudina (Jimma University Clinical Trial Unit, Jimma University Institute of Health), Kira Elsbernd (Division of Infectious Diseases and Tropical Medicine, LMU University Hospital), Andreas Wieser (Division of Infectious Diseases and Tropical Medicine, LMU University Hospital), Arne Kroidl (Division of Infectious Diseases and Tropical Medicine, LMU University Hospital), Jan Hasenauer (Life and Medical Sciences (LIMES) Institute and Bonn Center for Mathematical Life Sciences, University of Bonn) and Covicis Team (Centre Hospitalier Universitaire Vaudois)

Despite the end of COVID-19 as a global public health emergency in 2023, SARS-CoV-2 remains a threat due to its rapidly changing nature. Several variants have emerged after 2023 with some being classified as variants of concern or interest by the WHO. Thus, it is crucial to understand SARS-CoV-2 variant-specific immunity to enable the quantification of infection risk with newer variants. We established a mathematical model to quantify an individual's SARS-CoV-2 infection risk [1]. The model is based on a standard survival modeling framework. It depends on an individual's neutralization capacity against different SARS-CoV-2 variants, which is described by a mixed effects model, and the current number of infected individuals. We calibrated this model on individual neutralization and infection data from Munich and used it to predict the risk of infection following a COVID-19 vaccination. Following our previous work, we now aim to investigate the immune landscape of 202 individuals from the CoVICIS cohort (https://covicis.eu/) [2]. The dataset encompasses four rounds of sample collection between November 9, 2020, and October 30, 2023 from healthcare workers and community members in Jimma and Addis Ababa, Ethiopia. For each sample the neutralizing antibody titers against eleven SARS-CoV-2 variants were assessed. Building on both the available data from CoVICIS and our model, we extend the model to incorporate newer SARS-CoV-2 variants and predict the SARS-CoV-2 infection risk for each individual over time. [1]: Ahmed, M. I. M., Einhauser, S., Peiter, C., Senninger, A., Baranov, O., Eser, T. M., Huth, M., Olbrich, L., Castelletti, N., Rubio-Acero, R., Carnell, G., Heeney, J., Kroidl, I., Held, K., Wieser, A., Janke, C., Hoelscher, M., Hasenauer, J., Wagner, R., & Geldmacher, C. (2024). Evolution of protective SARS-CoV-2-specific B and T cell responses upon vaccination and Omicron breakthrough infection. In iScience (Vol. 27, Issue 6, p. 110138). Elsevier BV. https://doi.org/10.1016/j.isci.2024.110138 [2]: Merkt, S., Ali, S., Gudina, E. K., Adissu, W., Gize, A., Muenchhoff, M., Graf, A., Krebs, S., Elsbernd, K., Kisch, R., Betizazu, S. S., Fantahun, B., Bekele, D., Rubio-Acero, R., Gashaw, M., Girma, E., Yilma, D., Zeynudin, A., Paunovic, I., … Wieser, A. (2024). Long-term monitoring of SARS-CoV-2 seroprevalence and variants in Ethiopia provides prediction for immunity and cross-immunity. In Nature Communications (Vol. 15, Issue 1). Springer Science and Business Media LLC. https://doi.org/10.1038/s41467-024-47556-2

Ten Simple Rules for Developing Robust Pipelines for Public Health Applications

Kevin Libuit (Theiagen Genomics), Gerald Mboowa (Broad Institute of Harvard and MIT), Sarai Varona (Instituto de Salud Carlos III & Universidad Nacional de Educación a Distancia (UNED)) and Jamie Southgate (Public Health Alliance for Genomic Epidemiology)

Bioinformatics pipelines play a crucial role in public health, particularly regarding pathogen surveillance and outbreak investigations. For these pipelines to be effective, they should be accessible, transparent, and sustainable. The following paper presents ten fundamental principles for developing robust bioinformatics pipelines for public health applications. These principles align with best practices outlined by the Public Health Alliance for Genomic Epidemiology (PHA4GE) and emphasize open-source development, workflow standardization, and testing methodologies. Key recommendations include making code publicly available, adopting open-source licensing, implementing semantic versioning, and integration into workflow management systems. Additionally, the guidelines further stress the importance of containerization, standardized file formats, validation frameworks, and clear and concise documentation to enhance reproducibility and usability. In following these principles, software developers can create pipelines that not only meet immediate public health needs but also assist global efforts in outbreak preparedness and health security. The framework provides actionable guidance for developers and reviewers, ensuring that pipelines remain both reliable and impactful in addressing current and emerging public health challenges.

Phylogenetic analysis of extensively drug-resistant Tuberculosis strains circulating in Western Cape Province: 2010 to 2019

Justice Tresor Ngom Ngom, Johannes Loubser, Zainab Kashim-Bello, Marisa Klopper, Robin Warren and Elizabeth Streicher (Division of Molecular Biology and Human Genetics, Faculty of Medecin and Health Sciences, Stellenbosch University)

Background: Extensively drug-resistance tuberculosis (XDR-TB) is the most expensive, complex to diagnose, and deadliest form of drug-resistant TB. It is a public health concern in the Western Cape Province. Little is known about the phylogenetic relationships of Western Cape XDR-TB strains. Understanding the evolutionary history and relationships among prevalent strains in the province will allow targeted public health intervention. We aim to determine resistance mutations and assess the phylogeny of lineage 2.2.1 and lineage 2.2.2 strains in the Western Cape province over a period of ten years. Method: Our study included 729 whole genome sequences from bacteriologically confirmed XDR-TB. Genomic drug resistance profiles were identified using the TB profiler pipeline. Multi-sequence alignments (MSA) were produced through the MTBseq pipeline and used for phylogenetic tree reconstruction using IQ-TREE2 v2.2.0.3. We visualized the evolutionary relationships between genotypes by using iTOL. Results: 590 of 729 (80.93%) isolates were genomically classified as XDR-TB. The Lineage 2.2.2 strains (58.79%, n=378) were predominant, followed by the Lineage 2.2.1 strain (40.43%, n=260). Lineage 2.2.2 strains had diverse resistance variants compared to the Lineage 2.2.1 strains. Mapping resistance mutations on the phylogenetic tree revealed that the spread of Lineage 2.2.2 strains occurred following the evolution of pre-XDR (MDR+SLIDs), with subsequent evolution to XDR. In contrast, Lineage 2.2.1 strains have spread as MDR with subsequent evolution to XDR. Conclusion: Both lineages are now endemic, and strains are frequently transmitted within our setting. The dynamics of transmission differ by strain type, even between lineages that are closely related.

Leveraging Artificial Intelligence for Predicting Human-Viral Protein-Protein Interactions: A Benchmarking Study to Address Key Challenges

Chaima Hkimi, Selim Kamoun, Oussema Khamessi and Kais Ghedira (Laboratory of Bioinformatics, Biomathematics and Biostatistics (LR20IPT09), Pasteur Institute of Tunis)

Introduction: Viral infections pose significant global health challenges, with human-viral protein-protein interactions (HV-PPIs) playing a central role in infection mechanisms and host immune responses. While experimental methods for studying HV-PPIs are resource-intensive, computational approaches, particularly machine learning (ML), offer scalable and efficient alternatives. Methodology: Here, we present a benchmarking study evaluating the performance of various ML models in predicting HV-PPIs, focusing on three viruses: West Nile Virus (taxon ID: 11082), HIV-1 (taxon ID: 11676), and SARS-CoV-2 (taxon ID: 2697049). We curated positive and negative interaction datasets from six public databases and employed five sequence-based feature encoding methods to represent protein sequences. Six ML classifiers, including SVM and RF, were trained and evaluated using metrics such as accuracy and F1-score. Results: Our results reveal that dataset imbalance significantly impacts model performance, with balanced datasets (1:1 positive-to-negative ratio) yielding more reliable predictions, emphasizing the value of techniques like SMOTE for handling imbalanced real-world data. Encoding methods significantly influence outcomes, with pseudo-amino acid composition (PAAC) (type I), quasi-sequence-order (QSO), and conjoint-triad (CT) encodings showing better generalization for taxon ID "11676". Overfitting was observed in models like GBM, particularly for specific taxonomy IDs, underscoring the need for practices like limiting tree depth and hyperparameter tuning. The primary goal of HV-PPI models is to identify novel interactions. In this study, the SVM model using combination-set features identified 333 human-SARS-CoV-2 interactions, including 75 shared with experimental studies and 82 newly predicted ones. Although SARS-CoV-2 interacts with various host receptors, including ACE2, NRP-1, AXL, CD147, and heparan sulfate, as well as host proteases like FURIN, TMPRSS2, and cathepsins, our interactome revealed potential interactions between the spike (S) protein and TLR4, suggesting a role in antiviral immunity. Additionally, TRIM7 was predicted to interact with NSP12 and NSP7, possibly targeting them for ubiquitination and degradation, which could suppress viral replication. Another key finding was the predicted interaction between ACTN4 and ORF6, which may counteract the antiviral effects of ACTN4-NSP12 binding and facilitate immune suppression and viral replication. Conclusion: These findings highlight the potential of ML in uncovering new HV-PPIs, offering insights into viral pathogenesis and therapeutic targets. However, challenges such as overfitting and small dataset sizes underscore the need for further refinement of ML models and exploration of alternative learning approaches to enhance predictive accuracy and generalizability.

Ethics and governance of AI-powered genomics

Tendayi Mutangadura, Nchangwi Munung and Nicola Mulder
(University of Cape Town)

The integration of artificial intelligence (AI) into genomics promises substantial advancements in personalised medicine, diseases prediction, gene editing but it also presents critical ethical and governance challenges. This study explores these challenges by addressing three main research questions: (1) What are the primary ethical concerns related to AI applications in genomics, including privacy, consent, and bias? (2) How are current governance structures addressing or failing to address these issues? and (3) How can effective governance frameworks be established to ensure responsible, equitable, and transparent use of AI in this field? Using a mixed-methods approach that includes a systematic literature review, expert interviews, and case analysis, the study examines the ethical risks and governance gaps in AI-driven genomic research. Findings indicate significant concerns around data privacy, potential misuse of genetic information, and the exacerbation of existing health disparities due to biased data and algorithms. Additionally, existing regulatory frameworks lack sufficient guidelines to address algorithmic accountability, data ownership, and inclusive representation within genomic datasets.The study concludes by recommending a multi-stakeholder governance model that emphasizes transparency, fairness, and adaptability. This framework would involve guidelines for data handling, bias mitigation, and global collaboration among governments, private sectors and global health organizations. It provides actionable steps to establish ethical oversight in the evolving landscape of AI-driven genomics.These recommendations aim to enhance public trust and ensure that AI's role in genomics aligns with ethical standards that protect individual rights and foster equitable health outcomes.

The Gene catalogue and functional analysis of the gut microbiome of lions in Etosha National Park

Carl Belger (School of Animal, Plant and Environmental Sciences, University of the Witwatersrand), Robyn Hetem (chool of Biological Sciences, University of Canterbury) and Scott Hazelhurst (Sydney Brenner Institute for Molecular Bioscience, University of the Witwatersrand)

The gut microbiome is important for the health of all animals. Very little is known about the microbiome of large carnivores, and lion in particular. This study presents the first comprehensive microbiome classification of African lions (Panthera leo melanochaita). Our study used shotgun metagenomics data (Illumina short-reads) of DNA extracted from faecal samples from 20 lions from Etosha National Park, Namibia. Three of the lions were sampled twice, in different seasons. In addition, 10 of the samples were sent for long-read sequencing (Oxford Nanopore). Our findings illuminate potential connections between gut microbiome composition and social structure, diet, and pack-hunting in carnivores, with potential implications for wildlife conservation and veterinary medicine. We discovered distinct microbial profiles in the African lion gut, dominated by the genera Bacteroides and Phocaeicola. In particular, we note similar abundances of Bacteroides in other pack-hunting carnivores such as black-backed jackals, wolves and dholes. Solitary hunters like cheetahs on the other hand, have a relatively low abundance of Bacteroides. The high abundance of this genera is possibly caused by the high interaction (and therefore transmission of bacteria) between pack-hunters compared to solitary carnivores. Alternatively, Bacteroides abundance could be attributed to differences in diet: solitary hunters consume the most nutritious portions of prey immediately, while pack hunters usually distribute resources based on social hierarchies. Moreover, links were drawn between pregnancy and inflammation to the gut microbiome of female lions via the genus Fusobacterium. This genus is seen in high abundance in post-natal pigs and is also linked to gut inflammation in humans and pigs, indicating that postnatal lions may experience similar gut inflammation during and after pregnancy. For comparison, we analyzed three additional samples from Asiatic lions (Panthera leo leo) collected from previous research in India and found that these sub-species had similar abundances of bacterial phyla but differences in bacterial genera and species. We attribute the similarities in bacterial phyla to common evolutionary ancestry and the differences in bacterial genera to allopatric separation causing minor changes in bacterial composition over time. Finally, a large proportion of DNA in the lion gut was unclassified, representing new species of microorganisms not present in current databases. We were able to create 272 metagenome assembled genomes (MAGs) the majority of which represent new species which will contribute to current knowledge. The identification of novel microbial species highlights the importance of expanding microbial databases and the need for further research into host-microbe interactions in wildlife conservation contexts. Our plan for future research is to leverage long-read data to supplement databases and improve microbial classification.

Machine learning-based prediction of antibiotic resistance in Mycobacterium tuberculosis clinical isolates from Uganda

Sandra Babirye (African Center of Excellence in Bioinformatics and Data-intensive Sciences (ACE)), David Kateete (Makerere University), Gerald Mboowa (African Center of Excellence in Bioinformatics and Data-intensive Sciences (ACE)), Ronald Galiwango (African Center of Excellence in Bioinformatics and Data-intensive Sciences (ACE)), Mike Nsubuga (University of Bristol) and Charles Batte (Lung Institute)

Efforts towards the global tuberculosis (TB) control are challenged by the emergence and spread of Mycobacterium tuberculosis (MTB) resistance to existing anti-TB drugs. Despite the growing application of machine learning (ML) in antimicrobial resistance (AMR) prediction for TB, most of this has been applied with whole genome sequence (WGS) data and in the high-income countries with limited research done in the Low- and Middle-Income Countries (LMICs) like Uganda. This study therefore aimed to leverage ML algorithms to predict drug resistance to four anti-TB drugs: rifampicin (RIF), isoniazid (INH), streptomycin (STM), and ethambutol (EMB) using clinical data variables and WGS data for MTB Uganda isolates. The ML algorithms including Logistic regression (LR), Decision Trees, Extra Trees Classifier, Random Forest, Support Vector Machines, Multilayer Perceptron, Extreme Gradient Boosting (XGBoost), Gradient Boosting (GBC), CatBoost and Adaptive Boosting were implemented using Scikit-learn library. These ML algorithms were trained on a combined dataset comprising of 182 MTB isolates from Uganda with 4994 variants across the entire MTB genome and clinical data variables such as age, sex and HIV status as predictor variables and the phenotypic drug susceptibility testing (DST) data as the outcome variable. The best model was selected based on the highest Mathews Correlation Coefficient (MCC) and Area Under the Receiver Operating Characteristic Curve (AUC) score. LR excelled in predicting RIF (MCC: 0.83 (95% confidence intervals (CI) 0.73–0.86) and AUC: 0.96 (95% CI 0.95–0.98) and STM (MCC: 0.44 (95% CI 0.27–0.58) and AUC: 0.80 (95% CI 0.74–0.82), XGBoost for EMB (MCC: 0.65 (95% CI 0.54–0.74) and AUC: 0.90 (95% CI 0.83–0.96) and GBC for INH (MCC: 0.69 (95% CI 0.61–0.78) and AUC: 0.91 (95% CI 0.88–0.96). Compared to LR, the boosting classifier models (XGBoost and GBC) generalized well on the SA dataset. The benchmarking results revealed that LR for RIF was very sensitive and the GBC for INH and XGBoost for EMB were very specific on the Uganda dataset compared to TB profiler. Moreover, on the SA dataset, TB profiler outperformed all the best models. Among the identified mutations were those in known drug resistance-associated genes, such as rpoB, katG, and rpsL, as well as novel mutations in other genes that require further investigation. HIV status was also identified among the top significant features in predicting drug resistance. Our work shows the relevance of ML algorithms in predicting AMR in pathogens, offering a promising avenue to support robust surveillance systems and advance precision medicine to curb the rising threat of AMR. Additionally, the work demonstrates that integration of diverse data types such as genomic, transcriptomic, proteomics and clinical data could improve resistance predictions while using ML.

Assessing Techniques for the Accurate Linkage of Genomic and Clinical Data in the Western Cape

Themba Mutemaringa (Computational Biology, IBMS, UCT; PHDC, Western Cape Government Health) and Nicki Tiffin (South African National Bioinformatics Institute, University of the Western Cape)

Background Integrating genomic and clinical data is essential for advancing precision medicine and improving patient outcomes [1]. However, in low- and middle-income settings such as the Western Cape Province of South Africa, incomplete electronic medical records present significant challenges to data integration [2,3]. Provincial Health Data Centre (PHDC) facilitates clinical data consolidation using a unique patient identifier for province-wide public health [4]. Still, many datasets lack this identifier, necessitating record linkage techniques. Moreover, integrating genomic data further complicates the process due to the need for precise patient matching. Objective This study aims to (1) construct a comprehensive manually curated dataset to evaluate record linkage algorithms, and (2) systematically compare different record linkage techniques to determine their effectiveness, particularly in handling real-world data challenges such as spelling errors, swapped date formats, missing values, and keyboard key proximity errors. Methods We compiled a labeled dataset of 4,513 records, classifying them into matches (2,327) and non-matches (2,186) based on discrepancies in names, dates of birth (DOB), sex, and other unique identifiers. Name-based mismatches were analyzed using string similarity and phonetic methods, while DOB errors were categorized into common reporting inconsistencies (e.g., swapped day and month, minor numerical differences). Additional complexities, including twin identification and placeholder text issues, were also examined. We evaluated multiple record linkage algorithms using standard performance metrics, including sensitivity, specificity, false positive rate, recall, precision-recall curves, and ROC-AUC scores. Results Preliminary findings reveal that name-related discrepancies account for the highest proportion of mismatches (1,412 minor name variations and 65 cases with entirely different names). DOB inconsistencies were a major challenge, with 292 matched records exhibiting minor errors and 457 cases having substantial mismatches. Twin records posed additional difficulties, as identical birth dates often led to incorrect matches. Initial algorithmic evaluations showed significant variability in performance, with ROC-AUC scores ranging from 0.45 to 0.80, and precision scores between 0.31 and 0.86, highlighting the need for improved linkage methods tailored to local data characteristics. Conclusion Our results underscore the need for flexible record linkage strategies that account for real-world data entry errors while maintaining high accuracy. By systematically identifying the strengths and weaknesses of different algorithms, this study provides critical insights for improving patient data integration in the Western Cape, ultimately supporting better healthcare delivery and genomic research. Future work will refine algorithmic approaches to address region-specific challenges and enhance linkage reliability for precision medicine applications.

