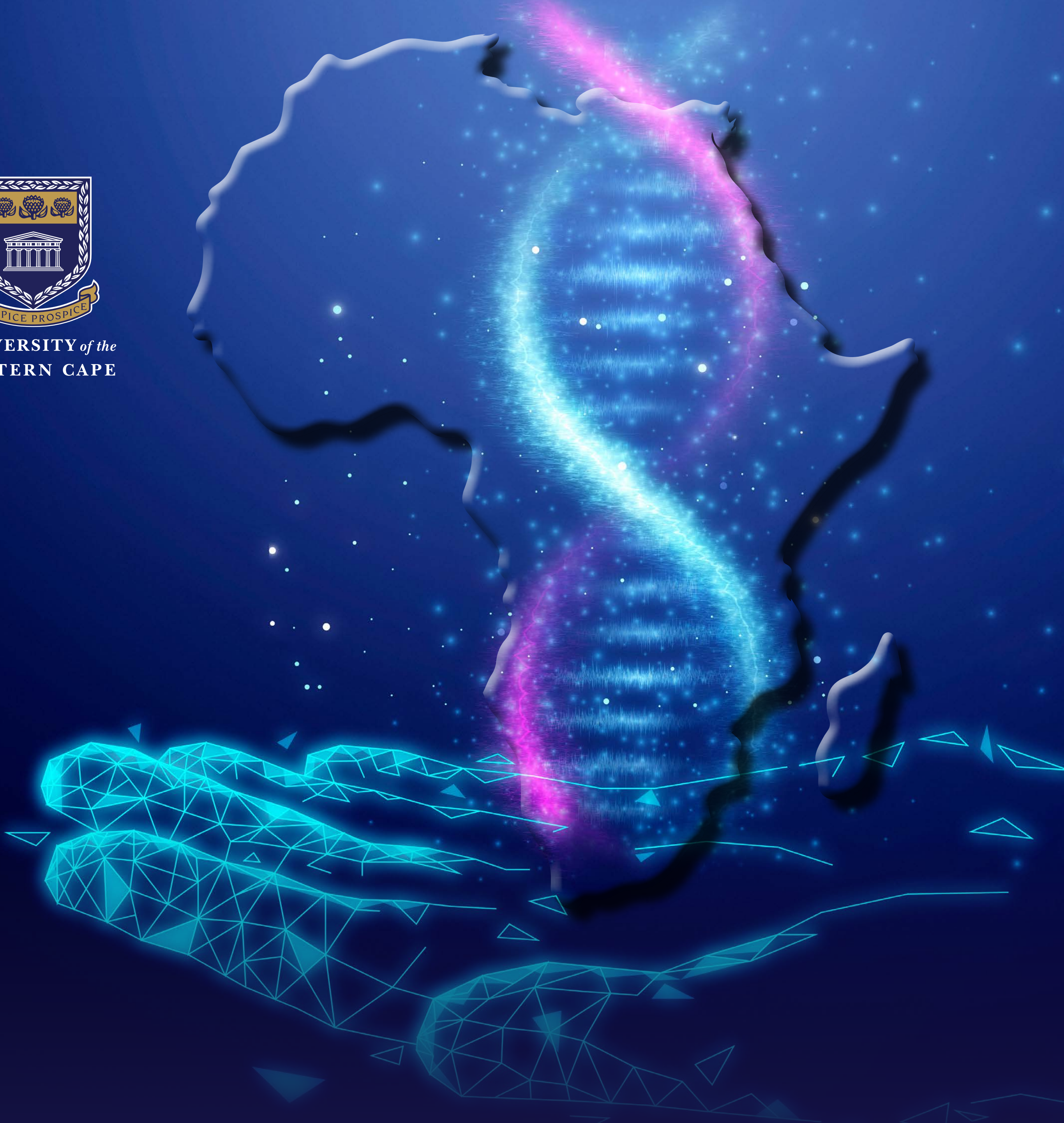
UNIVERSITY *of the*
WESTERN CAPE



SANBI
South African National
Bioinformatics Institute

ANNUAL REPORT 2023

CONTENTS

About SANBI

Bioinformatics is a specialist discipline straddling the fields of biology, mathematics and computer sciences and it is integral to modern biological research.

WHO WE ARE

The South African National Bioinformatics Institute (SANBI) is situated at the University of the Western Cape (UWC) in Cape Town. Our primary focus is the development and implementation of computational methodologies that allow biomedical researchers to accelerate their genomics data analyses. SANBI aims to heighten awareness of bioinformatics in South Africa and to assist the country in making optimal use of bioinformatics tools. As the leading bioinformatics entity in Africa, we continue to foster local and regional collaborations on health-related topics that cover both communicable and non- communicable diseases.

SANBI provides a focus for biological research located in Africa and as such, is dedicated to:

- the development of online specialised resources for genomics and genome informatics;
- capacity development in genomics and bioinformatics in Africa; and the development and implementation of genome annotation methods.

OUR VISION

- To be a global leader in computational biology, achieving the highest level in biomedical research and education in the global, African and South African context.

OUR MISSION

- To conduct cutting edge bioinformatics and computational biology research relevant to South African, African and global populations.
- To develop human resources in bioinformatics and computational biology by educating and mentoring scientists.
- To increase awareness of, and access to, bioinformatics and computational biology resources.

OUR GOALS

- To generate and publish high quality, relevant biomedical research.
- To train and graduate competent and productive researchers.
- To add value to the academic programme of UWC.
- To enhance other research fields through collaborative projects.
- To establish sources of renewable funding to pursue the mission of the institute.

POLICY MANDATES

National Strategic Plan (NSP) for HIV, TB and STIs (2023 - 2028)

The vision and mission of SANBI align with the NSP 2023 - 2028 that outlines how the country will respond to the prevention and treatment of HIV, TB and STIs. The NSP aims to *"...build resilient systems for HIV, TB and STIs that are integrated into systems for health, social protection and pandemic response..."*.

National Research Foundation (NRF) Strategy 2025

The vision and mission of SANBI aligns with the NRF's Strategy 2025, specifically "promoting globally and competitive research and innovation".

The research and development at SANBI aligns with the Department of Science and Innovation White Paper on Science, Technology and Innovation 2018 - 2028 by contributing to *"the development of human capabilities, knowledge expansion and innovation performance..."*.

The SA Medical Research Council (SAMRC) Act (Act 58 of 1991)

As an extramural unit of the SAMRC, SANBI falls under the legislative mandates of the SAMRC. At Section 3, this Act states that the Legislative Mandate of the SAMRC is: *"...through research, development and technology transfer, to promote the improvement of the health and quality of life of the population of the Republic, and to perform such functions as may be assigned to the SAMRC by or under this Act..."*.

Director's MESSAGE

The past year has been demonstration of a range of activities that have reached maturity including training programmes and conceptual frameworks that have been published. In this annual report we highlight the work done by all academic staff and international partners. In the context of sharing data to respond to disease threats, we are excited by the data sharing framework published by PHA4GE and a continental prototype for a sharing and analysing pathogen data.

The annual report highlights research findings in drug discovery, tool development for oncology diagnostics, pathogen evolutionary dynamics, and teaching strategies for data science capacity development.

We have delivered on our mandate to equip South Africa and the wider continental bioinformatics community as evidenced by 7 PhD and 5 MSc graduates for the past year.

I wish to congratulate staff and students for the quality and innovation that have characterised our scientific outputs.

Professor Alan Christoffels
PhD, M.ASSAf
Director & DSI/NRF Research Chair in Bioinformatics and Health Genomics
Director of SAMRC Bioinformatics Unit

South African National Bioinformatics Institute
University of the Western Cape

YEAR IN REVIEW

RESEARCH HIGHLIGHTS:

SANBI's academic excellence is once again reflected in the 2023 research outputs achieved:
- 19 journal publications
- 7 PhD graduates
- 5 MSc graduates

The Africa PGI (Pathogen Genomics Initiative) Data Management and Exchange Platform

During the past two years we have led the conceptualisation of a data archive for disease outbreak datasets in partnership with the Africa CDC. In 2023 we launched the prototype of the continental data platform called Agari.

Data sharing for secondary use of sensitive data for research

We describe secondary data sharing strategies that may accommodate some of the challenges associated with sharing sensitive health data for research.

A Benefit Sharing Framework was developed by the PHA4GE Ethics and Data Sharing working group (Chair N Tiffin) to promote more equitable practices for capacity development and benefit sharing in health research, particularly in LMICs. This framework provides a tool to operationalise benefit sharing and aims to make it easier for researchers without experience in benefit sharing to be able to design a benefit sharing plan for their research programmes.

Standardisation of clinical Phenotype data collections

Phenotype Harmonisation Working Group (PHWG) of the Human Heredity and Health in Africa (H3Africa) consortium adapted exisiting phenotype data standads for use in LMICS.

Data Science Training

SANBI staff have been developing and implementing data science training. This report shows the culmination of a global effort to shape a teaching framework through the use of the Galaxy platform.

SANBI has hosted regional training workshops as part of their efforts to respond to disease outbreaks in the region. We report on the most comprehensive bioinformatics training initiative led by the Africa CDC in Africa to support pathogen disease surveillance.

Development of genomic surveillance laboratory methods

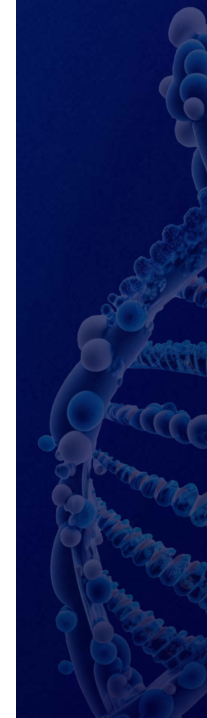
In partnership with the International Livestock Research Institute in Kenya, we have developed a multiplex PCR test to rapidly identify Rift Valley Fever Virus.

Application of machine learning techniques to oncology diagnostics

Gene expression data was used to identity genetic signatures that can improve the classification of intermediate-risk prognosis in AML patients, or reveal new cancer subtypes in Kidney renal clear cell carcinoma.

Drug targets for *Mycobacterium tuberculosis*

In collaboration with the UWC School of Pharmacy we describe a subtractive genomics approach to prioritise drug targets for *M.tuberculosis*.



New grants/funding awarded

Global Health European and Developing Countries Clinical Trials Partnership 3 (Global Health EDCTP3) Joint Undertaking works to deliver new solutions to reduce the burden of infectious diseases in sub-Saharan Africa and to strengthen research capacities to prepare and respond to re-emerging infectious diseases. The Public Health Alliance for Genomic Epidemiology is leading a global initiative to drive a world wide genomic epidemiology network funded by the EDCTP3 programme.

Evolutionary dynamics of SARS-CoV-2

In collaboration with colleagues at KU Leuven, Belgium, we developed a phylogeographic approach to estimate and compare the introduction and dispersal dynamics of the main SARS-CoV-2 variants that circulated in the New York City area between 2020 and 2022.

Public Health Alliance For Genomic Epidemiology (PHA4GE) Conference

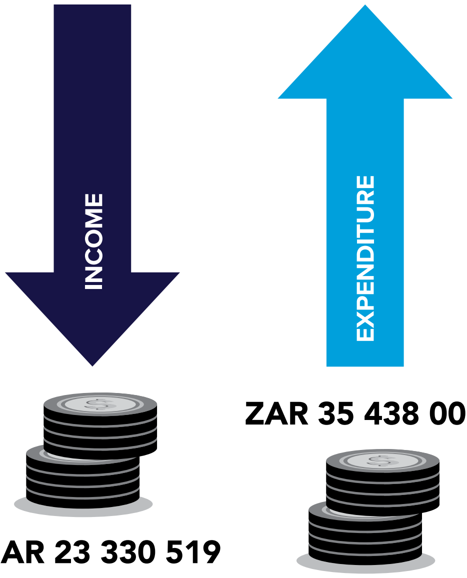
The PHA4GE secretariat, hosted at SANBI, organised and hosted 150 attendees at the first PHA4GE Conference held over 3 days from 30 October – 1 November.

Keep up to date on PHA4GE activities by subscribing to the newsletter:

(https://mailchi.mp/pha4ge/pha4ge-feb24-newsletter)

SANBI
South African National
Bioinformatics Institute

2023
IN SUMMARY



INCOME: ZAR 23 330 519

EXPENDITURE: ZAR 35 438 006

DIVERSITY PROFILE

STAFF
21

STUDENTS
27

GENDER

STAFF: 10 MALE
11 FEMALE

STUDENTS: 12 MALE
15 FEMALE

NATIONALITY

STAFF: 20 SOUTH AFRICAN

STUDENTS: 16 SOUTH AFRICAN

EDUCATION LEVEL

STAFF:

POSTDOCTORAL	8
MSC	4
BSC	2
DIP	8

STUDENTS:

POSTDOCTORAL	3
PHD	14
MSC	10

RESEARCH OUTPUTS

JOURNALS	19
GRADUATIONS 7 PHD | 5 MSC	12
AWARDS/ HONOURS	4
EXPERT PANEL/ COMMITTEE MEMBERSHIP	24
INTERNATIONAL CONFERENCE	1
INVITED TALKS	9
WORKSHOPS & COURSES	8
CONFERENCE PRESENTATIONS	16
VISITING SEMINARS	4
RESEARCH PROJECTS	41
COLLABS IN 42 INSTITUTIONS	42
ALUMNI IN 14 COUNTRIES	14

STAFF

The SANBI staff complement comprises a highly skilled, diverse group of research scientists, technical and administrative staff who all contribute to a dynamic productive working environment. As SANBI is a research institute, the Director reports through the faculty of Natural Sciences at UWC.

The research staff at SANBI focus on the development of analytical tools to address health challenges facing South Africa and the African continent. Through a series of national and international collaborations we have developed methods and used these methods to gain insight into disease mechanisms or to reduce the barrier to entry for researchers in resource-limited environments.

Dominique Anderson, Dr
Senior Lecturer

*UWC

Anja Bedeker
Research Associate

*B&MGF

Hocine Bendou, Dr
Senior Researcher

*SAMRC

* Funder

Alan Christoffels, Prof
Director,
SARChI Bioinformatics,
SAMRC Unit Director

*DSI/NRF Research
Chairs Programme

Ruben Cloete, Dr
Senior Lecturer

*UWC

Judit Kumuthini, Dr
Senior Researcher

*H3A BioNetwork

Gordon Harkins, Prof
Associate Professor

*UWC

Nicki Tiffin, Prof
Deputy Director

*UWC

Highlights

On 16 May 2023, Nicki Tiffin was appointed Deputy Director of SANBI by the UWC Joint Appointments and Promotions Committee of Senate and Council (SA2023/04) for the period 01 June 2023 to 31 May 2026.

Technical staff develop and maintain the computing infrastructure at SANBI.

Quinton Coert
Software Developer

*Baobab LIMS

Campbell Rae
Web Developer
(part time)

*DSI/NRF Research
Chairs Programme

Peter van Heusden
Senior Systems
Developer

*DSI/NRF Research
Chairs Programme

Shadley Wentzel
Senior Systems
Administrator

*SAMRC

* Funder

SANBI **administrators** ensure the smooth running of daily operations.

Saajidah Beghardien
Administrator

*SAMRC

Fungiwe Mpithi
PA/Administrator

*SAMRC

Ferial Mullins
Finance Administrator

*UWC

Junita Williams
Marketing Administrator
(part time)

*DSI/NRF Research
Chairs Programme

The **PHA4GE Secretariat** hosted at SANBI provide administrative and project management support to the PHA4GE working groups.

