Bioinformatics and Public Health

Duncan MacCannell PhD
Senior Advisor, NCEZID/OD

Georgia Tech Bioinformatics Program
January 9, 2013
Biology has at least fifty more interesting years.

James D. Watson (1984)
OUTBREAK INVESTIGATION

IDENTIFY
- Same pathogen?
- Does it fit the clinical syndrome?
- Is it present in all your cases?
- Is it absent in your controls?
- Biology/ecology of the organism?
...

SUBTYPE
- Same strain? How different?
- Does the clustering fit the hypothetical scenario? Eg: Timeframe? Spatial? Chain of transmission?

CHARACTERIZE
- Important features?
- Are there characteristics that could explain emergence or affect public health response?

CULTURE/PHENOTYPIC MOLECULAR
- Timeframe: DAYS

MOLECULAR PHENOTYPIC
- Timeframe: DAYS/WEEKS

MOLECULAR

GENOMICS, Other -OMICS

RESPONSE

INFORM PREVENTION STRATEGIES

Human Animal Environmental Samples
Drivers for Innovation and Change

- Faster time-to-answer.
  - More relevant, actionable information for public health response.

- Higher resolution/accuracy.
  - More specific, useful information from your data.

- Automate-able or standardize-able.
  - Different criteria for outbreak response/surveillance applications.

- Objective and reproducible results.

- Epidemiologic/clinical concordance.

- Cost effectiveness.
  - Comprehensive data capture can reduce downstream testing cost.
  - Cost per test for many of these technologies is extremely competitive, relative to conventional methods, and is decreasing.
MALDI-ToF Microbial Identification

- Results within minutes/seconds
- $0.25-$1/sample
- Automatable
- Simple workflow

12-24hr culture (bacteria)
24-48hr culture (myco/fungi)

Direct spotting or simple extract
Overlay with matrix solution
Air dry

Load and Run.

- Widely-established in European clinical/public health laboratories.
- Increasingly common in North America.

For identification purposes only; does not constitute endorsement by CDC or HHS.
Tuned m/z profile of specific high-abundance proteins (e.g., ribosomal)
Results within seconds. Similar workflows for a wide range of microorganisms. Databases.
Roche 454 PTP plate, Ion Torrent 314, Pacific BioSciences SMRTcells (x 3)
Rapid Whole-Genome Sequencing for Investigation of a Neonatal MRSA Outbreak

Claudio U. Köser, B.A., Matthew T.G. Holden, Ph.D., Matthew J. Ellington, D.Phil.,
Edward J.P. Cartwright, M.B., B.S., Nicholas M. Brown, M.D.,
Amanda L. Ogilvy-Stuart, F.R.C.P., Li Yang Hsu, M.R.C.P.,
Claire Chewapreecha, B.A., Nicholas J. Croucher, M.A.,
Simon R. Harris, Ph.D., Mandy Sanders, B.Sc., Mark C. Enright, Ph.D.,
Gordon Dougan, Ph.D., Stephen D. Bentley, Ph.D., Julian Parkhill, Ph.D.,
Louise J. Fraser, Ph.D., Jason R. Betley, Ph.D., Ole B. Schulz-Trieglaff, Ph.D.,
Geoffrey P. Smith, Ph.D., and Sharon J. Peacock, Ph.D., F.R.C.P.
shorter sequence-read lengths, faster protocols could be used that would reduce the time period to under a day. The approximate cost of all the materials for whole-genome sequencing is $150 per isolate, including sample preparation, library quality control (quantification and size assessment), and sequencing, which is roughly equivalent to the cost of two PCR tests (e.g., Cepheid Xpert) used to screen for MRSA carriage. Given the competition between current and emerging sequencing platforms, the price and turnaround time will
1992: 500 basepairs/day

2002: 50,000 basepairs/day

2012: 50,000,000,000 basepairs/day

Human genome: 3,000,000,000bp
Input: DNA/RNA

Source:
- Genomic
- Amplicon
- Whole sample
- Host/vector/pathogen/environment

... 

NGS

Workflow:
- Platforms
- Chemistry
- Perf. char.
- Labor/TaT
- Cost

Depends on Objectives

Bioinformatics

Workflow:
- Hardware/software
- Specialized skillsets
- Algorithms/pipelines
- Pathogen databases
- Data analysis/interpret/Integration/visualization

Depends on Objectives

Output: Information From Sequence Data

Comparative Genomics
HR Strain typing/Subtyping
Cluster identification
Molecular evolution
Genotypic characterization
Virulence, AR, signatures
Functional annotation
Diagnostic dev/validation

Metagenomics
Pathogen identification/discovery
Culture-independent diagnostics
Microbial ecology/diversity

Increasingly Universal Workflows
Established sequencing workflows for a wide range of pathogens.