The African Population Ontology (AfPO): Building a Framework for representing African Populations

Melek Chaouch (Laboratory of BioInformatics, bioMathematics and bioStatistics (BIMS) Institut Pasteur de Tunis), Anita R. Caron (European Bioinformatics Institute (EMBL-EBI), Wellcome Genome Campus), Afrigen-D The African Population Ontology Project (AfriGen-D), Nicola Mulder (Computational Biology Faculty of Health Sciences University of Cape Town), Danielle Welter (Luxembourg National Data Service) and Alia Benkahla (Laboratory of BioInformatics, bioMathematics and bioStatistics (BIMS) Institut Pasteur de Tunis)

Africa is woven together by population movements, ethno-linguistic diversity, and a unique genetic heritage. Recently, a number of genetics/genomics projects have arisen on the continent, such as those developed in the framework of the Human Heredity and Health in Africa (H3Africa) initiative. With significant ethnic diversity in the continent, researchers are faced with difficulties in defining a population or unit that represents a social group in a standardized format in Africa. Here we developed an African Population Ontology (AfPO) framework that aims to structure this knowledge in a harmonized and standardized way, in order to describe African populations and sub-populations. We used publicly available data related to African populations and their demography, geographic localization, spoken language and genetic background. The Webprotégé platform was used to design and implement this ontology. The country of origin and the populations were selected as descriptive classes. The AfPO was validated by the OBO foundry community and is available in both Github and EBI-OLS. The AfPO enables the annotation of African population groups, and brings together knowledge accumulated about existing populations with their genetic fingerprint in a standardized format; it can be employed to comprehensively annotate African participants in research studies. It can also be used to describe participants of past studies, by mapping them to population identifiers or synonyms. The ontology produced is essential to the study of the history of African populations and their genetics, and is therefore invaluable in addressing public health issues, promoting cultural preservation and fostering a more nuanced appreciation of Africa's unique place in human history.

Using explainable machine learning to enhance breast cancer biomarker discovery in precision medicine

Ouso Daniel (Centre for Research Training in Genomics Data Science, University College Dublin), Annabelle Nwaokorie (Accenture Labs Bioinnovation) and Luca Costabello (Accenture Labs Bioinnovation)

Breast cancer remains the leading cause of cancer mortality in women despite available interventions; however, information on its major genetic drivers is incomplete. We aimed to identify and quantify the impact of the critical genes in breast cancer (BC) pathology for tailored patient management. Numerous signal transduction networks (STNs) in BC have cross-cutting associations. Until now, survival studies and management interventions often consider a few STNs or genes, thus missing a global perspective integral to BC understanding. We hypothesised that integrating STNs information across major BC networks can improve disease understanding and provide application in precision medicine. We included all known major BC pathology STNs to maximise disease heterogeneity in identifying the critical genes. A bi-directional Kaplan-Meier (KM) survival scanning with log-rank statistics was used to triage genes by their expression patterns and select a statistically significant subset of all pathway genes. Moreover, we evaluated the triaged genes, including clinical features, by modelling overall survival (OS) using Cox's proportional hazard (CPH) regression – 79.2% accuracy for the best model. The SHapley Additive exPlanations (SHAP) then quantified feature contributions to model overall survival risk (OSR) predictions. The result is 28 most impactful genes, ranked by relevance, from three gene sets corresponding to the different expression patterns. The top three genes per category were validated through literature and databases. Among them were relatively less-studied but potentially critical genes in BC pathology. For example, DKK4 and KREMEN1. Both belong to families negatively regulating the growth-promoting Wnt/β-catenin pathway. A broadened scope of BC heterogeneity was captured by including all known major networks. Ultimately, we demonstrated important implications in BC clinical management by showcasing a quick, intuitive, and robust overview of patient monitoring for potential healthcare applications.

Prediction of universal stress proteins (USPs) in Mycobacterium tuberculosis

Henry Njoku and Angela Makolo
(University of Ibadan Bioinformatics Group)

The adaptation of Mycobacterium tuberculosis (Mtb) to varied environmental stresses is a fundamental aspect of its pathogenesis and survival. Universal Stress Proteins (USPs) have emerged as pivotal players in this adaptive response, with their expression triggered by an array of stressors. We seek to address the role that USPs play in the adaptation and pathogenicity of Mtb under diverse environmental stress conditions, as well as the significance of USPs in understanding the mechanisms of Mtb survival, and how predictive models can aid in developing targeted interventions against tuberculosis. A novel approach was proposed to predict USPs using a Support Vector Machine (SVM) model, aiming to enhance our comprehension of Mtb stress adaptation mechanisms. The USPs of Mtb play a pivotal role in the bacterium's ability to withstand diverse stress conditions and establish persistent infections in the host. There is a need to provide a data-driven, scalable approach to predict USPs from Mtb based on learned patterns and features extracted from protein sequences and properties. Protein sequences, that are annotated with both USPs and non-USPs, of 5,900 amino acids were obtained from the UniProt and NCBI protein database within the range of ten years. 3,082 of the amino acids were from the 600 non-USPs while 2,818 were from the 58 USPs. Data preprocessing was performed on both class of dataset and feature extraction techniques was used to transform raw protein sequences into numerical representations suitable for SVM training. Such evaluation matrices as Precision, Recall, F1-Score, cross-validation and Accuracy were used to evaluate the model performance. The model was able to predict 83% accuracy of USP sequences from Mtb which show practical implications of the SVM model's performance in predicting USPs and its potential for supporting further research on Mtb's stress response mechanisms.

AGVD: enhancing access to African genomic variation

Ayton Meintjes, Wilson Mudaki and Nicola Mulder
(University of Cape Town)

African populations exhibit the highest genetic diversity worldwide but remain underrepresented in global genomic datasets. Additionally, African data are often merged and represented as a single continent, but frequencies may vary significantly between African countries and regions. The African Genome Variation Database (AGVD) - a web resource - addresses this challenge by collating and improving the utility of available genomic variant data from African and African-ancestry cohorts. AGVD enables users to search, visualize, and prioritise genetic variants with a focus on African data. It provides allele frequency insights by study, ethnic group, and individual variants, incorporating annotations from both internal and external sources. Frequency-based queries (common in rare disease variant research) are supported, with an emphasis on clinical relevance. The initial AGVD release includes allele frequencies from roughly 4,000 samples. Future expansions will integrate additional genotypic data from targeted, exome, whole-genome sequencing, and array-based genotyping, enhancing the granularity of population-specific data and clinical insights. With its focus on groups with African-ancestry, AGVD advances the understanding of genomic diversity on the continent. Researchers are encouraged to leverage AGVD as a critical resource in bridging gaps in global genomic knowledge.

Cost-effective variant calling with DRAGEN using Illumina Connected Analytics

Regan Cannell, Shaun Aron and Scott Hazelhurst
(University of the Witwatersrand)

The analysis of whole genome and exome sequence data is a critical application area in bioinformatics. Variant calling is computationally expensive, and efficiency and accuracy are critical. A number of algorithms and pipelines have been developed to call genetic variants from whole genome sequencing data, however, not many have proved to scale well to accurately process thousands or tens of thousands of samples efficiently on modest computing infrastructure. Illumina's Dynamic Read Analysis for GENomics (DRAGEN) pipeline is highly regarded for its accuracy, and as it uses Field Programmable Gate Arrays (FPGAs), is very computationally efficient. DRAGEN has been shown to accurately and comprehensively call a range of variant types at scale. While DRAGEN is available via the purchase of a preconfigured server and usage licence, the most practical method for using DRAGEN is by acquiring credits on Illumina's cloud-based Connected Analytics (ICA) platform. ICA provides an interface to run various DRAGEN pipelines to call variants from genomic data utilising cloud resources, however, this can prove to be costly as users are charged for compute time and storage of data in the cloud. Long-term storage of large genomic datasets in the cloud may not be a viable option for many users, so the cost-effectiveness of DRAGEN will in many cases depend on optimal storage of data in the cloud. To explore the use of DRAGEN for large scale data analysis we have developed a reliable and robust workflow that supports the use of DRAGEN through ICA in a cost-effective manner. Our workflow is built with Nextflow, and is aimed at optimizing the upload, download, and deleting of files that are required for the execution of the DRAGEN variant calling pipeline, thereby reducing data storage costs on the ICA platform. While the ICA platform does provide a user graphical interface, to improve automation, the Nextflow workflow consists of several individual processes that execute Bash commands in addition to commands from the ICA command line interface (icav2). Altogether, the workflow automates the process of uploading raw genomic data (fastq, BAM, CRAM), initiating the DRAGEN variant calling workflow, downloading results, and deleting remote copies of all the data on completion of the analysis. Nextflow's ability to run several processes concurrently allows us to upload multiple files concurrently, and to trigger several pipeline runs so that the analyses take place in parallel. This reduces the total amount of time required to run a large batch of samples.

Metabolic Network Modeling of a Leishmania Infected Macrophage

Feriel Guennich (Laboratory BioInformatics, bioMathematics and bioStatistics (BIMS), Institut Pasteur de Tunis, Tunis, Tunisia), Mariem Hanachi (Laboratory BioInformatics, bioMathematics and bioStatistics (BIMS), Institut Pasteur de Tunis, Tunis, Tunisia), Oussema Souiai (Laboratory BioInformatics, bioMathematics and bioStatistics (BIMS), Institut Pasteur de Tunis, Tunis, Tunisia) , Bruno Sá (Centre of Biological Engineering, Campus Gualtar, University of Minho), Alexandre Oliveira (Centre of Biological Engineering, Campus Gualtar, University of Minho), Ezekiel Adebiyi (Department of Computer and Information Sciences, Covenant University), Miguel Rocha (Centre of Biological Engineering, Campus Gualtar, University of Minho), Lamia Guizani-Tabbane (Laboratory of Medical Parasitology, Biotechnology and Biomolecules (PMBB), Institut Pasteur de Tunis) and Alia Benkahla (Laboratory BioInformatics, bioMathematics and bioStatistics (BIMS), Institut Pasteur de Tunis)

Objective: Advancements in computational biology and bioinformatics have significantly enhanced our capacity to analyze complex biological systems, offering significant potential for the study of infectious diseases such as leishmaniasis. Caused by the Leishmania parasite, this disease remains a major public health burden, particularly in tropical and subtropical regions. The challenge of effective disease management is hindered by the absence of vaccines and the increasing resistance to treatments. It is therefore crucial to understand the functional biological mechanisms underlying the disease, to explore alternative therapies and optimize vaccines. This study aims to investigate the intricate metabolic interactions between Leishmania major and infected murine macrophages. It focuses on how Leishmania infection manipulates macrophage metabolism and the consequent impact on both the host's immune response and the parasite's survival. We leverage the COBRApy package, a powerful Python-based tool designed for constraint-based optimization and systems biology modeling, to investigate the interplay between Leishmania major and infected mouse macrophages. Methodology: In the first step, COBRApy was used to combine existing metabolic models of murine macrophages (iMM1865) and Leishmania (iAS556) into a host-pathogen model (HPM). Then, to explore the impact of transcriptional changes on macrophage metabolism during infection, a Dual-RNA-seq set from infected mouse macrophages (PMID: 27165796) was integrated into the HPM using RIPTiDe (PMID: 32298268), resulting in a Leishmania-infected host-macrophage model (LIHMΦ). Thereafter, a Flux Balance Analysis (FBA) was performed on the HMΦLIM to predict changes in metabolic flux. FBA identifies altered metabolic pathways during infection, highlighting crucial changes in nutrient uptake, energy production, and biosynthetic pathways that favor parasite survival. Finally, an analysis of the metabolic flux changes in the LIHMΦ was performed to pinpoint key pathways significantly altered during infection in host or pathogen. Results: By examining the metabolic shifts in the LIHMΦ, our study reveals, without relying on prior literature input, how Leishmania manipulates host cell metabolism to optimize its environment for survival, thus impairing macrophage function and altering immune responses. The findings also pinpoint potential metabolic vulnerabilities within the host-parasite interaction, which could serve as targets for future therapeutic strategies aimed at disrupting the parasite's adaptation mechanisms. The obtained model also found key metabolic rewiring occurring in infected macrophages, opening new pathways for developing novel treatment options based on targeting metabolic disturbances.

An AI Chatbot to support pathogen genomics trainees

Perceval Maturure, Oliver De Sousa, Kirsty Lee Garson, Nicola Mulder, Tony Li and Siddiqah George (University of Cape Town)

The NGS Academy Help-desk ticketing system, available at https://ngs-helpdesk.h3abionet.org/, was deployed in late 2022 as a support platform designed to address a wide range of needs across bioinformatics and computational infrastructure This tool not only resolves technical and computational errors post-training but also serves as a robust knowledge base and mentoring resource for researchers and trainees. The helpdesk supports diverse areas such as malaria, fungal pathogens, parasites, and other infectious diseases, while also addressing challenges in next-generation sequencing (NGS), phylogenetics, antimicrobial resistance (AMR), and related fields. A significant amount of time is often spent resolving tickets due to differences in time zones among the support teams and a lack of specialists in certain categories. Additionally, incorrectly routed tickets contribute to longer waiting periods before resolution. A pretrained Large Language Model (LLM) has been fine-tuned using web-scraped data and integrated into an application to develop a chatbot. The chatbot will not replace the helpdesk ticketing system; instead, both will coexist as a hybrid solution, complementing each other to enhance overall support with the aim to reduce NGS ticket response times. In this first phase, the chatbot was trained on a small dataset using the pretrained Flan-T5 model. The goal was to test technical feasibility, not to measure final performance. Early results show that the chatbot can generate relevant responses, suggesting its potential as a support tool for pathogen genomics. Quantitative testing highlights its expected to clear weaknesses at this stage. An F1 score 0.2272 shows poor word-for-word accuracy. ROUGE-1 (0.2383) and ROUGE-2 (0.0527) reveal difficulty in forming coherent phrases. Still, the cosine similarity score (0.7492) suggests the chatbot captures meaning correctly, even when phrasing is off. These results do not reflect real-world performance, as a small and non-specialized dataset was used during model training. Future work will focus on training with larger, domain-specific data. The hybrid approach, combining chatbot and human support, is a scalable, sustainable solution. It boosts help desk efficiency and ensures complex issues are escalated appropriately. Future work could expand the knowledge base, improve accuracy, and explore more AI-driven features to optimize support.

Detectable episodic positive selection in the virion strand
A-strain maize streak virus genes may have a role in its host adaptation.

Kehinde Oyeniran (International Center for Genetic Engineering and Biotechnology (ICGEB)) and Mobolaji Tenibiaje (Bamidele Olumilua University of Education, Science & Technology Ikere)

Maize streak virus (MSV) has four genes: cp, encoding the coat protein; mp, the movement protein; and repA and rep, encoding two distinct replication-associated proteins from an alternatively spliced transcript. These genes play roles in encapsidation, movement, replication, and interactions with the external environment, making them prone to stimuli-driven molecular adaptation. We accomplished selection studies on publicly available curated, recombination-free, complete coding sequences for representative A-strain maize streak virus (MSV-A) cp and mp genes. We found evidence of gene-wide selection in these two MSV genes at specific sites within the genes (cp 1.23% and mp 0.99%). Positively selected sites have amino acids that are 60% hydrophilic and 40% hydrophobic in nature. We found significant evidence of positive selection at branches (cp: 0.76 and mp :1.66%) representing the diversity of MSV-A strain in South Africa, which is related to the MSV-A-matA isolate (GenBank accession number: AF329881), well disseminated and adapted to the maize plant in sub-Saharan Africa. In the mp gene, selection significantly intensified for the overall diversities of the MSV-A sequences and those more related to the MSV-Mat-A isolate. These findings reveal that despite predominantly undergoing non-diversifying selection, the detectable diversifying positive selection observed in these genes may play a major role in MSV-A host adaptive evolution, ensuring sufficient pathogenicity for onward transmission without killing the host.

Genomic Analysis of Salmonella Typhi from a Typhoid Conjugate Vaccine Trial

Belson Kutambe (Malawi Liverpool Wellcome Trust, Phillip Ashton (University of Oxford,Pandemic Sciences Institute)), Priyanka Patel (Malawi Liverpool Wellcome Trust) and Melita Gordon (Malawi Liverpool Wellcome Trust)

Background Salmonella Typhi is the causative agent of typhoid fever, a significant public health burden in many low- and middle-income countries. The introduction of Vi polysaccharide typhoid conjugate vaccine (Vi-TCV) aims to reduce this burden. Methods To understand the short-term evolutionary dynamics of S. Typhi in response to Vi-TCV introduction, we conducted Illumina genome sequencing of isolates from a vaccine trial in Malawi. Whole-genome sequencing was performed on S. Typhi isolates from both the intervention (Vi-TCV) and control (Meningococcal capsular group A conjugate, Men A) vaccine arms. We analysed the S. Typhi sequencing data to determine i) the lineage, ii) antimicrobial resistance profiles, iii) presence of mutations in the Vi-capsule encoding genes, and iv) to construct a phylogenetic tree. Results We sequenced 94 S. Typhi isolates, 15 from TCV arm, 79 from MenA arm. All the 94 isolates from both arms were lineage 4.3.1.1 and phylogenetic analysis showed no clear separation between isolates from the Vi-TCV and Men A arms. All 94 (100%,) isolates encoded blaTEM-1, dfrA7, catA1, sul1, sul2, associated with resistance to ampicillin, cotrimoxazole, and chloramphenicol. Four isolates (2/79 from MCV arm and 2/15 from TCV, Fisher exact test P = 0.24) had non-synonymous mutations encoding amino acid changes in quinolone resistance determining regions; three with GyrA S83F and one with GyrB S464F. Mutations in genes responsible for Vi production and expression were seen in 9/79 (11%) MCV isolates and 1/15 (7%) (Fisher exact test P > 0.99) isolates from the Vi-TCV arms. Four patients were hospitalised in the MenA arm, of whom one had mutations in Vi previously associated with changes in virulence. There was no significant association between Vi mutations and hospitalisation (Fisher exact test P=0.8405). Conclusions We observed no significant genetic differences between S. Typhi infections in participants vaccinated with the typhoid conjugate vaccine and the MenA vaccine. However, our study had a relatively short follow up time, and longer term surveillance should be carried out.

Using Machine Learning and Bioinformatics Analysis to Identify Gene Expression Patterns and Predict the Risk of Preterm Birth in Pregnant Women

Emmanuel Biryabarema and Dr. Sinkala Musalula
(University Of Cape Town)

Introduction Preterm birth, defined as babies born alive before 37 weeks of pregnancy or fewer than 259 days since the first day of a woman's last menstrual period, is the leading cause of death in children under the age of 5 years. For the past decade, the rates of preterm birth have not considerably changed over the last decade despite the substantial focus on routine health data systems globally. Significance and Justification of the study Traditional methods of predicting women at risk for preterm birth, such as obstetric history, tocometry, biochemical markers, and ultrasonography of the cervix and interventions based on these findings have not lowered the rate of preterm birth Identification and treatment of women at risk of having a preterm birth. Diagnosis of preterm birth also presents another challenge as the initial symptoms and signs of preterm birth (contractions and cervical dilation) are often mild and may occur in normal pregnancies and many healthy women will report such symptoms during routine prenatal visits, and some women at increased risk of having a preterm birth may dismiss these early warning signs as normal in pregnancy making the clinical diagnosis very difficult. The use of a machine learning whole-blood gene expression-based model to predict the risk of preterm birth in pregnant women could solve this challenge providing a more accurate and personalised (patient specific) prediction. General objective Machine learning and bioinformatics analysis were used to identify gene expression patterns associated with preterm birth and develop a model that predicts the risk of preterm birth in pregnant women. Specific objectives 1) To analyse whole blood gene expression data from pregnant women, identifying differential gene expression signals and enriched pathways associated with preterm birth. 2) To develop and train machine learning models using gene expression data to predict the risk of preterm birth and provide interpretability for model predictions. 3) To validate and assess the predictive performance of the machine learning model using established metrics. Methods We carried out secondary data analysis of publicly available datasets and were obtained as part of the DREAM Preterm Birth Prediction Challenge through Synapse (syn18380862) platform to identify differential gene expression signals and enriched pathways associated with preterm birth and then developed and trained machine learning models using gene expression data to predict the risk of preterm birth and provide interpretability for model predictions. Ethics and informed consent requirements This study was carried out in accordance with the ethical principles stated in the Declaration of Helsinki (1996) and applicable guidelines on good clinical practice. We obtained ethical approval of the protocol from the UCT Faculty of Health Sciences Human Research Ethics Committee. We adhered to the guideline and policies of the data providers and made every effort to protect participant's privacy and confidentiality of the data, and to ensure that the study is fair and ethical.