Based at SANBI, Alan Christoffels leads the Secretariat of the Public Health Alliance for Genomic Epidemiology (PHA4GE), an organisation comprising more than 100 members based at academic institutions around the globe. This international network has worked hard to develop analytical tools and ethics frameworks to support public health interventions.

Michael Bridger
Business
Development

*B&MGF

Rangarirai Matima
Communication
Specialist

*B&MGF

**Nawaal
Nacerodien-Weitz**
Administrator

*B&MGF

Alecia Naidu, Dr
Technical
Manager

*B&MGF

Jamie Southgate
Communications
Officer

*B&MGF

* Funder

SANBI academic staff are actively involved in translating research into policy. In 2023, more than half of the academic staff were appointed on national and international committees and expert panels that develop research frameworks to accelerate data science and biomedical research.

Expert Panel, Committee Membership and Conference Review Panels

NAME	INSTITUTE	ROLE AND PURPOSE
Dominique Anderson	**Academy of Science of South Africa**	Standing Committee on POPIA - to guide the development of the POPIA Code of Conduct for Research
	National Scientific R Collections Platform	Medical Biobanks Working Group Co-Chair
	Mastercard Fellowship	UWC Selection Committee
	UWC Technology Transfer Office	Seed Fund Committee
Anja Bedeker	**PHA4GE Consortium**	Member of the Steering Committee and Co-Chair of the Ethics and Data Sharing Working Group
Alan Christoffels	**African Society for Bioinformatics and Computational Biology**	Governing Council Member
	Academy of Science of South Africa	Member - Standing Committee on Biosafety and Biosecurity and Standing Committee on POPIA
	NHLS	Member of Biospecimen and Data Access Ethics Committee
	Sydney Brenner Institute for Molecular Bioscience, Wits	Research Advisory Committee
	Eastern Africa Network of Bioinformatics Training	Independent Scientific Advisory Board
	Global Emerging Pathogens Consortium	Member - host an annual conference on emerging pathogens as part of our mandate to create awareness of biosecurity threats across Africa, and to empower African scientists to respond to disease outbreaks.
	Africa CDC	Senior Advisor on Pathogen Genomics
	Applied Bioinformatics & Public Health Microbiology Conference	Scientific Programme Committee
Ruben Cloete	**South African Society for Bioinformatics**	Secretary
	African Society for Bioinformatics and Computational biology	Programme Committee
Gordon Harkins	**Centre for the AIDS Programme of Research in South Africa (CAPRISA)**	Appointed Director of the Centre of Excellence in HIV Prevention at UWC
Nicki Tiffin	**University of Cape Town, Computational Biology Division**	Honorary Professor

	International Common Disease Alliance	Member - Executive Committee Co-Lead - Global Equity Working Group
	International 100k+ Cohorts Consortium	Member - Executive Committee and Steering Committee Co-Chair - Policy and Data Sharing Working Group
	African Population Cohorts Consortium	Independent Advisory Group
	PHA4GE Consortium	Member - Steering Committee; Chair - Ethics and Data Sharing Working Group
	Global Health EDCTP3 Joint Undertaking	Member of the Scientific Committee
Peter van Heusden	**PHA4GE Consortium**	Infrastructure Working Group
	National Institute of Allergy and Infectious Diseases (NIAID)	Global Council of NIAID's ACE (African Centers of Excellent in Bioinformatics and Data-Intensive Science) **(https://ace.niaid.nih.gov/about/global_council)**

Journal Editing and Reviews

NAME	JOURNAL
Alan Christoffels	Editorial Board Member - Data Journal Reviewer - BMC Genomics; BMC Bioinformatics; Bioinformatics; PLoS ONE; Trends in Genetics; Genome Research.
Ruben Cloete	Reviewer - Heliyon
Nicki Tiffin	Journal Associate Editor - Nucleic Acids Research Genomics and Bioinformatics Reviewer - Health Policy and Technology Journal, International Journal of Population Genomics, BMJ Global Health
Peter van Heusden	Reviewer - Microbial Genomics

Thesis Examination at other Institutions

EXAMINER	INSTITUTION	DEGREE
Dominique Anderson	**University of Pretoria**	MSc
Alan Christoffels	**University of Antwerp**	PhD
	Rhodes University	MSc
Ruben Cloete	**Stellenbosch University**	PhD
Nicki Tiffin	**University of New South Wales, Sydney**	PhD
	Stellenbosch University	PhD

Throughout the year, SANBI staff are encouraged to update their skills by attending relevant training interventions.

Staff Development

NAME	INSTITUTION	DATE	COURSE NAME AND PURPOSE OF COURSE
Junita Williams	SAMRC Knowledge and Information Management Services Division	10 March	RIMS Training and Standard Operating Procedure for Quarterly Reporting
Nicki Tiffin	Calestous Juma Fellowship leadership training	6 - 12 March	Writing workshop for: Ethical, Legal and Social Issues Related to Gene Therapy: A South African Perspective: Consensus Study
	Calestous Juma Fellowship leadership training	6 - 12 March	Leadership training week (Nairobi) as part of the Calestous Juma Fellowship programme
	Calestous Juma Fellowship network meeting	1 - 7 May	B&MGF Seattle, Networking meeting as component of CJ Fellowship Programme
Ruben Cloete	UWC	7 September	Dean's Mentoring workshop
	UWC	20 November	NRF Rating workshop

Nicki Tiffin 2nd from Right with Calestous Juma Fellows attended workshops at Grand Challenges 2023.

SANBI STAFF IN THE MEDIA

African scientists are working to pool data that decodes diseases – a giant step

Published: May 11, 2023 7.28am SAST

Alan Christoffels, University of the Western Cape, Sofonias Kifle Tessema

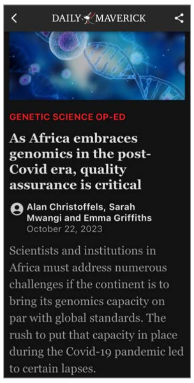
African scientists need a central repository where the genomic data they capture can be uploaded and shared. iStock/Getty Images Plus

A first for Africa: pathogen genomics platform to be available soon

The data-sharing platform on Africa's diseases will assist scientists and the medical fraternity online across the continent. Picture: Amir Cohen

By Ntsika Majiba | Published May 27, 2023

A first for Africa pathogen genomics platform, led by the SAMRC/SANBI/UWC Bioinformatics Capacity Development Research Unit, is set to be a crucial step in

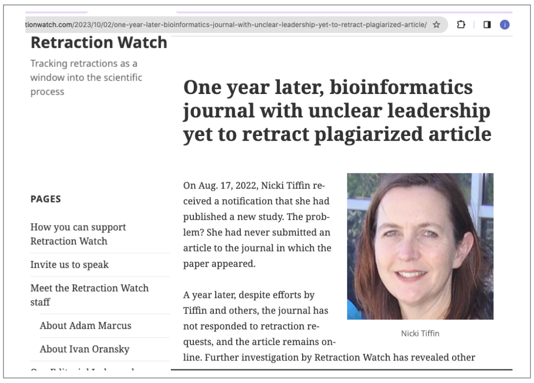


GENETIC SCIENCE OP-ED

As Africa embraces genomics in the post-Covid era, quality assurance is critical

Alan Christoffels, Sarah Mwangi and Emma Griffiths
October 22, 2023

Scientists and institutions in Africa must address numerous challenges if the continent is to bring its genomics capacity on par with global standards. The rush to put that capacity in place during the Covid-19 pandemic led to certain lapses.



Retraction Watch

Tracking retractions as a window into the scientific process

PAGES

How you can support Retraction Watch

Invite us to speak

Meet the Retraction Watch staff

About Adam Marcus

About Ivan Oransky

One year later, bioinformatics journal with unclear leadership yet to retract plagiarized article

On Aug. 17, 2022, Nicki Tiffin received a notification that she had published a new study. The problem? She had never submitted an article to the journal in which the paper appeared.

A year later, despite efforts by Tiffin and others, the journal has not responded to retraction requests, and the article remains online. Further investigation by Retraction Watch has revealed other

Nicki Tiffin

The ground breaking research at SANBI regularly features in print and online media. Above are some articles by SANBI researchers which appeared during 2023.

Capacity DEVELOPMENT

SANBI offers training programmes which are in keeping with its vision of becoming a centre of excellence in biomedical research and education in the global, African and South African context.

UNDERGRADUATE TRAINING PROGRAMME

Students who are interested in Bioinformatics as a career path are encouraged to take a combination of relevant subjects in Life or Health Sciences, Statistics, Computer Science and Mathematics during their undergraduate degree.

Bioinformatics Module (BTN 315)

Each year the UWC undergraduate Bioinformatics Module is taught to third-year Biotechnology students by the SANBI faculty. Topics covered include Introduction to Bioinformatics and databases, Protein structure prediction, Sequence alignments and Phylogenetics.

POSTGRADUATE TRAINING PROGRAMME

Postgraduate training at SANBI is well-established and alumni are now working all over South Africa and at bioinformatics research sites around the world.

Bioinformatics Training Course (BITC)

For this annual postgraduate Bioinformatics training course there were 9 participants who attended both face to face meetings and virtual pre-recorded lecturers. The topics covered included UNIX Command Line Language, Python Programming, How to use a Cluster, Using the Galaxy Pipeline and Introduction to Statistics using R. Study Design was also incorporated into the course. Students were given assignments to assess whether they were able to understand the concepts and apply the knowledge to their research project.

Honours Programme

Although SANBI does not have an Honours programme, students who attain a pass rate of >60% can include a bioinformatics component to their Honours project, branched off the BSc (Hons) Biotechnology programme with regards to the thesis project.

Masters' Programme

SANBI offers a research MSc in Bioinformatics by research thesis. Candidates with a BSc Honours degree or equivalent in a related scientific subject area may apply. The MSc degree is usually completed within two years.

Doctoral Programme

Candidates should be in possession of a MSc degree in Bioinformatics or in a related scientific field subject areas such as Computer Science, Mathematics, Biochemistry and Engineering. The PhD degree must be completed within five years.

Postdoctoral Programme

Postdoctoral fellows are admitted to the research programme after consultation with a potential SANBI supervisor. Fellowships may last for a maximum of three years.

SANBI STUDENTS

The SANBI student complement comprises a cohort of diverse and multi-talented researchers. In total there are 3 Postdoctoral Fellows, 14 PhD and 10 MSc candidates who are supervised by the SANBI academic staff.

POSTDOCTORAL FELLOWS

Eddie Lulamba

*Tiffin

Catherine Rossouw

*Christoffels

Tsaone Tamuhla

*Tiffin

DOCTORAL STUDENTS

Peter Abiodun

*Anderson

Adetutu Akinnuwesi

*Cloete

Abiola Babajide

*Anderson

Jeremy Burgess

*Cloete

Tracy Calvert-Joshua

*Christoffels

Sarah DeRaedt

*Christoffels

Nasr Eshibona

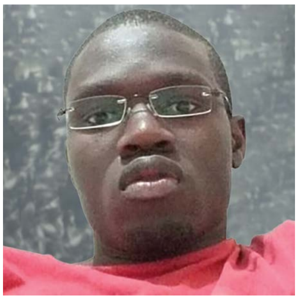
*Bendou

Mohammed Hassan

*Cloete

*Primary Supervisor

DOCTORAL STUDENTS



John Juma

*Christoffels

Michelle Livesey

*Bendou

Raziki Darius Martin

*Cloete

Josephine Peka

*Harkins

Nikkita Pillay

*Christoffels

Sohail Simon

*Tiffin

MASTERS' STUDENTS

Keagan Brown

*Cloete

Farzana Diedricks

*Christoffels

Kriheska Francis

*Harkins

Ridaa Fredericks

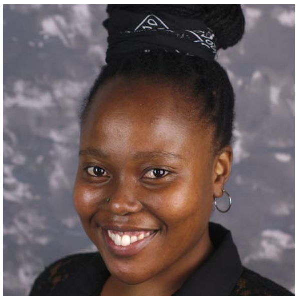
*Christoffels

*Primary Supervisor

MASTERS' STUDENTS

Wardah Jassiem

*Bendou



Siphosetu Mazitshana

*Harkins

Tatenda Mujuru

*Anderson

Lunathi Ndlondlo

*Bendou

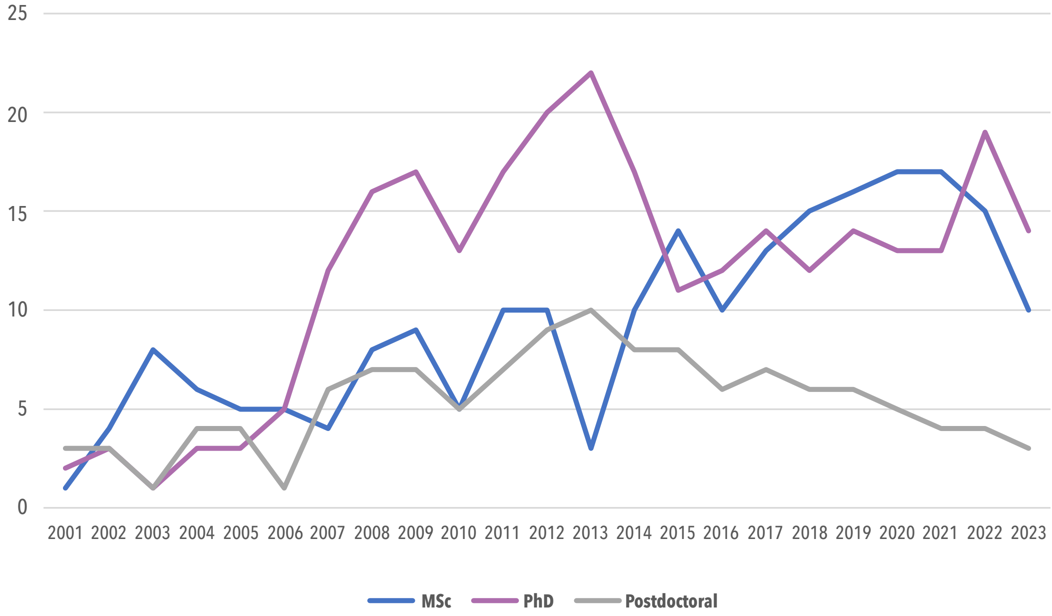
Nkosazana Nyembezi

*Harkins

Peter van Heusden

*Christoffels

SANBI POSTGRADUATE REGISTRATION TRENDS 2001 – 2023

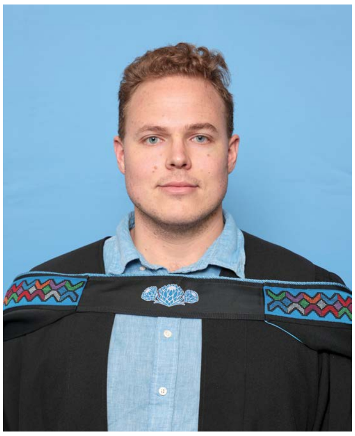


*Primary Supervisor

2023 SANBI GRADUATIONS

SANBI has great pleasure in presenting the 2023 graduants:

NAME	DEGREE	THESIS TITLE	PRIMARY SUPERVISOR	SECONDARY SUPERVISOR
Tracey Calvert-Joshua	PhD	Development of an Operon Detection Algorithm to Analyse Gene Regulation in Drug Resistant *Mycobacterium tuberculosis.*	Alan Christoffels	
Sarah DeRaedt	PhD	Investigating genetic diversity and microRNA of Hermetia illucens (the black soldier fly) to breed for mass production of a novel.	Alan Christoffels	Carmen Richards
Nasr Eshibona	PhD	Novel genomic biomarkers for pediatric and adult Acute Myeloid Leukaemia.	Hocine Bendou	Junaid Gamieldien
John Juma	PhD	Genomic epidemiology of Rift Valley fever in East Africa.	Alan Christoffels	Samuel Oyola
Riziki Martin	PhD	The identification and characterisation of DNA aptamers for application in diagnosis of infectious diseases.	Mervin Meyer	Abram Madiehe, Ruben Cloete
Nikita Pillay	PhD	Next Generation Sequencing Approaches for Novel Gene Discovery in South African Parkinson's Disease Families.	Alan Christoffels	Soraya Bardien (SUN)
Tsaoane Tamuhla	PhD (UCT)	Exploring new methodologies to identity disease-associated variant in African population through the integration of patient genotype data and clinical phenotypes derived from routine health data: A case study for type 2 Diabetes mellitus in patients in the Western Cape province South Africa.	Nicki Tiffin	Nicola Mulder
Keagan Brown	MSc *Cum Laude*	The development of an automated computational workflow to prioritise potential resistance variants in HIV Integrase Subtype C.	Ruben Cloete	
Susan Alicia Fernol	MSc	An investigation into the genetic basis of Osteogenesis imperfecta in a South African family of mixed ancestry.	Alan Christoffels	Manorgari Chetty
Wardah Jassiem	MSc	Identification of insertion-induced enhansers linked to gene drivers within noncoding DNA using a pipeline for diffuse large B-cell Lymphoma H3K27AC CHIP-SEQ DATA.	Hocine Bendou	
Lunathi Ndlondlo	MSc	Molecular modelling, docking and simulations studies to identify novel inhibitors against *Mycobacterium tuberculosis* drug target, Rv2194 (QcrC).	Ruben Cloete	
Christianah Omotoso	MSc	Homology Modeling and Molecular dynamics of three-dimensional protein structures of Selected Cyp3a4 variants with Lidocaine.	Samuel Egiyeh	Alan Christoffels



PHA4GE *Conference*

Over 150 geneticists, bioinformaticians and tech developers attended the inaugural conference of the Public Health Alliance for Genomic Epidemiology (PHA4GE) in Somerset West from 30 October - 1 November. The conference explored a range of topics that are crucial to public health and genomics.

Workshops were held at SANBI before the conference start. On Day 1, over 50 participants attended the Data Structures, Global Alliance for Genomics and Health Hackathon, and Bioinformatics Solutions workshops. Day 2 continued with sessions on using metagenomics to survey vector-based pathogens. The last day allowed for collaboration opportunities with focused working group meetings.

"In many ways, genomics is still an undervalued part of countries' – and the world's – public health response to disease outbreaks," said Alan Christoffels. "But Covid-19 illustrated how it can help inform how, when and where countries and regions can address diseases of concern."

Sign up for the PHA4GE quarterly newsletter at:

(https://mailchi.mp/pha4ge/pha4ge-feb24-newsletter)

WORKSHOPS AND COURSES

Africa CDC Pathogen Genomics workshop to support Cholera data analysis

During March 2023 the 2-week bioinformatics course held at SANBI was attended by 19 participants representing 13 African countries.

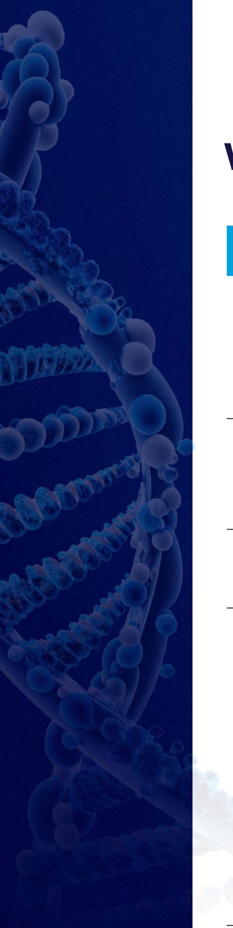
(https://www.uwc.ac.za/news-and-announcements/news/amid-recurring-outbreaks-sanbis-latest-training-programme-to-zero-in-on-cholera)

Workshop attendees in March 2023.

Introduction to Pathogen Bioinformatics workshop

This workshop **(https://pabix.sanbi.ac.za/)** ran over two weeks (26 - 30 June and 3 - 7 July) with 19 participants from the public health sector and spanned the whole SADC region. The course offered a practical introduction to Pathogen Bioinformatics for Genomic Epidemiology, using both Galaxy and command line interface (CLI). It covered analysis of sequencing data using SARS-CoV-2 as an example. Additionally, the course covered handling metadata to add context to the sequencing data being discussed and to illustrate techniques in genomic epidemiology.

Attendees of the workshop held in June-July 2023.



Workshops Organised/Taught By SANBI

TAUGHT BY	WORKSHOP/COURSE DETAILS	DATE
Nicki Tiffin	African Data and Biospecimen Exchange Stakeholders workshop for health genomics and biobanking researchers. Topics covered: User testing and stakeholder engagement for ADBEx functionality and user interface.	10 - 11 July
Nicki Tiffin and Dr Segun Fatumo, MRC Uganda/LSHTM	Research Open Day at MRC Uganda/UVRI: Finding solutions for equitable sharing of African data and biospecimens.	12 July
Anja Bedeker	Introduction to Research Ethics taught to UWC Medical Bioscience Honours degree class.	14 August
Peter van Heusden	Galaxy Smörgåsbord global online training event providing training on binformations to a global community of learners using the Galaxy platform.	22 - 26 May
	Mycobacterial Tuberculosis NGS made easy. Developed and facilitated together with the Swiss Tropical and Public Health Institute (Swiss TPH) and IRCSS Ospedale San Raffaele (Italy) and the global diagnostics NGO FIND. 20 learners participated virtually from research institutes and public health laboratories in Latin America, Africa, Asia and Europe.	6 September
Alan Christoffels	Africa CDC laboratory network writers workshop. 20 participants from National Public Health Laboratories were taught various aspects of scientific writing and publishing by a trainers from SANBI, KEMRI-Wellcome Trust and the Africa CDC.	5 - 7 December

Visiting Speaker Seminars at SANBI

TAUGHT BY	WORKSHOP/COURSE DETAILS	DATE
Aleksander Radovanovic, Quantum Computing Institute, Cape Town	Towards Quantum Bioinformatics. Exploring new technology and how bioinformatics can benefit from the information processing capabilities of quantum computers for better understanding and modelling of biological data.	14 February
Joshua Levy, Scripps Research California, USA	Wastewater Sequencing for Genomic Surveillance. With environmental sampling from wastewater sources playing an increasingly important role in public health pathogen genomics, interesting insights were provided.	5 May
Melissa-Rose Abrahams, UCT	Characterisation of the latent HIV reservoir in African women. Strengthening SANBI-UWC interaction with our neighbouring UCT researchers.	22 August
Placide Mbala-Kingebeni, Institut National de Recherche Biomédicale (INRB) and the University of Kinshasa	Experience of pathogen sequencing in the DRC. Strengthening the collaboration between SANBI-UWC and INRB in the DRC. We have a project with INRB that is led by Peter van Heusden.	2 November

Computational RESOURCES

SANBI's IT and bioinformatics software development team supports the work of the institute through software development and engagement with other computing providers to ensure that SANBI researchers are able to access appropriate computing resources.

Strategic computational infrastructure partnership with CHPC

As part of our role in leading the R&D for a continental data platform, we have established a partnership with CHPC. The cloud specialists at CHPC are contributing their skills to develop a proof-of-concept of sharing across geographical borders for pathogen data management. The HPC cluster supports the Ilifu cluster and cloud resources. Ruben Cloete's research group makes extensive use of both the CHPC and Ilifu for molecular dynamic simulations and docking simulation between candidate drugs and target proteins. Alan Christoffels and Nicki Tiffin use virtual machines hosted on the Ilifu cloud to host database resources.

Research and Development driving SANBI's computational infrastructure deployment

PRINCIPAL INVESTIGATOR	THEME	PROJECT
Dominique Anderson, Alan Christoffels	African Pathogen Data Archive	The Agari pathogen archive **(apaport.sanbi.ac.za)** is a pan-African infrastructure project that built a platform to curate and offer a pathogen analysis toolkit for use by national public health laboratories and academic partners. This work includes extending functionality to a federated model of data governance.
Nicki Tiffin	Hosting a public platform	The African Data and Biospecimen Exchange - Ilifu.
Peter van Heusden	Research Computing Infrastructure	Studying the deployment of research computing infrastructure and development of associated skills on the African continent through RSSE Africa.

Research Software and Systems Engineering Africa (RSSE Africa)

Initiated by Peter van Heusden with support from Talarify (a startup founded by ex-SANBI student Anelda van der Walt that supports researchers who want to develop digital and computational skills), a series of 6 online presentations and community discussions were held from May - October 2023. The aim of these events is to stimulate the growth of an African community of practice around research software and system engineering practices. By facilitating and promoting collaboration between African RSSEs and other regions, there is benefit from sharing innovative solutions implemented in different contexts while at the same time highlighting the work and skills on the African continent. The series alternated between webinars and community meetups to provide the African community with an opportunity to learn and share their own experiences.

Student Cluster Competition

SANBI supported the training of the UWC1 Student Cluster Competition that took part in this 2023 Centre for High Performance Computing Student Cluster Competition. The team qualified for the national round of the competition, which was held in December 2023 at the Kruger Gate Hotel on the outskirts of the Kruger Park.

Student Cluster Competition team practising at SANBI. Below in the pic are: (back row l to r): Lance, Vutomi and Ntanganedzeni (Ntanga) (front row): Mujtabaa and Peter.

Research OUTPUTS

SANBI's profile of excellence is reflected in its ability to publish high impact scientific articles in international publications. Bioinformatics is an interdisciplinary field and evidenced by our collaborative publication outputs. SANBI staff and students occupy leading roles in at least 40% of our outputs for the review period. Details of the institute's outputs are described in the tables that follow.

Journal Publications

SANBI AUTHORS TITLE PUBLICATION DETAILS DOI	PUBLISHED DATE
Hiltemann S, Rasche H, Gladman S, Hotz HR, Larivière D, Blankenberg D, Jagtap PD, Wollmann T, Bretaudeau A, Goué N, Griffin TJ, Royaux C, Le Bras Y, Mehta S, Syme A, Coppens F, Droesbeke B, Soranzo N, Bacon W, Psomopoulos F, Gallardo-Alba C, Davis J, Föll MC, Fahrner M, Doyle MA, Serrano-Solano B, Fouilloux AC, **van Heusden P**, Maier W, Clements D, Heyl F, Galaxy Training Network, Grüning B, Batut B ***Galaxy Training: A powerful framework for teaching!*** PLoS Comput Biol 2023;19(1): e1010752. **(https://doi.org/10.1371/journal.pcbi.1010752)**	9 January
du Bruyn E, Stek C, Daroowala R, Said-Hartley Q, Hsiao M, Schafer G, Goliath RT, Abrahams F, Jackson A, Wasserman S, Allwood BW, Davis AG, Lai RP, Coussens AK, Wilkinson KA, de Vries J, **Tiffin N**, Cerrone M, Ntusi NAB, HIATUS consortium, Riou C, Wilkinson RJ. ***Effects of tuberculosis and/or HIV-1 infection on COVID-19 presentation and immune response in Africa.*** Nature Communications 2023; 14(1):188. **(doi: 10.1038/s41467-022-35689-1)**	12 January
Juma J, Konongoi SL, Nsengimana I, Mwangi R, Akoko J, Nyamota R, Muli C, Dobi PO, Kiritu E, Osiany S, Onwong'a AA, Gachogo RW, Sang R, **Christoffels A**, Roesel K, Bett B, Oyola SO. ***Using Multiplex Amplicon PCR Technology to Efficiently and Timely Generate Rift Valley Fever Virus Sequence Data for Genomic Surveillance.*** Viruses 2023;15(2), 477. **(https://doi.org/10.3390/v15020477)**	9 February
Eshibona N, Livesey M, Christoffels A, Bendou H. ***Investigation of distinct gene expression profile patterns that can improve the classification of intermediate-risk prognosis in AML patients.*** Front Genet. 2023;14:1131159. **(doi: 10.3389/fgene.2023.1131159)**	14 February
Mutemaringa T, Heekes A, Smith M, Boulle A and **Tiffin N**. ***Record linkage for Routinely Collected Health Data in an African Health Information Exchange.*** International Journal of Population Data Science, 8(1). **(doi: 10.23889/ijpds.v8i1.1771)**	2 March
Patten V, Parker MI , Hendricks D, **Bendou H**, Mathew C, Chen C ***1795: Aberrant expression and somatic mutations in the MUC3A gene of South Africans with Esophageal Squamous Cell Carcinoma.*** Journal of Biological Chemistry 299 (3S). **(https://doi.org/10.1016/j.jbc.2023.103698)**	13 April

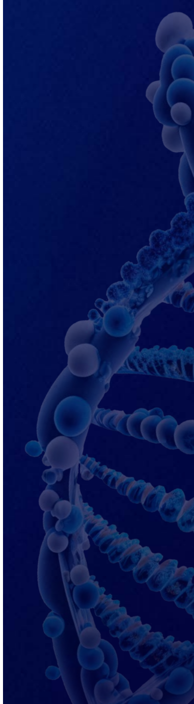
Dellicour S, Hong SL, Hill V, Dimartino D, Marier C, Zappile P, **Harkins GW**, *et al.* ***Variant-specific introduction and dispersal dynamics of SARS-CoV-2 in New York City–from Alpha to Omicron.*** PLoS Pathogens 2023;19(4): e1011348. (**https://doi.org/10.1371/journal.ppat.1011348)**	18 April
Livesey M, Rossouw SC, Blignaut R, **Christoffels A, Bendou H.** ***Transforming RNA-Seq gene expression to track cancer progression in the multi-stage early to advanced-stage cancer development.*** PLoS ONE 2023;18(4): e0284458. **(https://doi.org/10.1371/journal.pone.0284458)**	24 April
Christoffels A, Mboowa G, **van Heusden P**, Makhubela S, Githinji G, Mwangi S, Onywera H, Nnaemeka N, Amoako DG, Olawoye I, Diallo A, Mbala-Kingebeni P, Oyola SO, Adu B, *et al.* ***A pan-African pathogen genomics data sharing platform to support disease outbreaks.*** Nature Medicine 2023. **(10.1038/s41591-023-02266-y)**	9 May
Kassanjee R, Davies M-A, Ngwenya O, Osei-Yeboah R, Jacobs T, Morden E, Timmerman V, Britz S, Mendelson M, Taljaard J, Riou J, Boulle A, **Tiffin N**, Zinyakatira N. ***COVID-19 among adults living with HIV: correlates of mortality among public sector healthcare users in Western Cape, South Africa.*** Journal of the International AIDS Society 2023; 26, 6, e26104 **(https://doi.org/10.1002/jia2.26104)**	20 June
Tamuhla T, Lulamba ET, Mutemaringa T, **Tiffin N.** ***Multiple modes of data sharing can facilitate secondary use of sensitive health data for research.*** BMJ Glob Health 2023; 8:e013092. **(doi:10.1136/ bmjgh-2023-013092)**	12 September
Akinnuwesi A, Egieyeh S, **Cloete R.** ***State-of-the-art strategies to prioritise Mycobacterium tuberculosis drug targets for drug discovery using a subtractive genomics approach.*** Frontiers in Drug Discovery, Section. *In silico* Methods and Artificial Intelligence for Drug Discovery 2023; 3:1254656. **(https://doi.org/10.3389/fddsv.2023.1254656)**	18 September
Zass L, Johnston K, Benkahla A, Chaouch M, **Kumuthini J**, Radouani F, Mwita LA, Alsayed N, Allie T, Sathan D, Masamu U, Tchamga MSS, **Tamuhla T**, Samtal C, Nembaware V, Gill Z, Ahmed S, Hamdi Y, Fadlelmola F, **Tiffin N**, Mulder N. *Developing Clinical Phenotype Data Collection Standards for Research in Africa.* Global Health, Epidemiology and Genomics 2023; 6693323 **(https://doi.org/10.1155/2023/6693323)**	19 September
Onywera H, Ondoa P, Nfii F, Ogwell A, Kebede Y, **Christoffels A**, Tessema SK ***Boosting pathogen genomics and bioinformatics workforce in Africa.*** The Lancet Infectious Diseases 2023; (1473-3099) **(https://doi.org/10.1016/S1473-3099(23)00394-8)**	28 September

