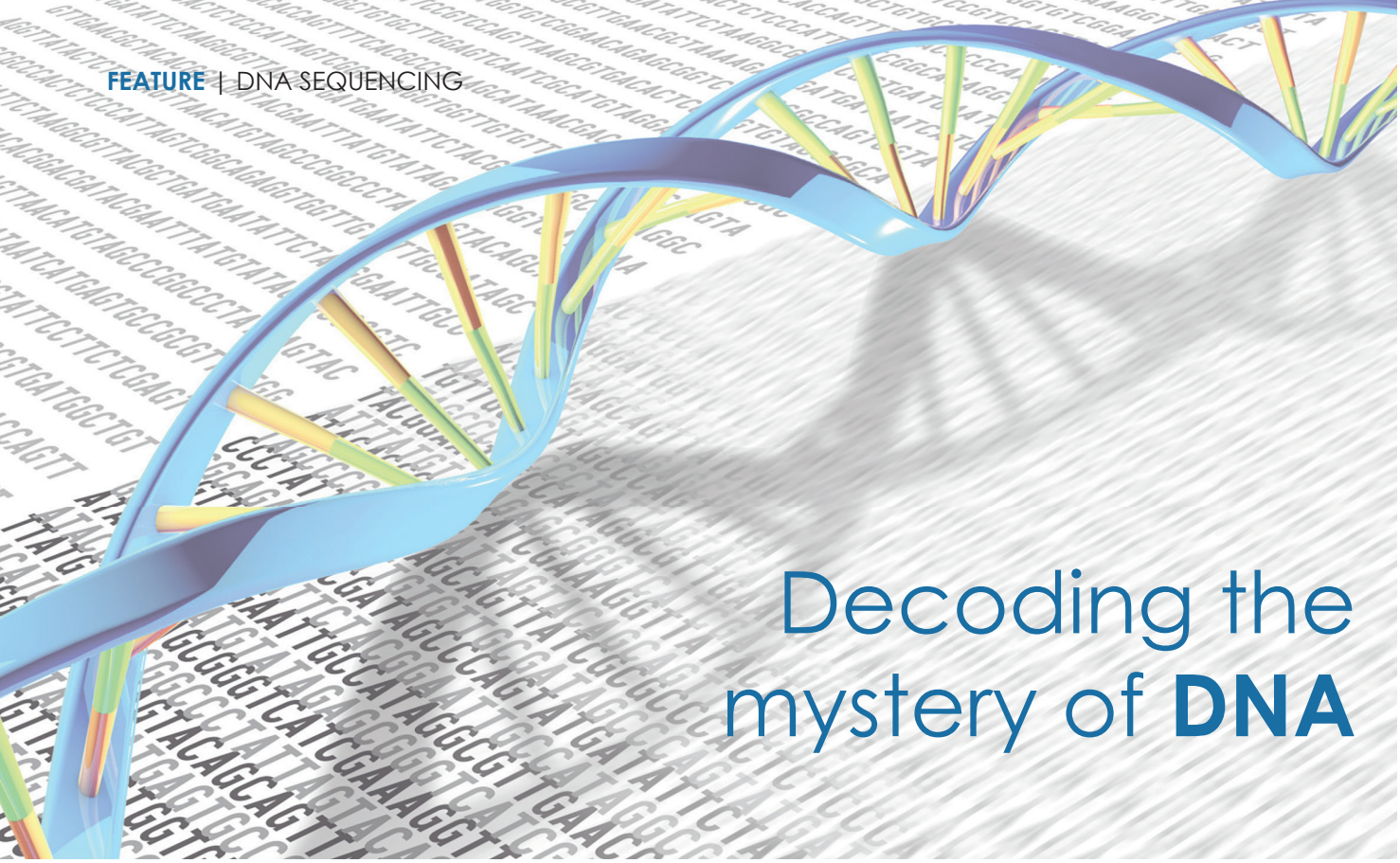


Decoding the mystery of **DNA**

Darryl Leja, NHGRI

By René Veikondis and Alvera Vorster

At the start of this century, on 26th June 2000, the completion of the first draft of the human genome sequence was announced by President Bill Clinton of the United States and Prime Minister Tony Blair of the United Kingdom. This announcement ushered in an era that has been an exciting time to be a molecular biologist. The first human genome, consisting of three billion nucleotide base-pairs, was sequenced by researchers from six countries over a period of 13 years and at a cost of US$3 billion. Today, we have the technology to read the entire genetic code of almost any living organism on Earth, in less than a week and at a cost of no more than US$1 000.