Developing Clinical Phenotype Data Collection Standards for Research in Africa

Katherine Johnston and Lyndon Zass (University of Cape Town)

Modern biomedical research is characterised by its high-throughput and interdisciplinary nature. Multiproject and consortium-based collaborations requiring meaningful analysis of multiple heterogeneous phenotypic datasets have become the norm; however, such analysis remains a challenge in many regions across the world. An increasing number of data harmonisation efforts are being undertaken by multistudy collaborations through either prospective standardised phenotype data collection or retrospective phenotype harmonisation. In this regard, the Phenotype Harmonisation Working Group (PHWG) of the Human Heredity and Health in Africa (H3Africa) consortium aimed to facilitate phenotype standardisation by both promoting the use of existing data collection standards (hosted by PhenX), adapting existing data collection standards for appropriate use in low- and middle-income regions such as Africa, and developing novel data collection standards where relevant gaps were identified. Ultimately, the PHWG produced 11 data collection kits, consisting of 82 protocols, 38 of which were existing protocols, 17 were adapted, and 27 were novel protocols. The data collection kits will facilitate phenotype standardisation and harmonisation not only in Africa but also across the larger research community. In addition, the PHWG aims to feedback adapted and novel protocols to existing reference platforms such as PhenX.

Metagenomic profiling of the gut microbiome in African populations: The AWI-Gen 2 Microbiome Study

Luicer Anne Ingasia Olubayo (University of the Witwatersrand), Dylan G. Maghini (University of the Witwatersrand), Ovokeraye H. Oduaran (University of the Witwatersrand), Natalie Smyth (University of the Witwatersrand), Furahini Tluway (University of the Witwatersrand), Michelé Ramsay (University of the Witwatersrand), Jakob Wirbel (Stanford University), Ami Bhatt (Stanford University), Scott Hazelhurst (University of the Witwatersrand), Carl W. Belger (University of the Witwatersrand) and Jane A. Cook (Stanford University)

Background: Despite its critical role in human health, the gut microbiome remains understudied in underrepresented populations, particularly in low- and middle-income countries. Large-scale gut microbiome research has historically focused on high-income, industrialized populations, limiting our understanding of microbial diversity, adaptation, and health implications across different environmental and lifestyle contexts. The AWI-Gen 2 Microbiome Project addresses this gap by investigating how geography, industrialization, lifestyle, and health status shape gut microbiome diversity in six study sites across Burkina Faso, Ghana, Kenya, and South Africa. Methods: A total of 1,801 women aged 41–84 years were enrolled from rural, semi-rural, and urban communities spanning distinct environmental and socioeconomic settings. Shotgun metagenomic sequencing was performed to generate high-resolution taxonomic and functional profiles of gut microbial communities. Metagenome-assembled genomes (MAGs) were reconstructed to expand microbial reference catalogues and uncover novel species. Statistical analyses assessed associations between microbiome composition, dietary patterns, antibiotic use, and disease states, including HIV infection. Results: Geography was the primary driver of microbiome composition, with distinct microbial transitions observed along an industrialization gradient. Rural populations exhibited higher microbial diversity, with a notable enrichment of Treponema species, while urban populations showed reduced Treponema and Cryptobacteroides abundance alongside a relative increase in Bifidobacterium species. Nairobi's informal settlements exhibited a unique hybrid microbiome signature, reflecting a mix of rural and urban microbial traits, challenging conventional rural–urban microbiome models. The study significantly expanded global microbial reference datasets, identifying 1,005 novel bacterial species and 40,135 previously uncharacterized viral genomes. The absence of Treponema succinifaciens in urban populations correlated with higher antibiotic exposure and lower dietary fiber intake, suggesting that antimicrobial-driven microbiome shifts may be occurring in transitioning populations. Additionally, a distinct HIV-associated microbiome signature was characterized, featuring taxa not previously linked to HIV in high-income cohorts, including Dysosmobacter welbionis and Enterocloster species. These findings underscore the need for population-specific microbiome research to better understand host-microbiome interactions in infectious diseases. Conclusion: This study provides critical insights into the diversity and adaptation of the gut microbiome in African populations, challenging existing models of industrialization-driven microbial shifts. By leveraging shotgun metagenomics, this work contributes to a more representative and equitable global microbiome atlas, expanding the known diversity of bacterial and viral species. These findings highlight the need for inclusive microbiome research that reflects diverse global populations and informs precision medicine approaches. Beyond advancing microbiome science, this study prioritizes community engagement, participant education, and the dissemination of findings. Future work will integrate participant feedback and explore the implications of microbiome shifts for public health. Ongoing research will investigate longitudinal microbiome dynamics and microbiome-host interactions, while planned follow-up analyses will assess microbiome stability over time in previously sampled participants.

Reference Genome and Pangenome Construction of Wild Spotted Hyenas (Crocuta crocuta) from the Kruger National Park

Ansia van Coller (SAMRC Genomics Platform), Brigitte Glanzmann (SAMRC Genomics Platform & Stellenbosch University Division of Molecular Biology and Human Genetics)), Nadia Carstens (SAMRC Genomics Platform & WITS Department of Human Genetics), Victoria Cole (SAMRC Genomics Platform), Craig Kinnear (SAMRC Genomics Platform & Stellenbosch University Division of Molecular Biology and Human Genetics), Tanya Kerr (Stellenbosch University Division of Molecular Biology and Human Genetics, SAMRC Centre for Tuberculosis Research), Giovanni Ghielmetti (Stellenbosch University Division of Molecular Biology and Human Genetics, SAMRC Centre for Tuberculosis Research), Wynand Goosen (Stellenbosch University Division of Molecular Biology and Human Genetics, SAMRC Centre for Tuberculosis Research) and Michele Miller (Stellenbosch University Division of Molecular Biology and Human Genetics, SAMRC Centre for Tuberculosis Research)

The spotted hyena (Crocuta crocuta) is a highly social carnivore with complex behavioural and ecological functions, making it an important model for studying genetic diversity, adaptation, and evolution. However, previous draft genomes for C. crocuta have been incomplete and derived from captive individuals, limiting insights into natural genetic variation. Here, we present a high-quality de novo genome assembly and the first pangenome of wild spotted hyenas sampled within the Kruger National Park in South Africa. Using Oxford Nanopore Technologies (ONT) long-read sequencing, we generated a reference genome for a male individual, achieving a total assembly size of 2.39 Gb with a scaffold N50 of 19.6 Mb. Assembly completeness, assessed with BUSCO, revealed 98.5% completeness against the Mammalia_odb10 database and 98.2% against the Carnivora_odb10 database, confirming a high-quality assembly. This assembly is more contiguous and complete than previously published hyena genomes, which had lower N50 values and only 95% completeness. Additionally, our assembly is derived from a wild individual, providing a more ecologically relevant reference compared to those from captive specimens. To explore population-level genomic diversity, we sequenced ten additional free-ranging individuals using MGI short-read sequencing, achieving an average depth of 32X (two individuals) and 10X (eight individuals). We identified an average of 4 million single nucleotide polymorphisms (SNPs) and 1 million INDELs across the ten individuals. To capture the full spectrum of diversity, we constructed a pangenome using the Progressive Genome Graph Builder (PGGB). Consensus sequences were extracted from all ten hyena samples and used as input for PGGB, resulting in a pangenome that revealed both conserved genomic regions and loci exhibiting variation, potentially associated with immune function, behaviour, and environmental adaptation. Our analysis uncovered notable genetic differences among individuals, reflecting fine-scale population structure and potential local adaptation. The graph-based genome enables examination of structural variations, including insertions, deletions and duplications, which may be overlooked in standard reference-based approaches. This study represents a major advancement in the genomic resources available for C. crocuta, providing the first wild-derived genomic reference and first pangenome for the species. Our findings contribute to a deeper understanding of spotted hyena genetic diversity and evolutionary history, with important applications in conservation genetics, behavioural ecology, and comparative genomics. This work also highlights the importance of utilising wild populations for genomic studies to better reflect the natural genetic landscape of a species.

Association analyses reveal susceptibility variants linked to Parkinson's disease in the South African population

Kathryn Step (Division of Molecular Biology and Human Genetics, Stellenbosch University), Thiago Peixoto Leal (Genomic Medicine, Lerner Research Institute, Cleveland Clinic Foundation), Emily Waldo (Genomic Medicine, Lerner Research Institute, Cleveland Clinic Foundation) , Lusanda Madula (Division of Molecular Biology and Human Genetics, Stellenbosch University), Yolandi Swart (Division of Molecular Biology and Human Genetics, Stellenbosch University), Carlos F. Hernández (Universidad del Desarrollo, Centro de Genética y Genómica, Facultad de Medicina Clínica Alemana), Sara Bandres-Ciga (Center for Alzheimer's and Related Dementias (CARD), National Institutes of Health, Bethesda) , Jonggeol Jeffrey Kim (Department of Molecular and Human Genetics, Baylor College of Medicine), Ignacio F. Mata (Genomic Medicine, Lerner Research Institute, Cleveland Clinic Foundation) and Soraya Bardien (Division of Molecular Biology and Human Genetics, Stellenbosch University)

Stemming from a complex etiology that includes a strong genetic component (1), Parkinson's disease (PD) is a neurodegenerative disorder characterized by a wide range of both motor and non-motor symptoms (2). The burden of PD is increasing rapidly within the aging Sub-Saharan African populations, ranking as the 11/12th most prevalent nervous system disorder in the region (3). Despite the rise in disease prevalence, the representation of African populations in PD genetic research remains limited. Allele frequencies vary across genomes due to factors such as natural selection, genetic drift, and differing exposures to environments and pathogens. These variations in allele frequencies can help identify population-specific disease risk variants in admixed individuals while simultaneously uncovering risk variants relevant to multiple populations (4). Genome-wide association studies (GWAS) have successfully identified susceptibility variants linked to PD (5,6). However, the majority of these studies have focused on European cohorts with few including diverse ancestries. Using genotyped and imputed data from 1,516 South African participants, we conducted a GWAS using SAIGE software, which includes the genetic relationship matrix as a random effect, allowing for the inclusion of related individuals. Moreover, we inferred global and local ancestry for the cohort to better understand the genetic admixture in the South African population and further investigate the GWAS results. Our GWAS findings were replicated using a Latin American cohort. The ancestry inference showed the South African cohort to be five-way admixed between the European (EUR; 56%), African (AFR; 18.8%), indigenous Khoe-San Nama ancestry (NAMA; 13%), South Asian (SAS; 6.9%), and Malaysian (MAL; 5.2%) ancestries. The GWAS identified one variant with a genome-wide significance and 351 variants with a suggestive significance. Of these, 14 variants replicated in the Latin American cohort. In the local ancestry window containing the top GWAS hit, 86.7% of the variant carriers were inferred to have AFR, 11% NAMA, and 2.2% MAL ancestries. No carriers exhibited EUR or SAS inferred ancestry. This suggests that the variant is ancestry-specific and highlights the value of including populations previously underrepresented in PD genetic research to reveal novel susceptibility variants. Our findings contribute to a global understanding of the complex genetic etiology of PD. 1. Trevisan, L. et al. Genetics in Parkinson's disease, state-of-the-art and future perspectives. Br. Med. Bull. 149, 60–71 (2024). 2. Armstrong, M. J. & Okun, M. S. Diagnosis and treatment of Parkinson disease: A review: A review. JAMA 323, 548–560 (2020). 3. GBD 2021 Nervous System Disorders Collaborators. Global, regional, and national burden of disorders affecting the nervous system, 1990-2021: a systematic analysis for the Global Burden of Disease Study 2021. Lancet Neurol. 23, 344–381 (2024). 4. Swart, Y. et al. Local ancestry adjusted Allelic association analysis robustly captures tuberculosis susceptibility loci. Front. Genet. 12, 716558 (2021). 5. Nalls, M. A. et al. Identification of novel risk loci, causal insights, and heritable risk for Parkinson's disease: a meta-analysis of genome-wide association studies. Lancet Neurol. 18, 1091–1102 (2019). 6. Kim, J. J. et al. Multi-ancestry genome-wide association meta-analysis of Parkinson's disease. Nat. Genet. 56, 27–36 (2024).

Participant-Centred Research: Personal Experiences with Implementing Tiered Informed Consent for Genomics Health Studies using a Qualitative Iterative Approach

Irene Muchada (University of the Western Cape)

Background: Advancing genomics health studies requires a research framework that respects participant autonomy while ensuring robust data collection. True participant-centred research goes beyond passive involvement to actively include participants as partners in designing their experience and engagement. The tiered informed consent model offers a flexible and participant-centred approach, allowing individuals to choose their preferred level of engagement and data sharing. A disease-agnostic population cohort that links participants' routine health data to their genomic data is being piloted in the Western Cape province of South Africa. A tiered consent process with tiers for each study component has been implemented. Methods: A genetic counsellor engaging with participants and implement the tiered informed consent model is using a qualitative iterative approach to refine recruitment and consent processes, drawing on personal experiences and case examples. Areas of reflection include participant responses, engagement, understanding, communication, ethical considerations and implications for practice. Results: Participants expressed that they felt empowered by the tiered consent process to make informed decisions without feeling overwhelmed or coerced. However, some said that they felt pressured to consent or confused due to institutional authority. Emergent themes include participants' perceptions of autonomy and control, emotional responses to consent discussions, challenges in comprehension and decision-making, and ethical considerations related to privacy and data sharing. The iterative approach resulted in rewording some sections of the consent documentation. Conclusion: Our step-by-step process showed that continuously collecting and using feedback from participants made our informed consent procedures much clearer and more responsive to their needs. Adjustments based on real-world experiences improved participant understanding and trust, and fostered a more collaborative research environment. The tiered informed consent framework, refined through a qualitative iterative process, effectively balances the demands of comprehensive data collection with the need to prioritize participant autonomy and well-being.

Machine Learning for Hypertension Genetics in African Populations

Peace Bassey Osim, Blessing Ekpenyong, Anita Yemi-Odae Nelson, Bede Anwan and Mary E. Kooffreh (University of Calabar)

Hypertension is a significant public health concern worldwide, affecting over 1.4 billion people and contributing to cardiovascular diseases, stroke, and kidney failure (World Health Organization, 2021). In Africa, the prevalence of hypertension has been increasing, with estimates suggesting that nearly 40% of adults are hypertensive, largely due to genetic predispositions and environmental factors (Adeloye et al., 2021). Despite its growing burden, the genetic underpinnings of hypertension remain poorly understood, especially in African populations that are underrepresented in global genomic studies. This study leverages machine learning algorithms to identify genetic variants associated with hypertension in African populations. We analyzed genomic data from 1,000 African individuals diagnosed with hypertension and 1,000 normotensive controls. Genotyping was conducted using the Illumina OmniExpress array, capturing approximately 700,000 single nucleotide polymorphisms (SNPs). Various machine learning models, including random forest, support vector machine (SVM), and gradient boosting, were implemented to identify key genetic variants predictive of hypertension. Our results demonstrate that machine learning models effectively predict hypertension risk based on genetic information, with the random forest model achieving the highest classification accuracy of 85.2%, outperforming both gradient boosting (82.7%) and SVM (79.5%). Notably, the analysis identified several hypertension-associated variants, particularly within the NOS3, AGT, and ACE genes, which have well-established roles in blood pressure regulation. These findings underscore the utility of artificial intelligence in detecting complex genetic patterns that contribute to hypertension susceptibility. The study highlights the potential of integrating machine learning with genomic research to enhance disease risk prediction and inform personalized medicine strategies tailored to African populations. Unlike traditional genome-wide association studies (GWAS), which primarily focus on linear associations, machine learning algorithms can capture complex, nonlinear interactions among genetic variants, enabling more robust disease modeling. The clinical implications of this research suggest that incorporating machine learning-driven genetic risk assessment into public health frameworks could improve hypertension prevention and treatment strategies, particularly in resource-limited settings. However, further research is necessary to validate these findings using larger, more diverse datasets and functional analyses of the identified variants. Future studies should explore how environmental factors interact with genetic predispositions to influence hypertension risk and evaluate the translational potential of these predictive models in clinical settings. Additionally, the inclusion of multi-omics data, such as transcriptomic and epigenomic profiles, may further enhance the accuracy of hypertension risk prediction. Overall, this study underscores the transformative role of artificial intelligence in genomic medicine and emphasizes the need for increased representation of African populations in genetic research. By leveraging machine learning approaches, researchers can uncover novel genetic markers of hypertension and contribute to the development of targeted therapeutic interventions that address the unique genetic architecture of African populations.

Capacity building in bioinformatics and data science among African trainers

Sindiswa Lukhele and Nicola Mulder
(Computational Biology Division, Institute of Infectious Diseases, Faculty of Health Sciences, University of Cape Town)

Africa has made substantial progress in providing genomics training. However, despite an increase in training initiatives, there remains a notable shortage of trainers proficient in bioinformatics and data science, a shortfall that hinders the continent's capacity to fully harness the potential of these rapidly evolving fields. The lack of expert trainers is especially urgent considering the growing accessibility of extensive and complex genomic datasets that demand a skilled workforce able to carry out advanced analysis and interpretation. To effectively address this gap, it is essential to focus on strategic investments in initiatives aimed at building capacity in developing and delivering competency-based training. Several initiatives, including H3ABioNet and eLwazi ODSP, have collaborated to enhance trainer expertise in developing educational programs in bioinformatics and data science. The program created a strategy designed to establish a network of trainers, focused on fostering and cultivating collaborative training partnerships among educators from various scientific disciplines who seek to integrate bioinformatics and data science training into their educational initiatives. This presentation aims to highlight the current activities in developing trainer capabilities in bioinformatics and data science, distinguishing between enhancing experience in organising training programs and developing teaching skills. The initiative has helped to create an active train-the-trainer community.

Evaluation of the Use of Count-Based Methods for Alternative Splicing Analysis in Monocyte-to-Macrophage Differentiation

Palesa Lesole, Vanessa Meyer and Nikki Gentle (School of Molecular and Cell Biology, University of the Witwatersrand)

Alternative splicing (AS) occurs in over 90% of multi-exonic genes, significantly increasing transcript and protein diversity. This tightly regulated process plays a crucial role in cell development and differentiation, and its dysregulation has been linked to various autoimmune diseases and cancers. Consequently, understanding AS is necessary to elucidate its role in normal cellular function and disease development. Advances in high throughput sequencing technologies, particularly RNA-seq, have led to the development of various tools and software packages for AS analysis. AS analysis is broadly categorised into isoform-based and count-based methods. Isoform-based approaches rely on full transcript reconstruction and assess differential transcript expression between conditions. Count-based methods (further divided into exon- and event-level) provide a more straightforward approach to AS analysis and are better suited for short-read RNA-seq data. These approaches detect and quantify AS events and measure differential expression of transcript features (exons and junctions) between conditions by constructing genes as disjointed counting bins inferred from isoform-specific reads. In this study, we employed three key count-based methods - differential exon usage (DEU), differential transcript usage (DTU) and differential AS events (DAS) - to investigate the global AS landscape during monocyte-to-macrophage differentiation. DEU was identified using DEXSeq (adjusted p-value < 0.05, |Log2FoldChange| > 1), DTU was identified using DRIMSeq (stage-wise analysis using stageR at overall FDR < 0.05), and DAS events were identified using rMATS-turbo (FDR < 0.05, |ΔPSI| > 0.1) in monocytic THP-1 cells and those treated with PMA for 96 hours to induce differentiation. We identified 451 differentially used exons, 679 differentially used transcripts and 238 differential AS events affecting 160, 211 and 196 genes, respectively. Despite the relatively large number of AS events detected, there was limited overlap across methods, with only one gene detected by all three approaches. GO enrichment analysis using enrichGO revealed that DEU-associated genes were enriched in ribosomal development and ECM remodelling, whereas DTU-associated genes were enriched in GTPase regulator activity and negative regulation of cell killing. These findings suggest that each approach captures distinct aspects of AS regulation and contributes uniquely to understanding AS-driven changes in cellular differentiation. The limited overlap between differentially spliced genes highlights the challenges of integrating count-based methods due to statistical and methodological differences among tools. However, these approaches provide unique and complementary insights at the exon, transcript and event levels, and uncover distinct regulatory mechanisms that shape cellular function during differentiation.