Fatumo S, Sathan D, Samtal C, Isewon I, Tamuhla T, Soremekun C, Jafali J, Panji S, Tiffin N and Fakim YJ. Polygenic risk scores for disease risk prediction in Africa: current challenges and future directions Genome Medicine 2023; 15:87 **(https://doi.org/10.1186/s13073-023-01245-9)**	30 October
Juma J, Oyola SO, Konongoi S, Nsengimana I, Mwangi RK, Akoko J, Nyamota R, Muli C, Kiritu E, Dobi P, Osiany A, Onwong'a A, rwanja8, Sang R, Christoffels A, Roesel K, Bett B, Oyola SO. ***Amplicon multiplex PCR sequencing of Rift Valley fever virus (RVFV) on Illumina MiSeq.*** protocols.io **(https://protocols.io/view/amplicon-multiplex-pcrsequencing-of-rift-valley-fckb2usqe)**	14 November
Bendou H, Livesey M, Eshibona N. ***Assessment of the progression of Kidney renal clear cell carcinoma using transcriptional profiles revealed new cancer subtypes with variable prognosis.*** Frontiers Genetics Section Computational Genomics 2023; Volume 14. **(doi: 10.3389/fgene.2023.1291043)**	13 November
LeFevre A, Welte O, Moopelo K, Tiffin N , Mothoagae G, Ncube N , Gwiji N, Shogole M, Slogrove AL, Moshani N, Boulle A, Goudge J, Griffiths F, Fairlie L, Mehta U, Scott K, Pillay N. ***Preferences for onward health data use in the electronic age among maternity patients and providers in South Africa: a qualitative study.*** Sexual and Reproductive Health Matters 2023; 31(4):2274667. **(https://doi.org/10.1080/26410397.2023.2274667)**	20 November
Tamuhla T, Raubenheimer P, Dave JA, Tiffin N. ***Routine health data describe adherence and persistence patterns for oral diabetes medication for a virtual cohort in the Khayelitsha sub-district of Cape Town, South Africa.*** PLoS Global Public Health 2023; 3(12): e0002730. **(https://doi.org/10.1371/journal.pgph.0002730)**	21 December

PHA4GE secretariat at the ABPHM23 Conference at the Wellcome Genome Campus, UK.

Invited Talks Presented

PRESENTER NAME	CONFERENCE NAME AND VENUE	TITLE	DATE
Alan Christoffels	Roche, Cape Town	Role of sequencing during the pandemic and its role in real-time disease surveillance in Africa	16 March
Anja Bedeker	GA4GH Connect 2023, Royal Society of London, UK	Ethical preparedness for future pandemics and epidemics	21 April
Alan Christoffels	PathRED2023: Pathology Research & Development Conference, Johannesburg	Harnessing the power of bioinformatics to transform biology	1 September
Peter van Heusden	Global Microbial Identifier, Simon Fraser University Wosk Centre for Dialogue, Vancouver, Canada. **(https://gmi13.org/)**	1. Data custodianship in Africa: Towards an African Data Sharing and Archive Platform (Alan Christoffels - presented by Peter van Heusden). 2. PHA4GE: bridging the gap between public health and bioinformatics (Alan Christoffels - presented by Peter van Heusden). 3. Wastewater sequencing for genomic surveillance: Infrastructure implications and options (Peter van Heusden).	13 - 15 September
Anja Bedeker	GA4GH 11th Plenary, San Francisco, USA	Ethical Preparedness for Future Pandemics and Epidemics	19 September
Nicki Tiffin	NeoNet Consortium Meeting	Presentation: NeoNet-Africa: a multi-disciplinary network to reduce neonatal deaths from bacterial infections in sub-Saharan Africa.	21 September
	Global Microbial Identifier Meeting GMI13, Vancouver, Canada	Facilitating Ethical Microbial Data Sharing with Innovative Tools	22 September
	Malaria Genomics Convening, Dakar, Senegal	Facilitating Equitable Data and Biospecimen Sharing	12 October
	MalariaGen Meeting	The African Data and Biospecimen Exchange	18 October



Conference Participation

PRESENTER NAME	CONFERENCE NAME AND VENUE	TYPE AND TITLE	DATE
Nicki Tiffin	Collaboration of AIDS Vaccine Discovery, Cape Town	Panelist: Democratising Access to Bioinformatics	14 February
Alan Christoffels	African Laboratory Operational Framework Review & Finalisation, Nairobi, Kenya	Technical Working Group Panelist: Laboratory Operational Framework Review	28 February - 2 March
Nicki Tiffin	ICTS Programme "Machine Learning for Health and Disease"	Presenter and attendee: International Centre for Theoretical Sciences of the Tata Institute of Fundamental Research (ICTS-TJFR), Bengaluru	24 July - 4 August
Tsaone Tamuhla	G2MC 7th International Conference, Geneva	Presenter: **Won the Best Podium Presentation Award: Early Career Investigators talk:** Implementation of a genotyped virtual African population cohort: A feasibility study in the Western Cape Province, South Africa	7 - 10 October
Nicki Tiffin	B&MGF Grand Challenges Meeting, Dakar	Speaker: Key Ingredients for R&D Ecosystems and Accelerating Innovation in LMICs: African Voices on the Vision for African Science	9 October
Alan Christoffels	B&MGF Grand Challenges Meeting, Dakar	Speaker: Genomics Epidemiology Network	10 October
Nicki Tiffin	B&MGF Grand Challenges Meeting, Dakar	Panelist: Understanding African governments' contributions towards competitive R&D ecosystems	10 October

Postdoctoral Fellow, Tsaone Tamuhla, receiving the Best Podium Presentation at the G2MC 7th International Conference in Geneva.

PhD student, Adetutu Akinnuwesi was awarded 3rd prize in the doctoral division at the CHPC National Conference.

Alan Christoffels	Malaria Genomics Convening 2023, Senegal, Dakara	Speaker: African data sharing and analysis platforms: Opportunities for Malaria	11 - 13 October
Nicki Tiffin	PHA4GE Conference, Cape Town	Ethics and Data sharing Working Group Session facilitator: Working group data sharing discussion and report back	1 November
Anja Bedeker	Global Genomics Conference	Expert panellist: How can we create a more equitable approach to benefit sharing in global genomics?	9 November
Nicki Tiffin	CPHIA 2023 3rd International Conference on Public Health in Africa	Panelist: Breaking Barriers: Repositioning Africa in the Global Health Architecture	28 November
Keaghan Brown	Centre for High Performance Computing (CHPC) National Conference 2023 Skukuza	MSc poster presentation: The development of an automate computational workflow to prioritise potential resistance variants in HIV Integrase Subtype C	4 - 7 December
Adetutu Akinnuwesi	Centre for High Performance Computing (CHPC) National Conference 2023 Skukuza	**3rd prize in the doctoral division of the 2023 research student poster competition:** Computational methods to identify novel compounds against *Mycobacterium tuberculosis* target Acetylglutamate kinase.	4 - 7 December
Peter van Heusden	Centre for High Performance Computing (CHPC) National Conference 2023 Skukuza	Attendee: Mentor for the UWC Student Cluster Competition team	4 - 7 December
Dominique Anderson	PHA4GE conference, Cape Town	Attendee	30 October - 1 November
	ASLM2023, Cape Town	Attendee	13 - 15 December

Research LABORATORIES

PROF ALAN CHRISTOFFELS

RESEARCH PROJECT THEMES

Pathogen genomics
- In collaboration with the Tygerberg MRC Unit, we are developing methods to analyse high throughput sequencing data for microbial genomes.
- Identification of novel drug targets in pathways known to contain drug resistant genes.
- Predicting the interaction networks between human and mycobacteria
- Identification of operons & non-coding RNAs to understand gene regulation in pathogen genomes.

Pathogen surveillance
- Developing systems to support reproducible analyses
- Development of global standards for data sharing and integration
- Public Health Alliance for Genomic Epidemiology (PHA4GE)

Drug discovery
- Identifying drug molecules against multi-drug resistant *M.tuberculosis*
- In partnership with the School of Pharmacy at UWC, delineate the pharmacokinetics of drugs in response to host genetic factors
- In partnership with the School of Pharmacy at UWC, use of machine learning techniques to prioritise molecules in natural compounds.

HIGHLIGHTS OF 2023

Four PhD and 1 MSc students graduated during 2023. This work included a computational tool to predict operons. This methodology has been submitted for publication. Understanding the genetic networks of genes that are co-regulated provide new targets for drug resistance.

The microbial genetics work is integrated into genomic surveillance. We are adapting our data platform to handle other data types as part of our collaborative work with neighboring countries.

Members of our group are active in the global PHA4GE consortium. Specifically, our efforts have been directed at strengthening national public health labs in resource limited settings in the data analytics space. I have engaged various international forums to promote our public health Bioinformatics work.

RESEARCH PROJECTS

The projects below underpin our translational work:

Tuberculosis

A comprehensive research programme is underway that includes investigating genetics determinants in both host (Human) and pathogen (*Mycobacterium tuberculosis*) to understand drug resistance, and protein structure determination to inform patient-centric drug design. These findings inform the development of a scientific workflow management system to support reproducible high throughput computational experiments. One of the projects focus on predicting operons. Operons may be important drug targets for the development of effective anti-microbials to combat the emerging, global drug resistance challenge. However, there is a shortage of known *Mycobacterium tuberculosis (Mtb)* operons. We developed COSMO, an algorithm that removes the limitations imposed by using the constraints of one specific organism's genome and exploits raw RNA-seq data instead. The code and examples of input and output files for testing and validation, are available at the project home page at **(https://github.com/SANBI-SA/COSMO)**

Biobanking

In collaboration with Dominique Anderson at SANBI, we continue to strengthen our biobank software capabilities so that genetics data and biospecimens can be curated and stored.

Applications of Machine Learning methods to protein-protein interactions

Recently we used quadruplet amino acids as a feature in training model for protein interactions. We are extending this work to improve on the accuracy of our models with small datasets.

RESEARCH COLLABORATIONS

1. Analysing genetic networks in M.tuberculosis

Collaborating Parties:
Prof Rob Warren - University of Stellenbosch
Dr Hocine Bendou - SANBI, UWC
Nature and Purpose:
Identification of Operons: We rewrote an algorithm previously published by a team at Johns Hopkins University.
Identification of non-coding RNAs.
Output in the last 12 months:
Our Operon detection algorithm has been tested with different metrics, and is currently benchmarked against two international methods.
Using a sequence conversation approach and a RNA folding approach, we are identifying putative small RNAs.

2. Identification of novel drug targets for drug resistant tuberculosis

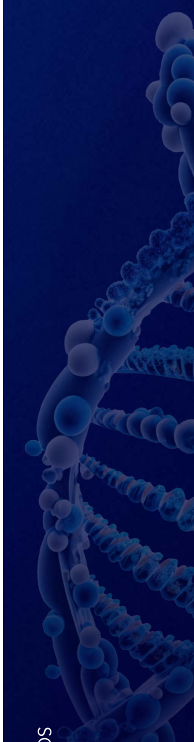
Collaborating Parties:
Profs Samantha Sampson and Rob Warren - University of Stellenbosch
Dr Ruben Cloete - SANBI, UWC
Prof Sarel Malan - School of Pharmacy, UWC
Prof Valery Danilenko and Dr Dmitry Maslov - Vavilov Institute of General Genetics, Russia
Nature and Purpose:
Current TB drugs are more than 30 years old and have unacceptable efficacy and safety profiles, emphasising the need for new drugs. In this regard we published a paper on kinase inhibitors in *M.tuberculosis*, and efflux pump inhibitors.
Output in the last 12 months:
We continue to refine our computational screening system. And submitted a funding application via the BRICS-country funding mechanism.

3. Public Health Alliance for Genomic Epidemiology (PHA4GE)

Collaborating Parties:
Africa CDC, Oxford University, Washington University, University of Melbourne, USA CDC, University of Birmingham UK, University of British Columbia and BC Centre for Disease Control, Broad Institute in Boston, and H3Africa.
Nature and Purpose:
The PHA4GE consortium was established to bring Bioinformatics closer to Public Health - to build data standards.
Output in the last 12 months:
PHA4GE data standards working group developed a metadata standard for COVID-19 biospecimen collection.

4. African Genomics Archive

Collaborating Parties:
Africa CDC
Nature and Purpose:
To establish a data platform in Africa to manage disease surveillance data with a view to strengthen public health institutes.
Output in the last 12 months:
A series of workshops to gather information on various use cases.



Team Members:
Postdocs: Tsaone Tamuhla, Eddie Lulamba
PhD students: Sohail Simon, Florence Phelanyane and Themba Mutemaringa
Researcher: Irene Muchada
Administration: Saajidah Beghardien

PROF NICKI TIFFIN

HIGHLIGHTS OF 2023

Students

In 2023 we celebrated the graduation with PhDs in Bioinformatics by Dr Tsaone Tamuhla and Dr Richard Osei-Yeboah, and the graduation of Pierre Dane with his Masters in Public Health Degree. Themba Mutemaringa and Florence Phelanyane continue part-time PhD degrees under my supervision at the University of Cape Town, with Themba having his first research chapter accepted for publication in a peer-reviewed journal (in press in 2023). We also celebrated the completion and submission of the PhD thesis of Ryan Aylward, who I co-supervise at the University of Cape Town/Bristol University and who received his PhD degree. These students have a variety of projects working with large routine health datasets as well as conducting pilot studies to link genomic and routine health data.
We are delighted to welcome Ms Irene Muchada to the team to work on the VCAMM project. She brings to the team her credentials as a genetic counsellor, as well as her strong community engagement ethos and her interest in the experiences and wellbeing of participants who join research studies. She will be a key player in overseeing how we return genetic results to the health service in a responsible way.
We are also excited to have a new PhD student, Mr Sohail Simon, join the team in 2023. Sohail will be working on Machine Learning approaches for working with routine health data.

