Johnbosco **Tayebwa**

BIOINFORMATICIAN

School of Allied Health Sciences, Kampala International University (KIU), Uganda

☎ +256 784362367 | ✉ johnboscotayebwa@kiu.ac.ug | 0000-0002-2355-8475 | biocodebreaker | Johnbosco Tayebwa

I am a Bioinformatician with a background in Biomedical Laboratory Technology. I have more than 12 years of experience in the management and analysis of high-throughput data from Next-Generation Sequencing (NGS) and Gene Expression Microarrays. I have worked in multidisciplinary research teams in a collaborative environment; as a Bioinformatician at the Department of Clinical Genetics, Lund University Hospital, Sweden, and as a Visiting Postgraduate Researcher at the School of Biosciences, University of Exeter, United Kingdom. I am currently an Assistant Lecturer/Research Scientist at the School of Allied Health Sciences, Kampala International University, Uganda. Recently, I was the Head of Department of Medical Laboratory Sciences, Kampala International University. My research interests include the study of Immune Cell Infiltration in the Tumor Microenvironment (TME) in Gastric Cancer, Cervical Cancer, and other cancers using bioinformatics tools. I am also interested in integrating Multi-Omics datasets to develop next-generation immunotherapies, Computational Modeling to develop novel prognostic models using AI/ML and T-cell-based immunotherapies (combination immune checkpoint blockade-CAR-T cell therapy). I am eager to learn and acquire new skills in Mass Spectrometry, Metabolomics and other techniques that I am not experienced with where necessary. Throughout my academic career, I have won several Scholarly Awards such as the Swedish Institute Study Scholarship, the Muljibhai Madhvani Foundation scholarship and I was a beneficiary of the NARO/WORLD BANK PROJECT Millennium Science Initiative Research Grant. I am a GitHub-verified Educator and a participant in the Global-Campus-Teachers (Github Education) Community. I am also a member of the International Society for Computational Biology (ISCB).

Education

Master of Science in Bioinformatics
2011-2012

UNIVERSITY OF SKOVDE *Sweden*

Bachelor of Biomedical Laboratory Technology, Second Class Honours – Upper Division
2005-2008

MAKERERE UNIVERSITY *Uganda*

Professional Experience

Assistant Lecturer/Research Scientist

SCHOOL OF ALLIED HEALTH SCIENCES, KAMPALA INTERNATIONAL UNIVERSITY *Current*

- Uganda

Head of Department

DEPARTMENT OF MEDICAL LABORATORY SCIENCES, KAMPALA INTERNATIONAL UNIVERSITY *2021-2023*

- Uganda

Bioinformatician

DEPARTMENT OF CLINICAL GENETICS, LUND UNIVERSITY HOSPITAL *2012-2014*

- Sweden

Visiting Postgraduate Researcher

SCHOOL OF BIOSCIENCES, UNIVERSITY OF EXETER *2010-2011*

- United Kingdom

Research Focus Area

My broad interest is integrating various multi-dimensional omics datasets such as genomics, transcriptomics, proteomics, metabolomics etc to develop next-generation immunotherapy approaches for cancer patients. I am also interested in the application of Single-Cell RNA-Seq for personalized immunotherapy.

My Research Focus area is Cancer Immunotherapy especially immune cell infiltration in the Tumor Microenvironment (TME). I am also interested in using AI/ML for computational modeling to develop better prognostic models for Cancer patients. This includes biomarker discovery.

Research Experience (Recent and ongoing projects I am leading as a Principal Investigator)

Bioinformatics study of the Tumor Microenvironment (TME) in Gastric Cancer [Manuscript Submitted for Review].

I have conducted a study on immune cell infiltration in the Gastric Cancer (GC) Tumor Microenvironment (TME) using the TCGA Gastric Cancer dataset. The study aimed to identify Immune-Related Genes (IRGs) that have a significant impact on the survival of GC patients. These IRGs that can predict the prognosis of GC patients can act as biomarkers for predicting patient outcomes. They can also be targeted for cancer immunotherapy.

=> Gene expression data (mRNA RNA-Seq) of 448 Gastric Cancer (Stomach Adenocarcinoma) cases were downloaded from the Genomic Data Commons Portal TCGA-STAD project.

=> mRNA expression data (FPKM) was used for estimation of stromal and immune scores using the ESTIMATE R package.