First Continental African Genome-Wide Association Study Identifies Novel Genetic Loci Associated with Blood Urea Nitrogen and Kidney Function

Gloria Kirabo (MAKERERE UNIVERSITY), Opeyemi Soremekun (HELMONTZ-MUNICH) and Segun Fatumo (QUEEN MARY UNIVERSITY OF LONDON)

Chronic kidney disease (CKD) is a critical global health concern with high mortality rates and severe complications, particularly in Africa, yet the underlying molecular mechanisms remain poorly understood. We conducted a Genome-Wide Association Study (GWAS) using Blood Urea Nitrogen (BUN) levels, a key biomarker of kidney function, in 5,910 Ugandan participants to identify single nucleotide polymorphisms (SNPs) associated with CKD risk. Our analysis identified 13 SNPs reaching a suggestive significance threshold ($p < 5\times10^{-7}$), refined to five independent lead SNPs through LD clumping. Notably, rs73309776 in the GALNT6 gene suggests potential pathways linking breast cancer and kidney function, while rs145326389, an intronic variant in LOC105374218, is associated with traits related to the RAAS pathway and blood pressure regulation. rs142038911 is a synonymous variant in TRIM11, TRIM17, and LOC124904537, and may play a role in regulating serum creatinine and protein binding, which are crucial in kidney disease. Bayesian fine mapping highlighted rs1286795408 on chromosome 7 as a strong candidate with a posterior probability of 84% with a 99% credible set, warranting further investigation. Functional annotations using MAGMA and GTEx revealed gene expression in the pituitary gland and kidney medulla, though these did not reach statistical significance. Replication in European, East Asian, and Latin American populations validated associations with genes such as HOXD11, BCAS3, and TFCP2L1, which are involved in kidney development and function, emphasizing shared genetic factors across ancestries. Rigorous quality control measures, including filtering for Hardy-Weinberg equilibrium, sex discrepancies, and minor allele frequency, ensured robust results. This study, the first GWAS of BUN in a continental African population, underscores the importance of inclusive genetic research and contributes to understanding CKD's genetic underpinnings, paving the way for precision medicine and potential targeted treatments for underrepresented populations.

A proteolipidomic signature for hair curvature

Michelle Mukonyora, Henry Adeola and Nonhlanhla Khumalo
(University of Cape Town)

Background and Objectives Scalp hair is increasingly being used as a non-invasive testing substrate in diagnostic and forensic medicine. Preliminary evidence suggests curly hair has more structural lipids, which may result in the incorporation of higher concentrations of lipid-soluble molecules. We aim to investigate for the first time, in the same samples, whether there is a correlation between hair curliness, proteins, and lipids. Our objective is to use bioinformatics tools to investigate the correlation between hair curvature and proteins/lipids extracted from geometrically classified hair. Methods A retrospective cohort of 90 virgin scalp hair samples was classified into six geometric hair types. Structural proteins and lipids were extracted from 1mg of hair and were identified and semi-quantified using mass spectrometry (LC-MS/MS for proteins and GC-MS/MS for lipids). Lipid spectral data was analysed using MS-DIAL software, while protein spectral data was analysed using MaxQuant and PeptideShaker software. Hair proteolipidomic models for geometric hair curvature were built using unsupervised and supervised machine learning tools. Results Hair proteolipidomic models built for the first time, revealed a hair proteolipidome comprising several heterogeneous subclusters. The hair proteolipidomic models showed no correlation between proteins and hair curvature (poor/moderate ROC scores 0.6 to 0.8), whereas lipids correlated well with hair curvature (excellent ROC scores 0.96 to 1). The lipids responsible for hair curvature grouping include ceramides, cholesterol, and fatty acids. Discussion/Conclusion This study demonstrates that lipids correlate with hair curvature. Future research should investigate whether this influences the incorporation of lipid-soluble drugs/biomarkers into hair, which may affect the interpretation of their concentrations in hair.

Biphasic Middle East respiratory syndrome coronavirus incidence in dromedary camels in Northern Kenya, 2022 - 2023

Brian Ogoti (University of Nairobi, Center of Epidemiological Modelling and Analysis), Victor Riitho (University of Nairobi, Center of Epidemiological Modelling and Analysis), Johanna Wildemann (Charité –Universitätsmedizin Berlin), Nyamai Mutono (University of Nairobi, Center of Epidemiological Modelling and Analysis), Julia Tesch (Charité –Universitätsmedizin Berlin), Jordi Rodon (Charité –Universitätsmedizin Berlin0, Kaneemozhe Harichandran (Charité –Universitätsmedizin Berlin), Jackson Emanuel (Charité –Universitätsmedizin Berlin), Elisabeth Möncke-Buchner (Charité –Universitätsmedizin Berlin), Stella Kiambi (Food and Agriculture Organization), Julius Oyugi (University of Nairobi, Center of Epidemiological Modelling and Analysis), Marianne Mureithi (University of Nairobi), Victor Corman (Charité –Universitätsmedizin Berlin), Christian Drosten (Charité –Universitätsmedizin Berlin), Samuel Thumbi (University of Nairobi, Center of Epidemiological Modelling and Analysis) and Marcel Müller (Charité –Universitätsmedizin Berlin)

Introduction Middle East respiratory syndrome coronavirus (MERS-CoV) is endemic in dromedary camels from the Arabian Peninsula and Africa with comparably high seroprevalence of >75%. High camel population density and the loss of maternal antibodies in farmed camel calves are linked to acute MERS-CoV outbreaks. Investigations into MERS-CoV outbreak patterns in nomadic camels are challenged by limited infrastructures in remote and resource-restricted camel migration regions. Study Objective We performed a continuous 12-month study at an abattoir hub for nomadic camels in Northern Kenya. We investigated MERS-CoV incidence in migrating camels and determined genomic diversity of contemporary MERS-CoV variants. Methods We collected nasal swabs from 10-15 camels 4-5 days per week from September 2022 to September 2023, totalling 2711 camels sampled during the period in the main abattoir in Isiolo County, Kenya. The samples were tested for MERS-CoV RNA using UpE and ORF1a RT-qPCR. Genomic diversity was assessed using Illumina next-generation sequencing (NGS) and ORF1ab domain assembly for RNA samples with >1x106 genome copies/ml. Results MERS-CoV RNA was detected in 36/2711 (1.3%) nasal swabs. MERS-CoV incidence was biphasic with detection peaks in the respective first week of October 2022 (7/60, 11.7%) and February 2023 (7/58, 12.1%). The cumulative MERS-CoV RNA positivity rate was higher in September–October 2022 with 19/381 (5.0%) compared to 17/727 (2.3%) in January–March 2023. For 9/36 MERS-CoV RNA-positive samples ORF1ab sequences were obtained, and phylogenetic analysis were performed. The sequences formed a distinct clade from other Clade C viruses but clustered with Clade C2.2, mostly prevalent in East Africa. The 9 ORF1ab sequences were highly similar (>99.93% nucleotide identity) and had 99.75–99.78% nucleotide identity with the closest MERS-CoV relative identified in Akaki, Ethiopia, in 2019. Conclusion The biphasic MERS-CoV incidence in nomadic camels may be linked to seasonality factors, such as the biannual alternating wet and dry seasons in Northern Kenya. Interestingly, camel calves are primarily born during the two wet seasons and maternal antibody loss coincides with the observed two MERS-CoV RNA detection peaks. Phylogenetic analysis suggests that we identified at least 3 MERS-CoV clusters over 3 different weeks in dromedaries originating from different locations.

Admixture and Evolutionary Variation: Genetic Insights into Body Composition in a Malagasy cohort

Iman Hamid (Variant Bio, Inc), Severine Nantenaina Stephie Raveloson (University of Antananarivo, Mention Anthropobiologie et Développement Durable), Germain Jules Spiral (University of Antananarivo, Mention Anthropobiologie et Développement Durable), Soanorolalao Ravelonjanahary (University of Antananarivo, Mention Anthropobiologie et Développement Durable), Brigitte Marie Raharivololona (University of Antananarivo, Mention Anthropobiologie et Développement Durable), José Mahenina Randria (University of Antananarivo, Mention Anthropobiologie et Développement Durable), Mosa Zafimaro (University of Antananarivo, Mention Anthropobiologie et Développement Durable), Tsiorimanitra Aimée Randriambola (University of Antananarivo, Mention Anthropobiologie et Développement Durable) , Rota Mamimbahiny Andriantsoa (University of Antananarivo, Mention Anthropobiologie et Développement Durable), Tojo Julio Andriamahefa (University of Antananarivo, Mention Anthropobiologie et Développement Durable), Bodonomena Fitahiana Laza Rafidison (University of Antananarivo, Mention Anthropobiologie et Développement Durable), Mehreen Mughal (Variant Bio, Inc), Anne-Katrin Emde (Variant Bio, Inc), Melissa Hendershott (Variant Bio, Inc), Sarah LeBaron von Baeyer (Variant Bio, Inc), Kaja A. Wasik (Variant Bio, Inc), Jean Freddy Ranaivoarisoa (University of Antananarivo, Mention Anthropobiologie et Développement Durable), Laura Yerges-Armstrong (Variant Bio, Inc), Stephane E. Castel (Variant Bio, Inc) and Rindra Rakotoarivony (University of Antananarivo, Mention Anthropobiologie et Développement Durable)

The Malagasy population represents a unique case of human admixture, shaped by historical interactions between African- and Austronesian-ancestry groups. This complex demographic history has likely influenced patterns of genetic variation and trait evolution, yet the genetic basis of phenotypic diversity in this population remains understudied. Here, we investigate population structure and the genetic architecture of body composition using whole-genome sequencing(WGS). We analyzed mid-pass WGS data from 214 Malagasy individuals across three localities across Madagascar, integrating anthropometric measurements to explore body composition variation. Population structure was inferred using principal component analysis (PCA) and ADMIXTURE to characterize African and Austronesian ancestry proportions. GWAS, conducted with Hail, identified genetic variants associated with body composition traits. Our results reveal fine-scale variation in African and Austronesian ancestry across Madagascar, reflecting the population's complex demographic history. GWAS identified novel variants influencing body composition, implicating genes involved in metabolic and skeletal pathways. These findings suggest that Malagasy populations harbor unique genetic variants, potentially shaped by natural selection, drift, and historical admixture events. By combining WGS and GWAS, this study provides new insights into the evolutionary dynamics of an admixed population, shedding light on how genetic variation and demographic history contribute to phenotypic diversity. Our findings emphasize the importance of studying underrepresented populations to better understand human genetic diversity and evolution.

Genomic insights into virulence and antimicrobial resistance in Xanthomonas oryzae: implications for plant pathogen management

Onyewuchi Henry Njoku
(University of Ibadan)

Plant pathogenic bacteria pose significant threats to agricultural sustainability and public health, necessitating genomic investigations to understand their pathogenic mechanisms. Xanthomonas oryzae, a major bacterial pathogen of rice, exhibits considerable genomic diversity, influencing its virulence and antimicrobial resistance (AMR) profiles. This study aims to analyze the whole-genome sequences of 80 X. oryzae strains to identify virulence factors, AMR determinants, and mobile genetic elements contributing to pathogenicity and adaptation. Whole-genome sequencing data were processed through quality control measures, functional genome annotation, and comparative genomic analyses, including resistome and virulome profiling. Chromosomal properties such as genome size, GC content, and plasmid composition were examined alongside integrases, transposases, prophages, and DNA polymerases associated with genomic mobility. The results revealed significant variations in virulence-associated genes, with distinct AMR profiles and phylogenetic clustering patterns across strains. The presence of mobile genetic elements highlighted mechanisms of genetic adaptability and resistance evolution. These insights enhance our understanding of bacterial adaptation and pathogenicity, providing a foundation for improved risk assessment strategies in plant disease management and potential mitigation of agricultural and public health risks.

High KIR diversity in Uganda and Botswana children living with HIV.

John Mukisa (Makerere University College of Health Sciences), Samuel Kyobe (Makerere University College of Health Sciences), Marion Amujal (Makerere University College of Health Sciences), Daudi Jjingo (Makerere University College of Health Sciences), Graeme Mardon (Baylor College of Medicine), Matshaba Mogomotsi (Botswana-Baylor Children's Clinical Centre of Excellence), David Patrick Kateete (Makerere University, College of Health Sciences), Moses L. Joloba (Makerere University, College of Health Sciences), Neil Hanchard (National Human Genome Research Institute) and Jill A Hollenbach (University of California San Francisco)

Abstract Killer-cell immunoglobulin-like receptors (KIRs) are essential components of the innate immune system found on the surfaces of natural killer (NK) cells. The KIRs encoding genes are located on chromosome 19q13.4 and are genetically diverse across populations. KIRs are associated with various disease states including HIV progression, and are linked to transplantation rejection and reproductive success. However, there is limited knowledge on the diversity of KIRs from Uganda and Botswana HIV-infected paediatric cohorts, with high endemic HIV rates. We used next-generation sequencing technologies on 312 (246 Uganda, 66 Botswana) samples to generate KIR allele data and employed customised bioinformatics techniques for allelic, allotype and disease association analysis. We show that these sample sets from Botswana and Uganda have different KIRs of different diversities. In Uganda, we observed 147 vs 111 alleles in the Botswana cohort, which had a more than 1 % frequency. In the Ugandan cohort, we also found significant deviation towards homozygosity for the KIR3DL2 gene for both rapid (RPs) and long-term non-progressors (LTNPs). The frequency of the Bw4-80I ligand was also significantly higher among the LTNPs than RPs (109Vs 52, P-value: <0.001). In the Ugandan cohort, KIR2DS4*001 (OR: 0.671, 95 % CI: 0.481-0.937, FDR adjusted Pc=0.142) and KIR2DS4*006 (OR: 2.519, 95 % CI: 1.085-5.851, FDR adjusted Pc=0.142) were not associated with HIV disease progression after adjustment for multiple testing. Our study results provide additional knowledge of the genetic diversity of KIRs in African populations and provide evidence that will inform future immunogenetics studies concerning human disease susceptibility, evolution and host immune responses.

Integrating multi-omics datasets with machine learning algorithms in developing clinical decision support systems for cancer management

Itunuoluwa Isewon (Department of Computer and Information Sciences), Emmanuel Alagbe (Department of Computer and Information Sciences), Solomon Rotimi (Department of Biochemistry) and Jelili Oyelade (Department of Computer and Information Sciences) (Covenant University)

Multi-omics strategies hold great promise for disease prognosis and diagnosis, offering a more comprehensive understanding of biological systems than single-omics approaches. By integrating multiple layers of biological information, multi-omics analyses enable better identification of disease mechanisms, biomarker discovery, and personalized treatment strategies. Machine learning (ML) algorithms are increasingly applied to these datasets to extract meaningful insights, improve disease detection, predict treatment responses, and identify biomarkers inferring susceptibility to diseases. However, despite the growing interest in multi-omics and ML integration, there is a lack of systematic investigation into how different combinations of omics datasets affect ML model performance in clinical decision support systems. This study explores the integration of ML algorithms with multi-omics datasets to predict prostate cancer (PCa) treatment outcomes and biochemical recurrence (BCR) using The Cancer Genome Atlas (TCGA) dataset. We evaluated the predictive performance of nine ML algorithms across 63 possible omics combinations, incorporating six omics data types: single nucleotide variation (SNV), copy number variation (CNV), DNA methylation, RNA sequencing (RNA-seq), microRNA sequencing (miRNA-seq), and reverse-phase protein array (RPPA) datasets. To rank these models and omics combinations, we developed a multi-criteria decision scoring system based on key performance metrics such as accuracy, precision, recall, F1-score, and area under the receiver operating characteristic curve (AUC-ROC). Our results demonstrate that selective multi-omics integration outperforms indiscriminate aggregation. For PCa treatment outcome prediction, the best-performing combinations were CNV + SNV + DNA methylation + miRNA-seq, SNV + DNA methylation, and CNV + DNA methylation + RPPA. For BCR prediction, SNV + DNA methylation ranked highest, followed by SNV + DNA methylation + miRNA-seq and CNV + SNV + DNA methylation + miRNA-seq + RPPA. Notably, while multi-omics generally improved ML model performance compared to single-omics, the combination of all six omics datasets did not yield the best predictive power. Instead, targeted integration of specific omics types proved more effective. XGBoost (xGB) algorithm consistently outperformed other ML models across both tasks. Feature selection (FS) using elastic-net penalized regression yielded superior results compared to feature extraction (FE) via autoencoders. To validate our methodology, we applied the same ML framework to TCGA breast cancer (BRCA) multi-omics datasets for PAM50 subtyping. The best-performing omics combinations for BRCA were SNV + DNA methylation + miRNA-seq, CNV + SNV + DNA methylation + RPPA, and SNV + DNA methylation + miRNA-seq + RPPA. Notably, SNV + DNA methylation alone ranked 30th, reinforcing the importance of carefully integrating complementary omics layers. The BRCA validation confirmed that while multi-omics strategies enhance predictive power, dataset size and phenotype balance play crucial roles. XGBoost emerged as the best-performing algorithm, followed by gradient boosting and support vector machines. In conclusion, this study provides a large-scale investigation into multi-omics data integration with ML for precision oncology. It highlights the need for careful omics selection rather than arbitrary multi-omics aggregation and underscores the importance of addressing class imbalance and feature representation challenges in clinical ML applications. Our findings contribute to the development of more reliable and interpretable AI-driven clinical decision support systems for cancer management.

Feasibility analysis of RNA sequencing data for mutation discovery compared to whole exome sequencing

Magdalene Namuswe (African Centres of Excellence in Bioinformatics and Data Intensive Sciences), Syrus Semawule (African Centres of Excellence in Bioinformatics and Data Intensive Sciences), Gloria Nakabiri (African Centres of Excellence in Bioinformatics and Data Intensive Sciences), Fredrick Kakembo (African Centres of Excellence in Bioinformatics and Data Intensive Sciences) , Daudi Jjingo (African Centres of Excellence in Bioinformatics and Data Intensive Sciences) and Edus Warren (Fred Hutchinson Cancer Center)

Whole exome sequencing (WES) is a genomic technique that sequences all the protein-coding regions of genes in a genome, making it an efficient method for identifying genetic variants across an individual's genes. However, it is limited by its focus on exonic regions, leaving non-coding regions unexamined. RNA sequencing (RNA-Seq), another next-generation sequencing (NGS) approach commonly used for gene expression profiling, has emerged as a complementary method for mutation detection, offering several advantages over WES. The problem lies in determining whether RNA-Seq can reliably identify mutations with similar accuracy and comprehensiveness as WES. Therefore, the aim of this study was to evaluate the feasibility of RNA-Seq for mutation discovery in comparison to WES, given RNA-Seq's increasing application in identifying genetic variants. This study involved a comparison across six samples with both RNA-Seq and WES FASTQ files to detect the number of common and unique variants called by each method. RNA-Seq consistently identified 2 to 25.8 times more variants than WES. The percentage of common variants (those detected by both RNA-Seq and WES) relative to total RNA variants ranged from 0.1% to 11.3%, while for total DNA variants, it ranged from 16.5% to 20.3%. The low overlap observed in this study likely reflects the inherent differences between RNA-Seq and WES methodologies. RNA-Seq captures transcribed RNA molecules, including splicing isoforms, alternative transcripts, and non-coding RNAs, which are specific to the transcriptional state of the cell at the time of sampling, while WES focuses exclusively on the protein-coding regions of the genome. Therefore, integrating both RNA-Seq and WES provides a more comprehensive understanding of genetic variation, rather than suggesting that RNA-Seq could completely replace WES.