Funding

In 2023 we completed our second year of funding with the Bill & Melinda Gates Foundation, provided through the Calestous Juma Science Leadership Fellowship award to myself. This fellowship provides five years of funding for building the African Data and Biospecimen Exchange (ADBEx) as well as opportunities for networking and training in non-scientific expertise for the fellows.
We were also delighted to get going with the seed funding from UKRI/MRC which is supporting our VCAMM pilot programme to build a virtual genotyped multi-morbidity population cohort through the linkage of routine health data to genotype data for consenting participants in the Western Cape, South Africa. The research programme started in June 2023, and Tsaone has started to pilot these processes with a proof-of-concept study together with our service partners at the Central Analytical Facility at Stellenbosch led by Carel Van Heerden, and the sterling contributions on Irene's part.
Finally, we were co-investigators on a successful funding application to the Bill & Melinda Gates Foundation for a collaborative programme together with partners from Centro de Integração de Dados e Conhecimentos para Saúde (CIDACS) in Brazil and the Provincial Health Data Centre at the Western Cape Department of Health and Wellness, to work together to share learnings, develop a common data model, and undertake analysis of drivers of disease outcomes in pregnant women with syphillis and TB. This work will start in early 2024 with reciprocal working meetings in Brazil and Cape Town.

RESEARCH PROJECTS AND COLLABORATIONS

1. Multimorbidity in the Western Cape

The focus of this project is the management, integration and analysis of routine health data from health care clients attending government health facilities in the Western Cape.

1.1 Establishing a virtual population cohort using routine health data for epidemiological analyses

Nature and purpose: We are analysing anonymised and perturbed routine health data from a virtual cohort of all health care clients who attended any government health facility in the Khayelitsha sub-district in Cape Town during 2016 and 2017. Analysis of these data provides insights into multimorbidity in this population group that suffers from a high burden of both infectious and non- communicable diseases. In addition, we have continued to work in collaboration with the Provincial Health Data Centre (PHDC) at the Western Cape Department of Health in epidemiological and service-delivery projects working with routine health data from the Western Cape.

Researchers: Tsaone Tamuhla; Sohail Simon, Richard Osei-Yeboah, Florence Phelanyane

Outputs in the past 12 months:

Tamuhla T, Dave JA, Raubenheimer P, Tiffin N. Routine health data describe adherence and persistence patterns for oral diabetes medication for a virtual cohort in the Khayelitsha sub-district of Cape Town, South Africa. In Press, PLoS Global Public Health. 2023

Banda GT, Bosire E, Bunn C, *et al*. (including Tiffin N, Tamuhla T). Multimorbidity research in Sub-Saharan Africa: Proceedings of an interdisciplinary workshop. Wellcome Open Research, 2023, 8 e110. **(https://doi.org/10.12688/wellcomeopenres.18850.1)**

Future Directions:

In this project, we will continue to build on appropriate methodologies for analysis of complex routine health data in order to inform evidence-based care. Sohail Simon will be working within this project to further explore analysis using latent factor analysis as well as machine learning approaches to analyse risk factors for kidney disease in this population group.

1.2 Piloting a virtual genotyped population cohort linking genotype and routine health data:

Nature and purpose: With informed consent from health care clients, it is possible to link individuals' genotype data with their routine health data in order to establish a genotyped virtual cohort with complex clinical phenotype data that can be updated into the future using routine health records. We are piloting this approach by establishing the informed consent process, sample and data workflows to create, in the first instance, a virtual genotyped cohort for patients with diabetes in Cape Town. Funding from UKRI/MRC will fund a pilot study for proof-of-principle establishment of a virtual genotyped population cohort using this approach, as well as establishing the infrastructure to return actionable findings from the analysis to clinicians at the Western Cape Department of Health. We also continue to work with collaborators across a variety of projects as new approaches to genomic data analyses are applied to genomic data from African populations.

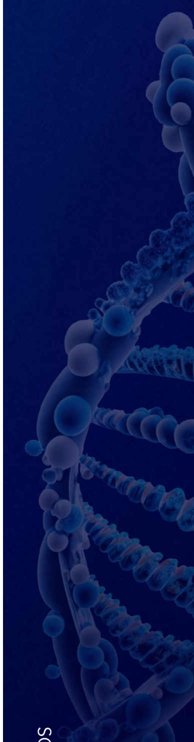
Researchers: Tsaone Tamuhla, Nicki Tiffin.

Outputs in the past 12 months:

Fatumo S, Sathan D, Samtal C, Isewon I, Tamuhla T, Soremekun C, Jafali J, Panji S, Tiffin N, Fakim YJ. Polygenic risk scores for disease risk prediction in Africa: current challenges and future directions. Genome Med. 2023 Oct 30;15(1):87. doi: 10.1186/s13073-023-01245-9.

Esterhuizen A, Tiffin N *et al*. Precision medicine for developmental and epileptic encephalopathies in Africa–strategies for a resource-limited setting. Journal of Clinical Investigation 2023,25, 2

Future Directions: A proof-of-concept pilot project is ongoing to establish the analysis pipeline with an appropriate informed consent process, collection of buccal swabs and preparation of high quality DNA sample, generation and analysis of genotype data, and linkage to routine health data for phenotype definition (Tsaone Tamuhla). In addition we will be building infrastructure to ensure that actionable genetic findings from these studies can be appropriately returned to the Western Cape Department of Health to inform evidence-based care.



1.3 Optimising data linkage of African health client data:

Nature and purpose:

When integrating individualised data from a variety of sources, different linkage algorithms can be used to ensure the most accurate linkage and deduplication of records that originate from the same individual. Most of these algorithms, however, have been developed using Euro-centric datasets and do not necessarily perform as well in African contexts, and we are working on a systematic approach to improve data linkage for routine health data. In 2022 we completed a descriptive analysis of the current iteration of the linkage algorithm at the PHCD at the Western Cape Department of Health.

Researcher: Themba Mutemaringa

Collaborators: Andrew Boulle, Alexa Heekes, Western Cape Department of Health

Outputs in the past 12 months:

Mutemaringa T, Heekes A, Smith M, Boulle A, Tiffin N. Record linkage for Routinely Collected Health Data in an African Health Information Exchange. International Journal of Population Data Science, 2023 Vol. 8 No. 1. DOI: 10.23889/ijpds.v8i1.1771

Future Directions: In this project**,** Themba will continue to lead the validation of existing linkage and de-duplication algorithms, and will develop and test new methodologies and algorithms to best achieve de-duplication and data linkage. In addition, he is working on a thorough review of the governance protocols and ethical requirements for this ongoing work using routine health data at the PHDC.

2. Facilitating ethical and equitable sharing of African data and biospecimens: the African Data and Biospecimen Exchange - ADBEx

Nature and purpose:

Towards the end of 2021 I established a programme, funded by the Calestous Juma fellowship from the Bill & Melinda Gates Foundation, to build an online platform to facilitate ethical and equitable secondary sharing of African data and biospecimen resources. The query-able platform will store meta-data about sample and data collections, submitted by researchers who hold those resources, and will enable direct sharing partnerships to be established between resource providers and resource consumers without centralisation of samples or data. It will also provide online infrastructure for drawing up and recording data-sharing, benefit-sharing and collaborative agreements as well as building a repository of data and biospecimen guidelines and governance resources.

We have developed two related projects that are linked to ADBEx and are currently under development. The first is being led by Tsaone, to create a stand-alone application that can be used to build fit-for-purpose data- and biospecimen-sharing agreements. Whilst this functionality will be incorporated into the ADBEx platform we believe it may also be useful for a wider audience. The second project is to build a Data Sharing license. Drawing inspiration from the Creative Commons licensing structure, we are in the process of developing a similar type of license that can be used to effectively and unambiguously communicate the nuanced sharing requirements for individual datasets.

We have also embarked on a collaborative project with PHA4GE to create a Data Sharing Accord - a document that outlines the basic "common sense" principles for sharing microbial data, that end users can adopt as a common ground for general accepted principles for sharing data for secondary use. We are engaging with multiple stakeholders through a variety of meetings and forums in order to build a community consensus on what this document should include.

Researchers: Nicki Tiffin, Eddie Lulamba, Tsaone Tamuhla.

Collaborators: Hominum Global, www.hominum.global - User Interface and Software Specification and Development; MethodLab, **(https://methodlab.io/),** - Software development

PHA4GE collaborators: Peter van Heusden, Emma Griffiths, Michelle Nichols, Anja Bedeker

Outputs in the past 12 months:

Tamuhla T, Lulamba ET, Mutemaringa T, Tiffin N. Multiple modes of data sharing can facilitate secondary use of sensitive health data for research. BMJ Glob Health. 2023 Oct;8(10):e013092. doi: 10.1136/bmjgh-2023-013092.

Future Directions: We anticipate that the platform will be completed by June 2023, and we will be launching the platform and undertaking a series of roadshows to introduce the functionality to various user groups and help to get individuals uploading their sample collections' and datasets' metadata, to begin the process of advertising their resources that are available for collaborative analysis and/or onward sharing.

In Nairobi with Calestous Juma Fellows Iruka Okeke, Mainga Hamaluba and Annettee Nakimuli.

The Calestous Juma Fellowship

(https://gcgh.grandchallenges.org/article/nicki-tiffin)

In November 2021 the Bill & Melinda Gates Foundation launched the Calestous Juma Fellowship programme for scientists permanently located at African research organisations. I was one of fourteen recipients of these fellowships, and I have been participating in a series of leadership trainings, networking opportunities and skills development as part of this programme.

In 2023 I participated in two weeks of leadership training provided by the African Leadership Institute, in Stellenbosch in March 2023, and in Nairobi in November 2023. In addition I travelled to the Bill & Melinda Gates Foundation in Seattle together with the other CJ Fellows for a week of networking and engagement with the Foundation. I also participated as a CJ Fellow in the B&MGF Grand Challenges Meeting in Dakar, Senegal, in October. Once again I was able to bring along a team member to this meeting as a benefit of the fellowship, so Themba also came to Dakar for an extremely busy and exciting meeting. I additionally participated in the Malaria Convening which was a side meeting following on from the main Grand Challenges Meeting.

As always, the cohort of CJ Fellows are extremely inspirational, and it is an honour to work together with this amazing group of African scientists.

3. Ethics and governance to promote ethical and equitable health research in Africa

Nature and purpose:
In order to undertake ethical research in Africa and other regions of the global South, there are multiple considerations around risks, benefits, informed consent and equitable research. We explore different ways to better do research and to ensure that health research in Africa is ethical and equitable through the development of tools and frameworks. I also participate in multiple ethics and governance working groups of international consortia including PHA4GE (Chair: Ethics and Data Sharing Working Group), ICDA (co-Lead: Global Equity Working Group) and IHCC (co-Chair: Systems and Policy Working Group). I have continued to lead the Ethics Working Group Forum, which brings together ELSI working groups from the many different large health genomics research networks and consortia in order to meet and share information about their activities, challenges and resources.

Outputs in the past 12 months:
LeFevre A, Welte O, Moopelo K, **Tiffin N**, Mothoagae G, Ncube N, Gwiji N, Shogole M, Slogrove AL, Moshani N, Boulle A, Goudge J, Griffiths F, Fairlie L, Mehta U, Scott K, Pillay N. Preferences for onward health data use in the electronic age among maternity patients and providers in South Africa: a qualitative study.Sex Reprod Health Matters. 2023 Dec;31(4):2274667. **(doi: 10.1080/26410397.2023.2274667)**.

Future Directions:
We are currently developing a data sharing license to capture and communicated nuances of data sharing permissions, and to facilitate data sharing that respects conditions of sharing.
We are also collaborating with PHA4GE to develop a Data Sharing Accord, that can similarly assist with defining fundamental principles and conditions for data sharing that are commonly recognised for sharing of microbial data.

HIGHLIGHTS OF 2023

- Lead in Data Curation Capacity Development and User Design
- Supervisor of 1 PhD student and 1 MSc students
- Co-supervisor of 1 PhD student, 1 MSc student

DR DOMINIQUE ANDERSON

The research group is investigating several areas with an interest in metagenomics, data security and privacy, biomedical data governance, informatics of biobanking and application of AI in *OneHealth*. I have an ongoing role in a team focused on informatics solutions for biobanking with continued development, enhancement, and training in the open-source Baobab laboratory information management system. Information management and quality management remains a core area of activity in my research group with hopes to expand the collaborative network of developers and researchers. A recent collaboration has been established with the NHLS, to evaluate data workflows and metadata standards with a view to improve data quality in the public health setting. In relation to additional interests, I am involved in community engagement and effective science communication and training and am an active committee member of the DSI Medical Biobank Cluster, the TTO seed-fund committee and the POPIA code of conduct committee.

With regards to data privacy and governance, my collaborators and I combine expertise in biobanking, bioinformatics and big data, and law to investigate how both regulations and best practices in ICT influence the biomedical data and data sharing landscape in Africa. I am also currently working towards establishing networks within the agricultural, veterinary, medical, and environmental sphere in order to intensify research in the area of *OneHealth* and machine learning and investigating field-able molecular biology and '-omics' sequencing for pathogen genomic surveillance, as well as investigating biodiversity bioprospecting and the associated governance of indigenous knowledge resources.

RESEARCH PROJECT THEMES

Informatics Tools

- Biobank informatics - multiple collaborations for biocollection sample quality management and e-infrastructure with a focus on Baobab LIMS.
- Pathogen genome databases for impact in public health and pathogen surveillance.

Biomedical data governance

- Collaboration with biobank and law experts to evaluate data governance of biological diversity and human biomedical collections.
- Cybersecurity and data management in Africa.
- Metadata standardisation and impacts of data management in public health.

OneHealth

- *In silico* research focussed on OneHealth.

RESEARCH COLLABORATIONS

1. Data governance and the POPI Act

Collaborating Parties:
ASSAf
Nature and purpose: To investigate the impact of local and international regulation on the sharing of biomedical research data in Africa. Drafting committee member for the POPIA code of conduct for research.
Output in the last 12 months: Submission of draft code for public comment and evaluation by the Information regulator.
Future Direction: Ongoing engagement and amendments to the code of conduct.

2. Bioprospecting, biodiversity and biobanking

Collaborating Parties:
Prof Bryan Maritz - UWC
Dr Ciara Staunton - EURAC
Dr Carmen Swanepoel - SUN
IMBM UWC
Medical Biocollections cluster - MBirSA
Nature and purpose: Facilitating the development of a network of medical biorepositories in South Africa
Output in the last 12 months: Business case approved for MBirSA.
Future directions: Ongoing engagements and collaboration for the Medical biorepository network and establishing biodiversity bioprospecting and data governance with respect to indigenous knowledge resources.

3. Biobank informatics and data quality management

Collaborating Parties:
Prof Alan Christoffels - SANBI, UWC
NHLS
UCT
Africa CDC
PHA4GE Consortium
Nature and purpose: Enhancement and customisation of Baobab LIMS for LMICs. Standardisation of e-infrastructure. Evaluating the impact of data standardisation in public health and mechanisms to improve data quality across workflows. Development of Data platforms for sharing and archiving pathogen data.
Output in the last 12 months: Successful completion of Phase 1 of the African Pathogen Data Sharing and Archive platform.
Future Direction: New collaborations within Africa as well as training initiatives on the continent. Maintenance and enhancement of Baobab LIMS. Entrepreneurship focused on hybrid social and sustainability models for QM software. User design thinking in software development as well as building a data curation cohort with regional representation on the continent. Phase 2 of the APA.

