Genomic Epidemiology

Ole Lund
“The time has come to close the book on infectious diseases. We have basically wiped out infection in the United States.” Surgeon General of the United States of America, William Stewart, 1967

- 335 cases of Emerging infectious diseases 1940-2003

Whole genome sequence based Diagnostics

Infectious diseases are responsible for >25% of all global deaths

An increasing number of infectious diseases have a global epidemiology (e.g. SARS, avian flu, influenza, *Salmonella* etc.)

Resistance to antimicrobial drugs are emerging faster than new drugs can be developed
Threat by Anti Microbial Resistance (AMR)

• In a recent report from April 2014, WHO concludes
  – “A post-antibiotic era – in which common infections and minor injuries can kill – far from being an apocalyptic fantasy, is instead a very real possibility for the 21st Century”.

http://www.who.int/drugresistance/documents/surveillancereport/en/
Projected number of deaths globally

Deaths from selected sources (adapted from O’Neill, 2016)
What to do about it?

• Develop better drugs
  – May not keep up with development of resistance
  – Risky only to gamble on this being possible

• Limit spread
  – Limit use of antibiotics
  – Better information on what can/should be used
  – **Interventions against spread of organisms with resistance**
How can whole genome sequencing help?

- Is being developed to give a fast, cheap, accurate and easy way of getting information about resistance
  - Sequencing currently cost ~100$
- Gives the ultimate resolution (genome) typing to track spread
- At the same time many other types/phenotypes may be determined
Did you know that 80% of all antibiotics are given to healthy food-producing animals, and only 20% for treating sick people?

DATA: US Food & Drug Administration (FDA)
Routine microbial diagnostic

- Sample: 1-2 days
- Culturing: 1-2 days
- ID: 1-2 days
- Antibiotic resistance: 1 – several weeks
- Typing
The Whole Genome Sequencing (WGS) Process

WGS is a laboratory procedure that determines the order of bases in the genome of an organism in one process. WGS provides a very precise DNA fingerprint that can help link cases to one another allowing an outbreak to be detected and solved sooner.

1. DNA Extraction

Scientists take bacterial cells from an agar plate and treat them with chemicals that break them open, releasing the DNA. The DNA is then purified.

2. DNA Shearing

DNA is cut into short fragments of known length, either by using enzymes “molecular scissors” or mechanical disruption.

3. DNA Library Preparation

Scientists make many copies of each DNA fragment using a process called polymerase chain reaction (PCR). The pool of fragments generated in a PCR machine is called a “DNA library.”

4. DNA Library Sequencing

The DNA library is loaded onto a sequencer. The combination of nucleotides (A, T, C, and G) making up each individual fragment of DNA is determined, and each result is called a “DNA read.”

5. DNA Sequence Analysis

The sequencer produces millions of DNA reads and specialized computer programs are used to put them together in the correct order like pieces of a jigsaw puzzle. When completed, the genome sequence containing millions of nucleotides (in one or a few large pieces) is ready for further analysis.
Whole genome sequencing

• Advantages
  – Give the full blueprint of how to make a given bacteria
    • Antimicrobial resistance
    • Best resolution data to track epidemics

• Disadvantages
  – Generates ~200mb of data per bacteria in form of a “puzzle” that needs to be assembled to a ~5mb assembled genome
Bacterial genomics

• Sequencing a bacterial genome cost ~$100 (on a desk top sequencer)
• Equipment will cost less than $100 000
• In Denmark 1 million clinical microbiology isolates are handled each year
  – EU/USA ~100 million
  – Globally ~ 1 billion (10 billion needed)
• Future limiting factor will not be sequencing but handling the sequences
Whole genome sequence based diagnostic

- Sample: 1-2 days
- Culturing: ½-1 day
- ID Resistance Typing + Much more
Welcome to the Center for Genomic Epidemiology

The cost of sequencing a bacterial genome is $50 and is expected to decrease further in the near future and the equipment needed cost less than $150,000. Thus, within a few years all clinical microbiological laboratories will have a sequencer in use on a daily basis. The price of genome sequencing is already so low that whole genome sequencing will also find widespread application in human and veterinary practices as well as many other places where bacteria are handled. In Denmark alone this equals more than 1 million isolates annually in 15-20 laboratories and globally up to 1-2 billion isolates per year. The limiting factor will therefore in the future not be the cost of the sequencing, but now to assemble, process and handle the large amount of data in a standardized way that will make the information useful, especially for diagnostic and surveillance.

