SEQUENCING AN HISTORIC BACTERIAL COLLECTION FOR THE FUTURE

SEQUENCE WITH CONFIDENCE

PACBIO®

A Mission to Preserve and Grow

The UK's National Collection of Type Cultures (NCTC) is a unique collection of more than 5,000 expertly preserved and authenticated bacterial cultures, many of historical significance. Founded in 1920, NCTC is the longest established collection of its type anywhere in the world, with a history of its own that has reflected – and contributed to – the evolution of microbiology for more than 100 years.

But the NCTC is far from stuck in the past. In fact, it has been at the forefront of advances in the field, continually implementing the latest technologies in order to provide the best, most comprehensive resources to support microbiology research institutes worldwide charged with a broad range of missions.

In the 1930s, this meant the introduction of freeze-drying strains to ensure longevity and to streamline storage and shipment. In 1949, the NCTC began a 10-year effort to characterize every organism in the collection, generating records of colony morphology, biochemical test results, and freeze-drying status. In 1965, the bench-top staple Cowan and Steel's Manual for the Identification of Bacteria, co-authored by NCTC curator Samuel Cowan, was published, and the NCTC began to develop computer-based bacterial identification methods.

> "If you're trying to generate reference genomes that are going to be valuable to as many people as possible, with as much information in them as possible, then Pacific Biosciences has the edge in terms of generating more complete data."
>
> **– Professor Julian Parkhill**
> Wellcome Sanger Institute

With the advent of the genomics era, NCTC embraced the use of DNA-based methods for species identification. During the 1980s, Curator L. R. Hill introduced GC content analysis, DNA-DNA hybridization and restriction fragment length polymorphism (RFLP) analyses for more accurate species identification. 16S ribosomal RNA gene sequence analysis was introduced in the 1990s and NCTC strains continue to be routinely identified and authenticated using this method.

Currently, new strains are characterized for morphology, nutritional requirements, enzyme activity and subjected to serotyping, mass spectrometry, 16S sequencing, and, most recently, whole genome sequencing (WGS).

NCTC 3000 Project: A Resource for Scientists Worldwide

In 2014, NCTC launched an ambitious five-year project, together with the Wellcome Sanger Institute, to generate high-quality reference genomes for 3,000 bacterial strains. The method of choice? Single Molecule, Real-Time (SMRT®) Sequencing. According to Professor Julian Parkhill of the Wellcome Sanger Institute, PacBio® technology was selected because it produces long-read genome assemblies with the highest consensus accuracy and uniform coverage.

"Whole genome sequencing is set to revolutionize medical microbiology. Genomic characterization of expertly-authenticated type and reference culture collection strains will contribute significantly to a wide range of clinical and research applications as we progress further into this new genomic era," Prof Parkhill said. "SMRT Sequencing delivers long read lengths, generating the most comprehensive *de novo* assemblies."

Julie E. Russell, Head of Culture Collections at Public Health England, added: "If NCTC is to continue to supply relevant authentic bacteria for use in scientific studies, then the quality of our own characterization and authentication data must be outstanding. Combining sequences, strain metadata and links to other resources in the public domain will ensure that this resource provides a unique comprehensive source of data to underpin microbial research and improve the provision of diagnostics and public health interventions for medically important bacteria and viruses."

Julian Parkhill is a Professor at the Wellcome Sanger Institute

A Brief History of Microbiology

PACBIO®

Our understanding of microbiology has evolved enormously over the last 150 years. Few institutions have witnessed our collective progress more closely than the National Collection of Type Cultures (NCTC). In fact, the collection itself is a record of the many milestones microbiologists have crossed, building on the discoveries of those who came before.

To date, 60% of NCTC's historic collection now has a closed, finished reference genome, thanks to PacBio Single Molecule, Real-Time (SMRT) Sequencing. We are excited to be their partner in crossing this latest milestone on their quest to improve human and animal health by understanding the microscopic world.

NCTC National Collection of Type Cultures
Operated by Public Health England

Laying the Foundation

1886: Theodore Escherich describes a bacterium which he called "bacterium coli commune" and which was later to be called *Escherichia coli*. A strain he isolated in 1886 is added to the collection upon its founding (**NCTC 86**).

1887: Julius Petri invents the agar-coated glass dish for culturing bacteria; earlier attempts at culturing involved potato slices and gelatin.