The DNA Sequencing Unit at Stellenbosch University's Central Analytical Facilities (CAF) makes use of the rapid DNA sequencing method that was used to sequence the first human genome. The method, known as Sanger dideoxy sequencing, was introduced in 1975 to sequence the human mitochondrial genome, and this work was published in 1977 by a group of scientists in Frederick Sanger's laboratory. For this effort Frederick Sanger – a British scientist – received his second Nobel Prize in chemistry, which he shared with two American scientists in 1980.

Though the Sanger method is still considered the most accurate way to sequence a single DNA fragment, next-generation sequencing methods are the most cost-effective to read the entire genetic code of an organism.

At the CAF we offer massively parallel sequencing on the Ion Torrent platforms. With this method, a genome is broken up into millions of fragments, which are all sequenced at the same time. The data that is produced in parallel from all the fragments are pieced together by high-performance computing clusters, with the aim of delivering a single piece of DNA that encodes the instructions for the organism's life.

With these methods the staff at the CAF provide an affordable, cutting-edge sequencing service, delivering the highest amount of quality data to South African researchers in the shortest possible time frame. The

Staff at the DNA Sequencing Unit at Stellenbosch University's Central Analytical Facilities celebrated the 21st anniversary of its inception in 2018. Back row: René Veikondis, Carel van Heerden, Alvera Vorster. Front row: Sinead Robberts, Annette Laten, Marianna Retief.

success of this service relies on three stakeholder groups: the Department of Science and Innovation, which funds our sequencing instruments, the manufacturers and distributors of sequencing reagents, and the researchers, who hail primarily from the agricultural and medical fraternities. Here we celebrate the researchers who support DNA sequencing in the local context.

Mitochondrial DNA (mtDNA) sequencing

While it took Frederick Sanger's laboratory more than two years to determine the genetic code for one human mitochondrial genome, the advances in sequencing methodologies over the past two decades have enabled us to produce the DNA sequence for 59 of these genomes in less than two weeks on the Ion Torrent S5 next-generation sequencing platform.

The genetic determinants of Parkinson's disease

The sequencing data mentioned above forms part of the research that a doctoral candidate, Amica Muller-Nedebock, is conducting at Tygerberg Hospital. Her research is focused on South Africans with Parkinson's disease, an incurable movement disorder characterised by the loss of dopamine-producing and highly energy-dependant neurons.