The aim of the center is to provide the scientific foundation for future internet-based services where a central database will enable simplification of whole genome sequence information and comparison to all other sequenced including spatial-temporal analysis. We will develop algorithms for rapid analysis of whole genome DNA-sequences, tools for analysis and extraction of information from the sequence data and internet-based interfaces for using the tools in the global scientific and medical community. The activity is being expanded to also include other organisms, such as viruses and parasites as well as metagenomic samples.
How to see if two sequences are similar

• Alignment/mapping sequences
  – Slow
  – accurate

• K-mers
  – split sequences up in parts of length K
  – Fast
  – Robust
  – Sometimes less accurate
K-mers

• K-mer: piece of DNA of length K
• Longer sequences can be broken up in K-mers
• Overlap of K-mers may be used as a measure of similarity

<table>
<thead>
<tr>
<th>50 mers</th>
<th>Strains</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>A</td>
</tr>
<tr>
<td>2</td>
<td>A, B</td>
</tr>
<tr>
<td>3</td>
<td>C</td>
</tr>
<tr>
<td>4</td>
<td>D, E, F</td>
</tr>
<tr>
<td>5</td>
<td>B</td>
</tr>
</tbody>
</table>
K-mer based method works well for species identification

K-mers: Not a new idea

Classification of methanogenic bacteria by 16S ribosomal RNA characterization

(comparative oligonucleotide cataloging/phylogeny/molecular evolution)

GEORGE E. FOX*, LINDA J. MAGRUM*, WILLIAM E. BALCH‡, RALPH S. WOLFE‡, AND CARL R. WOESE‡

Departments of *Genetics and Development and ‡Microbiology, University of Illinois, Urbana, Illinois 61801