A Moving Target
Rapidly evolving technology space.
Changing hardware and COTS/OSS capabilities. Lots of choice, but lack of consistent standards. BIG DATA. New workforce and skillset.

Objective, “Future-Proof” Data
Intrinsic quality metrics. Ability to back-test retrospective sequence data in silico for genes/markers identified at a future date.
WGS and Pathogen Genomics: Advantages

- **It’s universal…**
  - DNA/RNA sequencing workflows and approaches can be applied to a wide range of pathogenic organisms.

- **It’s fundamental…**
  - Genomics is a cornerstone for other “omic” approaches
  - Sequence databases starting point for assay devel./validation.

- **It’s objective…**
  - Sequence-based methods avoid subjectivity of phenotypic or fragment-based approaches. Volume of data → internal controls.

- **It’s (relatively) future proof…**
  - Comprehensive sequencing captures the features you know about, and those you don’t. Quality may change, but the sequence will not.
  - This makes it possible to back-test future approaches/targets on the data you collect today.
WGS and Genomic Epidemiology: Limitations

- **It lacks standardization…**
  - WGS is a rapidly-evolving technology space, both in terms of sequencing and analytics.
  - Standards and mechanisms for data/metadata analysis, storage and exchange remain under active debate and development.

- **Comprehensive databases are still being built…**
  - Without a useful baseline understanding of pathogen features/diversity, interpretation may be limited.
  - Need curated, and comprehensive epi-linked reference databases.

- **Many analyses require specialized bioinformatics infrastructure and staff.**
  - Bioinformaticists, DBAs, programmers, system administrators, etc.
  - Technical and computational complexity of tasks can vary widely.

- **Data management, retention and release. Storage. LIMS.**
Laboratory Methods for Molecular Epi: A Brief History

- Serotyping
- MEE
- REA
- PFGE
- RAPD
- MLST
- Phage typing
- Bacteriocin typing
- AFLP
- MLVA
- DNA Microarrays
- SNPs
- Mass Spec
- Whole Genome Sequencing
- Biotyping
- Phage typing
- Bacteriocin typing
- Antibiogram Sequencing
- Plasmid profiles
- Ribotyping
- Rep-PCR
- PAMPs
- Phenotypic
- Molecular (mostly genotypic)

* Slide adapted from: Peter Gerner-Smidt
Reference-Guided (Mapped) Assembly

UNMAPPED READS
1. Sequences not present in the reference.
2. Plasmids or other extrachromosomal.
3. DNA Structural Variation/Rearrangement

ADVANTAGES: Relatively fast, well-suited to highly-conserved genomes.
DISADVANTAGES: Issues with high diversity, mobile elements, linear reference
**ADVANTAGES:** Reference agnostic: assembles all the reads it can into contigs.

**DISADVANTAGES:** Doesn’t always get things right. Repeat sequences etc.
Whole Genome SNP Typing (WGST)

Reference Sequence/Genome

1. ACTAGA
2. ACTAGT
3. TCTACT
K-mer Based Sequence Typing

1. ATCGCTATCGTAGGGCATTTTACTGTTGCTGTGTTGCAGATGCAGA
2. ATCGGTATCGTAGGGCAATTACTGTAAGCTGATGCGATGAGA

Advantages: Extremely fast, once k-mer library is established. Doesn’t need assembled sequence to work (can be performed directly from short reads)… although assembly helps with the context and accuracy.
**wgMLST / Binary Typing**

- **SuperMLST**
  - Same concept as conventional MLST, but with 20+ loci
  - Each sequence is matched against a database of numbered alleles.
  - Example: (1,1,2,1,4,1,5,5,5,1,2,3,4,1,2,3,1,4,5,6,7,2,3,2,2,1)

- **Binary typing**
  - Presence or absence of specific virulence markers/sequences.
  - Example: Binary matrix (0,1,1,0,1,1,0,1,1,1,1,0)
## Comparative Genomics: A Critical Starting Point

<table>
<thead>
<tr>
<th>Characterization</th>
</tr>
</thead>
<tbody>
<tr>
<td>▪ Virulence factors</td>
</tr>
<tr>
<td>▪ Resistance genes</td>
</tr>
<tr>
<td>▪ Phenotypic markers</td>
</tr>
<tr>
<td>▪ Eng. signatures.</td>
</tr>
<tr>
<td>▪ GWAS</td>
</tr>
</tbody>
</table>