1KSA - Decoding South Africa's Biodiversity

Christina Meiring and The One Thousand Genomes For South African Biodiversity Project Consortium Diplomics (DIPLOMICS)

South Africa is one of the most biodiverse countries in the world with many institutions researching and documenting local biodiversity. However, South African scientists often conduct genetic research overseas to take advantage of competitive prices internationally. This contributes to a drain of skills, data, knowledge and opportunity out of South Africa. To mitigate this challenge and build capacity for designing and carrying out biodiversity genomics experiments in country, a South African Biodiversity genomics program called 1KSA (www.1kSA.org.za) was launched, in 2023, by DIPLOMICS, a Genomics, Proteomics, Metabolomics and Bioinformatics Research Infrastructure program based in South Africa and supported by the Department of Science, Technology and Innovation's SARIR program (South Africa Research Infrastructure Roadmap). Following a successful species nomination application, South African sample contributors submit DNA for species of interest to one of several 1KSA partner labs. Whole genome sequencing takes place using Oxford Nanopore Technology PromethION devices and a draft genome assembly is generated using the 1KSA pipeline on the Centre for High Performance Computing. All lab protocols are documented in the DIPLOMICS 1KSA Zenodo community while the 1KSA draft genome assembly pipeline is recorded on GitHub (https://github.com/DIPLOMICS-SA/Genome-Assembly-Pipeline-Nextflow). The resulting data are stored using the Data Intensive Research Initiative of South Africa; are made known via the generation of species information cards on the 1KSA website; and access can be requested through the 1KSA Data Access Committee. Sequenced genomes of biodiversity and economically important species are tools for population genomics studies, conservation efforts, management of the impacts of climate change, and the identification of novel compounds with spin off potential for development of the bioeconomy. An overview of the 1KSA project will be presented, highlighting the outputs of the 1KSA draft genome assembly pipeline and how the metrics are recorded on the 1KSA species information cards published on the 1KSA website.

Discovery of Synergistic Drug Combinations For E. Coli from Drug Information, Pathogen Response and Disease Microenvironment Data.

Racheal Claire Kyomukama (African Centres of Excellence in Bioinformatics and Data Intensive Sciences), Ronad Galiwango (African Centres of Excellence in Bioinformatics and Data Intensive Sciences) and Moses Ainembabazi (Makerere University)

The global rise of antimicrobial resistance (AMR) poses a significant threat to public health, as pathogens become resistant to drugs, rendering many standard infection treatments increasingly ineffective. Among these pathogens, Escherichia coli (E. Coli), a common bacterium listed on the WHO Bacterial Priority pathogens list, has shown significant resistance to multiple antibiotics, necessitating the development of innovative therapeutic strategies. Current efforts are focused on identifying novel compounds that inhibit unexplored molecular targets. However, the process of commercially introducing new drugs is complex, costly, and time-consuming. One promising approach to overcoming this challenge is combination therapy, which involves using two or more drugs simultaneously to treat an infection. Combination therapy can enhance treatment efficacy, and reduce the likelihood of resistance development due to the synergistic effect of drug combinations compared with a single drug. However,identifying effective drug combinations is a complex and challenging task due to the vast number of potential drug pairs and interactions. To address this issue, we propose a novel approach that leverages machine learning (ML) to predict synergistic drug combinations. Unlike previous approaches, our methodology which we apply to E. coli integrates multiple factors that influence drug efficacy, including drug structure information, pathogen response, and disease microenvironment data. Drug structure information, which includes the specific substructures present in certain drugs, provides insights into potential interactions at the molecular level. Pathogen response data, particularly chemogenomic data that measures the sensitivity of single-gene knockout strains to various drugs, offers valuable information on the genetic basis of resistance and susceptibility. Disease microenvironment data is crucial, as antibiotic efficacy can fluctuate significantly under different metabolic conditions, reflecting the complexity of real-world infection scenarios. We analyzed the respective datasets and trained various ML algorithms on the augmented data to predict drug interactions and identify combinations that are likely to be synergistic. The models included random forest (RF) and gradient boost (GB) classifiers, which are known for their robustness and accuracy in handling complex datasets. Model performance was evaluated using several key metrics, including accuracy, precision, and recall. The GB model performed best, with an accuracy of 70%, a precision of 66%, and a recall of 68%. These preliminary results, pending hyperparameter tuning, indicate that the model has a reasonable predictive capability, although there is room for improvement. Nevertheless, the preliminary findings are promising, as the model has identified several drug pairs that could be more efficacious against E. coli than individual drugs. These drug pairs can potentially serve as candidates for further experimental validation and clinical testing. In conclusion, our study demonstrates the potential of ML algorithms to guide the design of effective combination therapies for combating AMR. By incorporating diverse datasets and leveraging advanced computational techniques, we can accelerate the discovery of synergistic drug combinations, ultimately contributing to developing more robust treatments for antibiotic-resistant bacterial infections. Future work will focus on refining the model, expanding the dataset, and conducting experimental validations to confirm the predicted drug interactions.

PREMEDIT: A Centralized Platform for Genetic Diseases and Therapeutic Solutions in Tunisia

Nessrine Mezzi (Laboratory of Biomedical Genomics and Oncogenetics, Institut Pasteur de Tunis, Tunis El Manar University), Imen Abdallah (Laboratory of Biomedical Genomics and Oncogenetics, Institut Pasteur de Tunis, Tunis El Manar University), Maroua Louati (Technologies et Services de l'Information, Technoriat), Nejla Abassi (Laboratory of Biomedical Genomics and Oncogenetics, Institut Pasteur de Tunis, Tunis El Manar University), Abdelwaheb Sifi (MO2 Advisory & Invest), Ridha Mrad (Department of Congenital and Hereditary Diseases, Charles Nicolle Hospital), Mediha Trabelsi (Department of Congenital and Hereditary Diseases, Charles Nicolle Hospital) and Lilia Romdhane (Laboratory of Biomedical Genomics and Oncogenetics, Institut Pasteur de Tunis / Faculty of Sciences of Bizerte)

Genetic diseases (GDs) pose a significant public health challenge as they are a leading cause of morbidity and premature death. In Tunisia, there are 589 reported GDs, with more than 60% being autosomal recessive. In 40% of these cases, the molecular etiology is unknown, highlighting the urgent need for advanced genetic research and diagnostic tools. Addressing this gap is crucial for improving patient outcomes and guiding therapeutic interventions. A combination of manual and AI-assisted text mining from the literature is used to collect complex genetic data on GDs in Tunisia, ensuring data integrity. Python and R scripts are employed for data validation and biological database queries. Bioinformatic approaches, including AI, are being utilized for in-silico drug (re)discovery. Cutting-edge technologies support the development of the PREMEDIT platform. To date, more than 600 GDs have been identified in Tunisian patients, with approximately 1,000 mutations, representing the most comprehensive mutatome of the Tunisian population. Data analyses revealed the scarcity of epidemiological data and treatments for rare GDs. The genetic, epidemiological, and pharmaceutical data have been integrated into a centralized platform: PREMEDIT. By consolidating comprehensive data on genetic mutations and their correlation with specific treatments, PREMEDIT aims to enhance diagnosis and tailor therapeutic strategies for the Tunisian population. The integration of AI not only refines data accuracy but also facilitates the efficient identification of complex genetic patterns, empowering the platform to provide more precise diagnostic and therapeutic recommendations. This platform serves as a crucial resource for healthcare professionals, researchers, and policymakers, bridging the gap between genetic research and clinical practice. PREMEDIT will contribute to innovation across biomedical communities as well as pharmaceutical companies, improving the quality of life for patients.

Epidemiological insights from genomic analyses of the SARS-CoV-2 infection wave in coastal Kenya, 2023-24

Arnold Lambisia (KEMRI-Wellcome Trust Research Programme), Esther Katama (KEMRI-Wellcome Trust Research Programme), Edidah Moraa (KEMRI-Wellcome Trust Research Programme), John Mwita (KEMRI-Wellcome Trust Research Programme), George Githinji (KEMRI-Wellcome Trust Research Programme), Isabella Ochola-Oyier (KEMRI-Wellcome Trust Research Programme), Edward Holmes (School of Medical Sciences, University of Sydney) and Charles Agoti (KEMRI-Wellcome Trust Research Programme)

Between November 2023 and March 2024, coastal Kenya experienced a new wave of severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) infections. Herein, we report the clinical and genomic epidemiology of the wave following analysis of 185 positive samples from 179 individuals residing in the Kilifi Health and Demographic Surveillance (KHDSS) area (~900 km2) during this wave. Sixteen lineages within three sub-variants (XBB.2.3-like (58.4%), JN.1-like (40.5%) and XBB.1-like (1.1%)) were identified. Symptomatic infection rate was estimated at 16.0% (95% CI 11.1%-23.9%) based on community testing regardless of symptom status, and did not differ across the major sub-variants (p = 0.13). The most common infection symptoms in community cases were cough (49.2%), fever (27.0%), sore throat (7.3%), headache (6.9%), and difficulty in breathing (5.5%). Genomic analysis confirmed repeat infections among five participants under follow-up (median interval 21 days, range 16-95 days); in four participants, the same lineage was involved in both the first and second infections, while one participant had a different Omicron subvariant in the second infection compared to the first. Phylogenetic analysis including >18,000 contemporaneous global sequences estimated that at least 38 independent virus introduction events occurred into the KHDSS area during the wave, the majority likely originating in Europe. Our study highlights that in this post-public health emergency of international concern era for coronavirus disease 2019, coastal Kenya, like most other localities, continues to face new SARS-CoV-2 waves characterized by the circulation of new variants, multiple lineage importations and reinfections. Locally the virus may circulate unrecognized as most infections are asymptomatic, highlighting the need for sustained surveillance to inform appropriate public health responses.

Targeting aldose reductase using natural African compounds as promising agents for managing diabetic complications

Miriam Lawson Gakpey (Department of Clinical Pathology, Noguchi Memorial Institute for Medical Research, University of Ghana), Shadrack Aidoo (Department of Virology, Noguchi Memorial Institute for Medical Research, University of Ghana), Toheeb Jumah (School of Collective Intelligence, University Mohammed VI Polytechnic), George Hanson (Department of Parasitology, Noguchi Memorial Institute for Medical Research, University of Ghana), Siyabonga Msipa (Department of Integrative Biomedical Science, Faculty of Health Sciences, University of Cape Town), Florence Mbaoji (Department of Pharmacology and Toxicology, Faculty of Pharmaceutical Sciences, University of Nigeria), Bukola Omonijo (Faculty of Basic Medical Sciences, Ladoke Akintola University of Technology), Palesa Tjale (Department of Computational Biology, Faculty of Health Sciences, University of Cape Town), Mamadou Sangare (African Center of Excellence in Bioinformatics, University of Science, Techniques and Technologies of Bamako), Heidia Tebourbi (Higher Institute of Biotechnology of Sidi Thabet) and Olaitan Awe (African Society for Bioinformatics and Computational Biology)

Background: Diabetes remains a leading cause of morbidity and mortality due to various complications induced by hyperglycemia. Inhibiting Aldose Reductase (AR), an enzyme that converts glucose to sorbitol, has been studied to prevent long-term diabetic consequences. Unfortunately, drugs targeting AR have demonstrated toxicity, adverse reactions, and a lack of specificity. This study aims to explore African indigenous compounds with high specificity as potential AR inhibitors for pharmacological intervention. Methodology: A total of 7,344 compounds from the AfroDB, EANPDB, and NANPDB databases were obtained and pre-filtered using the Lipinski rule of five to generate a compound library for virtual screening against the Aldose Reductase. The top 20 compounds with the highest binding affinity were selected. Subsequently, in silico analyses such as protein-ligand interaction, physicochemical and pharmacokinetic profiling (ADMET), and molecular dynamics simulation coupled with binding free energy calculations were performed to identify lead compounds with high binding affinity and low toxicity. Results: Five natural compounds, namely, (+)-pipoxide, Zinc000095485961, Naamidine A, (−)-pipoxide, and 1,6-di-o-p-hydroxybenzoyl-beta-d-glucopyranoside, were identified as potential inhibitors of aldose reductase. Molecular docking results showed that these compounds exhibited binding energies ranging from −12.3 to −10.7 kcal/mol, which were better than the standard inhibitors (zopolrestat, epalrestat, IDD594, tolrestat, and sorbinil) used in this study. The ADMET and protein-ligand interaction results revealed that these compounds interacted with key inhibiting residues through hydrogen and hydrophobic interactions and demonstrated favorable pharmacological and low toxicity profiles. Prediction of biological activity highlighted Zinc000095485961 and 1,6-di-o-p-hydroxybenzoyl-beta-d-glucopyranoside as having significant inhibitory activity against aldose reductase. Molecular dynamics simulations and MM-PBSA analysis confirmed that the compounds bound to AR exhibited high stability and less conformational change to the AR-inhibitor complex. Conclusion: This study highlighted the potential inhibitory activity of 5 compounds that belong to the African region: (+)-Pipoxide, Zinc000095485961, Naamidine A, (−)-Pipoxide, and 1,6-di-o-p-hydroxybenzoyl-beta-d-glucopyranoside. These molecules inhibiting the aldose reductase, the key enzyme of the polyol pathway, can be developed as therapeutic agents to manage diabetic complications. However, we recommend in vitro and in vivo studies to confirm our findings.

The African Human Microbiome Portal: a public web portal of curated metagenomic metadata

Anmol Kiran (Malawi-Liverpool-Wellcome Trust, Blantyre 3), Mariem Hanachi (Laboratory of Bioinformatics, Biomathematics and Biostatistics , Institute Pasteur of Tunis), Nihad Alsayed (Kush Centre for Genomics and Biomedical Informatics, Biotechnology Perspectives Organization), Meriem Fassatoui (Laboratory of Biomedical Genomics & Oncogenetics, Institut Pasteur de Tunis), Ovokeraye H Oduaran (The Sydney Brenner Institute for Molecular Bioscience, University of the Witwatersrand), Imane Allali (Laboratory of Human Pathologies Biology, Department of Biology, Faculty of Sciences), Suresh Maslamoney (Department of Integrative Biomedical Sciences and Institute of Infectious Disease and Molecular Medicin), Ayton Meintjes (Department of Integrative Biomedical Sciences and Institute of Infectious Disease and Molecular Medicine), Lyndon Zass (Department of Integrative Biomedical Sciences and Institute of Infectious Disease and Molecular Medicine), Jorge Da Rocha (The Sydney Brenner Institute for Molecular Bioscience, University of the Witwatersrand), Rym Kefi (Laboratory of Biomedical Genomics & Oncogenetics, Institut Pasteur de Tunis), Alia Benkhla (Laboratory of Bioinformatics, Biomathematics and Biostatistics , Institute Pasteur of Tunis), Kais Ghedira (Laboratory of Bioinformatics, Biomathematics and Biostatistics , Institute Pasteur of Tunis), Sumir Panji (Department of Integrative Biomedical Sciences and Institute of Infectious Disease and Molecular Medicine), Nicola Mulder (Department of Integrative Biomedical Sciences and Institute of Infectious Disease and Molecular Medicine), Faisal M Fadlelmola (Kush Centre for Genomics and Biomedical Informatics, Biotechnology Perspectives Organization) and Oussema Souiai (Laboratory of Bioinformatics, Biomathematics and Biostatistics, Institute Pasteur of Tunis)

Metadata in public repositories often lack consistency and completeness, limiting the potential for re-use of data in new research contexts. In addition to being underrepresented in research data, African populations face an additional challenge, the existing data rarely reflect their unique genetic, socio-cultural, and demographic diversity. Despite the microbiome's recognized importance in human health, this gap is especially pronounced in metagenomic data. To address this, we developed the African Human Microbiome portal (AHMP) focused on microbiome metadata from African populations, guided by the FAIR principles (findability, accessibility, interoperability, and reproducibility). The portal was supported by the Pan-African Bioinformatics Network for H3Africa (H3ABioNet) and currently offers access to 14,889 curated records across 70 bioprojects, linked to 72 peer-reviewed studies. Recognizing that metadata in repositories is often incomplete, we supplemented it with critical details from the original publications. This includes participant demographics (such as ethnicity, health status, common diets, and age), geographic information (urban or rural context, region), sampling (body site, sample type), sequencing parameters (hypervariable region, assay type, platform), and study design (sample count and links to publications and repositories). Health status terms are standardized via the Disease Ontology, while ethnicity information is cross-referenced with the African Population Ontology (AfPO), a resource developed by the consortium. Through its interactive query interface and data dashboard, the AHMP portal allows users to search and filter by parameters like body site, assay type, country, and disease status. Data visualizations, maps, and charts enhance the ability to refine searches and download results. This resource is freely accessible to the scientific community at https://microbiome.h3abionet.org/, supporting diverse and inclusive research on the African microbiome. This portal is intended to be expanded and maintained through various African initiatives.

An assessment of the genomic structural variation landscape in Sub-Saharan African populations

Zane Lombard (University of the Witwatersrand), Scott Hazelhurst (University of the Witwatersrand), Emma Wiener (University of the Witwatersrand), Laura Cottino (University of the Witwatersrand), Gerrit Botha (University of Cape Town), Oscar Nyangiri (Makerere University), Harry Noyes (University of Liverpool), Annette MacLeod (University of Glasgow), David Jakubosky (University of California San Diego), Clement Adebamowo (University of Maryland), Philip Awadalla (Ontario Institute for Cancer Research. University of Toronto), Guida Landoure (University of Sciences, Techniques and Technology of Bamako), Mogomotsi Matshaba (Botswana-Baylor Children's Clinical Center of Excellence. Baylor College of Medicine), Enock Matovu (Makerere University), Michele Ramsay (University of the Witwatersrand), Gustave Simo (University of Dschang Dschang), Martin Simuunza (University of Zambia), Caroline Tiemessen (National Institute for Communicable Diseases, National Health Laboratory Services. University of the Witwatersrand), Ambroise Wonkam (Johns Hopkins University), Venesa Sahibdeen (University of the Witwatersrand) and Amanda Krause (National Health Laboratory Service. University of the Witwatersrand)

Structural variants (SVs) contribute significantly to human genomic diversity and are implicated in both common and rare diseases. As with most genomic data in the public domain, there is limited representation of SV datasets derived from African populations, creating a critical gap in our understanding of global genomic diversity. To address this underrepresentation, this H3Africa collaboration analysed 1,091 high-coverage African whole genomes, including 546 previously unanalysed genomes for structural variants. We employed an ensemble approach for detecting SVs in whole genome sequencing data, combining five SV detection tools and then merging datasets jointly called through SURVIVOR. This conservative methodology identified 67,795 structural variants across the genome, with SVs observed to impact on 10,421 gene regions. By SV subtype, our analysis revealed 75% deletions, 19% duplications, 4% insertions and 2% inversions, though these proportions reflect algorithmic detection biases. There was significant novelty in the data, 10% being obersved for the first time in this cohort of African individuals. Variants were distributed throughout the genome with 42% occurring in introns, 4% in coding regions and 53% in intergenic regions. Size distribution analysis showed that a third of SVs detected are over 800bp in length. We observed a higher proportion of common variants (17% occurring at >10% frequency) than previously reported in non-African populations, potentially representing a distinctive feature of African structural variant patterns. The potential functional impact of the SVs detected were assessed according to ACMG/AMP classification guidelines. This analysis indicated that the majority of SVs (68%) were classified as variants of uncertain significance. A small portion of SVs were classified as likely pathogenic (0.2%) and only 15 pathogenic variants were identified. Of the latter, the majority (60%) were known African variants that were previously linked to disease. The variants described as pathogenic for the first time (5/15) require further investigation. This study highlights the technical challenges in SV research, including computational intensity and the limitations of short-read sequencing technologies. Different detection algorithms showed complementary strengths across various SV types and sizes, reinforcing the value of ensemble approaches despite their computational demands. Our work provides a valuable resource for population genetics and health-related research, addressing the critical need for high-quality baseline data on structural variant diversity in African populations. This dataset will enhance interpretation of potentially pathogenic variants and improve our understanding of genetic diseases in understudied populations, contributing to more equitable genomic medicine.