4. OneHealth

Collaborating Parties: Seeking collaborative partners.
Nature and purpose: Multidisciplinary approaches to understanding infectious diseases. Development of metadata standards, novel AI based tools for surveillance and species cross-over and translational research in *OneHealth*.Examining field-forward technologies bringing molecular biology and '-omics' sequencing to the sample.
Output in the last 12 months: MSc graduate: Mr Rudolph Serage, *Cum Laude*
Future Direction: Expanding on the ML algorithm developed in the MSc project and examining model prediction efficacy with new data inputs.

HIGHLIGHTS OF 2023

Graduation in September 2023 of my PhD student Nasr Eshibona, with thesis title *"Novel Genomic Biomarkers for Paediatric and Adult Acute Myeloid Leukemia (AML)"*. Through principal component analysis and hierarchical clustering of RNA-seq gene expression counts from paediatric acute myeloid leukaemia samples, found that samples carrying the *FLT3-ITD* mutation (poor prognosis) were grouped, in the same cluster, with samples carrying the *NPM1* and the *CEBPA* mutations (favourable prognosis). This surprising result led to further bioinformatics analyses and the discovery of three genes potentially responsible for the poor prognosis in *FLT3-ITD* mutated patients (a research manuscript is under review).

DR HOCINE BENDOU

Submission of an NIH grant application with my postdoctoral student, Catherine Rossouw, aimed at the early-stage development of informatics technologies for cancer research and management.

My research group identified a methylation CpG signature capable of predicting the amplification of *MYCN*, a predictor of poor prognosis, in the early stages of neuroblastoma. This achievement has the benefit of helping clinicians design better treatment for patients with *MYCN* amplification. The work was published in the journal *Future Oncology* with an impact factor of 3.4 (2020).

My PhD student Abdulazeez Giwa, submitted his thesis entitled "Computational analysis of multi-omic data for the elucidation of molecular mechanisms of neuroblastoma" for examination.

Launch of a collaborative research project with Dr Carmen Pheiffer from the SAMRC to study the effect of physical training on the expression of miRNAs in gluteal adipose tissue in women with obesity.

RESEARCH PROJECT THEMES

Cancer Bioinformatics

- Novel genomic biomarkers for Pediatric and Adult Acute Myeloid Leukemia
- Computational analysis of multi-omic data for the elucidation of molecular mechanisms of neuroblastoma.

Pipeline development

- Development of a Nextflow pipeline for the detection of enhancer-associated noncoding variants using ChIP-seq data.

Machine learning

- Prediction of synthetic lethal gene interaction in human cancer using supervised machine learning.

Gene regulatory networks

- Gene regulatory network study of diffuse large B-cell lymphoma.

Viral Study

- Identify impact of viral inserts in human esophageal cancer. A collaborative work with the University of Cape Town.

RESEARCH COLLABORATIONS

1. Molecular mechanisms of neuroblastoma

Collaborating Parties:
Dr Azeez Fatai - Department of Biochemistry, Lagos State University, Lagos, Nigeria
Abdulazeez Giwa - SANBI, UWC
Nature and purpose: Elucidate the molecular mechanisms of neuroblastoma in high-risk patients using multi-omics data and machine learning approaches
Output in the last 12 months: Publication article in Future Oncology (3.4 impact factor): Predicting amplification of *MYCN* using CpG methylation biomarkers in neuroblastoma.

2. Molecular Biology and Genetics of Oesophageal Cancer

Collaborating Parties:
Prof Iqbal Parker - University of Cape Town
Nature and purpose: Provide bioinformatics expertise and PhD student co-supervision.
Output in the last 12 months: Whole genome sequencing analysis on paired tumour-normal samples from 60 participants. Prepare and publish a manuscript.
Future Direction: Wet lab work for validation of the identified somatic mutations.

3. Effect of physical training in miRNA expression in gluteal adipose tissue in women with obesity

Collaborating Parties:
Dr Carmen Pheiffer - South African Medical Research Council
Nasr Eshibona - SANBI UWC
Nature and purpose: Provide bioinformatics expertise.
Output in the last 12 months: Manuscript submitted to an international journal with a high impact factor.

RESEARCH PROJECT THEMES

Pathogen resistance
- NGS sequencing of Tuberculosis isolates and HIV plasma samples allowed the identification of novel mutations.
  To identify mutations associated with drug resistance we develop structural computational pipelines to further our understanding of HIV-1 and tuberculosis drug resistance.
- To identify drug resistance pathways in HIV-1 using Phylogenetics

Drug discovery
- Previous pathway mapping and gene prioritisation methods allowed us to identify known and novel *Mycobacterium tuberculosis* drug targets. Here we exploit the drug targets using virtual screening methods to identify novel drugs to treat drug resistant tuberculosis. The drugs identified are tested for activity against *Mycobacterium tuberculosis* using a bioluminescent reporter assay system. *In silico* discovery of compounds with activity against the novel SARS-coronavirus and *in vitro* testing.

Human disease genomics
- Whole genome sequencing identified variants that co- segregate within South African families with Parkinson's disease. Here we use various software tools to prioritise these variants to identify causal variants that might be associated with Parkinson's disease.

Machine learning approaches
- Using machine learning methods to identify putative compounds with activity against *Mycobacterium tuberculosis* and HIV.

Diagnostic tool development
- Develop structural methods to aid in the identification of biomarkers useful for Tuberculosis and Ebola virus diagnosis.

DR RUBEN CLOETE

HIGHLIGHTS OF 2023

- Currently, I supervise three PhD students, Mohammed Hassan, Jeremy Burgess and Adetutu Akinnuwesi and I co-supervise a PhD student, Dewald Schoeman, Medical Biosciences, UWC.
- My UWC Masters student, Keaghan Brown, graduated in 2023.
- **Grants/Funding Awarded:** Co-Principal investigator on a SAMRC RCDI grant with Prof Megan Shaw for 2021-2024.
- Attended the 2023 Fulbright Pre-Departure Orientation in Pretoria from June 8-9.
- Presented a seminar at MBS Department, UWC Computational structure-based methods to identify novel drugs and cheminformatics techniques to prioritise drugs to treat drug resistant forms of *Mycobacterium tuberculosis*,18 May 2023.

RESEARCH PROJECTS

My primary research interests focus on understanding drug resistance in *Mycobacterium tuberculosis* and Human Immunodeficiency Virus (HIV-1) integrase protein and the identification of causal variants in Parkinson's disease development. Tuberculosis and HIV-1 drug resistance is a huge problem in South Africa and calls for the identification of newer drugs to curb the spread of these diseases. The use of computational methods to speed up the process of drug discovery can reduce the cost and time spent pursuing drugs that later fail in clinical trials.

We have developed computational pipelines within my laboratory to interrogate drug targets to screen for drugs that inhibit enzyme targets that can be purchased and tested *in vitro* for activity against *Mycobacterium tuberculosis*. We also use computational methods to validate mutations and their effects on protein structure and function. This is also very useful to determine if a drug remains within an enzyme active pocket and if the mutation results in reduced binding or no binding. The result of this research may provide improve treatment regimens for South African populations to combat infectious disease and non-communicable diseases like Parkinson's. Recently, my group also embarked on identifying novel drugs to treat SARS-CoV-2 coronavirus infections using computational methods. Other collaborative projects involve studying drug transporter proteins associated with Diabetes Mellitus and investigating structural differences between human Coronavirus envelope proteins to understand pathogenicity between virulent and non-virulent strains.

RESEARCH COLLABORATIONS

1. Novel drug discovery against *Mycobacterium tuberculosis*

Collaborating Parties:

Prof Samuel Egieyeh – School of Pharmacy, UWC
Prof Samantha Samson and Dr Melanie Grobbelaar - Stellenbosch University
Prof Andrej Sali - UCSF, USA
Prof John Irwin - UCSF, USA

Nature and purpose:

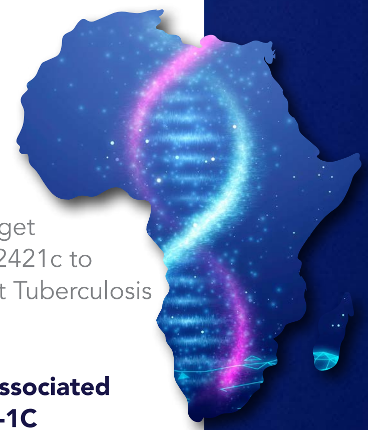
To identify drugs with a new mode of action against *Mycobacterium tuberculosis* protein targets.

Output in the last 12 months:

Akinnuwesi, A., Cloete, R. and Egieyeh, S.A., State-of-the-art Strategies to Prioritise *Mycobacterium tuberculosis* Drug Targets for Drug Discovery using a Subtractive Genomics Approach. Frontiers in Drug Discovery, 3, p.1254656. (2023).

Future Direction:

Draft two manuscripts one targeting *Mycobacterium tuberculosis* drug target Rv2196 and a follow up article on Rv2421c to identify novel drug molecules to treat Tuberculosis infections.



2. Structural impact of resistance associated mutations in the South African HIV-1C integrase protein

Collaborating Parties:

The late Dr Graeme Jacobs - Stellenbosch University (deceased)

Nature and purpose:

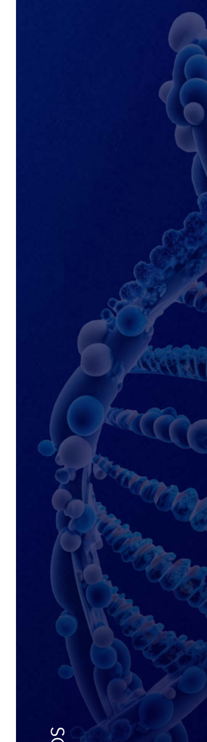
Firstly, to understand genetic diversity in HIV-1 subtype C integrase gene in South African HIV-1 infected patients and recombinant subtype AG in Cameroonian patients. Secondly, to determine if second-line integrase inhibitors will be a viable option for South African and Cameroonian patients infected with HIV-1.

Output in the last 12 months:

None

Future Direction:

One manuscript in preparation focusing on the development of an automated pipeline to investigate the effects of mutations on HIV-1C and HIV-1 AG recombinant Integrase structure.



3. Prioritising mutations identified in South African Parkinson's disease patients using structural methods

Collaborating Parties:

Prof Soraya Bardien - Stellenbosch University

Nature and purpose:

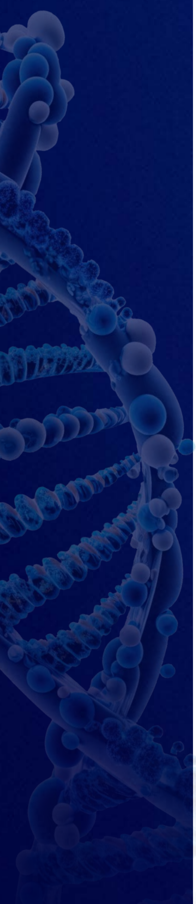
To identify novel genes associated with Parkinson's disease development using Whole Exome sequencing and using structural computational methods to understand the impact of mutations on protein structure and function.

Output in the last 12 months:

Cuttler, K., de Swardt, D., Engelbrecht, L. *et al.* Neurexin 2 p.G849D variant, implicated in Parkinson's disease, increases reactive oxygen species, and reduces cell viability and mitochondrial membrane potential in SH-SY5Y cells. *J Neural Transm* 129, 1435–1446 (2022). **(https://doi.org/10.1007/s00702-022-02548-8)**

Future Direction:

Two manuscripts under preparation for publication following up on neurexin and neuroligin interaction associated with PD and another on a novel gene candidate identified in a South African family with PD using exome sequencing.



4. Haplotype variation within South African Xhosa population and its effect on diabetic treatment

Collaborating Parties:

Prof Mongi Benjeddou - Biotechnology Department, UWC

Nature and purpose:

Understand the effect of haplotype variation on SLCAA2 transporter protein and the binding of diabetic drugs to SLCAA2 within a South African Xhosa population

Output in the last 12 months:

None

Future Direction:

None to report.

5. Structural studies of more and less virulent coronavirus envelope proteins to understand human host interaction and severity of disease

Collaborating Parties:

Prof Burtram Fielding - Department of Medical Biosciences, UWC

Nature and purpose:

Perform topology predictions, structural modelling as well as simulation studies to understand the structural differences between four coronavirus envelope proteins and their interaction with the human host PALS1 protein.

Output in the last 12 months:

None

Future Direction:

Manuscript in preparation.

6. The identification and characterisation of DNA aptamers for application in diagnosis of infectious diseases

Collaborating Parties:

Prof Mervin Meyer - Biotechnology Department, UWC

Nature and purpose:

The aim of the research is to develop multiplex lateral flow devices (LFDs) for the detection of serum human biomarker proteins for TB and Ebola diagnosis using DNA aptamers.

Output in the last 12 months:

Martin, D.R., Mutombwera, A.T., Madiehe, A.M., Onani, M.O., Meyer, M. and Cloete, R., 2024. Molecular modeling and simulation studies of SELEX-derived high-affinity DNA aptamers to the Ebola virus nucleoprotein. *Journal of Biomolecular Structure and Dynamics*, pp.1-18.

Future Direction:

To build an mysql database containing 3D structures of DNA aptamers for diagnostic purposes.

7. Ebola virus 3D database (EBOV-3D): Understanding the Ebola virus proteome and evaluating possible drug targets

Collaborating Parties:

Dr Arun Pandurangan - Department of Medicine, University of Cambridge.

Prof Placide Mbala- Kingebeni - National Institute of Biomedical Research, Democratic Republic of the Congo, University of Kinshasa.

Prof/Sir Tom Blundell - Department of Medicine, University of Cambridge.

Dr Hocine Bendou - Division of Computational Biology, University of Cape Town.

Dr Adetayo Emmanuel Obasa - Faculty of Medicine and Health Sciences, Stellenbosch University

Nature and purpose:

The aim of this project is to provide a user-friendly and easily accessible 3D webserver, so that end-users can navigate, inspect and download the 3D structural proteome data, visualise modelled oligomeric complexes, analyse pockets of modelled structures, and investigate Ebola virus human protein interactions, mutations and protein-ligand docking interfaces.

Output in the last 12 months:

None

Future Direction:

Build the Neo4j database and prepare a manuscript.

HIGHLIGHTS OF 2023

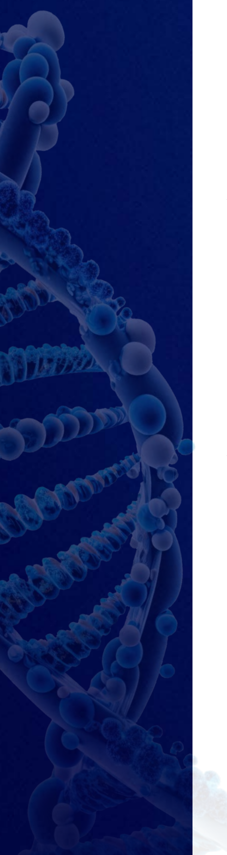
Appointed as Director of the Centre of Excellence (CoE) in HIV Prevention at UWC that is funded by the South African Department of Science and Innovation (DSI) and National Research Foundation (NRF).

South African Medical Research Council (SAMRC) funding to predict and monitor T cell immune escape mutations at HLA-binding anchor motifs found in SARS-CoV-2 epitopes from genomes isolated in South Africa and sub-Saharan Africa. PI: Nobubelo Ngandu.