Communicated by H. A. Barker, August 10, 1977

Table 1. Oligonucleotide cataloging for 16S rRNA of 10 methanogens

<table>
<thead>
<tr>
<th>Oligonucleotide sequence</th>
<th>Present in organism number</th>
<th>Oligonucleotide sequence</th>
<th>Present in organism number</th>
<th>Oligonucleotide sequence</th>
<th>Present in organism number</th>
</tr>
</thead>
<tbody>
<tr>
<td>6-mers</td>
<td></td>
<td>6-mers</td>
<td></td>
<td>6-mers</td>
<td></td>
</tr>
<tr>
<td>CCCCG</td>
<td>1-10</td>
<td>CAACAG</td>
<td>1</td>
<td>AACCUCG</td>
<td>3</td>
</tr>
<tr>
<td>CCCCG</td>
<td>10</td>
<td>CAACAG</td>
<td>6</td>
<td>AACCUCG</td>
<td>8</td>
</tr>
<tr>
<td>CCCCG</td>
<td>9</td>
<td>CAACAG</td>
<td>1-6,7-9</td>
<td>AACCUCG</td>
<td>3-9</td>
</tr>
<tr>
<td>ACCCG</td>
<td>10</td>
<td>CAACAG</td>
<td>7-8-9</td>
<td>ACCCG</td>
<td>3-10</td>
</tr>
<tr>
<td>ACCCG</td>
<td>10</td>
<td>CAACAG</td>
<td>7-8</td>
<td>ACCCG</td>
<td>3-10</td>
</tr>
<tr>
<td>ACCCG</td>
<td>9</td>
<td>CAACAG</td>
<td>1-6,8-9</td>
<td>ACCCG</td>
<td>3-10</td>
</tr>
<tr>
<td>ACCCG</td>
<td>10</td>
<td>CAACAG</td>
<td>1-10</td>
<td>ACCCG</td>
<td>3-10</td>
</tr>
<tr>
<td>ACCCG</td>
<td>10</td>
<td>CAACAG</td>
<td>1-10</td>
<td>ACCCG</td>
<td>3-10</td>
</tr>
<tr>
<td>7-mers</td>
<td></td>
<td>7-mers</td>
<td></td>
<td>7-mers</td>
<td></td>
</tr>
<tr>
<td>CCAAC</td>
<td>9</td>
<td>CAACAG</td>
<td>7-8</td>
<td>ACCCG</td>
<td>3-10</td>
</tr>
<tr>
<td>CCAAC</td>
<td>9</td>
<td>CAACAG</td>
<td>1-6,8-9</td>
<td>ACCCG</td>
<td>3-10</td>
</tr>
<tr>
<td>CCAAC</td>
<td>1-10</td>
<td>CAACAG</td>
<td>1-10</td>
<td>ACCCG</td>
<td>3-10</td>
</tr>
<tr>
<td>AACCG</td>
<td>1-10</td>
<td>CAACAG</td>
<td>1-10</td>
<td>ACCCG</td>
<td>3-10</td>
</tr>
<tr>
<td>AACCG</td>
<td>1-10</td>
<td>CAACAG</td>
<td>1-10</td>
<td>ACCCG</td>
<td>3-10</td>
</tr>
<tr>
<td>AACCG</td>
<td>1-10</td>
<td>CAACAG</td>
<td>1-10</td>
<td>ACCCG</td>
<td>3-10</td>
</tr>
<tr>
<td>AACCG</td>
<td>1-10</td>
<td>CAACAG</td>
<td>1-10</td>
<td>ACCCG</td>
<td>3-10</td>
</tr>
<tr>
<td>AACCG</td>
<td>1-10</td>
<td>CAACAG</td>
<td>1-10</td>
<td>ACCCG</td>
<td>3-10</td>
</tr>
<tr>
<td>AACCG</td>
<td>1-10</td>
<td>CAACAG</td>
<td>1-10</td>
<td>ACCCG</td>
<td>3-10</td>
</tr>
<tr>
<td>AACCG</td>
<td>1-10</td>
<td>CAACAG</td>
<td>1-10</td>
<td>ACCCG</td>
<td>3-10</td>
</tr>
<tr>
<td>ACGCG</td>
<td>1-10</td>
<td>CAACAG</td>
<td>1-10</td>
<td>ACCCG</td>
<td>3-10</td>
</tr>
<tr>
<td>ACCCG</td>
<td>1-10</td>
<td>CAACAG</td>
<td>1-10</td>
<td>ACCCG</td>
<td>3-10</td>
</tr>
<tr>
<td>ACCCG</td>
<td>1-10</td>
<td>CAACAG</td>
<td>1-10</td>
<td>ACCCG</td>
<td>3-10</td>
</tr>
<tr>
<td>ACCCG</td>
<td>1-10</td>
<td>CAACAG</td>
<td>1-10</td>
<td>ACCCG</td>
<td>3-10</td>
</tr>
</tbody>
</table>
Antimicrobial resistance (AMR) prediction

- **ResFinder**
  - Find acquired resistance genes (often on plasmids) where presence within a similarity threshold normally signifies resistance

- **PointFinder**
  - Find point mutations/indels in chromosomal regions which give resistance

PointFinder: a novel web tool for WGS-based detection of antimicrobial resistance associated with chromosomal point mutations in bacterial pathogens
E Zankari, R Allesøe, KG Joensen, LM Cavaco, O Lund... - Journal of Antimicrobial ..., 2017
Camilla Hundahl, unpublished work
PointFinder - Flow diagram