OSOC – Open Science for Omics Community

Anelda Van der Walt (Talarify), Aleksandra Pawlik (Independent), Mireille Grobbelaar (Independent), Mhlekazi Molatoli (DIPLOMICS) and Shane Murray (DIPLOMICS)

Open Science refers to the practice of making scientific research, data, and knowledge freely accessible to the public. This facilitates greater transparency, collaboration, and reproducibility in research. It encompasses various aspects, including open access to publications, open data sharing, open-source software, open peer review and more. The goal is to break down barriers that prevent wider dissemination of scientific knowledge, ensuring that research is available to everyone, regardless of institutional affiliation, geographic location or academic status. Many institutions in South Africa wish to embrace Open Science but are unsure how to start. DIPLOMICS, a Genomics, Proteomics, Metabolomics and Bioinformatics (‘omics) Research Infrastructure program based in South Africa and supported by the Department of Science, Technology and Innovation’s SARIR program (South Africa Research Infrastructure Roadmap), commissioned Talarify, a digital and computational capacity development company in South Africa, to design a series of workshops for the Omics community. Branded “Open Science for Omics Community” OSOC, the virtual workshops ran between January and March 2025 and were set out as follows: - Workshop 1: An Introduction to Open Science - Workshop 2: Open Science in the Omics Research Laboratory - Workshop 3: Open Science for Computational Omics Research An accompanying workbook, “Open Science Essentials for South Africa”, was also developed and is available on Zenodo (https://zenodo.org/records/14891809). The structure of the workshops will be presented, along with the results of the post-course evaluation survey.

CLARITY: A Bioinformatics Marketplace

Patricia Swart and Distributed Platform In Omics Diplomics (DIPLOMICS / CPGR)

What comes after sequencing and generating Omics data? Data transfer, analysis, interpretation, storage, management and more. These factors are often not thought of (or budgeted) during the planning stages of an Omics project. Further, inadequate sample size and batch effects, due to poor project design, lead to data that cannot be used to answer the research question. Early engagement with bioinformaticians can significantly improve project success by optimizing project design and budgeting for critical analytical needs. However, where can one access bioinformatics services and what does a bioinformatics service model entail? CLARITY, a bioinformatics marketplace powered by DIPLOMICS (DIstributed PLatform in OMICS, a South African Research Infrastructure Program), aims to address these gaps by highlighting the importance of bioinformatics in any Omics project, enabling access to bioinformatics expertise, and creating work opportunities for bioinformaticians in South Africa. CLARITY services include consultations on project and experimental design, bioinformatics analytical support, and project-specific bioinformatics training. Projects vary in terms of research area, Omics technology, client expertise, training requirements and infrastructure availability. Therefore, there is no one size fits all bioinformatics service model. CLARITY strives to provide personalised bioinformatics solutions that researchers can utilise in their own environments for future projects. The initial consultation is free, with additional services offered at a cost, but steep discounts are made possible by DIPLOMICS. This approach makes bioinformatics support more accessible to South African researchers and highlights the costs associated with bioinformatics, in terms of management, compute, storage and people time. DIPLOMICS' CLARITY initiative is learning by doing and developing a bioinformatics service platform, fit for purpose in the South African research landscape, enabling the success of both small- and large-scale Omics projects and saving researchers valuable time and money.

Design of a Multi-epitope Vaccine Against Drug-Resistant Mycobacterium tuberculosis and Mycobacterium bovis using Reverse vaccinology

Eva Akurut (African center of Excellence in Bioinformatics and Data intensive sciences), Yahaya Gavamukulya (Department of Biochemistry and Molecular Biology, Faculty of Health Sciences, Busitema University), Julius Mulindwa (Department of Biochemistry and Sports Science, College of Natural Sciences, Makerere University), Moses Isiagi (Department of Medicine, Division of Pulmonology, University of Cape Town), Ronald Galiwango (African center of Excellence in Bioinformatics and Data intensive sciences), Mudashiru Bbuye (Makerere University Lung Institute, College of Health Sciences), Ibra Lujumba (African center of Excellence in Bioinformatics and Data intensive sciences), Davis Kiberu (African center of Excellence in Bioinformatics and Data intensive sciences), Patricia Nabisubi (African center of Excellence in Bioinformatics and Data intensive sciences), Grace Kebirungi (African center of Excellence in Bioinformatics and Data intensive sciences), Andrew Kambugu (Infectious Disease Institute), Barbara Castelnuovo (Infectious Disease Institute), Gyaviira Nkurunungi (Medical Research Council/Uganda Virus Research Institute and London School of Hygiene and Tropical Medicine), Daudi Jjingo (African center of Excellence in Bioinformatics and Data intensive sciences), Brenda Oketch (International AIDS Vaccine Initiative), David Patrick Kateete (Makerere University) and Gerald Mboowa (African center of Excellence in Bioinformatics and Data intensive sciences)

The increasing global burden of Mycobacterium tuberculosis (M. tb), alongside the rise of drug-resistant strains, despite vaccination efforts calls for the creation of novel and more effective vaccines. This study employed an in-silico approach to design a multi-epitope vaccine targeting a conserved PE_PGRS 16 protein, a key virulence factor that is present across multiple Mycobacterium species, including drug-resistant strains. PE_PGRS 16 was identified as a promising vaccine candidate due to its extracellular localization, strong adhesion properties, and virulence characteristics. Epitopes for B-cells, Cytotoxic T Lymphocytes (CTL), and Helper T Lymphocytes (HTL) were selected based on antigenicity, non-toxicity, and their potential to elicit a robust immune response. These epitopes were computationally linked to form a stable, flexible multi-epitope vaccine construct. The designed construct demonstrated favourable properties, including high antigenicity, solubility, and stability, with the lowest instability index of -31.31 with the lowest binding energy (-44.566) on molecular docking with TLR4 (PDB ID: 4G8A), suggesting its potential role in immune activation, though further experimental validation is needed to confirm its immunostimulatory effects. The incorporation of Griselimycin as an adjuvant further boosted the vaccine's immunogenicity, as confirmed by simulations using the C-ImmSim algorithm, which predicted optimal immune system interactions. This study presents a novel multi-epitope vaccine leveraging the conserved PE_PGRS 16 protein, with the potential to target multiple TB strains, including drug-resistant forms. Further validation through in vitro, in vivo testing, and clinical trials will be crucial to establish the vaccine’s efficacy and safety.

Developing a Genome-Scale Metabolic Model of European Seabass (GEM) for Immune Response Analysis

Alia Benkahla (Laboratory BioInformatics, bioMathematics and bioStatistics (BIMS), Institut Pasteur de Tunis,), Philipp Schneider (Department of Molecular Biology, Massachusetts General Hospital), Oussema Souiai (Laboratory BioInformatics, bioMathematics and bioStatistics (BIMS), Institut Pasteur de Tunis), Mariem Hanachi (Laboratory BioInformatics, bioMathematics and bioStatistics (BIMS), Institut Pasteur de Tunis), Feriel Guennich (Laboratory BioInformatics, bioMathematics and bioStatistics (BIMS), Institut Pasteur de Tunis), Ridha Bouallegue (École Supérieure des Communications de Tunis (SUP'COM), Tunis), Nadia Cherif (Aquaculture Laboratory, National Institute of Sea Sciences and Technologies), Balkiss Bouhaouala-Zahar (Laboratory of Venoms & Therapeutic Molecules, Institut Pasteur de Tunis) and Haitham Sghaier (Laboratory Energy and Matter for Development of Nuclear Sciences, CNSTN)

Objective: The European seabass (Dicentrarchus labrax) is an economically important aquaculture species. Understanding its metabolic response to vaccination is crucial for improving disease prevention strategies. This work outlines ongoing efforts to develop an accurate and comprehensive genome-scale metabolic model (GEM) of the European seabass using a zebrafish (Danio rerio) GEM as a template. The Zebrafish GEM, downloaded from GitHub, was chosen for its inclusion of immune response enzymes, a crucial factor for this study. This model served as the basis for developing Seabass GEM v1.2. To ensure FAIR (Findable, Accessible, Interoperable, Reusable) compliance and promote data interoperability, the model has to be developed using standardized BIGG identifiers and descriptions for metabolites, reactions, as well as standard gene description extracted from the seabass gene catalog available on ENSEMBL. Methods: Specifically, the Zebrafish GEM's metabolite list was extracted from Zebrafish GEM and annotated with BIGG identifiers and descriptions. Since the Zebrafish GEM's gene list contained NCBI gene names, these were first mapped to corresponding ENSEMBL zebrafish gene IDs, then their orthologous in the current ENSEMBL seabass gene catalog were identified, thus enabling the transfer of information from the zebrafish model to the seabass model. These seabass gene annotations were integrated into Seabass GEM v1.2, enhancing its annotation and accuracy. Results and perspectives: This standardized annotation of the Seabass GEM v1.2's genes and metabolites ensures the model adherence to BIGG database standards. Furthermore, it will facilitate the integration of omics expression data (RNA-Seq and proteomics) to identify key metabolic pathways modulated by vaccination. Thereafter, a Flux Balance Analysis (FBA) will be employed to predict metabolic fluxes and investigate the effects of vaccination on seabass metabolism. Ultimately, the final model will be a valuable tool for understanding and manipulating the metabolic response of seabass to vaccination. Acknowledgements: This work was funded in parts by the MEHSR Ministry and the TUN5032 project funded by IAEA organization.

Increasing Africa's Literacy in Bioinformatics through the Introduction to Bioinformatics Training Course

Sindiswa Lukhele (Computational Biology Division, Institute of Infectious Diseases, Faculty of Health Sciences, University of Cape Town), Shaun Aron (Sydney Brenner Institute for Molecular Bioscience, University of the Witwatersrand), Tshinakaho Malesa (Computational Biology Division, Institute of Infectious Diseases, Faculty of Health Sciences, University of Cape Town), Anwani Siwada (Computational Biology Division, Institute of Infectious Diseases, Faculty of Health Sciences, University of Cape Town), Pertunia Mutheiwana (Computational Biology Division, Institute of Infectious Diseases, Faculty of Health Sciences, University of Cape Town), Sumir Panji (Computational Biology Division, Institute of Infectious Diseases, Faculty of Health Sciences, University of Cape Town), Suresh Maslamoney (Computational Biology Division, Institute of Infectious Diseases, Faculty of Health Sciences, University of Cape Town) and Nicola Mulder (Computational Biology Division, Institute of Infectious Diseases, Faculty of Health Sciences, University of Cape Town)

Bioinformatics is a crucial discipline for the analysis and interpretation of biological data. In Africa, where health challenges, agricultural development, and environmental conservation are critical issues, bioinformatics offers a remarkable opportunity to transform research and innovation. However, with limited access to basic bioinformatics training at many institutions across the continent, there is a general lack of literacy in the use of bioinformatics tools for life science research. To address this gap, in 2016, H3ABioNet initiated the Introduction to Bioinformatics Training (IBT) course, pioneering a novel remote classroom approach to ensure broad participation across Africa. IBT continues to be one of the most popular introductory bioinformatics programs in Africa among life science researchers, with more than 9,000 participants having enrolled in the course over the years. Here, we will highlight the progress made by IBT, its role in building bioinformatics capacity and approaches to ensuring sustainable training in the region. From 2016 to 2024, participant enrolment in IBT increased from 356 to 1546. Participants were enrolled across 40-50 classrooms in 23 African countries. Subsequent to the successful IBT training, we launched the H3ABioNet IBT-Model Train the Trainer initiative, established the Train the Trainer community, and implemented the IBT regional coordinator program. The H3ABioNet IBT-Model Train the Trainer program trained 31 hosts and staff of the IBT in planning and executing regional training, with a focus on blended learning models. The IBT-Model Train the Trainer program resulted in the initiation of the Regional Coordinator program aimed at strengthening the expertise of IBT staff in training coordination. So far, six regional coordinators have been selected to oversee IBT logistics across East, West, and North Africa. Additionally, participants of the IBT-Model Train the Trainer were incorporated into the Train the Trainer Community to enhance the capabilities of trainers in bioinformatics and data science. The IBT program also led to activities around FAIR training and increased access to training materials through the translation of training materials to French and Arabic. These initiatives have improved the basic bioinformatics knowledge and skills of thousands of course participants across Africa while consistently advancing training efforts across the continent and beyond.

Menopause-related changes in the gut microbiome and their association with cardiometabolic diseases in women from four sub-Saharan African countries

Phehello Chauke (Sydney Brenner Institute for Molecular Biosciences, University of the Witwatersrand), Luicer Olubayo (Sydney Brenner Institute for Molecular Biosciences, University of the Witwatersrand), Dylan Maghini (Sydney Brenner Institute for Molecular Biosciences and Stanford University Department of Hematology) and Scott Hazelhurst (Sydney Brenner Institute for Molecular Biosciences, University of the Witwatersrand)

Background: The menopausal transition has been associated with changes in the gut microbiome (GM). This compositional shift is likely related to changing hormone levels during menopause: the GM harbours bacterial taxa that can deconjugate estrogen and other sex hormones, allowing reabsorption of sex hormones. Estrogen also maintains gut homeostasis by influencing the intestinal barrier function and microbial composition. The altered GM composition accompanying the decreased hormone levels may be partially responsible for the onset of menopause-related health conditions, including cardiometabolic diseases (CMDs). However, it is unclear which microbiome features are associated with menopause-related health outcomes particularly in the context of African populations. Aim and objectives: This study investigated compositional differences in the GM between pre- and postmenopausal women in sub-Saharan Africa and their association with CMDs by characterising alpha and beta microbial diversity patterns, identifying differentially abundant bacterial taxa between menopausal groups and determining how these microbial taxa may be linked to CMDs. Methods: The cross-sectional analysis included 1,801 women from Burkina Faso, Ghana, Kenya and South Africa that were selected from the Africa-Wits INDEPTH partnership for Genomic studies (AWI-Gen) wave 2. Shotgun metagenomic sequencing was performed on DNA extracted from faecal samples using Illumina technology. The metagenomic reads underwent quality control processing and alignment, followed by taxonomic profiling. Microbial diversity and composition were assessed using Inverse Simpson index and Bray-Curtis dissimilarity between menopausal groups. Linear discriminant analysis Effect Size was used to identify differentially abundant taxa between menopausal groups. Results: Postmenopausal women had lower microbial diversity than premenopausal women. Women with CMDs showed reduced microbial diversity regardless of menopausal status, suggesting CMD status as a stronger determinant of microbial diversity than menopause alone. Our taxonomic analysis revealed that premenopausal women were enriched with beneficial short-chain fatty acid (SCFA) producers including Faecalibacterium prausnitzii, Bacteroides fragilis, and Prevotella. In contrast, postmenopausal women showed increased abundance of species with varying metabolic implications, including both potentially beneficial (Ruminococcus champanellensis, Fusicatenibacter saccharivorans) and potentially harmful (Collinsella bouchesdurhonensis) bacteria. This study is the first investigation of menopause-related changes in the GM and their association with CMDs in African women. Conclusions: The taxonomic shifts observed in postmenopausal women likely reflect the relationship between estrogen and gut microbiota. While some beneficial SCFA producers decrease, others may increase as compensatory mechanisms. However, the enrichment of potentially harmful species suggests these compensatory changes may be insufficient to maintain optimal metabolic health. This altered microbial ecosystem, combined with reduced estrogen-mediated regulation of intestinal barrier function, could create conditions favouring CMD development in postmenopausal women. Overall, this study highlights significant shifts in GM composition associated with menopausal status, revealing key differences in microbial diversity, taxonomic abundance, and microbial profiles and the influence of CMDs, reinforcing the need for further research to elucidate their long-term implications.

FPGA Acceleration of GWAS Permutation Testing

Yaniv Swiel (University of the Witwatersrand, Max Planck Institute for Evolutionary Anthropology), Jean-Tristan Brandenburg (University of the Witwatersrand), Mahtaab Hayat (University of the Witwatersrand), Wenlong Carl Chen (University of the Witwatersrand), Mitchell Arij Cox (University of the Witwatersrand)and Scott Hazelhurst (University of the Witwatersrand)

Background: Genome-wide association studies (GWASs) analyse genetic variation over the genomes of many individuals in an attempt to identify single nucleotide polymorphisms (SNPs) associated with complex phenotypes. To capture a large amount of genetic variation and increase the chance of detecting associated SNPs, modern GWASs include millions of SNPs sampled from thousands of individuals. Including many SNPs in a GWAS gives rise to a phenomenon known as the multiple testing problem, whereby the cumulative probability of incorrect associations increases with the number of SNPs included in the analysis. A GWAS, therefore, needs to compensate for multiple hypothesis testing by controlling the number of incorrect associations but many of the methods used to control for multiple testing are too strict and can miss valid associations. Permutation testing is a straightforward and accurate method of controlling the false positive rate, but it is very computationally expensive (and slow) so there is a need for a permutation testing accelerator that can process modern GWAS datasets in reasonable time. FPGAs (Field-Programmable Gate Arrays) are reconfigurable integrated circuits which provide a high level of parallelisation that can be harnessed to accelerate GWAS permutation testing. Results: This work presents an accessible FPGA-based tool (designed to run on a cloud-based AWS EC2 FPGA instance) that significantly accelerates GWAS permutation testing for continuous phenotypes. The tool implements two known GWAS permutation testing algorithms: maxT permutation testing and adaptive permutation testing. The speed of the FPGA-based tool was compared to the speed of PLINK (a popular CPU-based tool) running on 40 Intel Xeon 4114 CPU cores using an imputed breast cancer dataset of 13.7 million SNPs sampled from 3652 individuals. For 1000 maxT permutations, the FPGA-based algorithm's run time was 22 minutes while PLINK's run time was almost 7 days; for 100 million adaptive permutations, the FPGA-based algorithm's run time was 325 minutes and PLINK's run time was about 8.5 days. For 700 million adaptive permutations of the same dataset (an almost infeasible workload for PLINK running on a 40-core CPU in terms of the computational run time, which can be extrapolated to be at least one month) the run time of the FPGA-accelerated algorithm was 33 hours. The tool is designed to run on an AWS AMI, requiring no knowledge of FPGAs to use. Conclusions: FPGAs allow permutation testing to be highly parallelised, resulting in performance at least an order of magnitude faster than a conventional software implementation on a cluster. Our tool makes use of cloud-based FPGA instances, which enables an accessible solution to rapid permutation testing without any requirement for FPGA-expertise or cluster access.

