Bioinformatics and Public Health

Scott Sammons
Technology Officer
Office of Advanced Molecular Detection (detail)
Team Lead, Bioinformatics
NCEZID/DSR/BCFB

January 12, 2015
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
Human
Animal
Environmental
Samples

INFORM PREVENTION STRATEGIES
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.
Cost per Raw Megabase of DNA Sequence

Moore's Law

VOLUME OF RAW DATA

NIH - National Human Genome Research Institute

genome.gov/sequencingcosts
**Input: DNA/RNA**

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

**NGS**

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

**Bioinformatics**

**Workflow:**
- Hardware/software
- Specialized skillsets
- Algorithms/pipelines
- Pathogen databases
- Data analysis/interpretation/integration/visualization

**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
- Phage typing
- Bacteriocin typing
- Ribotyping
- Plasmid profiles
- Antibiogram Sequencing
- Rep-PCR
- DNA Microarrays
- SNPs
- Mass Spec
- Whole Genome Sequencing

- PFGE
- MLST
- MLVA
- AFLP
- PAMPs
- MEE
- RAPD
- REA

* Slide adapted from: Peter Gerner-Smidt
Applying Pathogen Genomics to Public Health Microbiology
MALDI-ToF Microbial Identification

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

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

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

Direct spotting or simple extract
Overlay with matrix solution
Air dry

Load and Run.

For identification purposes only; does not constitute endorsement by CDC or HHS.
Tuned m/z profile of specific high-abundance proteins (eg: ribosomal)
Results within seconds. Similar workflows for a wide range of microorganisms. Databases.
Next-Generation Sequencing Applications

Shotgun / Whole Genome Sequencing: random shearing of DNA, even sequence coverage over entire genome.

Amplicon or Deep Sequencing: ‘Massively parallel’ sequencing not only produces throughput, it provides sequences of potentially millions of individual molecules (instant cloning). By sequencing a PCR reaction it allows the detailed search for low expression quasi-species or mutations which may signal growing drug or vaccine resistance.

Metagenomic Sequencing: Complex clinical, laboratory or environmental samples can be sequenced to provide a diagnostic ‘snapshot’ of the resident organisms.
Reference-Guided (Mapped) Assembly

ADVANTAGES: Relatively fast, well-suited to highly-conserved genomes.

DISADVANTAGES: Issues with high diversity, mobile elements, linear reference

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

2
A C T A G T T

3
T C T A C T

ACTAGA
ACTAGT
TCTACT
K-mer Based Sequence Typing

1. ATCGCTATCGTAGGGCATTTTACTGTGCTGTGTCGATGCGAGA
2. ATCGGTATCGTAGGGCAATTACCTGTAGCGTGATGCGATGAGA

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 (0,1,1,0,1,1,0,1,1,1,1,1,0)
NCHHSTP/DTBE: Whole Genome Sequencing TB

72 Isolates
102 SNPs
41 Genomic types
4 Clusters

Branch length = Number of differences between 2 genomic types
Circle size = Number of isolates within a genomic type

Slide courtesy: James Posey (NCHHSTP/DTBE; CDC Public Health Grand Rounds)
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.
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

**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 samples/year

x

2 to 3 GB raw sequence + 5-10 GB intermediate

~0.9 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.
Identification of Yellow Fever by next generation sequencing

Dr. Laura McMullan and Dr. Stuart Nichol (DHCPP/VSPB)

November 23, 2010
Mystery Ebola-Like Illness Kills 13 in Uganda

Gregory Branch, Global Post

KAMPALA, Uganda — Thirteen people have died in northeastern Uganda from a mysterious disease. The World Health Organization and the Centers for Disease Control are investigating the outbreak of the disease that appears to be similar to the Ebola and Marburg hemorrhagic viruses.

Twenty cases of the frightening illness have been reported in the remote Abim district. Due to the number of cases reported, the district has been labeled the epicenter of what many are calling it the “Abim disease.”

Of the 20 cases in Abim, eight have died, two are quarantined and another 10 have been discharged after their conditions improved, according to Emmanuel Okech, a health official in Abim. Five more died in the neighboring Agago district.

The victims usually...

TAGGED: mystery illness, uganda, Ebola virus

Bubonic Plague Suspected as Cause of Ugandan Outbreak That Has Killed 38

By Fred Ojiambo - Oct 3, 2010 11:14 AM ET

Bubonic plague is suspected as the cause of a disease outbreak that has killed 38 people in northeastern Uganda in the past month, the Ministry of Health said.