Assembly
- Velvet

Blast
- AMR genes

Parsing
- Mutation list
## PointFinder - Species

<table>
<thead>
<tr>
<th>Species</th>
<th>Status</th>
<th>Availability</th>
</tr>
</thead>
<tbody>
<tr>
<td>Campylobacter</td>
<td>Published</td>
<td>Online</td>
</tr>
<tr>
<td>E. Coli</td>
<td>Published</td>
<td>Online</td>
</tr>
<tr>
<td>Salmonella</td>
<td>Published</td>
<td>Online</td>
</tr>
<tr>
<td>N. gonorrhea</td>
<td>Validation ongoing</td>
<td>Online</td>
</tr>
<tr>
<td>M. tuberculosis</td>
<td>Validation ongoing</td>
<td>Online</td>
</tr>
<tr>
<td>Klebsiella</td>
<td>Validation ongoing</td>
<td></td>
</tr>
<tr>
<td>P. Falciparum</td>
<td>Validated, unpublished</td>
<td>Online soon</td>
</tr>
<tr>
<td>E. Faecalis</td>
<td>Planned</td>
<td></td>
</tr>
<tr>
<td>S. Aureus</td>
<td>Planned</td>
<td></td>
</tr>
<tr>
<td>Actinobacter</td>
<td>Planned</td>
<td></td>
</tr>
<tr>
<td>PS</td>
<td>Planned</td>
<td></td>
</tr>
<tr>
<td>Clostridium difficile</td>
<td>Planned</td>
<td></td>
</tr>
<tr>
<td>Streptococcus pneumonia</td>
<td>Planned</td>
<td></td>
</tr>
</tbody>
</table>


Antimicrobial resistance

Campylobacter
E. Coli
Salmonella
M. tuberculosis
N. gonorrhea

To be added:
Klebsiella
P. Falciparum
...
From single analysis to pipeline analysis

To be added:
Upload to public repositories (SRA/ENA)
Multiple GB File upload via JavaScript
Break
Phylogeny

• Trees are traditionally made using aligned sequences of single genes or proteins
• Whole genome data may be used to create trees based on
  – SNP calling
  – K-mer overlap

A \quad B \quad C

D
What is a SNP (Wikipedia)

• A Single Nucleotide Polymorphism (SNP, pronounced snip; plural snips) is a DNA sequence variation occurring commonly* within a population (e.g. 1%) in which a Single Nucleotide — A, T, C or G — in the genome (or other shared sequence) differs between members of a biological species or paired chromosomes.

• Some prefer the term SNV
Reads -> Mapping -> SNP

Reference

Mapped sequence
Sequences of 4 strains

<table>
<thead>
<tr>
<th>Strain</th>
<th>Sequence</th>
</tr>
</thead>
<tbody>
<tr>
<td>Strain A</td>
<td>AT\textbf{T}CAGT\textbf{A}</td>
</tr>
<tr>
<td>Strain B</td>
<td>ATG\textbf{C}AG\textbf{T}C</td>
</tr>
<tr>
<td>Strain C</td>
<td>ATG\textbf{C}A\textbf{A}T\textbf{C}</td>
</tr>
<tr>
<td>Strain D</td>
<td>AT\textbf{T}CAG\textbf{T}C</td>
</tr>
</tbody>
</table>
Construct distance matrix

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>0</td>
<td>2</td>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td>B</td>
<td>2</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>C</td>
<td>3</td>
<td>1</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>D</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>0</td>
</tr>
</tbody>
</table>
Make Tree

Strain A  AT\textbf{TCA\textbf{GTA}}
Strain B  AT\textbf{GCA\textbf{GTC}}
Strain C  AT\textbf{GCA\textbf{ATC}}
Strain D  AT\textbf{TCA\textbf{GTC}}

A B C D
A 0 2 3 1
B 2 0 1 1
C 3 1 0 2
D 1 1 2 0
Relatedness of samples from whole genome sequencing data: Phylogeny of the isolates by the NDtree method.