Afrigen-D Imputation Service: A Comprehensive Platform for African-Specific Genotype Imputation and Polygenic Risk Score Calculation

Mamana Mbiyavanga, Nicola Mulder and Ayton Meintjes
(Afrigen-D, University of Cape Town)

Over the past decade, the Human Heredity and Health in Africa (H3Africa) initiative has driven the development of genomic research for human health in Africa through its bioinformatics network (H3ABioNet). Through collaborative efforts, H3ABioNet has established robust frameworks for data processing, quality control, and imputation pipelines specifically optimized for African populations. An African genotype imputation service with a comprehensive reference panel is indispensable for accurate genetic analyses tailored to the continent's diverse genetic landscape. We developed an imputation platform (Afrigen-D Imputation Service, https://impute.afrigen-d.org) that leverages the high-quality H3Africa reference panel, comprising 8,894 high-coverage haplotypes from 48 populations worldwide, with 50% of African ancestry. The service implements established guidelines and workflows while addressing data privacy challenges by maintaining genetic data within continental boundaries. It utilizes the validated software stack and workflow architecture of the Michigan Imputation Server and TopMed Imputation Service, ensuring methodological consistency and standardization of genetic imputation procedures. This enables the combination of genotype data after imputation with multiple reference panels. Additionally, the platform integrates an HLA reference panel and incorporates polygenic score (PGS) calculation capabilities, enabling automated standardized computation of genetic risk scores from imputed genotypes. The Afrigen-D Imputation Service facilitates efficient genotype imputation through a user-friendly interface, requiring minimal computational expertise and resources. The platform provides comprehensive preprocessing utilities for automated quality control and data preparation, adhering to established bioinformatics standards. Integration with population-specific reference panels and polygenic scoring capabilities provides a robust foundation for investigating complex diseases and genetic traits in African populations. Through ongoing development and community collaboration, this resource contributes significantly to advancing our understanding of African genetic diversity and its implications for health outcomes.

Acquisition and persistence of Extended-spectrum beta-lactamase (ESBL) and Carbapenem resistant (CRE) Escherichia coli carriage in hospitalized Kenyan children

Caroline Tigoi (KEMRI/Wellcome Trust), James Berkley (KEMRI/Wellcome Trust) and Nicole Stoesser (Nuffield Department of Medicine, Oxford University)

Introduction: Antimicrobial Resistance (AMR) and lack of new drugs poses a serious public health threat. Carriage of AMR may be important drivers of inpatient and post-discharge mortality risk in Low Middle-Income countries (LMICs) despite following guidelines. ESBL and CRE are important as proxies for broad multi-class resistance spread on mobile genetic elements that promote horizontal gene transfer intra- and inter-species in hospitals and communities. We hypothesise that intestinal colonisation and carriage is a possible means of transmission of AMR and a precursor to invasive disease. Methods: This was a prospective cohort study enrolling children admitted to 3 Kenyan hospitals followed for 6 months after discharge and well community controls. Detailed demographic, clinical, and antimicrobial use data were collected along with blood and rectal swab culture. We carried out short and long read whole genome sequencing of 486 E.coli isolates to detect AMR and virulence genes and assess genetic relatedness at gene, mobile genetic element, and strain level through core genome phylogeny. Results: Of the 804 inpatient participants, 291 (36%) carried ESBL-E at admission, 447/630 (71 %) at discharge, 199/455 (44%) at day 45, 152/457 (33%) at day 90 and 120/452 (27%) at 180 days post-discharge from hospital. The baseline ESBL-E carriage prevalence among healthy community participants was 65/404 (16%). Acquisition of ESBL carriage in hospital was associated with prior hospitalization, prior use of antibiotics, prolonged stay in hospital and antimicrobial classes use; and with outcomes of post-discharge death or readmission after adjusting for potential confounders. CPE of up to 26 (6%) and 4 (8%) during readmission were seen in Nairobi site. E. coli isolates were diverse across pathotypes with 12 of the 14 E. coli phylogroups identified globally present including those associated with invasive disease; D3, B1, B2 and D1. Sequence types linked to invasive disease like ST 131, ST 410 and ST 38 were also identified and concordance in ST types among invasive and carriage isolates seen. Several AMR genes cutting across all classes of antibiotics and virulence genes were identified with the leading ESBL gene being blaCTX-M-15 and CRE gene blaNDM-5. Conclusions: There was significant AMR acquisition before and during hospitalisation that took more than six months to return to community level. Carriage and invasive ST types were similar. Further genomic studies and antimicrobial trials to monitor changes on the whole microbiome and calculation of invasiveness of the ST types and phylogroups should be conducted for infection control.

MOLECULAR INVESTIGATION OF MARFAN SYNDROME IN TUNISIA: A MULTIDISCIPLINARY STUDY USING GENOMIC SEQUENCING, BIOINFORMATICS, AND AI-BASED APPROACHES

Imen Abdallah (Laboratory of Biomedical Genomics and Oncogenetics (LR20PT05), Institut Pasteur de Tunis), Nesrine Mezzi (Laboratory of Biomedical Genomics and Oncogenetics (LR20PT05), Institut Pasteur de Tunis), Wajdi Arfa (Pediatric Orthopedic Department, Kassab Institute, El Manar University), Sami Bouchoucha (Children's Hospital Béchir Hamza, Tunis Tunisia Pediatric Orthopedic Surgery Department), Thouraya Ben Younes (Pediatric Neurology Department. Mongi Ben Hmida National Institute of Neurology), Zied Jlailia (Pediatric Orthopedic Department, Kassab Institute, El Manar University), Ilhem Turki-Youssef (Pediatric Neurology Department. Mongi Ben Hmida National Institute of Neurology), Ichraf Kraoua (Pediatric Neurology Department. Mongi Ben Hmida National Institute of Neurology), Mourad Jenzri (Pediatric Orthopedic Department, Kassab Institute, El Manar University) and Lilia Romdhane (Laboratory of Biomedical Genomics and Oncogenetics (LR20PT05), Institut Pasteur de Tunis)

Genetic diseases represent a major public health challenge as they are a leading cause of chronic morbidity and premature death. In Tunisia, 589 genetic disorders have been reported, with 60% being autosomal recessive, 23% autosomal dominant, and 40% with an unknown molecular etiology. Among these, Marfan syndrome (MS) is a an autosomal dominant connective tissue disorder that primarily affects the cardiovascular, ocular, and skeletal systems, profoundly impacting patients' quality of life. As the molecular etiology of MS is unknown in the Tunisian population, we conducted a multicentric study in collaboration with referring clinicians. Initially, Sanger sequencing was performed to identify targeted FBN1 gene mutations, followed by whole exome sequencing (WES) and advanced bioinformatics analyses, molecular modeling, and docking to characterize new mutations. Genomic data analysis coupled with AI-driven models will uncover genetic patterns linked to MS, potentially revealing new therapeutic targets. Additionally, virtual screening will identify small molecules capable of modulating FBN1 mutations. To date, 15 families with suspected MS have been recruited, comprising 17 Tunisian patients, 7 of whom were born to consanguineous unions from different regions of Tunisia. Clinical investigation revealed heterogeneity. Sanger sequencing identified two FBN1 mutations in two families, while the remaining cases are candidates for WES. Raw high-throughput sequencing data will be processed using advanced bioinformatics tools integrating artificial intelligence to filter and prioritize variants followed by Sanger sequencing for variant validation and Mendelian segregation. Further molecular modeling and docking, will be performed to investigate how new mutations affect protein structure, function, and interaction. The goal is to provide an efficient molecular diagnostic tool, establish genotype-phenotype correlations, and offer valuable insights to clinicians for developing tailored management programs. Newly identified mutations and therapeutic molecules will be integrated into the genetic disease platform in Tunisia called PREMEDIT that is expected to accelerate the genetic disease diagnosis and in silico drug discovery through genomic data integration, bioinformatics, and AI.

INSIGHTS ON PATTERNS OF CLONAL SPREAD OF MULTI DRUG-RESISTANT TUBERCULOSIS IN THE WESTERN CAPE, SOUTH AFRICA

Zainab Kashim-Bello, Johannes Loubser, Justice Tresor Ngom, Robin Warren, Elizabeth Streicher and Marisa Klopper
(Division of Molecular Biology and Human Genetics, Faculty of Medicine and Health Sciences, Stellenbosch University)

Tuberculosis (TB) is a critical concern in public health on a global scale. Mycobacterium tuberculosis is the principal causative agent of the disease, and its severity has increased due to the development and spread of strains that are multi-drug resistant (MDR) or extensively drug resistant (XDR). The Western Cape Province in South Africa is heavily burdened with MDR-TB, yet recent studies that seek to carefully track the complex evolution and population dynamics of MDR-TB and understanding why they evolve to become outbreaks are scarce. We aimed to compare the resistance patterns of different clonal clusters of DR-TB in the Western Cape, South Africa, at the genomic level. Samples (582) belonging to MDR-TB and XDR-TB clusters from DRF150, West Coast and Beijing outbreaks in the Western Cape province were selected retrospectively. Isolates were sub-cultured and DNA extracted, followed by WGS on an Illumina platform. Bioinformatic analyses were carried out using some pipelines, including TB-Profiler, MTBseq, and IQTREE tool. Results confirmed that the isolates belonged to the East Asian lineage (2.2.2; Beijing) and Euro-American lineage (4.1.1.3; DRF150 and West Coast). The type of drug resistance (DR) within the samples was noted: Sensitive - 0.5%, HR-TB - 3%, MDR - 28%, Pre-XDR-TB - 64% and XDR – 4.5%. The most prevalent rifampicin resistance mutation was rpoB S450L in DRF150 and West Coast samples, while rpoB D435V was found in Beijing samples. Similarly, the resistance mutation was identified for other drugs including isoniazid, ethambutol, fluoroquinolones, ethionamide and streptomycin), which had some similar mutations, while others varied based on the outbreak sample. In conclusion, by using WGS to compare the samples, we were able to gain insight into the genomic diversity and drug resistance mutation circulating in the Western Cape. Further studies will focus on finding commonalities within the outbreaks and possible transmission chains, which will inform surveillance decisions to curb the spread of DR-TB in the Western Cape, South Africa.

Leveraging Machine Learning for COVID-19 Outbreak Prediction through Wastewater Surveillance in the Western Cape, South Africa

Setshaba Taukobong (South African Medical Research Council), Renee Street (South African Medical Research Council), Sam Scarpino (Northeastern University), Sizwe Nkambule (South African Medical Research Council), Rabia Johnson (South African Medical Research Council) and Craig Kinnear (South African Medical Research Council)

Introduction Wastewater-based epidemiology (WBE) has emerged as an effective complementary tool for COVID-19 surveillance, providing valuable insights into disease dynamics at community level, particularly in areas with limited testing capacity. It allows for the detection of pre-symptomatic and asymptomatic carriers and offers early warning signals of changes in pandemic trends. Several challenges do however persist in COVID-19 wastewater surveillance, including the difficulty in making accurate quantitative predictions of infected cases and potential surges due to fluctuations in SARS-CoV-2 RNA concentrations. Conventional predictive models, including regression approaches, have been employed to tackle these challenges, but these often incorporate very few parameters and features. This study proposes exploiting a multi-model architecture of long short-term memory, a transformer model, and a large language model, integrating RNA concentrations from wastewater, confirmed clinical cases, and wastewater sequencing data for surge predictions. Methods A total of 328 raw sewage samples were collected weekly, concentrated, and RNA extracted weekly, from 8 major wastewater treatment plants in the Western Cape, between August, 2021 to February, 2022, during the delta and omicron wave. SARS-CoV-2 levels in wastewater were detected by RT-PCR and positive samples were quantified as described in Johnson et al.,2022. Furthermore, the extracted nucleic acids were subjected to multiplexed amplicon-based whole genome sequencing as described in Johnson et al.,2022. Sequence reads were trimmed, filtered, and mapped to the reference genome (NC_045512.2) for variant calling. A long Short-Term Memory (LSTM), a transformer model, and a large language model (LLM) were used for surge predictions. Results Comparisons between wastewater and clinical data indicated that wastewater data is an indicator of COVID-19 surges. The performance of each model was evaluated using a set of metrics including accuracy, precision and mean squared error, with cross-validation used to ensure robustness. The LSTM model demonstrated a strong ability to capture temporal dependencies, while the Transformer model outperformed in handling complex relationships between different wastewater treatment plant locations and the clinical data. Furthermore, the models were trained to identify and classify "surge" events, defined as periods when the viral load in wastewater exceeds a threshold based on historical RNA concentration data and clinical case trends. Conclusion This research highlights the promise of combining several models for wastewater surveillance to improve early detection of SARS-CoV-2 outbreaks. Furthermore, it opens avenues for expanding the application of these models to other infectious diseases, improving pandemic preparedness and response strategies. References: Johnson, R., Sharma, J.R., Ramharack, P. et al. Tracking the circulating SARS-CoV-2 variant of concern in South Africa using wastewater-based epidemiology. Sci Rep 12, 1182 (2022). https://doi.org/10.1038/s41598-022-05110-4

Identifying genetic biomarkers of dilated cardiomyopathy using whole exome sequencing data from Southern African patients

Phelelani T. Mpangase (University of the Witwatersrand), Minenhle P. Mayisela (University of the Witwatersrand), Dineo Mpanya (University of the Witwatersrand), Megan Shuey (Vanderbilt University), Roy Zent (Vanderbilt University), Quinn Wells (University of the Witwatersrand) and Nqoba Tsabedze (University of the Witwatersrand)

The underlying genetic architecture of dilated cardiomyopathy (DCM) in Southern Africa has not been described despite the high prevalence of this condition in patients residing in this region. The availability of multiple “omics” techniques for genomics sequencing, including whole exome sequencing (WES), at reduced costs is slowly enabling the study of many diseases affecting African populations in under-resourced settings. This study was aimed at determining the underlying genetic aetiology of DCM in patients from Southern Africa using WES. A cohort of 100 unrelated patients with heart failure of unknown origin were recruited from Charlotte Maxeke Johannesburg Academic Hospital (CMJAH) and subjected to WES. The cohort consisted of participants of ages between 16 and 77 years (47 years average), of whom 67% were males and 92% identified as black. The median left ventricular ejection fraction was 26.5% (interquartile range: 16 – 37), and late gadolinium enhancement was visualised in 42% of participants. Variant calling was carried out on the WES data following the Genome Analysis Toolkit (GATK) Best Practices for WES data analyses, and the resulting variants annotated using Ensemble’s Variant Effect Predictor (VEP). Through various bioinformatics techniques, in combination with genetic- and clinical-guided interpretations, we identified and prioritised several genetic variants in BAG3, FLNC, DSP, MYH7 and TTN genes that have potential roles in the pathogenicity of DCM. This study not only presents potential DCM causal variants but also lays foundation for WES data analyses workflows for similar studies utilising WES to determine the underlying genetic aetiology of diseases in the under-resourced African settings.

Autophosphorylation and Ca2+-binding alter the conformational landscape of Plasmodium falciparum Ca2+ dependent protein kinase 1, impacting stability and ligand binding: contextualising PTM using computational modelling

Charmaine Chido Matimba, Marushka Soobben, Ikechukwu Achilonu and Okechinyere Achilonu
(Wits University)

Malaria remains a significant global public health issue, causing over 600,000 deaths annually. One promising research direction is disrupting crucial pathways, such as the cell signalling mechanisms enabling malaria parasites to grow and survive. By focusing on these pathways, scientists aim to develop a new generation of antimalarial drugs capable of effectively addressing drug resistance and improving treatment options for vulnerable populations worldwide. Post-translational modifications (PTMs), such as phosphorylation, can significantly change protein structures. This poses challenges for computational approaches like computer-aided drug design (CADD), where even slight structural changes can affect ligand binding and functionality. This study investigates how autophosphorylation and Ca2+ binding influence the conformational dynamics of PfCDPK1 using computational modelling, mainly through molecular dynamics simulations (MDS) and high-throughput virtual screening (HTVS). By analysing the changes in protein dynamics, the research may reveal important insights into the druggability of protein kinases, facilitating the design of more effective drugs. The results demonstrated notable variations in the dynamic behaviour of the four systems with or without the ligand (BKI-1294) based on metrics such as Cα-RMSD, Cα-RMSF, radius of gyration (RoG), and ligand properties. The findings suggest that Ca2+ binding alone results in structural changes in the conformity of the protein over time and Ca2+ binding and autophosphorylation enhances structural stability. While phosphorylation alone leads to significant structural deviations, with statistically significant differences observed amongst all systems. Phosphorylation, particularly autophosphorylation, and Ca2+ binding to CDPK1 may reshape the conformational landscape of the enzyme. Such structural changes could influence its functionality, including substrate binding and allosteric inhibition. Ultimately, this study elucidated how these modifications affect the structure and function of PfCDPK1, providing insights into the molecular mechanisms that regulate enzyme activity and calcium homeostasis in Plasmodium falciparum.

Prelude - Building Research Data Platforms Incrementally: A Guide for Teams of All Sizes

Mitchell Shiell, Jon Eubank, Justin Richardsson, Leonardo Rivera, Brandon Chan, Robin Haw, Lincoln Stein, Melanie Courtot and Overture Team (Ontario Institute of Cancer Research (OICR))

Large-scale data platforms enable researchers and the public to access, manage and study massive amounts of genomics data. While small research teams can generate these massive datasets, they often struggle to build the platforms needed for transparent and reproducible FAIR data management and sharing. We built Overture, a suite of reusable, open-source software to develop reliable data management systems quickly, flexibly and at multiple scales. Overture successfully underpins many large-scale international data platforms, including ICGC-ARGO which aims to store genomic and clinical data for over 100,000 participants and VirusSeq which hosts data for over 500,000 pathogen genomes. Behind these platforms are large organizations with large teams that plan, develop and deploy the Overture suite with relative ease. Yet, this can be prohibitively demanding for smaller research groups. How can we help them build data platforms more efficiently and with fewer resources? We address this challenge with Prelude, a tool that enables teams to incrementally build their data platforms by breaking down development into systematic phases. Prelude focuses on solving a specific challenge in platform adoption: the high technical overhead and configuration burden required during the planning and development stages. By breaking down data portal development into phased steps, teams can systematically verify requirements through hands-on testing, which provides clear insights into user workflows, data needs, and overall platform fit. Prelude guides teams through three progressive phases of data platform development. Each phase builds upon the previous one's foundation and can be deployed locally with a single command: - Phase one focuses on data exploration and theming, enabling teams to visualize and search their tabular data through a customizable portal UI; - Phase two expands capabilities to enable tabular data management and validation with persistent storage; - Phase three adds file management and object storage. Prelude also includes configuration generation services that validates and transforms your data into key configuration files; this greatly reduces time spent doing tedious manual configurations. With early adopters reporting significant reductions in configuration time, Prelude is enabling teams to transition through initial planning and development stages efficiently. Looking ahead, we are focused on enabling teams to independently transition to production settings. We are sharing this work to gather community feedback on our approach and learn from others' experiences. Prelude represents a practical step toward making data platform development accessible to research teams with limited resources, reducing initial barriers so teams can do more with less.