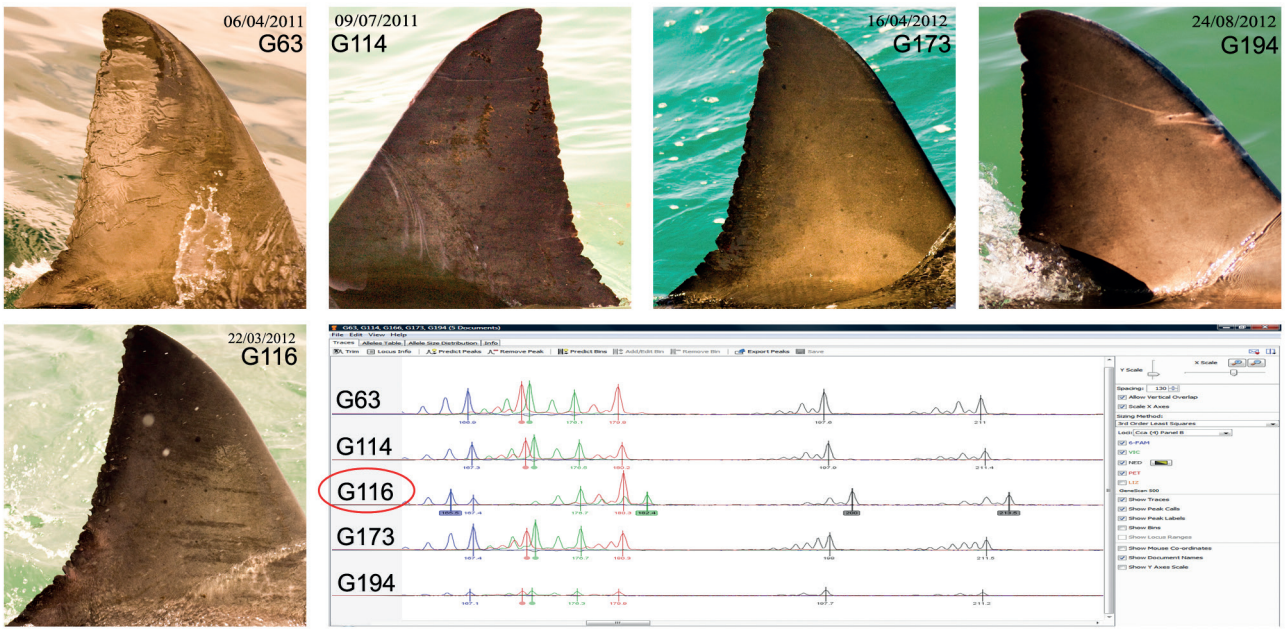
Amica explains that mitochondria are the powerhouses of cells and their genomes encode critical components that are required for energy production. She theorises that small variations in the mtDNA may account for a decrease in energy production of dopamine-producing neurons, and in turn cause deterioration of these highly energy-dependent cells. With this sequencing data, the Parkinson's disease research group at Stellenbosch University is analysing the mtDNA of individuals living with the disease to identify whether several of the mtDNA changes may collectively contribute to its development. Finding the underlying cause for this debilitating disease provides an opportunity to develop new treatment and disease-management strategies.

Molecular research on great white sharks

Great white sharks are magnificent creatures, but their elusive nature means very little is known about them. Questions exist about their lifespan, their numbers, and the location of their nurseries. To address these questions, Dr Sara Andreotti is investigating the population size of great white sharks along the South African coastline, with specific emphasis on the level of genetic diversity amongst the counted individuals. High genetic diversity confers some protection to a population in the event of a catastrophic change in the environment or disease outbreak, helping to ensure the continuity of the species.

Dr Andreotti started this study with a photographic database of dorsal fin notch patterns from more than 5 000 photographs. Analysis of this data identified 426 individuals, but to verify this photographic evidence, biopsy samples were collected from 233 sharks for genetic analysis. The molecular sizes of 14 genetic markers called microsatellites were determined at the CAF's DNA sequencing unit for all biopsy samples. The markers, which generate a unique genetic fingerprint for each great white shark, indicated a South African population of 333 individuals.

Dr Andreotti also used the Sanger sequencing method for mtDNA analysis. The results showed that the South African great white shark population has low genetic diversity, with only four maternally inherited lineages and 89% of sharks belonging to one of these, indicating significant inbreeding. This finding is concerning, as it implies that the species could become extinct in a very short period of time if faced with environmental changes or disease.



Correspondence between photo identification and genetic fingerprint of a great white shark (C_050606B) sampled on four different occasions over two years. Each genetic sample has been uniquely coded (G63, G114, G173 and G194) to allow for blind scoring of duplicates using a genetic fingerprint. The sample G116 belongs to a different individual (C_040705), as confirmed by the genetic profile and the different notch pattern on the dorsal fin (Andreotti et al. 2016, https://doi.org/10.3354/meps11744).

16S ribosomal RNA (rRNA) gene sequencing

At the turn of this century, 16S rRNA gene sequencing revolutionised bacterial isolate identification. Sections of the 16S rRNA gene are identical between bacterial species and are used as a common target to amplify the signal for all bacteria in one reaction tube. Subtle differences in the rRNA gene serve as a fingerprint, allowing us to distinguish the individual bacteria in a sample. With this approach, tedious bacterial culturing and costly biochemical identification techniques are avoided. Instead, DNA is extracted from bacteria in a sample and then sequenced to produce data that can be compared to existing public databases. One such database, GreenGenes, contains the DNA sequences serving as fingerprints for specific bacteria.