Salmonella Typhimurium DT104

SNPtree

Remote reference

CSIPhylogeny

Close reference

NDtree

Controlled evolution

Johanne Ahrenfeldt, Submitted
Phylogenetic tree using neighbor joining
Outbreak analysis of billions of strains: Real-time tracking of all microbial genomes

- Number of *earlier* isolates (from within the last year) with less than 10 SNP differences to the current isolate
  - Do not need to be updated
  - Search can/should be restricted to those that cluster to the same template
Evergreen - flowchart

Every 24 hours

Download all new data from SRA/ENA

For every sample

Typing
Find closest match in homology reduced database

Map to closest match

Find the phylogenetic distance to consensus sequences for the given reference. Cluster isolates with < 10 SNPs distance

Hobohm 1 homology reduction on new, non-clustered isolates

Databases

NCBI

Homology reduced database
Hobohm 1 clustering to 99% homology on all bacterial reference genomes from NCBI

Infer/recalculate phylogenetic tree
For each reference genome with new mapped samples. Infer tree from cluster representatives and add the redundant isolates to their respective nodes based on the clustering

Database of previous homology reduced isolates (representative consensus sequence) and redundant isolates
Evergreen phylogenetic trees

Constantly updated phylogenetic trees with publicly available data from the short sequencing read archives.

The template refers to the reference sequence to which the reads were mapped to get a consensus sequence. The time notes the last time a tree has been updated with new isolates. Visualisation is done with Phylocanvas.

<table>
<thead>
<tr>
<th>Template</th>
<th>Time</th>
</tr>
</thead>
<tbody>
<tr>
<td>Campylobacter coli_CVM_N29710_NC_023347_1</td>
<td>2017-10-01</td>
</tr>
<tr>
<td>Campylobacter coli_RMS611_NZ_CP007179_1</td>
<td>2017-10-01</td>
</tr>
<tr>
<td>Campylobacter coli_strain_BFR_CA_9557_NZ_CP011777_1</td>
<td>2017-10-08</td>
</tr>
<tr>
<td>Campylobacter jejuni_32488_NZ_021934_1</td>
<td>2017-10-08</td>
</tr>
<tr>
<td>Campylobacter jejuni_4531_NZ_022529_1</td>
<td>2017-10-09</td>
</tr>
<tr>
<td>Campylobacter jejuni_NZ_LNB31025_1</td>
<td>2017-10-01</td>
</tr>
<tr>
<td>Campylobacter jejuni_RM13221_NZ_003912_1</td>
<td>2017-10-08</td>
</tr>
<tr>
<td>Campylobacter jejuni_strain_CJ677CC519_NZ_CP010471_1</td>
<td>2017-10-08</td>
</tr>
<tr>
<td>Campylobacter jejuni_strain_F0124_NZ_CP017862_1</td>
<td>2017-10-09</td>
</tr>
<tr>
<td>Campylobacter jejuni_strain_HF5_4A_4_NZ_CP007188_1</td>
<td>2017-10-08</td>
</tr>
<tr>
<td>Campylobacter jejuni_strain_CD067_NZ_CP014744_1</td>
<td>2017-10-08</td>
</tr>
<tr>
<td>Campylobacter jejuni_strain_RM3194_NZ_CP014344_1</td>
<td>2017-10-08</td>
</tr>
<tr>
<td>Campylobacter jejuni_strain_TS1218_NZ_CP017860_1</td>
<td>2017-10-08</td>
</tr>
<tr>
<td>Campylobacter jejuni_subsp_jejuni_81_178_NZ_008787_1</td>
<td>2017-10-09</td>
</tr>
<tr>
<td>Campylobacter jejuni_subsp_jejuni.CG9421_NZ_CP005388_1</td>
<td>2017-10-08</td>
</tr>
<tr>
<td>Campylobacter jejuni_subsp_jejuni.F36011_NZ_CP006851_1</td>
<td>2017-10-01</td>
</tr>
<tr>
<td>Campylobacter jejuni_subsp_jejuni.NCTC_11168_ATCC_700819_chromosome_NC_002163_1</td>
<td>2017-10-09</td>
</tr>
</tbody>
</table>
Multistate Outbreak of Shiga toxin-producing Escherichia coli O157:H7...
https://www.cdc.gov/ecoli/2017/o157h7-03-17/index.html
Jun 5, 2017 - Multistate outbreak of L. monocytogenes associated with turkey deli meat. ... March 9, 2017 - The CDC announces it is working with the FDA to ... were reported in four states - Connecticut, Florida, New York, and Vermont.
Outbreak Investigations