=> Based on risk score calculation, established a prognostic risk model comprising of eight differentially expressed Immune-Related Genes (IRGs) which were the best at predicting Overall Survival at the 1-year, 3-year and 5-year time points (AUC ROC curves).

=> Using univariate and multivariate analyses, three (3) Immune-Related DEGs emerged as independent prognostic factors and it was shown that their risk scores were significantly correlated with the infiltration levels of some but not all immune cell types.

=> For validation of the prognostic Immune-Related DEGs, two microarray Gastric Cancer datasets (433 GC samples) were obtained from the Gene Expression Omnibus (GEO) database.

=> Immune Cell Deconvolution (determination of immune cell proportions) of the TCGA tumors was done using the CIBERSORTx Impute Cell Fraction module.

=> It was observed that among the 448 TCGA tumors, 10 out of the 22 immune cell types had significantly different immune cell proportions between tumors and normal samples in both the paired and unpaired comparisons. We suggested that these 10 immune cells whose proportions were consistently different in tumors compared to normal samples may constitute a unique Tumor Infiltrating Leukocyte Phenotype/signature which is enriched in tumors but not normal samples, which may not only have prognostic significance but also immune therapy implications, thus deserves further research [MANUSCRIPT SUBMITTED FOR REVIEW].

Study on the role of immune cell infiltration and oxidative stress in the Tumor Microenvironment (TME) of Cervical Cancer [Manuscript in Preparation]

Machine Learning Model to Predict the Prognosis of Cervical Cancer Patients based on Oxidative Stress-Related Genes (OS-Related DEGs) and Immune Cell Infiltration.

Study Justification: Recent studies have shown that immune cell infiltration and oxidative stress are key components of the tumor microenvironment (TME) that influence the progression and prognosis of cervical cancer. Effective prognostic models are crucial to predict patient outcomes, guide treatment strategies, and improve survival rates. Investigating genes associated with oxidative stress could provide valuable insights into the molecular mechanisms driving cervical cancer and identify potential therapeutic targets.

Study Methodology

=> Bulk tissue RNA-Seq expression data for cervical tissues were obtained from publicly available databases. 306 cervical cancer samples, consisting of 304 primary tumors and 2 metastatic samples, were obtained from The Cancer Genome Atlas (TCGA) through the GDC Portal. A total of 22 normal cervical tissue samples: 19 samples (endocervix, n = 10; and ectocervix, n = 9) were retrieved from the the Genotype-Tissue Expression (GTEx) Portal, and 3 normal cervix uteri samples were obtained from the TCGA-CESC project.

=> Oxidative stress-related genes were downloaded from the GeneCards website and the Molecular Signatures Database (MSigDB).

=> The limma R package was used to identify the OS-Related DEGs between the 306 Tumor and 22 Normal samples.

=> 65 OS-Related DEGs between Tumor and Normal samples in the TCGA-GTEx, GSE63514 and GSE39001 Cervical Cancer datasets were significant in the Univariate analysis.

=> The Multivariate Cox Proportional Hazards Model identified three (3) OS-Related DEGs as significant independent predictors of Overall Survival of Cervical Cancer patients.

=> A t-distributed Stochastic Neighbor Embedding (t-SNE) plot to visualize the distribution of the Tumor and Normal samples based on the expression of the 65 OS-Related DEGs was generated.

=> **MACHINE LEARNING**: TCGA, GTEx and GEO Datasets to be used in the Training cohort and Validation cohort were identified. Training Cohort: TCGA Primary tumors (306); TCGA-GTEx Normal (22); GEO (GSE39001, GSE63514); Validation cohort: GSE44001 (300 tumor samples).

=> The 'glmnet' and the 'randomForestSRC' machine learning R packages are being used in the development, training and validation of the prognostic risk model to select the most appropriate genes from the Prognostic OS-related DEGs.

SOME OF THE DATA SCIENCE PROJECTS I HAVE WORKED ON:

Reproducible Bioinformatics Research using R/RStudio and GitHub

While at Kampala International University (KIU), I observed overtime that most Early-Career Researchers in the Life Sciences in Uganda/Africa use SPSS and/or STATA for data analysis in their research. This tradition persists despite the availability of more powerful, flexible, and scalable data science platforms like Posit Cloud (formerly RStudio). When I became Head of Department (HoD) in the Department of Medical Laboratory Sciences at KIU, I introduced a series of workshops and seminars to train postgraduate students in the use of R for data analysis.