### Diagnostic dev/validation.

### Transcriptomics, proteomics and other ‘omics.

### Reference Databases
1. Known pathogens
2. Near neighbors
3. (Un)common Variants
4. Environmental
5. Commensal
6. Host

### Diversity and phylogenetics.
Applying Pathogen Genomics to Public Health Microbiology
NDM1 *Klebsiella pneumoniae* Outbreak Response

Dr. Brandi Limbago (NCEZID/DHQ) / Dr. Michael Farrell (NCEZID/DPEI)

- **Enterobacteriaceae** bearing New Delhi metallo-beta-lactamase (NDM) carbapenemases are rare in the US.
  - Typically associated with travel/overseas medical exposures.
  - Until recently, <20 CDC-confirmed cases in the United States.

- **A healthcare facility in Colorado identified two cases with no history of travel or clear route of transmission.**
  - Laboratory look-back identified one earlier case.
  - Active surveillance identified 5 more colonized patients.

---

### Timeline

- **Genomic DNA received by BRRATL:** Oct 25
- **Initial Sequencing Complete:** Oct 27
- **Initial Analysis Completed:** Oct 29

**Sequencing** **Bioinformatics** **START OF FIELD INVESTIGATION**
Outbreak Investigation: NDM1 CRE
Erin Epson (EISO) & Brandi Limbago/Alex Kallen (NCEZID/DHQI)

PFGE: 4 pattern types, over >90% similarity.
Whole Genome Sequence Analysis — 67 core SNPs
Adapted from: Epson et al.
## Antibiotic Resistance Genes (ARDB/CARD)

### ABX Markers

<table>
<thead>
<tr>
<th>Key</th>
<th>Culture_Date</th>
<th>blaNDM</th>
<th>Source</th>
</tr>
</thead>
<tbody>
<tr>
<td>S1</td>
<td>7/11/2012</td>
<td>POS</td>
<td>Blood</td>
</tr>
<tr>
<td>S2</td>
<td>6/29/2012</td>
<td>NEG</td>
<td>Blood</td>
</tr>
<tr>
<td>S3</td>
<td>8/11/2012</td>
<td>POS</td>
<td>Sputum</td>
</tr>
<tr>
<td>S4</td>
<td>6/17/2012</td>
<td>POS</td>
<td>Sputum</td>
</tr>
<tr>
<td>S5</td>
<td>6/7/2012</td>
<td>NEG</td>
<td>Sputum</td>
</tr>
<tr>
<td>S6</td>
<td>9/26/2012</td>
<td>POS</td>
<td>Rectum</td>
</tr>
<tr>
<td>S7</td>
<td>9/25/2012</td>
<td>NEG</td>
<td>Perit. cavity</td>
</tr>
<tr>
<td>S8</td>
<td>9/26/2012</td>
<td>POS</td>
<td>Rectum</td>
</tr>
<tr>
<td>S9</td>
<td>9/26/2012</td>
<td>POS</td>
<td>Rectum</td>
</tr>
<tr>
<td>S10</td>
<td>9/26/2012</td>
<td>NEG</td>
<td>Rectum</td>
</tr>
<tr>
<td>S11</td>
<td>9/26/2012</td>
<td>POS</td>
<td>Rectum</td>
</tr>
<tr>
<td>S12</td>
<td>9/26/2012</td>
<td>POS</td>
<td>Rectum</td>
</tr>
</tbody>
</table>

**ARDB:** [http://ardb.cbcb.umd.edu](http://ardb.cbcb.umd.edu)  
**CARD:** [http://arpcard.mcmaster.ca](http://arpcard.mcmaster.ca)
National network of laboratories that conduct standardized molecular typing of food-borne bacterial pathogens.

Permits rapid detection of clusters of indistinguishable strains.

PFGE is robust and proven, but low-throughput, labor-intensive, subjective, and dependent on cultures.