• Pionered by John Snow 1854
  – Find two or more linked cases
Outbreak Visualization
Color-based clusters

Influenza
Metagenomic based diagnostic

½-1 day

| Sample | ID Resistance Type + Every thing |
Phylogenetic tree of isolates and metagenomic samples
Metagenomic based diagnostic with non batch mode sequencing (nanopore technologies)
Global Surveillance
Meta-genomic analysis of toilet waste from long distance flights; a step towards global surveillance of infectious diseases and antimicrobial resistance.
Portable next generation sequencing workstation

• How to run whole genome sequence analysis from hospitals with limited internet access (in regional Tanzania and around Copenhagen)

• CGE tools (genomicepidemiology.org) on a Laptop
  – Example
    • Dell Precision M4700 Linux:
    • 750Gb 7200 rpm disk, 4 cores, 32Gb ram
    • DKK 14 000/€1 900/$2 500 ex vat

Easily deployed and returned for upgrades
More than enough for coping with 2x8Gbase weekly output from a MiSeq (42 CPU h/Gb)
Whole genome sequencing

• Is it a game changer in the combat against infectious diseases

• Game changer? - what is new with WGS?
  – Typing with ultimate resolution (bar epigenetics?)
    • Resolution = 1/mutation rate = 1 year
  – Can (soon) be done in a day
  – Instant deep pheno-typing (e.g. resistance/virulence genes)
  – With falling prices surveillance may be ubiquitous
    • Everything is under constant surveillance
      – People, animals, planes, places, doorhandles …
  – Information can be shared instantly around the globe
Transmission do not have to be zero

• But $R_0$:
  – The number of secondary infections that a case on the average give rise to

• Have to be below 1
Game changer?

• Can WGS + IT be used to set $R_0$ to less than 1 for some pathogens in some areas?
• Which are the best cases?
Thanks

DTU Systems Biology/CBS/Lund group
- Mette Voldby Larsen
- Martin Thomsen
- Johanne Ahrenfeldt
- Vanessa Jurtz
- Jose L. Bellod Cisneros
- Johanne Ahrenfeldt
- Anna Maria Malberg Tetzchner

Ex members
- Salvatore Consentiono

Student helpers
- Jamie Neubert Pedersen
- Valentin Ibanez
- Rosa Allesøe
- Camilla Lemvigh

DTU Systems Biology/CBS
- Dave Ussery
- Thomas Ponten
- Dhany Saputra
- Simon Rasmussen
- Thomas Nordahl Petersen

DTU DMAC
- Laurent Gautier
- Marlene Dalgaard

DTU Food
- Frank Aarestrup
- Henrik Hasman
- Rene S. Hendriksen
- Shinny Leekitcharoenphon
- Rolf Sommer Kaas
- Marlene Hansen
- Katrine Grimstrup Joensen
- Oksana Lukjancenko

Copenhagen University/CMP
- Thor Theander
- Michael Alifrangis
- Sidsel Nag

KCMC Moshi, Tanzania
- Gibson Kibiki
- Happiness Kumburu
- Tolbert Sonda