PROF GORDON HARKINS

My research focuses on the evolution and molecular epidemiology of DNA and RNA infectious viral pathogens where I seek to determine the evolutionary underpinnings of the emergence and spread of the numerous viral diseases that seriously threaten the health and food security of Africa and the rest of the developing world.

In early 2020, in response to the emergence and rapid early spread of COVID-19, we immediately shifted our research focus to tackle the unprecedented volume of SARS-CoV-2 genomic and spatial data that was being generated and shared with the scientific community to gain real-time insights into the virus transmission during a viral pandemic.



RESEARCH THEMES AND PROJECTS

A summary of some of the research projects that my laboratory has been involved in since 2020 is presented below.

SARS-CoV-2

The characterisation and clinical manifestation of the SARS-CoV-2 outbreak in Uruguay

Although South America was mostly spared in the early months of the SARS-CoV-2 pandemic, it was severely hit with the arrival of the autumn season in the Southern hemisphere where the virus continued to ravage Latin America. In contrast, Uruguay, a small country located south of Brazil, has become known for curtailing SARS-CoV-2 exceptionally well. In order to attain a better understanding of disease patterns and regional aspects of the pandemic in Latin America we combined spatial and genetic analyses to infer the global historical dispersal dynamics of the causative virus SARS-CoV-2. We showed that most virus introductions into Uruguay originated from other South American countries, with the earliest seeding of the virus occurring weeks before the borders were closed to all non-citizens and a partial lockdown implemented.

The evolutionary dynamics of SARS-CoV-2 lineages during the first epidemic wave in New York City.

During the first phase of the COVID-19 epidemic in the United States, New York City rapidly became an epicentre of the pandemic. In order to gain insights on the dispersal history and transmission dynamics of SARS-CoV-2 during the first months of the New York City epidemic, my collaborators and analysed the genomic and spatial data using a Bayesian phylogeographic approach to infer the dispersal history and dynamics of viral lineages at the state and city level. Our finding showed that peripheral samples likely correspond to distinct dispersal events originating from the central city area and that the borough of Queens was a relatively important transmission hub associated with higher local circulation and dispersal of viral lineages to the surrounding boroughs.

The role of natural selection in the emergence and ongoing evolution of the N501Y lineages (V1-V3) of SARS-CoV2.

The independent emergence and rapid rise in prevalence of three SARS-CoV-2 "501Y" lineages since October 2020 has raised renewed concerns about the evolutionary capacity of SARS-CoV-2 to adapt to both public health interventions such as vaccines and social distancing and, rising population immunity. In collaboration with researchers from the South African SARS-CoV-2 genomics task force we analysed the genomic and spatial data for the three existing 501Y lineages and found that they had likely independently acquired multiple unique and convergent mutations and that a major change in the selective forces acting on immunologically important SARS-CoV-2 lineages coincided with the emergence of the 501Y lineages.

Human immunodeficiency virus (HIV-1)

Investigation of the viral dynamics involved in HIV-1 subtype C latent reservoir formation, maintenance and evolution

The HIV-1 latent reservoir is established during acute infection and viral variants from both early infection time points as well as time points proximal to ART initiation can be identified in the cellular long-lived reservoir. However, it is unknown if different distinct compartments in the body contribute to the latent reservoir in different ways and to what extent the reservoir contributes to viral evolution during viremia by reseeding of infection. We are currently generating the necessary data to elucidate the contribution of viral variants from samples collected from the blood and cervix to the long-lived latent reservoir in a longitudinal sample of subtype C infected women from the CAPRISA 002 cohort. Our analyses will provide estimates of the timing of the establishment of the reservoir and the most probable source along with the dates, magnitude, and direction of the independent movements between the different tissue compartments within each individual over the course of infection.

Reservoir size determination and timing of entry of HIV-1 variants into the latent reservoir

Understanding the key determinants of HIV-1 latent reservoir establishment, size and maintenance is imperative to designing appropriate HIV cure interventions. We are making use of longitudinal next-generation sequence data from a population of subtype C infected women from the CAPRISA 002 cohort to elucidate viral evolution patterns and generate a pre-treatment evolutionary timeline with which to date viruses isolated from the long-lived latent reservoir. These analyses will aid in better understanding the establishment of viral reservoirs in individuals who initiate therapy. The proposed work will contribute to a larger study investigating reservoir dynamics that will also include measurement of reservoir size and viral promoter function and *nef* gene function.

Phytovirus Diversity

Emergent diseases of plants, a high proportion of which are caused by phytoviruses, are a significant burden on the food security and economic stability of society. However, a comprehensive view of the geographical distribution of phytovirus diversity does not exist, including both the numbers or richness of virus species and the evenness of their distribution in any individual environment on Earth. In collaboration with researchers from CIRAD France, the State University of Arizona and UCT we received funding from the French National Research Agency (ANR) in 2019 for a project that had three primary objectives
(i) test whether plant species richness influences phytovirus species richness in natural and cultivated areas (ii) experimentally study the effect of plant communities on phytoviral species richness
(iii) search for evolutionary footprints associated with emergence within phytovirus genomes.

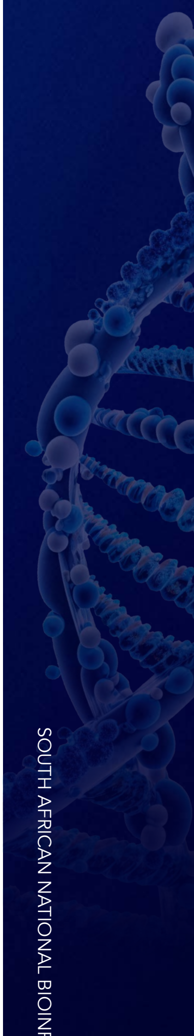
RESEARCH COLLABORATIONS

1. HIV-1 Latent Viral Reservoir Dynamics

Collaborating Parties:
Melissa-Rose Abrahams, Lynn Tyers, David Matten, Deelan Doolabh, Colin Anthony, Carolyn Williamson - Division of Medical Virology, Institute of Infectious Disease and Molecular Medicine, University of Cape Town
Salim Abdool Karim - Centre for the AIDS Programme of Research in South Africa, University of KwaZulu-Natal
Andrew Redd - Johns Hopkins University, School of Medicine in Baltimore, USA.
Siposethu Matzishana, Nkosazama Nyembezi - SANBI
Nature and purpose:
To study the viral dynamics involved in HIV-1 subtype C latent reservoir formation maintenance and evolution to better understand the determinants thereof.
Output in the last 12 months:
None.
Future Direction:
This project is funded by the National Institutes of Health (NIH) USA and the South African Medical Research Council and will run between 2020 and 2025.

2. Phytovirus Diversity

Collaborating Parties:
Philippe Roumagnac, Denis Filloux, Charlotte Julian and Emmanuel Fernandez - CIRAD, Montpelier, France
Pierre Lefeuvre, Frederic Chiroleux and Jean-Michel Lett - CIRAD, UMR PVBMT, Réunion Island, France
Thierry Candresse, Armelle Marais, Marie Lefebvre and Chantal Faure - INRA, UMR BFP, Villenave d'Ornon France
Darren Martin – UCT, South Africa
Arvind Varsani - Arizona State University, USA
Pascal Gentit and Benoir Remenant - ANSES France
Francois Mesleard and Hugo Fontes - La Tour du Valat, France
Christophe Levergne - CBN CPIE Mascarin.
Anne Duputie and Francois Massol - UMR Evo-Eco-Paleo, France
Nature and purpose:
While it is well established that emerging viruses generally originate in uncultivated hosts with which they have well-established interactions, we currently know almost nothing about the diversity and the distribution of phytoviruses circulating in either natural or managed ecosystems. This project proposes (i) to develop an integrated approach to reveal the phytoviral species richness of plant communities within several unmanaged and managed ecosystems, (ii) to experimentally study the effect of controlled plant communities on phytovirus species richness, and (iii) to search phytovirus nucleotide sequence data *in silico* evidence of the evolutionary footprints adaptation to a new host species.
Output in the last 12 months:
Plant samples were first collected in South Africa in 2019 but the field sampling trips planned for 2020 and 2021 was postponed until October 2022 and scheduled annually thereafter until 2025, pandemic conditions permitting. Efforts to detect phytoviruses in our plant samples are currently being performed in the laboratory of Philippe Roumagnac in Montpellier France.
Future Direction:
This is an ongoing project that is funded until 2023.



Dr Uljana Hesse 3rd from right with members of the Rooibos Genomics Programme.

DR ULJANA HESSE

South Africa is home to an exceptional biological resource - the unique flora of the Cape Floristic Region includes over 3000 plant species actively used in traditional medicine. These plants produce a wide range of medicinally active compounds, many of which are rare or even species specific. One of the more famous endemic South African medicinal plant species is rooibos, best known as a herbal tea. It only grows in the Cederberg Mountain Region and represents a key cash crop for the local farming communities. Rooibos is increasingly recognised as a potential phytopharmaceutical: the species produces a wide range of phenolic compounds, which are associated with diverse medicinal properties of the plant (including anti-diabetic, cardioprotective, antispasmodic and anti-aging effects). It therefore represents a suitable pilot plant species to initiate medicinal plant genomics research that focuses on the endemic flora of South Africa.

The Rooibos Genomics Programme aims to

- locally establish all methodologies essential for medicinal plant genome analysis,
- generate a high-quality assembly of the rooibos genome with extensive functional annotation of the rooibos genes,
- identify rooibos genes involved in medicinal compound production and plant stress tolerance, and
- develop machine learning tools that can be trained to mine plant genomes for diverse genomic features (specific protein families, gene clusters, non-coding RNA, etc.).

COLLABORATION WITH SANBI:

I have a longstanding partnership with SANBI. Since the initiation of the Rooibos Genomics Programme, all postgraduate students of my research team have received training in bioinformatics through the 1-month course offered by SANBI. Establishment of biocomputational data analysis procedures is conducted or initiated at SANBI.

PROGRESS 2023:

1. Published method for MinION sequencing and data analysis used to generate a high quality assembly of the 1.25Gbp rooibos genome.
2. Evaluated methods for repeat masking and completed repeat masking of the rooibos genome.
3. Optimised 3rd generation sequencing procedures for the analysis of plant transcriptomes using MinION from Oxford Nanopore.

4 Sequenced 12 transcriptomes from commercial and wild rooibos plants using MinION.

5. Established methods for gene prediction using hints from short and long read rooibos transcriptome data.
6. Established a genome browser, JBrowse, for visualisation of the rooibos genome.
7. Established a collection of leaf samples from 105 rooibos ecotypes for population studies (9 rooibos growth types; 1-3 populations per growth type; 5 plants per population; 3 harvests: Feb 2022, Oct 2022, Feb 2023) which will permit comparative gene expression analyses between different growth types, years and before/after summer to identify genes associated with different morphological characteristics, biochemical profiles and with drought/heat stress tolerance.
8. Analysed all plant samples from the Feb 2022 harvest using HPLC to identify concentrations of 12 different rooibos phenolic compounds.
9. Established a greenhouse experiment with rooibos seedlings for comparative transcriptomics and proteomics associated with drought stress.
10. Established and evaluated a Convolutional Neural Networks foundation model for identification and classification of novel plant proteins to predict their functions.

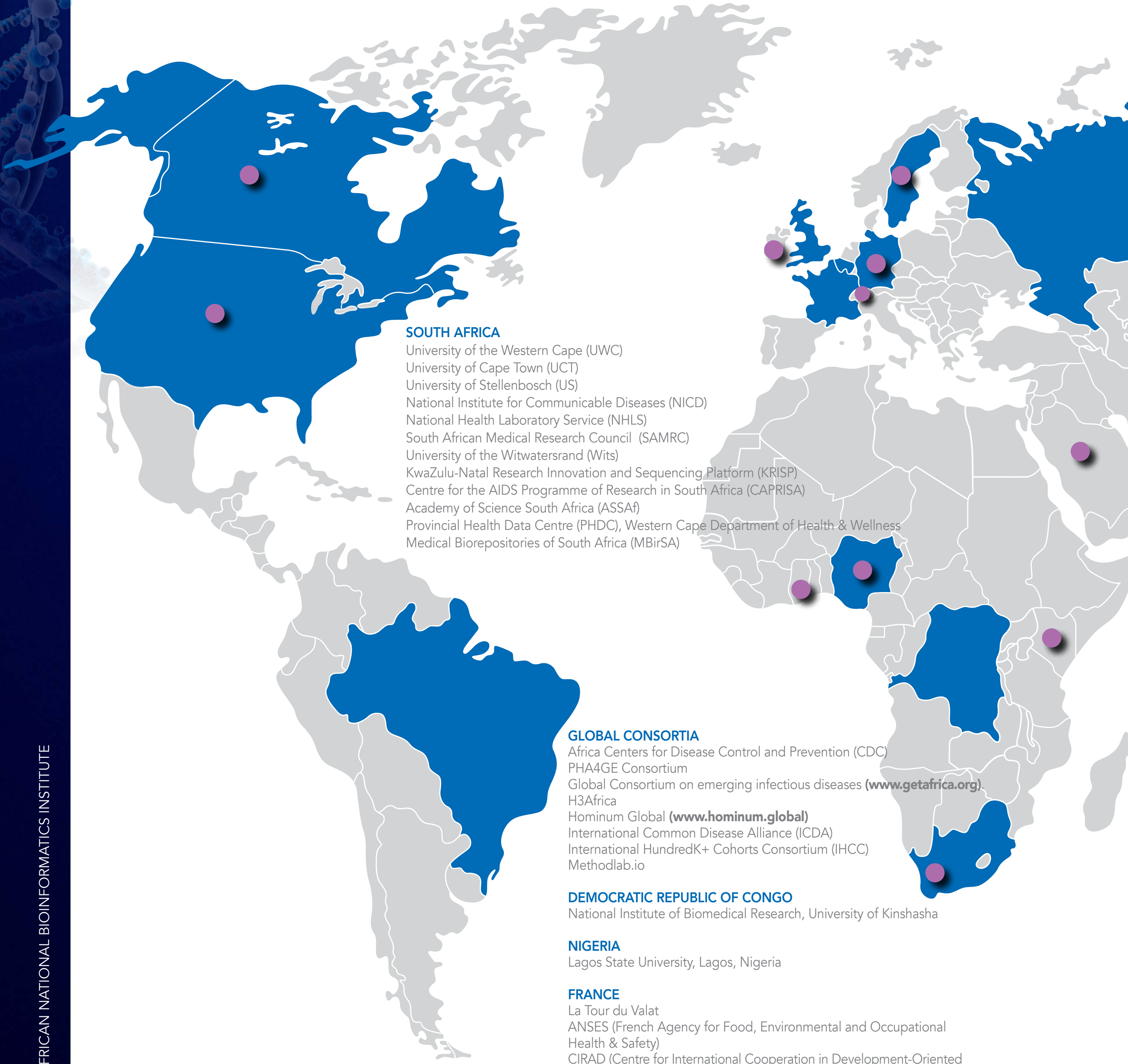
PUBLICATIONS AND PRESENTATIONS:

Hesse, U. (2023) “K-Mer-Based Genome Size Estimation in Theory and Practice”. In Plant Cytogenetics and Cytogenomics: Methods and Protocols (pp. 79-113). New York, NY: Springer US.