Non-culture based diagnostics for foodborne pathogens are increasingly common in clinical practice.
PFGE

O157:H7

Non-O157

Whole Genome Sequences

# of variations: ~106,872
Probing Diversity within Outbreak Clusters [1]

- Outbreak-associated cluster of O157:H7 STEC, childcare setting
- PFGE and MLVA: limited ability to discriminate within the cluster
39 parsimoniously informative SNPs identified. 
High-resolution discrimination of PFGE/MLVA-indistinguishable isolates.
Salmonella Heidelberg (May-July 2013)

- WGS-based analysis of putative multistate cluster of PFGE-indistinguishable Salmonella Heidelberg (XbaI: JF6X01.0022).
- Contemporary timeframe, similar presentations, possible link

![Genetic Tree Diagram]

**ALABAMA (Funeral)**
6-13 SNP differences

**NEW YORK (Childcare)**
10-13 SNP differences

**COLORADO (Family Gathering)**
6-9 SNP differences
Genomic Epidemiology on the Frontlines

Life Technologies
Ion Torrent PGM
(~$90,000)
Run time: 2hr-8hr
$1000-$1200/run

Illumina MiSeq
(~$125,000)
Run time: 27-40hr
$1250-$1500/run

For identification purposes only; does not constitute endorsement by CDC or HHS.
1. Begin with “area lab” concept.
2. Standardized sequencer platform(s).
3. Turnkey laboratory protocols and analytical pipelines.
4. Realtime data upload of raw or post-processed/extracted data to CDC/NCBI/Gov. Cloud.
5. Build capacity and scale up.

Cloud resources can rapidly become cost-prohibitive for heavy data access/manipulation.
Data Acquisition/Analysis Challenges

For PulseNet USA alone:

\[
>70,000 \text{ samples/year} \times 2 \text{ to } 3 \text{ GB raw sequence} + 5-10 \text{ GB intermediate}
\]

\[
\approx 0.9 \text{ petabytes of raw data/year}
\]

Transmission and storage?
Is better data compression the answer?
Distributed processing and extraction?
Is full WGS the right approach for large-scale surveillance?

Any solution must balance the advantages of WGS, with the costs of implementation.
Nationwide “Real-time” Listeriosis Surveillance using WGS

- Ongoing collaboration with FDA, NCBI, USDA and several state public health laboratories. Participation from several countries: France, UK, Denmark, Australia and Canada.

- Goal: Near “realtime” whole genome sequencing and analysis of isolates from all U.S. clinical cases of Listeria monocytogenes infection, as well as those from food/environmental sources.

- WGS within ~1 week of isolate receipt.

- Initiated: September 1, 2013.
Project Objectives

- **Improve the resolution and timeliness of cluster identification, and epidemiologic follow up.**
  - Faster, with fewer cases (e.g., increase sensitivity to 2-3 cases)
  - Strong hypotheses and better determination of scope

- **The use of WGS findings to help guide response efforts:**
  - Corroborate/refute alternative risk factors and exposures
  - Judicious use of resources during investigation/followup

- **Foundational infrastructure and methods for next generation PulseNet, and a generalizeable platform for genome-scale molecular epidemiologic surveillance.**
  - Develop and adapt epidemiologic tools to incorporate and visualize WGS-based data.
  - Determine optimal WGS-based strain typing strategy.
  - Assess technological and interpretive challenges.
Current Progress and New Challenges

**Public Health Impact:**
- Five confirmed clusters using WGS, including one that would likely have been missed or delayed by conventional methods (PFGE).
- Collaborative participation from state and federal partners.
- Real-time, high-resolution comparison of human, food and environmental isolates, interpreted within a sound epidemiological framework.

**New challenges:**
- Use and release of public health sequence data and associated metadata. Federal, state, local and international levels.
- Interpretive criteria: what defines a cluster?
- Data management: moving and analyzing data, standardizing informatics workflows, and integrating WGS data into investigation.
### MicrobeNet DNA Sequence Results

#### Sequence Data

Your query sequence has matched on the **GyrB locus**.

For my selection(s), use the

- GyrB (match) locus to
  - Align Sequences
  - BLAST

Include query sequence in alignment

Select Top… Clear

<table>
<thead>
<tr>
<th>Taxon</th>
<th>Strain #</th>
<th>Similarity Score</th>
<th>Identity %</th>
<th>E Value</th>
<th>Gyrb</th>
<th>16S</th>
<th>23S</th>
<th>rpoB</th>
<th>Actinomycetes Panel</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>Nocardi a arthriticid</em></td>
<td>DSM 44731T</td>
<td>2242.84</td>
<td>100%</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td><em>Nocardi a gamkensig</em></td>
<td>DSM 44956T</td>
<td>2033.65</td>
<td>97.6%</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td><em>Nocardi a asiatic</em></td>
<td>DSM 44083T</td>
<td>1705.44</td>
<td>93.7%</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td><em>Nocardi a beijingensis</em></td>
<td>W7678T</td>
<td>1533.3</td>
<td>92.8%</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td><em>Nocardi a amemensis</em></td>
<td>DSM 45056T</td>
<td>1633.3</td>
<td>92.8%</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td><em>Nocardi a aracensis</em></td>
<td>DSM 44720T</td>
<td>1584.61</td>
<td>92.3%</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td><em>Nocardi a asteroides</em></td>
<td>ATCC 19247T</td>
<td>1539.53</td>
<td>91.8%</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td><em>Nocardi a nealeonelondens</em></td>
<td>DSM 4471T</td>
<td>1521.49</td>
<td>91.3%</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td><em>Nocardi a brasiliensis</em></td>
<td>W287T</td>
<td>1400.88</td>
<td>91.3%</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
</tbody>
</table>
Nocardia asiatica