“Plague is not something new in the country,” James Kakozza, junior health minister, said today by phone from the capital, Kampala.

The ministry as ruled out Ebola, Marburg, Rift Valley, typhoid and Congo Crimean fevers as the sources of the illness, which was first reported on Nov. 10. The deaths, which took place up to Dec. 3, are among 91 cases recorded in the six districts in the north. More recent figures aren’t available, Kakozza said.

A team of Ugandan health officials and experts from the U.S. Centers for Disease Control and Prevention are working closely to contain the outbreak, Kakozza said. Drugs to treat bubonic plague are already available in the country because it has experienced the disease previously, he said.

The ministry said the outbreak is in Acholi, a region wracked by a two-decade insurgency by the Lord’s Resistance Army.

Bubonic plague affected the country’s northwestern region of West Nile in November 2008, according to the ministry.

To contact the reporter on this story: Fred Ojiambo in Kampala via the Johannesburg bureau at franrichardsen@bloomberg.net.

To contact the editor responsible for this story: Anthony Squazzin at asquazzin@bloomberg.net.
Amplicon (deep) sequencing project
Dr. Yu Li and Dr. Inger Damon - NCZEID/DVRD/PRB

- Clinical case of progressive vaccinia infection from smallpox vaccination of an immune compromised patient
- Pox antiviral ST-246 administered which targets pox gene F13L, a major envelope protein which mediates production of extracellular virus
- Oral ST-246 given daily and vaccination site sampled over 3 week period
Control swab prior to ST-246
2 weeks after ST-246

C → T
869

T → A
943
3 weeks after ST-246

C > T
869

T > A
943
Sequencing Capacity at CDC

- **DSR core sequencing lab**
  - Ion Torrent PGM
  - Illumina MiSeq
  - Illumina NextSeq 500
  - Illumina HiSeq 2500
  - 2 Pacific BioSciences RSII
  - Oxford Nanopore MinIon (early access)

- **OID program laboratories**
  - ~5 x Ion Torrent PGM
  - ~20 x Illumina MiSeq
  - 1 Pacific BioSciences RSII
  - 1 x Roche 454jr
High Performance Computing

- **OID Scientific Computing HPC**
  - 2 HPC clusters (ASPEN, BIGSKY)
    - ~1200 CPU cores, ~33000 GPU Cores; 20TB available RAM
  - 4 PB storage array
  - >50 physical servers / >100 virtual servers

- **Expanded access to scientific workstations, site licenses and training for Linux, HPC and specific bioinformatics apps.**
Planned System Outage for Scicomp Servers starting on Dec 6th at 4:00 pm to 9:00 pm Dec 8th

Posted on December 5, 2013 by Dhwani Bhatia

Time and date: 4:00 pm on December 6th 2013 – 9:00 pm Dec 8th 2013

Affected Servers: All biotech.cdc.gov servers

Contact: nextgensupport@cdc.gov

The Scicomp Servers will be undergoing some hardware upgrades that will affect the entire server farm. This upgrade is being performed to add redundancy to our network and will bring stability and improved performance. Over this period some systems will be available till the morning of Dec 8th.
Fellowship Opportunities at CDC

Bioinformatics in Public Health Fellowship Program

Preparing bioinformaticians for careers in public health and designing bioinformatics resources for laboratorians

The Bioinformatics in Public Health (BPH) Fellowship aims to train and prepare bioinformaticians to apply their expertise within public health and design tools to aid existing public health personnel in the use of bioinformatics. The BPH fellowship program was developed to meet the need for expertise in the analysis of the vast amounts of genetic data that are generated through molecular sequencing techniques and to harness that data to solve public health problems. In the fellowship’s current form, fellows will be paired with a complimentary CDC infectious disease laboratory which will serve as a host laboratory. In future iterations, the fellowship will be expanded to provide the opportunity for public health laboratories to serve as host laboratories.

http://www.aphl.org/mycareer/fellowships/bioinformatics/Pages/default.aspx
Metagenomics of the Healthcare Environment

Dr. Margaret Williams (NCEZID/DHQF)

- **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

* 16S V1/V2 regions; 454 pyrosequencing
Comparative Genomics: A Critical Starting Point

Characterization
- Virulence factors
- Resistance genes
- Phenotypic markers
- Signatures

Diagnostic dev/validation.

Transcriptomics and proteomics

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

Diversity and phylogenetics.
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