Hesse U. “The Rooibos Genome Programme: computational requirements” at the Centre for High Performance Computing 2023 National Conference; 4-7 December 2023; Nombolo Mdhluli Conference Centre (NMCC), Skukuza, South Africa.

COLLABORATIONS

SANBI researchers have established a vast network of partnerships and collaborations with peers all over the world (indicated in blue) in the map below:



SOUTH AFRICA

University of the Western Cape (UWC)
University of Cape Town (UCT)
University of Stellenbosch (US)
National Institute for Communicable Diseases (NICD)
National Health Laboratory Service (NHLS)
South African Medical Research Council (SAMRC)
University of the Witwatersrand (Wits)
KwaZulu-Natal Research Innovation and Sequencing Platform (KRISP)
Centre for the AIDS Programme of Research in South Africa (CAPRISA)
Academy of Science South Africa (ASSAf)
Provincial Health Data Centre (PHDC), Western Cape Department of Health & Wellness
Medical Biorepositories of South Africa (MBirSA)

GLOBAL CONSORTIA

Africa Centers for Disease Control and Prevention (CDC)
PHA4GE Consortium
Global Consortium on emerging infectious diseases **(www.getafrica.org)**
H3Africa
Hominum Global **(www.hominum.global)**
International Common Disease Alliance (ICDA)
International HundredK+ Cohorts Consortium (IHCC)
Methodlab.io

DEMOCRATIC REPUBLIC OF CONGO

National Institute of Biomedical Research, University of Kinshasha

NIGERIA

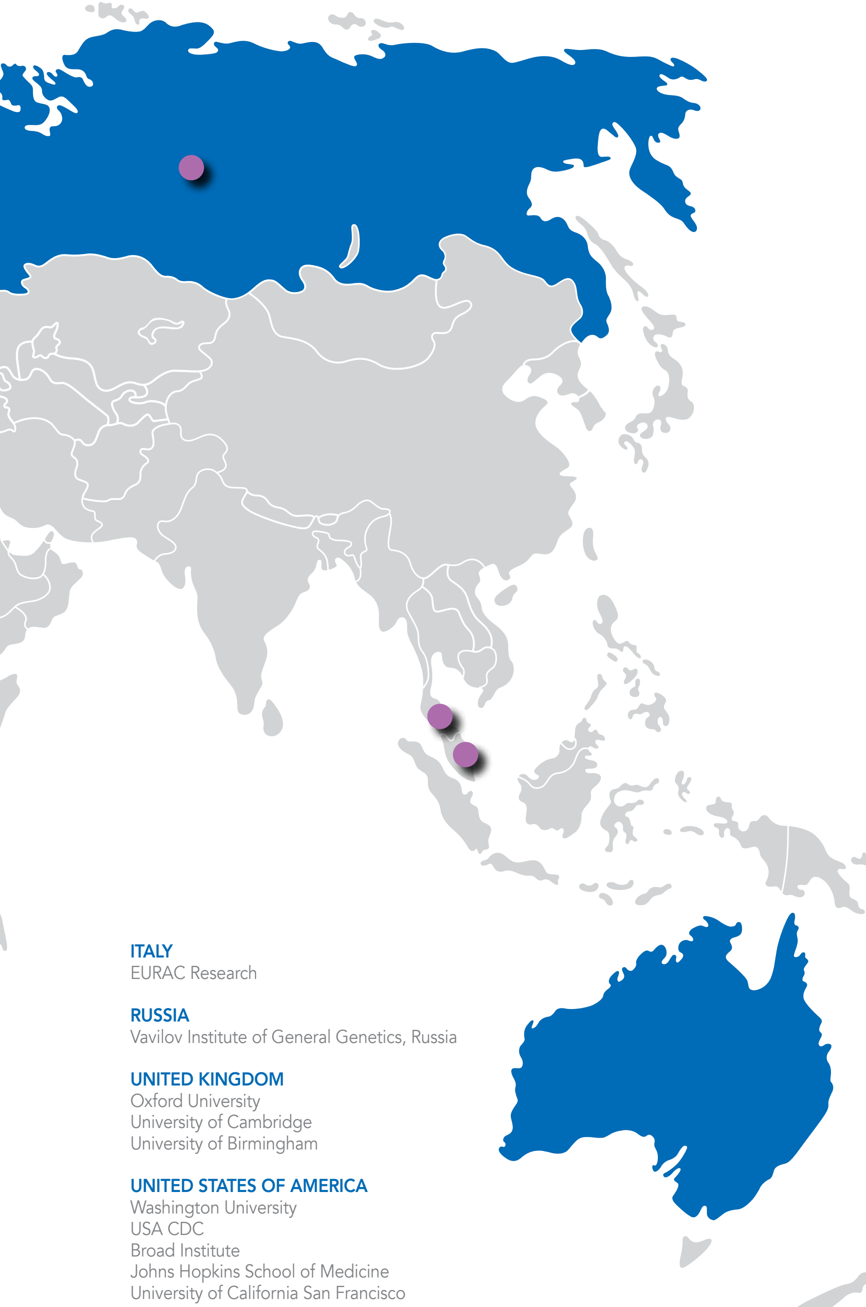
Lagos State University, Lagos, Nigeria

FRANCE

La Tour du Valat
ANSES (French Agency for Food, Environmental and Occupational Health & Safety)
CIRAD (Centre for International Cooperation in Development-Oriented Agronomical Research), Montpelier
CIRAD, UMR PVBMT (Plant Populations and Bio-aggressors in Tropical Ecosystems Joint Research Unit), Réunion Island
INRA (French National Agronomical Research Institute), UMR BFP (Fruit Biology and Pathology), University of Bordeaux
CBN CPIE (National Botanical Conservatory Permanent Centre for Environmental Studies), Mascarin.
UMR Evo-Eco-Paleo, University of Lille

ALUMNI

SANBI has produced many alumni over the years who have taken their research to institutions and corporates all over the world (indicated in purple in the map below).



ITALY
EURAC Research

RUSSIA
Vavilov Institute of General Genetics, Russia

UNITED KINGDOM
Oxford University
University of Cambridge
University of Birmingham

UNITED STATES OF AMERICA
Washington University
USA CDC
Broad Institute
Johns Hopkins School of Medicine
University of California San Francisco
Arizona State University

CANADA
University of British Columbia
BC Centre for Disease Control

BRAZIL
Centro de Integração de Dados e Conhecimentos para Saúde (CIDACS)

AUSTRALIA
University of Melbourne

SOUTH AFRICA
University of the Western Cape
University of KwaZulu-Natal
University of Cape Town
University of Stellenbosch
University of Pretoria
University of South Africa
University of the Free State
University of the Witwatersrand
SA Medical Research Council
National Institute of Communicable Diseases
Centre for Proteomic and Genomic Research
Centre for High Performance Computing
H3ABioNet

INDUSTRY
- Hyrax Biosciences
- IBM
- Roche
- B&M Scientific
- GENEdiagnostics
- City of Cape Town
- World Bank
- SA Government Administration
- KappaBiosystems
- Takealot
- Konga.com
- Monash/ The Independent Institute of Education MSA
- Entersekt

KENYA
Egerton University, Njoro
International Livestock Research Institute

GHANA
University of Ghana

NIGERIA
Lagos State University
University of Nairobi

SAUDI ARABIA
King Abdullah University of Science and Technology

GERMANY
Institute for Inorganic Chemistry, RWTH Aachen University
Max-Planck Institute for Evolutionary Anthropology, Leipzig

UNITED STATES OF AMERICA
Harvard School of Public Health
Princeton University
Washington University
Jackson State University
University of California, San Diego
Wayne State University, Detroit, Michigan
Pacific Northwest Diabetes Research Institute, Washington
Beth Israel Deaconess Medical Centre, Harvard Medical School
Johns Hopkins University

CANADA
University of Western Ontario
University of British Columbia

IRELAND
National University of Ireland, Galway

SWEDEN
Karolinska Institute

SWITZERLAND
Syngenta Crop Protein AG, Basel

RUSSIA
Institute of Cytology and Genetics, Novosibirsk

SINGAPORE
Nanyang Technology University
University of Singapore

MALAYSIA
Novocraft Technologies Research

FINANCIALS

SANBI's income and expenditure trends for 2023 are shown in this section.

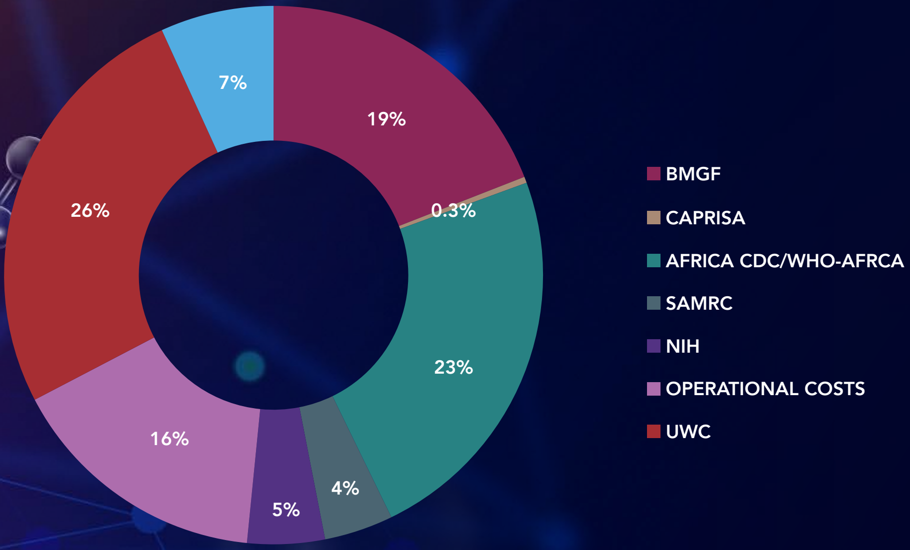


FIGURE 1. Distribution of income received from all sources 2023.

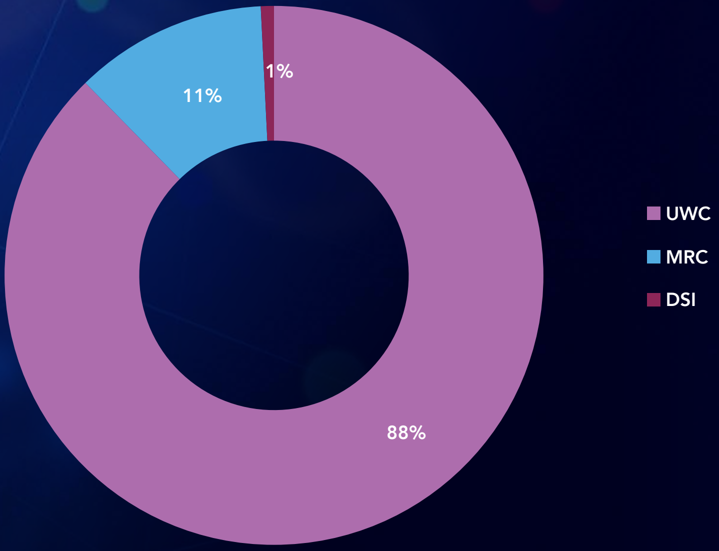


FIGURE 2. Distribution of income received from SA sources 2023.

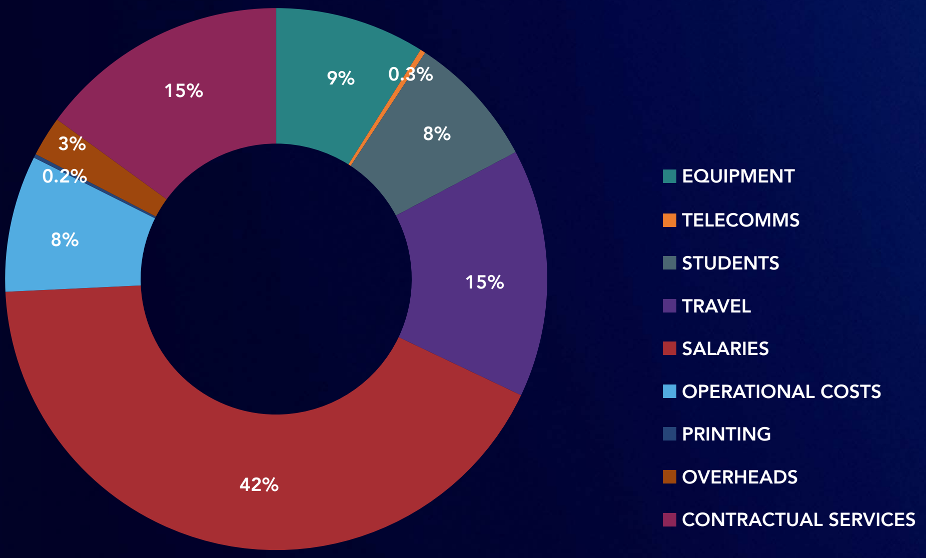


FIGURE 3. Expenditure in 2023.

FIGURE 4. Income vs expenditure 2013 – 2023.

Note: Funding received in a calendar year often reflects projects that span multiple years.

FUNDERS

SANBI
South African National
Bioinformatics Institute

UNIVERSITY of the
WESTERN CAPE

Research Institute at UWC since 1997

NRF
National
Research
Foundation

National Research Foundation funding since 1998

saMRC
advancinglife

South African Medical Research Council Bioinformatics Unit since 2002

science & innovation
Department:
Science and Innovation
REPUBLIC OF SOUTH AFRICA

Department of Science and Innovation National Research Foundation Research Chair in Bioinformatics and Public Health Genomics since 2007

BILL & MELINDA GATES foundation

Bill & Melinda Gates Foundation funding since 2014

UKRI
Medical
Research
Council

UK Research Innovation/ Medical Research Council (UKRI/MRC) funding since 2023

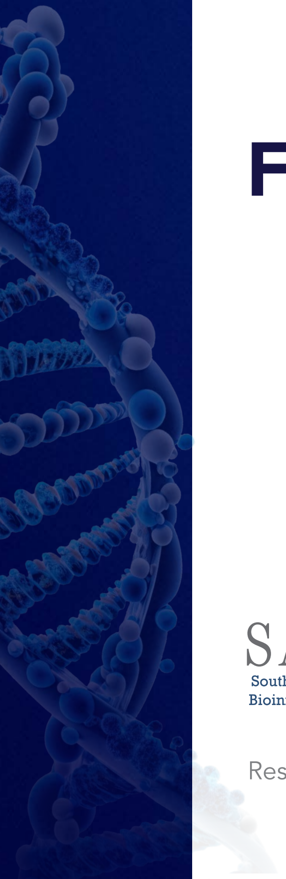
Funding since 2021:

World Health Organization

World Health Organization
REGIONAL OFFICE FOR Africa

African Union

AFRICA CDC
Centres for Disease Control and Prevention
Safeguarding Africa's Health



SANBI
South African National
Bioinformatics Institute

POSTAL ADDRESS:
South African National Bioinformatics Institute | University of the Western Cape
Private Bag X17 | Bellville 7535

PHYSICAL ADDRESS:
South African National Bioinformatics Institute | 5th Floor Life Sciences Building
University of the Western Cape | Robert Sobukwe Road | Bellville 7530

TELEPHONE: +27 (0)21 959 3645 | **EMAIL:** info@sanbi.ac.za | **WEBSITE:** www.sanbi.ac.za | **TWITTER:** @SANBI_SA