1890: German scientist Robert Koch provides proof of germ theory by injecting pure cultures of the *Anthrax bacilli* into mice.

1900: Almwroth Wright isolates **NCTC 160** *Salmonella enterica* subsp. enterica serotype Typhi from the spleen of a typhoid patient during the Boer War. His wartime experiences later lead him to persuade the armed forces to produce 10 million vaccine doses for WWI troops in northern France.

1915: Isolation of the very first bacterial strain registered in the collection. **NCTC 1** is a strain of *Shigella flexneri* recovered from Private Ernst Cable, a WWI soldier who died from dysentery. It is resistant to penicillin and erythromycin even though it was isolated before the discovery of antibiotics.

The Advent of Antibiotics

1920s: Selman Waksman and Albert Schatz lead a systematic effort to screen soil bacteria for antimicrobial compounds. NCTC later acquires the *Streptomyces griseus* strain (**NCTC 4523**) from which they isolated streptomycin.

1920: NCTC is established to "provide a trustworthy source of authentic bacteria for use in scientific studies." Frederick William Andrewes deposits the first cultures.

1928: Alexander Fleming accidentally discovers penicillin. He returns from vacation and notices that a culture plate left lying out had become overgrown with staphylococci colonies, except where mold was growing. He explores further after his former assistant Merlin Price reminds him, "That's how you discovered lysozyme." Over the next 20 years, Fleming deposits 16 samples with NCTC, including a sample of *Haemophilus influenza* isolated from his own nose in November 1935.

1930s: NCTC introduces freeze-drying of samples to ensure longevity and streamline storage and shipment.

1930s: Fritz Kauffman and Phillip White co-develop a scheme for classifying salmonellae by serotype.

Marshalling Science for Public Health

1941: Howard Florey and Ernest Chain begin mass production of penicillin with funds from the US and British governments after the bombing of Pearl Harbor. By D-Day in 1944, enough penicillin has been produced to treat all wounded Allied Forces.

1942: Florey and Chain contribute three Bacillus strains (**NCTC 6431, 6432,** and **6474**) thought to produce 'antibacterial substances active against the Staphylococcus,' demonstrating the researchers were even then seeking antibiotics beyond penicillin.

1947: Edward Tatum and Joshua Lederberg produce the first gene map of *E. coli* K12 (**NCTC 10538**). Despite being one of the most intensively studied organisms in the 20th century, no one definitively knows why it is called "K12".

1947: NCTC focus shifts from a general microbial collection to bacteria of medical or veterinary interest.

1949: NCTC begins a 10-year effort to characterize every organism in the collection.

1953: Pioneering food safety microbiologist Betty Constance Hobbs publishes a study establishing *Clostridium perfringens* as the cause of many outbreaks of food poisoning. She eventually deposits more than 20 NCTC strains of bacteria associated with food-borne illness.

1961: NCTC curator Samuel Cowan and Kenneth Steel publish 'Diagnostic Tables for the Common Medical Bacteria' in the Journal of Hygiene. Demand is so great the journal reprints and distributes them in pamphlet form. The work forms the basis of Cowan & Steel's Manual for the Identification of Medical Bacteria, first published in 1965 and a bench-top staple for years to come.

1969: Don Brenner and colleagues establish DNA hybridization as a more reliable basis for classifying clinical isolates of Enterobacteriaceae. He uses the new method to replace type strains with more representative specimens and identify numerous new microbial species, including *Moellerella wisconsensis* (**NCTC 12132**), *Leminorella grimontii* (**NCTC 12152**), *Enterobacter asburiae* (**NCTC 12123**), and *Citrobacter braakii* (**NCTC 13630**).

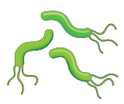
1977: Gilbert and Sanger independently develop methods to determine the exact sequence of DNA molecules.

1977: CDC researchers Joseph McDade and Charles C. Shepard isolate *Legionella pneumophilia* (**NCTC 11230** and **11192**) as the bacterial pathogen behind the outbreak of a new pulmonary disease at a convention in Philadelphia.

The Genomics Era

1981: The European Culture Collections' Organization, of which NCTC is a member, is established.

1982: Future Nobel Prize winner Barry Marshall drinks a culture of the *Helicobacter pylori* (**NCTC 11638** and **11639**) to prove his theory that most stomach ulcers are caused by bacteria.