Reproducibility is now one of the most desirable tenets of academic/scientific publishing yet authors typically do not provide the code alongside their manuscripts.
As Head of Department, I trained my postgraduate students to reproduce the work done in a recently published Cervical Cancer article so that they can appreciate principles of open-science and open-source tools used in bioinformatics research.

We selected a purely bioinformatics recently published Cervical Cancer article in PubMed (Zhao et al., 2022). The study used gene expression and clinical data from three independent clinical cohorts: TCGA-CESC, GSE44001 and GSE52903 We obtained the study information, datasets used and R packages used and followed the methods to reproduce the article. Our work can be found here: (`https://github.com/biocodebreaker/Reproduce-HMAG-Signature-`

Cervical cancer multi-omics study

Another project we are working on is a multi-omics study that integrates various omics datasets (genomics, transcriptomics, proteomics, metabolomics, etc) that uses best-practice recommendations like version control (github) [`https://github.com/biocodebreaker/cervical-cancer-multiomics-study`].

Some of my Research Skills

I have been part of multidisciplinary research groups, where the contributions from different members (clinicians, wet-lab biologists, technicians, students and bioinformaticians) bring to life original research publications in a collaborative environment.

Analysis of high-throughput data from public databases such as TCGA and GEO.

=> Experience with gene expression array and microarray (SNP-array) data analysis

=> Analyzed NGS data (Whole-Genome-Sequencing, Whole-Exome-Sequencing, TruSeq targeted panel, RNA-Seq) for bone and soft tissue tumor studies. The sequencing was done by BGI (formerly Beijing Genomics Institute), SciLifeLab (Uppsala) or in-house. => Developed pipelines for variant calling, fusion gene detection, and differential expression analysis.

=> Using the Genomic Data Commons (GDC) Data Transfer Tool to download TCGA data.

=> Data preprocessing using R packages such as TCGAbiolinks for TCGA data and GEOquery for GEO data.

=> Data Normalization and Removal of Batch Effects using R packages like sva, ComBat and merging the cohorts after batch effect removal.

=> Differential Expression analysis using R packages like limma, edgeR, DESeq2.

=> Using the clusterProfiler R package for Gene Ontology (GO) and Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway enrichment analysis.

=> Using the timeROC R package to compute the area under the curve (AUC) for receiver operating characteristic (ROC) curves.

=>Kaplan-Meier survival analysis using the survival R package.

=> ggplot2 R package for data visualization.

=> Univariate and multivariate Cox regression analysis using the survival R package.

=> Machine Learning using the the glmnet and the ‘randomForestSRC’ R packages.

High Performance Computing (HPC) and Cloud Computing.

=> Experience with High Performance Computing clusters, such as the Uppsala Multidisciplinary Center for Advanced Computational Science (UPPMAX) which is part of the Swedish National Infrastructure for Computing (SNIC).

=> Experience with cloud computing platforms such as Amazon AWS, Google Cloud Platform (GCP), and Microsoft Azure for data storage and analysis.

Conducting an Automated Literature Review.

=> Hypothesis generation and hypothesis testing using the Scientific Method.

=> Conducting an automated literature review to identify the Research Gap using PubMed, Medline, Scopus, Web of Science,Google Scholar, Dimensions, Semantic Scholar and other databases to find relevant articles.

=> Using the PubMed API to search for articles using keywords and boolean operators and download the results in a structured format (e.g. JSON, XML).

=> Using Python/R tools such as PubMed Parser to extract the metadata of the selected articles e.g. pmid, title, abstract, authors, journal, publication date, doi, etc.

=> Finding a suitable Research Topic/Research Question and writing a concise Research Concept.

=> Developing a Research Proposal based on the corpus of articles found in the automated literature review.

=> Generating an automated flow diagram (flow chart) showing the steps of the analysis.

=> Developing end-to-end data analysis workflows using R and GitHub.

=> Using Linux Terminal, bash/shell scripting for automation of data analysis tasks, e.g. GNU parallel for parallel processing of multiple files.

=> Using the R console/IDE, R Bioconductor or CRAN (Comprehensive R Archive Network) packages for data analysis.

=> Using Python packages/Libraries for data analysis tasks.

=> Developing a a FAIR-compliant, machine-readable, modern, data-driven GitHub-hosted project powered by RStudio, which can be easily updated when the underlying data in the RStudio project changes.