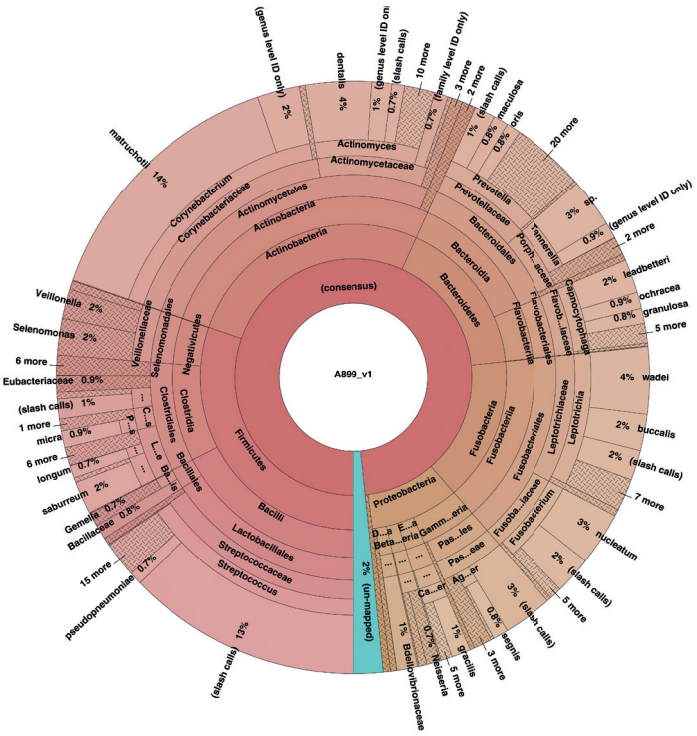
A probiotic for broiler chickens

There is a high demand for animal protein worldwide, and the poultry industry is ranked amongst the largest suppliers, but it incurs substantial losses due to bacterial and fungal infections. The main reason for this is because newly hatched chicks are not exposed to mature birds for long enough to obtain the beneficial bacteria from the gastrointestinal tract of hens. It takes two to four weeks for a chick to develop a balanced and stable gut microbiome, while a well-developed immune system takes even longer. Chicks raised in broiler houses are therefore more susceptible to infections that lead to weight loss, poor meat quality and even death.

As part of his doctoral research, Dr Deon Neveling focused his study on the development of a multi-species probiotic that can be added to feed, with the aim of improving growth performance and health. Bacteria were isolated from different sections of the gastrointestinal tract of healthy free-range broilers, identified to species level using Sanger sequencing at the CAF, and then screened for probiotic properties. Those that exhibited the desired probiotic effect were used by Dr Neveling to develop the probiotic, which was patented after being successfully tested as a feed additive.

USDA NRCS, CC BY 2.0

Sanger sequencing was used to develop a probiotic feed additive for broiler chickens.



A Korona plot representing the bacterial diversity (alpha diversity) within a single oral plaque DNA sample. The relative abundance of a bacterial species is shown as a percentage of the total sequencing data that was generated for the poly-bacterial DNA sample.

Diabetes and cardiovascular risk

In April 2019, 128 poly-bacterial DNA samples were sequenced at the CAF, using the 16S rRNA gene sequencing approach. This work forms part of Yvonne Prince's doctoral degree, which focuses on the effects of smoking and alcohol on the oral microbiome in persons with cardiometabolic risk factors. Her research is conducted under the leadership of Prof. Tandi Matsha and Prof. Glenda Davison at the South African Medical Research Council's Cardiometabolic Health Research Unit, based at Cape Peninsula University of Technology. Yvonne recognises that the incidence of diabetes mellitus, obesity and cardiovascular disease is becoming a growing problem in South Africa. Internationally, research has shown evidence of a bi-directional relationship between diabetes and oral disease – the presence and management of one influences the presence and management of the other – so individuals with diabetes are more likely to have an increased severity of oral disease.