Towards an AI-empowered mobile application for microscopy image analysis in low resources settings

Nesrine Fekih-Romdhane, Donia Driss, Rafeh Oualha and Emna Harigua (Institut Pasteur de Tunis)

Infectious diseases constitute a significant global health challenge, requiring accurate diagnosis and therapeutic solutions. Addressing such challenges requires research efforts and technological development that can be used and deployed in low resource settings, particularly in LMICs. We aim at developing an AI-based software that is able to quantify intracellular parasites, such as Leishmania amastigotes, based on light microscopy images acquired using a smartphone. The software will then be deployed as a mobile application for use in laboratories with limited access to advanced microscopy equipment and expertise. We collected Giemsa-stained microscopy images of macrophages infected with Leishmania parasites. We annotated all pixels within these images into four classes: host cells, nucleus, amastigotes and promastigotes. The annotated dataset was exported in the COCO format, used for segmentation tasks. Then, we employed a two-step deep learning-based segmentation approach using the U-Net architecture with a VGG16 backbone. The first step involved a binary U-Net model training and the second step leveraged a multi-class U-Net model to segment and classify the different classes. We trained the models under different conditions and we used the Dice score and the IoU as key metrics to evaluate its performances. To optimize segmentation performance, we employed Dice loss as the loss function for both models, for its effectiveness in handling class imbalances and improving segmentation performance by maximizing overlap between predicted and ground truth masks. We employed Adam as the optimizer with a learning rate of 0.001 to train the model. Finally, we optimized the models and assessed their performances in comparison to existing tools and software with comparable functionality. We were able to acquire and annotate 34 images, containing 974 amastigotes. The pixel distribution of the dataset revealed a background dominance (70.30%), with host cells (23.92%) and nuclei (5.04%) contributing significantly, while amastigotes and promastigotes accounted for only 0.65% and 0.04% of the pixels, respectively. We generated instance segmentation masks tailored for multiclass segmentation and split the dataset into training (26), validation (5) and test (3) sets. We then performed data augmentation to the training set, through introducing spatial variations by applying geometric transformations (e.g., flips, rotations) to improve the model's ability to generalize across different orientations. We added intensity variations to simulate changes in lighting or color, and simulate realistic conditions by introducing noise or blurs to replicate scenarios the model may encounter in real-world applications. The trained model achieved satisfactory performances of 95.34% Dice score and 82.58% MeanIoU. The comparison of our model's performances as compared to existing tools is still ongoing. Our project aims at automating the detection and the accurate counting of amastigotes in Giemsa-stained microscopy images using an adapted U-Net architecture. The pipeline will be further optimized on low resolution images (smartphone), to be then deployed as a mobile application. It will serve as a scalable user-friendly tool to reduce manual effort, thus enabling fast and accurate analysis of microscopy images in low-resources settings.

Predicting Lead Compounds against Mycobacterium tuberculosis using Machine Learning and Molecular Docking

George Hanson (Department of Parasitology, Noguchi Memorial Institute for Medical Research, University of Ghana), Joseph Adams (Department of Parasitology, Noguchi Memorial Institute for Medical Research, University of Ghana), Richmond K. Kudiabor (Department of Clinical Pathology, Noguchi Memorial Institute for Medical Research, University of Ghana), Daveson I. B. Kepgang (Department of Biochemistry, Faculty of Sciences, University of Douala), Miriam E. L. Gakpey (Department of Clinical Pathology, Noguchi Memorial Institute for Medical Research, University of Ghana), Maame E. Annor-Apaflo (Department of Clinical Pathology, Noguchi Memorial Institute for Medical Research, University of Ghana), Edgar Sungwacha (Department of Medical Microbiology, Jomo Kenyatta University of Agriculture and Technology), Clifford S. Yisufu (Department of Clinical Pathology, Noguchi Memorial Institute for Medical Research, University of Ghana) and Olaitan I. Awe (African Society for Bioinformatics and Computational Biology)

Introduction: Tuberculosis (TB), caused by Mycobacterium tuberculosis (M.tb), remains a major global health challenge and a leading cause of mortality worldwide. The rise of drug-resistant TB strains has significantly hindered treatment, highlighting the urgent need for novel therapeutics. This alarming resistance underscores the urgent need to discover and develop novel drug candidates to combat TB effectively. Machine learning (ML) and computational chemistry provide promising avenues for accelerating TB drug discovery by identifying potential inhibitors efficiently. Methods: This ongoing study integrates ML-based predictive modeling with computational drug discovery techniques. Five classifiers such as Logistic Regression (LR), Support Vector Machine (SVM), k-Nearest Neighbors (k-NN), Random Forest (RF), and Gaussian Naïve Bayes (GaussianNB) were developed and evaluated using a dataset of 211,606 compounds curated from PubChem (M.tb H37Rv, bioassay AID: 1626). Compounds predicted as inhibitors undergo molecular docking, molecular dynamics (MD) simulations, and Molecular Mechanics Poisson-Boltzmann Surface Area (MM/PBSA) calculations to assess binding interactions and stability. Toxicity and physicochemical properties are also being analyzed. Preliminary Results: LR and SVM demonstrated the highest predictive performance, identifying 2,307 potential inhibitors. These compounds are currently undergoing molecular docking against the energy-minimized DprE1 enzyme, a key target in M.tb cell wall biosynthesis, followed by downstream analyses. Conclusion and Future Work: This study demonstrates the potential of integrating ML and computational drug discovery techniques for TB therapeutics. Ongoing analyses will refine hits compounds and assess their binding efficacy by performing protein-ligand interactions and toxicity profiles. Additionally, MD simulation, combined with MM/PBSA, will be employed to evaluate the stability of the complexes. These findings highlight the power of computational methods in expediting TB drug discovery.

CAMRAH: A Cloud-enabled workflow for analysis and harmonization of antibiotic resistance gene predictions

Daniella Matute, Indresh Singh and Derrick Fouts
(J Craig Venter Institute)

CAMRAH is a customizable workflow developed as part of the Combatting Antimicrobial Resistance in Africa Using Data Science (CAMRA) initiative. It annotates antimicrobial resistance (AMR) genes in microbial genomes using the consensus of multiple sequence-based AMR detection methods. It can be run on cloud platforms or on any local device. The methods include BLAST-based searches, Hidden Markov Model profiling, structural gene analysis and kmer-based predictive classification. Since no single AMR prediction software tool can predict all AMR genes, CAMRAH runs six different AMR-tools and database permutations, mapping hits to CDS coordinates and harmonizing them to produce gene symbols and ontologies consistent with The Comprehensive Antibiotic Resistance Database (CARD). This results in a consensus-aware resistance profile of a bacterial isolate. A comparison of CAMRAH with individual AMR finders demonstrated that it predicts more AMR genes with higher accuracy, while reducing inconsistencies in gene annotation across different tools through the implementation of our harmonization algorithm.

Prostruc: An Open-source Tool for 3D Structure Prediction using Homology Modeling

Shivani Pawar (Department of Biotechnology and Bioinformatics, Deogiri College), Wilson S.K. Banini (Department of Theoretical and Applied Biology, Kwame Nkrumah University of Science and Technology), Shamsuddeen Musa (Faculty of Health Sciences, Department of Public Health, National Open University of Nigeria), Toheeb Jumah (School of Collective Intelligence, University Mohammed VI Polytechnic), Nigel Dolling (Department of Parasitology, Noguchi Memorial Institute for Medical Research, University of Ghana), Abdulwasiu Tiamiyu (School of Collective Intelligence, University Mohammed VI Polytechnic) and Olaitan Awe (African Society for Bioinformatics and Computational Biology)

Introduction: Homology modeling is a widely used computational technique for predicting the three-dimensional (3D) structures of proteins based on known templates,evolutionary relationships to provide structural insights critical for understanding protein function, interactions, and potential therapeutic targets. However, existing tools often require significant expertise and computational resources, presenting a barrier for many researchers. Methods: Prostruc is a Python-based homology modeling tool designed to simplify protein structure prediction through an intuitive, automated pipeline. Integrating Biopython for sequence alignment, BLAST for template identification, and ProMod3 for structure generation, Prostruc streamlines complex workflows into a user-friendly interface. The tool enables researchers to input protein sequences, identify homologous templates from databases such as the Protein Data Bank (PDB), and generate high-quality 3D structures with minimal computational expertise. Prostruc implements a two-stage vSquarealidation process: first, it uses TM-align for structural comparison, assessing Root Mean Deviations (RMSD) and TM scores against reference models. Second, it evaluates model quality via QMEANDisCo to ensure high accuracy. Results: The top five models are selected based on these metrics and provided to the user. Prostruc stands out by offering scalability, flexibility, and ease of use. It is accessible via a cloud-based web interface or as a Python package for local use, ensuring adaptability across research environments. Benchmarking against existing tools like SWISS-MODEL,I-TASSER & Phyre2 demonstrates Prostruc's competitive performance in terms of structural accuracy and job runtime, while its open-source nature encourages community-driven innovation. Discussion: Prostruc is positioned as a significant advancement in homology modeling, making high-quality protein structure prediction more accessible to the scientific community.

Comparative Transcriptomic Analysis of UHRF1 Expression Knockout in Different Cancer Types

Jonathan Kalami (Kwame Nkrumah University of Science and Technology), Benson Kidenya (Catholic University of Health and Allied Sciences), Miriam E.L. Gakpey (Noguchi Memorial Institute for Medical Research, University of Ghana), Sala Kotochi (Kwame Nkrumah University of Science and Technology), Benthai Benjamin Department of Pharmaceutical Microbiology and Biotechnology, University of Jos), Nana Yaa Karikari Agyeman (Noguchi Memorial Institute for Medical Research, University of Ghana) and Olaitan I. Awe (African Society for Bioinformatics and Computational Biology)

Background: Ubiquitin-like with PHD and RING Finger domains 1 (UHRF1) is an important epigenetic regulator that plays a pivotal role in modulating DNA methylation patterns and chromatin structure. Its role in cancer progression is significant, yet the extent of its impact on gene expression across different types of cancer remains poorly understood. This study focuses on a comparative transcriptomic analysis of UHRF1 expression knockout in four cancers—two solid tumors and two hematological cancers—following UHRF1 expression knockout. Methods: To understand the role UHRF1 plays in different cancers, we obtained transcriptomic data from NCBI Gene Expression Omnibus (GEO) database and performed comparative differential gene expression and gene set enrichment analysis of four distinct cancer cell lines: retinoblastoma (Y79), breast cancer (MCF-7), acute myeloid leukaemia (Kasumi-1) and acute monocytic leukaemia (THP-1). Results: Among the differentially expressed genes of each cancer cell line, 80 genes were commonly regulated following UHRF1 expression knockout. Gene set enrichment analysis of the 80 overlap genes revealed 179 GO terms significantly enriched, suggesting roles in processes like response to hypoxia, regulation of Wnt signaling pathway, and positive regulation of apoptotic processes with no KEGG pathways found to be enriched. Protein-protein interaction analysis of the 80 overlapping genes revealed a network with 79 nodes and 42 edges, identifying key hub genes GLUL, SOD2, GPI, TXNDR1, and HSPD1, with GLUL and SOD2 exercising strong functional interactions. Conclusion: Our comparative transcriptomic analysis reveals eighty (80) shared differentially expressed genes across solid and hematological cancers, with GLUL and SOD2 being key hub genes with strong functional interactions. Furthermore, our findings suggest that UHRF1 expression knockout leads to significant enrichment in processes such as the response to hypoxia, regulation of the Wnt signaling pathway, and apoptosis, thereby improving our understanding of UHRF1's critical role in pathways essential to cancer progression. This indicates that UHRF1 knockout influences vital cellular functions, potentially identifying novel targets for cancer therapy.

Identifying Therapeutic Biomarkers for HIV Infection Using Systems Biology Transcriptomics and Molecular Docking Approaches

Asmaa Reda (Computational Biology and Bioinformatics Division, Zoology Department, Faculty of Science, Benha University), Noella Okumu (Jomo Kenyatta University of Agriculture and Technology, Department of Biochemistry), George Hanson (Department of Virology, Noguchi Memorial Institute for Medical Research, University of Ghana), Jacob James (ioinformatics and Biotechnology, University of Nairobi), Olaitan Awe (African Society for Bioinformatics and Computational Biology), Armando Djiyou (Molecular Diagnostic Research Group (MDRG), Biotechnology Center, University of Yaounde), Eugene Adjei (University of Ghana, Department of Biomedical Engineering), Kelvin Mungai (Pwani University, department of Biotechnology and Bioinformatics) and Cateline Ouma (Jomo Kenyatta University of Agriculture and Technology, Department of Biochemistry)

Human immunodeficiency virus (HIV) remains a global health challenge due to its high pathogenicity and the complex host-pathogen interactions that drive disease progression. Transcriptomic data from microarray and RNA-sequencing in HIV-infected samples were analyzed to uncover novel therapeutic targets. This integrative approach identified a set of key hub genes [MKI67, NCAPG, SKA1, MELK, TPX2, CEP55, NEK2, CKAP2L, CDK1, and ESPL1] that exhibit significant dysregulation in the presence of HIV. These genes are predominantly involved in cell cycle regulation and kinase signaling pathways. To further explore the therapeutic potential of these targets, molecular docking, and molecular dynamics simulations were performed against a library of FDA-approved drugs, focusing on compounds known to modulate kinase activity and cell cycle processes. This computational approach enabled the rapid screening of drug-gene interactions, facilitating the potential repurposing of existing drugs to counteract HIV-associated cellular dysregulation. Furthermore, expression quantitative trait locus (eQTL) analyses were conducted to investigate how genetic variations influence the expression of these hub genes. This step is essential for understanding individual variability in disease susceptibility and drug response, thereby supporting personalized medicine. By linking genetic polymorphisms to gene expression changes, the eQTL findings emphasize the functional significance of these targets and provide deeper insights into host regulatory networks during HIV infection. Overall, this study provides a systems biology-driven framework for identifying novel drug targets in HIV infection and paves the way for repurposing existing FDA-approved drugs for effective antiviral intervention.

Multiscale Modeling of NAT2 Genetic Variants: Structural Stability and Functional Impact

Houcemeddine Othman (Sydney Brenner Institute for Molecular Bioscience, University of the Witwatersrand)

The N-acetyltransferase 2 (NAT2) protein plays a critical role in the metabolism of various drugs, including the first-line anti-tuberculosis (TB) treatment isoniazid and its hepatotoxic metabolite, hydrazine. Genetic polymorphisms in the NAT2 gene influence acetylation activity, impacting drug efficacy and toxicity. Africa exhibits remarkable genetic diversity in NAT2, with 27 star alleles identified in our previous study. Despite extensive genetic characterization, the functional impacts of most NAT2 alleles remain unknown. Experimental assays are effective for studying these variants but are resource-intensive and time-consuming. Multiscale computational modeling offers an efficient alternative to screen genetic variants for functional impact, enabling prioritization for further study. Here, we analyzed 18 missense variants resulting in single amino acid substitutions and 25 NAT2 star alleles identified from 1,079 high-depth African whole genomes. Using an in silico mutagenesis pipeline combined with molecular mechanics refinement, we modeled the structures of NAT2 variants and calculated free energy changes upon mutation (ΔΔG) using DDGun and FoldX approaches, incorporating vibrational entropy estimates. Among single variants, L24I, I10T, R64Q, I114T, R197Q, and Y208H exhibited significant ΔΔG variations. Notably, Y208H (*42) directly interacts with the catalytic site cofactor, potentially altering enzymatic activity. These variants are distributed across 17 star alleles, with *45 showing the most pronounced ΔΔG variation of -2.7 kcal/mol (ΔGRef - ΔGvariant).Thermodynamic analysis of star alleles highlighted a sub-functional group (*33, *31, *40, *5 and *16) associated with structural destabilization. These haplotypes are prevalent in African populations and may influence NAT2 activity. Our findings reveal interesting properties of several star alleles that merit further clinical investigation to understand their impact on TB treatment outcomes and toxicity risk. This study underscores the potential of computational approaches to accelerate pharmacogenomic research in diverse populations.

Bioinformatic characterisation of antimicrobial activity in snake venom

Veronica Recheal Wokibula (South African National Bioinformatics Institute, UWC), Dominique Anderson (South African National Bioinformatics Institute) and Ruben Cloete (SANBI)

Antibiotic resistance is increasing globally and poses a substantial risk to human and animal health, with estimates that drug-resistant diseases will result in 10 million human deaths by 2050, there is an urgent need to explore other sources of antimicrobials with different modes of action. Venomous animals occur in numerous phyla, distributed across geographies. These animals have evolved venoms, a chemical weaponry, as an instrument of predation or defense. Venoms produced by these animals have immense potential for scientific, medical, and pharmaceutical applications such as development of antivenoms, novel drugs, and bio-inspired materials. Venoms from specific snake species, for instance, are a source of pharmacologically active proteins that have been used to develop novel drugs to manage disorders such as hypertension, embolism, and myocardial infarction. Furthermore, research investigating venom proteomes has demonstrated antiviral, antiparasitic, antifungal and antibacterial activity, and as such, diversity in snake venoms makes them a feasible biological resource to explore alternative antimicrobial therapies. The study aims to investigate the molecular interaction between snake venom phospholipases A2 (PLA2) and the Mycobacterium tuberculosis (M. tb) cell wall using molecular docking. The svPLA2 sequences and structures considered are from viperid species endemic to the Africa continent namely Echis pyramidum, Echis ocellatus, Echis coloratus, Cerastes cerastes, Bitis arientans, and Bitis gabonica. The svPLA2 mycobacterium tuberculosis outer call wall protein component targets include CpnT, FecB2, SubI, MsPA, ProX, and lipoprotein LpqX. The study harnesses Haddock2 and Schrödinger software for the PLA2-M. tb protein docking. In addition, phospholipids: POPE, POPG, Cardiolipin, NAG, and Mycolic acid, will be docked to the snake venom PLA2 using Autodock and Vina software. The project further utilizes PyMol and Visual Molecular Dynamics (VMD) tools for visualization and interaction analysis. The study is expected to identify interactions between the identified M. tb cell wall components and svPLA2 in order to evaluate the potential of these molecules in drug discovery and design, contributing to the available knowledge on snake venoms as a source of antibiotic compounds against Mycobacterium tuberculosis.

Empowering Bioinformatics in Africa through Bioconductor: Expanding Training and Community Engagement

Maria Doyle (V), Umar Ahmad (Africa CDC / Bauchi State University), Aedin Culhane (University of Limerick), Laurent Gatto (de Duve Institute, UCLouvain), Zedias Chikwambi (African Institute of Biomedical Sciences and Technology), Charlotte Soneson (Friedrich Miescher Institute for Biomedical Research) and Trushar Shah (International Institute of Tropical Agriculture)

For over 20 years, Bioconductor (https://bioconductor.org/) has provided an open-source ecosystem for reproducible genomic data analysis, supporting over 2,000 R-based packages with 1M+ annual unique downloads worldwide. However, access to bioinformatics training remains a challenge in Africa, limiting the adoption of these tools. To bridge this gap, Bioconductor is expanding its training initiatives and community collaborations across the continent. Through the Bioconductor Carpentry program, we have trained 31 instructors in the last two years, developed structured training materials on R/Bioconductor, RNA-seq and single-cell analysis, and delivered 28+ workshops globally. To address the growing demand for in-person training, we have partnered with local bioinformatics leaders to launch on-site workshops in East and West Africa. In March 2025, we held a highly successful week-long workshop in Nairobi, Kenya, in collaboration with the International Institute of Tropical Agriculture (IITA) and international experts. The course provided hands-on training in genomic data analysis and RNA-seq workflows, enabling participants to apply Bioconductor tools to their own datasets. More details: training.bioconductor.org/workshops/2025-03-Nairobi . We plan to run a similar course in West Africa in October 2025. We aim to connect with and support the African bioinformatics community by: Expanding instructor networks to increase local training capacity. Identifying needs and gaps in bioinformatics training and tools to ensure our initiatives are needs-based. Developing region-specific curricula for One Health research, including pathogen and agricultural genomics. Collaborating with African bioinformatics institutions to scale training opportunities. This presentation will highlight progress to date, lessons learned, and future opportunities, inviting African researchers and educators to engage with Bioconductor's growing global community.