Overview

Nocardia asiatica (n. asiaticum of Asia, the source of the isolates). Aerobic, Gram-positive, partially acid-fast, non-motile actinomycetes forming a beige substrate. Substrate mycelia are fragmented. Aerial mycelia are straight to flexuous with a mean spore number of 2–20. Spores are cylindrical (0.3–0.5 μm x 1.5–1.7 μm). Colonies on BHI agar and on most other commonly used media, such as Sabouraud's dextrose agar, are very small compared with those of N. brasiliensis and N. asteroides type strains. The reverse side of colonies is yellowish. Colonies are 0.3–1.0 mm in diameter after 7 days at 30 °C on Mueller–Hinton II medium with 0.2% glucose. Glucose, acetate, rhamnose and citrate are utilized, but adipic acid, acetic acid, acetic acid, arabinose, arbutin, aesculin, erythritol, galactose, inositol, mannose and sorbitol are not. Adenine, casein, hypoxanthine, testosterone and xanthine are not hydrolysed. The type strain is a clinical isolate. It grows at 37 °C, but not at 45 °C; nitrate is reduced to nitrite. It does not have arylsulfatase activity. The G+C content of the DNA is 68.4–69.9 mol%. The type strain is strain IFM 0245T (=NBRC 100129T = JCM 11892T = DSM 44668). The five strains described were isolated from Japan and Thailand.

Roadmap:
1.0 Sequence-based searches (current)
2.0 Phenotypic searches and MLST (2012)
3.0 Whole genome comparisons (2013-4)
Fungal Meningitis (*Exserohilum spp.*)

Ana Litvintseva, Mary Brandt & Shawn Lockhart (NCEZID/DFWED)

- Not typically a human pathogen
- Strain typing methods for many fungi and moulds are limited
- Retrospective WGS: 23 outbreak and 7 comparator strains.
- ~34.7Mb genome

Image: http://phil.cdc.gov
What we’re doing:

- Assessing the importance of genetic drift on the antigenic and epitopic stability of measles hemagglutinin (H).
  - Analyzing 350 measles H sequences from PRC over a 17 year period.
  - Sequence alignment, comparison and structural modeling of measles hemagglutinin
  - Comparing mutational hotspots to functional domains, correlating to epi, vaccine coverage, etc.
Metagenomics of the Healthcare Environment
Dr. Margaret Williams (NCEZID/DHQP)

- **What we’re doing:**
  - Assessing the microecology of healthcare environmental samples and needleless catheter access devices using next generation sequencing
    - Bacterial: 16S V1 & V2 regions
    - Mycotic: Internal Transcribed spacer (ITS)
  - Comparing results to conventional, culture-based survey methods

- **Important implications for the prevention of HAIs, particularly catheter-associated infections.**
Metagenomic Survey of a Needleless Connector

Heterotrophic plate count: ~10CFU/mL → cultured *Paenibacillus & Bacillus* spp.

* 16S V1/V2 regions; 454 pyrosequencing
Bioinformatics in Public Health Fellowship

- Collaboration with APHL, Georgia Tech and Emory.
- In inaugural year, this program will place bioinformatics students and fellows from GT/Emory into CDC laboratory programs.
- In subsequent years, we hope to open this program to other schools, with placement at other CDC sites/PHL.
- Labs can apply for two categories of fellows:
  - Practical: a 4 month, semester-long project, possibility to extend.
  - Post doctoral: 12-24 month conventional postdoc.
- We expect to place 4 to 6 BPH fellows during 2014-5.
Questions?

For more information please contact Centers for Disease Control and Prevention

1600 Clifton Road NE, Atlanta, GA  30333
Telephone: 1-800-CDC-INFO (232-4636)/TTY: 1-888-232-6348
E-mail: cdcinfo@cdc.gov   Web: http://www.cdc.gov

The findings and conclusions in this report are those of the authors and do not necessarily represent the official position of the Centers for Disease Control and Prevention.