Because the oral microbiome is such a complex system, harbouring a rich and diverse population of more than 600 microflora that includes bacteria, viruses, fungi and protozoa, it makes it extremely difficult to identify and isolate the microorganisms present. With the help of independent molecular methods such as 16S rRNA sequencing, developed to easily group and identify the uncultivable microbiota that are present, this analysis has become possible. The data generated from the sequencing was used to search the Greengenes and MicroSEQ databases for comparison to 16S sequences from closely related species. The results are listed as highest or closest relatives to the species present within the sample. Finding signatures in the bacterial diversity that are specific to persons with diabetes would enable

early diagnosis and allow opportunities to promote health and wellness in pre-symptomatic persons.

The genetics, distribution and diversity of cat fleas

Fleas are a pesky problem for pet owners. Worldwide, there are about 2 500 species or subspecies of these tiny bloodsucking insects. They cause discomfort due to itching, and can transmit tapeworms and pathogens that may cause disease in humans and pets. Dr Luther van der Mescht is a postdoctoral researcher in the evolutionary ecology of parasites at Stellenbosch University. He is interested in the genetic and morphological differences within cat-flea species, particularly those that are found on the African continent. Genetic analyses of the fleas are completed by performing Sanger sequencing of similar regions in the cat-flea genome, allowing Dr van der Mescht to draw phylogenetic trees that assist with species classification. The aim of the study is to develop an accurate classification system that can be used to develop a pest control programme. Genetic analyses will also include 16S rRNA gene sequencing to determine the diversity of the bacterial population on the various cat-flea species.

Andrei Savitsky, CC BY 4.0

Whole genome sequencing

During the 13 years that it took to complete the first draft of the human genome, one of the main challenges facing researchers was computational power. To sequence an entire genome, the DNA is broken into pieces that are less than 600bp (base-pairs) in length. The sequence of each fragment is determined individually and the resulting data is pieced together by finding identical, overlapping regions. Today the computers that can piece together this genome puzzle are cheaper, smaller and faster – allowing a laboratory like ours to complete genome assembly in weeks instead of years.

Fungal resistance genes in grapevine

The CAF is situated in the Stellenbosch winelands region, which has some of the oldest vineyards in South Africa. The market demand for *Vitis vinifera*, the most common type of grapevine, places constant pressure on plant breeders to develop varieties that have improved characteristics, such as disease resistance. Using conventional breeding for disease resistance is a long and costly process, due to the relatively long growth cycle of grapevine. However, the introduction of molecular markers has made it possible to identify genes or genomic regions related to specific disease resistance much earlier in the breeding cycle. Extensive research has identified the disease-resistance genes in known resistant varieties, as well as microsatellite markers that are linked to these genes to enable marker-assisted selection (MAS). In other words, the progeny in a breeding experiment can be screened for microsatellite markers linked to desired traits, and those that do not carry the desired genes may be eliminated from the process, significantly reducing the costs involved in plant breeding.

DNA sequencing is used to study viruses affecting grapevines in the Stellenbosch winelands.

Prof. Gerhard Petersen uses both Sanger and next-generation sequencing methods as fundamental tools in his research on plant viruses. Amongst these are plant pathogens of *V. vinifera.* The data generated from the sequencing techniques are used in diagnosis, characterisation, phylogenetic analysis, population analysis and discovery of various plant viruses that are new to South Africa, or to science.

Poultry disease research

Prof. Celia Abolnik, the Research Chair for Poultry Health and Production at the University of Pretoria, applies next-generation sequencing in the diagnosis and study of avian diseases. A diverse group of bacterial and viral pathogens cause diseases in the national chicken flock, which is important for regional food security. Pathogens that are regularly investigated through next-generation sequencing at the CAF include *Mycoplasma* species, influenza A virus, avian avulavirus type 1, infectious bronchitis virus, fowl adenovirus and infectious laryngotracheitis virus. The next-generation sequencing data for *Mycoplasma gallinarum* and *M. gallinaceum* have been assembled and published as the first complete annotated genomes of the species.

For enquiries about DNA sequencing services offered by the Central Analytical Facilities at Stellenbosch University, please contact Ms René Veikondis (renev@sun.ac.za) for Sanger sequencing services or Ms Alvera Vorster (vorster@sun.ac.za) for next-generation sequencing services.

CURRICULUM CORNER

LIFE SCIENCES: GRADE 11
DNA: the code of life; Genetics and inheritance

LIFE SCIENCES: GRADE 12
Evolution by natural selection: artificial selection