1987: The first automated DNA sequencing instrument, invented by Lloyd Smith, is commercialized by Applied Biosystems.

1995: Craig Venter, Hamilton Smith, Claire Fraser, and colleagues at TIGR elucidate the first complete genome sequence of a microorganism, *Haemophilus influenza*, and submit the sequence to NCBI.

2003: Cornell University scientists led by Watt Webb and Harold Craighead publish the first report of using arrays of zero-mode waveguides for single-molecule sequencing.

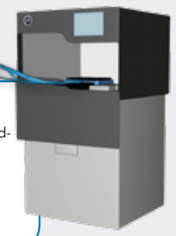
2011: PacBio ships its first commercial SMRT Sequencing system, introducing scientists to the long-read sequencing platform that will ultimately become the gold standard for generating complete, closed microbial genomes.

The largest recorded outbreak of foodborne hemolytic-uremic syndrome, eventually linked to German-grown sprouts, occurs in Europe. The organism responsible, a Shiga toxic *E. coli* (**NCTC 13562**).

2014: NCTC and Wellcome Sanger Institute (WSI) launch a five-year project to sequence 3,000 bacterial strains from the collection using PacBio sequencing technology.

Sanger scientists publish the genome of **NCTC 1**, generated with SMRT Sequencing, and compare it to other *S. flexneri* isolates collected in 1954, 1984, and 2002.

2018: NCTC scientists Sarah Alexander and Mohammed-Abbas Fazal complete the extraction of DNA from more than 3000 NCTC species and samples are delivered to WSI for sequencing using PacBio technology.



Trustworthy Biological Resources for the Future

Historical collections are not only important in preserving the past – they are vital in helping us understand current pathogens, and in the development of future medical advances.

"Knowing very accurately what bacteria looked like before and during the introduction of antibiotics and vaccines and comparing them to current strains from the same collection shows us how they have responded to these treatments," Prof Parkhill said. "This, in turn, helps us develop new antibiotics and vaccines."

Plague, cholera, streptomyces, and 250 strains of *E. coli*, were among the reference genomes created and released in June 2018, as well as several of the most important known drug-resistant bacteria, such as tuberculosis and gonorrhea. The genome sequences of these highly valuable strains are fundamental for developing methods to identify specific infections in people, including tests that can be used in the field to rapidly identify the source of an outbreak and help contain infections. Applications could also include detection of bioterrorism agents, such as anthrax.

Julie E. Russell is the Head of Culture Collections at Public Health England

"Having these reference collections would allow more accurate assessment of the source of any eventual biological threat," Prof Parkhill said.

The sequences will also be of great scientific significance. They include the 'type strains' of many bacteria in the collection – the first strains that describe the species and are used to classify them; 852 were bacterial species associated with human infection, and at least 298 of those type strains had no WGS data available in any public databases. Furthermore, the NCTC 3000 data set may reveal gaps in the range of the current collection, allowing NCTC to better curate the collection by identifying where strains are missing from clinically important lineages. "This provides great potential for using the data in phylogeny and populations genetics studies," Russell said.

The genomic data has been made publicly available through the European Bioinformatics Institute. An additional electronic portal that will bring together all the metadata associated the NCTC strains, including the raw and assembled genomic data is being created. The WGS information also benefits proteomic scientists who access the data to help to interpret the profiles generated by mass spectrometry.

> "If NCTC is to remain scientifically important, it is essential that we embrace WGS technology and provide accurate reference genome sequences."
>
> **– Julie E. Russell**
> Head of Culture Collections
> Public Health England

"Our collection was established by scientists with incredible foresight in recognizing the need for trustworthy biological resources, and we are committed to ensuring that it remains scientifically relevant for the emerging challenges of the 21st century," Russell said.

PacBio CSO Jonas Korlach said the project demonstrates the value of having complete genomes and praised the NCTC and Sanger Institute for making the most of its "living fossil" collection to bridge past, present and future.

"Rather than trying to understand and replicate an organism from the past by relying on a few bones, the NCTC has been able to get a much more complete picture from the living fossils in its collection, which will be of great benefit to all of us," Korlach added.

MICROBIOLOGY AND INFECTIOUS DISEASE

PN: CS122-061919