=> Using R Markdown and Quarto for publishing reproducible research reports and data-driven manuscripts similar to PapersWithCode.

=> Using Docker for containerization of the R environment: Packaging everything (scripts, dependencies, environment) into a Docker image for easy sharing and reproducibility across different platforms.

FAIR GUIDING PRINCIPLES

Our work inherently implements the FAIR Guiding Principles of Findability, Accessibility, Interoperability and Reusability which are in alignment with the open-science and open-source principles towards universal reproducibility of research.

Selected Publications

I have a dozen publications including one where I'm **First Author (shared)**, one in **Nature Genetics**, and a **Review Article** on the Methods used in Cancer Research.

[1] Mohajeri, A., Tayebwa, J., Collin, A., Nilsson, J., Magnusson, L., von Steyern, F. V., Brosjö, O., Domanski, H. A., Larsson, O., Sciot, R., Debiec-Rychter, M., Hornick, J. L., Mandahl, N., Nord, K. H., & Mertens, F. (2013). Comprehensive genetic analysis identifies a pathognomonic NAB2/STAT6 fusion gene, nonrandom secondary genomic imbalances, and a characteristic gene expression profile in solitary fibrous tumor. Genes, chromosomes & cancer, 52(10), 873–886. `https://doi.org/10.1002/gcc.22083`

[2] Nord, K. H., Lilljebjörn, H., Vezzi, F., Nilsson, J., Magnusson, L., Tayebwa, J., de Jong, D., Bovée, J. V., Hogendoorn, P. C., & Szuhai, K. (2014). GRM1 is upregulated through gene fusion and promoter swapping in chondromyxoid fibroma. Nature genetics, 46(5), 474–477. `https://doi.org/10.1038/ng.2927`

[3] Mertens, F., & Tayebwa, J. (2014). Evolving techniques for gene fusion detection in soft tissue tumours. Histopathology, 64(1), 151–162. `https://doi.org/10.1111/his.12272`

For a complete list of my publications, please visit my: Google Scholar Profile.

Professional Membership

GitHub-verified Educator

I joined GitHub Education and I am implementing GitHub Classroom for postgraduate students in my Department for the Research Methodology Course Unit. We are following best-practice recommendations like integrating Posit-Cloud with GitHub. We have been using the Cloud Instructor (for Educators and Students) Pricing Plan on Posit Cloud.

I'm also a participant in the Global-Campus-Teachers (Github Education) Community where I actively participate and contribute to Discussions on this Teachers' Email Forum.

I am also a Member of the International Society for Computational Biology (ISCB) and the African Society for Bioinformatics and Computational Biology (ASBCB). I joined ISCB's Communities of Special Interest (COSI), that is, Computational Systems Immunology (CSI) and Computational Mass Spectrometry (CompMS) to keep up with the latest advances in sequencing, proteomics, metabolomics and imaging.

Honors & Awards

Top 500 Scientists in Uganda - AD Scientific Index

Position 136 in Uganda; Position 16 at KIU; Category: Digital Health Technologies | Health Informatics *2022*

Swedish Institute Scholarship

Swedish Institute *2011-2012*

Millennium Science Initiative Research Grant

NARO/WORLD BANK PROJECT *2010-2011*

Muljibhai Madhvani Foundation scholarship

Muljibhai Madhvani Foundation *2006-2008*

References

1. **Anthony Mukwaya, PhD**
   Postdoctoral Research Fellow,
   Department of Ophthalmology, Harvard Medical School,
   Schepens Eye Research Institute, Boston, MA, USA.
   Email: amukwaya@meei.harvard.edu
   Tel: +1 617 4609884

2. **David Studholme, PhD**
   Associate Professor in Bioinformatics,
   School of Biosciences, University of Exeter,
   Exeter, UK.
   Email: D.J.Studholme@exeter.ac.uk
   Tel: +44 (0) 1392 724678

3. **Fredrik Mertens, MD, PhD**
   Professor, Senior Consultant,
   Department of Clinical Genetics, University Hospital
   Lund, Sweden.
   Email: Fredrik.Mertens@med.lu.se
   Tel: 0046 46 173387

4. **Kyobuhaire Christine, BPA**
   Administrator, Faculty of Biomedical Sciences,
   Kampala International University - Western Campus,
   Ishaka-Bushenyi, Uganda.
   Email: KyobuhaireChristine@kiu.ac.ug
   Tel: +256 703786082