National Collection of Type Cultures and the NCTC 3000 Whole Genome Sequencing Project

sarah.alexander@phe.gov.uk

@NCTC_3000
National Collection of Type Cultures

• A unique bacterial strain collection founded in 1920

• Clinical strains – veterinary and medical importance

• Dynamic collection – modern and historical strains

• Type and Reference - ~6200 strains

• Diverse Collection: Freeze dried, Lenticulated or DNA Format

• Awarded funding - Wellcome Trust Sanger Institute - Sequence 3000 NCTC strains
“Generate reference genomes for 3000 bacterial strains within the collection and embed these genomes in an accessible resource which will enhance the scientific value of the collection”

- Community Resource Project
- High Molecular Weight DNA – extracted from NCTC strains
- DNA quality profile checked on the TapeStation
- DNA sent to the WTSI - PacBio Sequencing
- Upload of assembled genomes – public databases
NCTC3000 – Data Sharing

- Data regularly uploaded on the WTSI website
- [http://www.sanger.ac.uk/resources/downloads/bacteria/nctc/#t_2](http://www.sanger.ac.uk/resources/downloads/bacteria/nctc/#t_2)
- NCBI – biosample code PRJEB6403
Collating Strain Data: Historical vs Modern

1943

2017
NCTC3000: Nucleic Acid Extractions

- NCTC3000 Project: DNA Extractions
  - 3230 strains
  - 852 species
  - 82 families represented
Sequencing Progress & Data Quality

- Genomes Sequenced:
  - 2700+ Strains Sequenced
  - 56.5% Assembled Single Contig
  - 92.6% Assembled <5 Contigs
  - 23.1% - Evidence of =>1 plasmids
Scientific Impact: Type Strains

• 852 (90%) NCTC Type Strains have been sequenced: 852 Species/subspecies

Comparison with Current Sequence Availability in the NCBI database:

• 12% (108/852) - No WGS for the species
• 30% - (298/852) – No WGS for the Type strain
• 44% (372/852) Only draft genomes available
• Key Type strains from pathogenic species that are missing from public databases
• NCTC3000 and the GCM2 Project – is filling these gaps
<table>
<thead>
<tr>
<th>Species Name</th>
<th>No. Genomes</th>
<th>Species Name</th>
<th>No. Genomes</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>Clostridium difficile</em></td>
<td>11</td>
<td><em>Shigella dysenteriae</em></td>
<td>22</td>
</tr>
<tr>
<td><em>Shigella flexneri</em></td>
<td>11</td>
<td><em>Streptococcus agalactiae</em></td>
<td>23</td>
</tr>
<tr>
<td><em>Enterococcus faecium</em></td>
<td>13</td>
<td><em>Vibrio cholerae</em></td>
<td>38</td>
</tr>
<tr>
<td><em>Escherichia alkaescens</em></td>
<td>13</td>
<td><em>Legionella pneumophila</em></td>
<td>27</td>
</tr>
<tr>
<td><em>Bordetella pertussis</em></td>
<td>14</td>
<td><em>Shigella boydii</em></td>
<td>27</td>
</tr>
<tr>
<td><em>Chromobacterium violaceum</em></td>
<td>14</td>
<td><em>Streptococcus pneumoniae</em></td>
<td>33</td>
</tr>
<tr>
<td><em>Citrobacter koseri</em></td>
<td>15</td>
<td><em>Clostridium perfringens</em></td>
<td>36</td>
</tr>
<tr>
<td><em>Haemophilus influenzae</em></td>
<td>16</td>
<td><em>Neisseria meningitidis</em></td>
<td>37</td>
</tr>
<tr>
<td><em>Klebsiella aerogenes</em></td>
<td>16</td>
<td><em>Neisseria gonorrhoeae</em></td>
<td>40</td>
</tr>
<tr>
<td><em>Campylobacter jejuni</em></td>
<td>18</td>
<td><em>Pseudomonas aeruginosa</em></td>
<td>56</td>
</tr>
<tr>
<td><em>Corynebacterium diphtheriae</em></td>
<td>18</td>
<td><em>Salmonella enterica subsp. enterica</em></td>
<td>131</td>
</tr>
<tr>
<td><em>Enterobacter cloaceae</em></td>
<td>18</td>
<td><em>Klebsiella pneumoniae</em></td>
<td>159</td>
</tr>
<tr>
<td><em>Proteus mirabilis</em></td>
<td>18</td>
<td><em>Streptococcus pyogenes</em></td>
<td>83</td>
</tr>
<tr>
<td><em>Enterococcus faecalis</em></td>
<td>21</td>
<td><em>Staphylococcus aureus</em></td>
<td>167</td>
</tr>
<tr>
<td><em>Acinetobacter baumannii</em></td>
<td>11</td>
<td><em>Escherichia coli</em></td>
<td>255</td>
</tr>
</tbody>
</table>
NCTC3000 Data Usage

- AMR genes in Enterobacteriaceae vs date of strain isolation
- Type strain equivalence
- Multiple external publications – analysing NCTC3000 dataset

NCTC9657
*K. pneumoniae*

Hinge: Long-Read Assembly Achieves Optimal Repeat Resolution. Govinda Kamath et al. *biorxiv*

Circlator Hunt et al. *Genome Biology* 2015 16:294

*Complete Genome Sequence of Neisseria weaveri Strain NCTC13585*
Sarah Alexander,† Mohammed-Abbas Fazal,† Edward Burnett,† Ana Deheer-Graham,‡ Karen Oliver,‡ Nancy Holroyd,‡ Julian Parkhill,¶ Julie E. Russell†

Received 22 June 2016. Accepted 29 June 2016. Published 25 August 2016

Culture Collections – a Public Health England Biological Resource
Conclusions

• NCTC3000 - generate reference genomes for the scientific community

• Encourage scientists to use NCTC3000 sequencing data and deposit clinically relevant strains into NCTC

• GCM2 & NCTC3000 – represent an opportunity to sequence all of the Type strains

• Type strain equivalence and comparison – extremely important

• Current Focus – assembling genomes & ensuring quality and scientific accessibility

• Availability of publically accessible strains, Nucleic Acid products & complete reference genomes enhance our understanding of these important pathogens
Acknowledgments

Culture Collections Team
- NCTC operational team
- Mohammed Abbas Fazal
- Julie E. Russell
- Ed Burnett
- Heena Shah

WTSL
- Karen Oliver
- Nick Grayson
- Julian Parkhill