Genomic Molecular Epidemiology Workshop Discussion Topics

Sept. 24-25

Establish a working group for each topic. Each group should have at least two chairs. Make a list of all members/chairs. Prioritize the most important obstacles that need to be solved. Provide a 6-month timeline for completing each goal (provide names and goals).

Topics

1) Implementation of sequence and meta-data publication standards. Now that we have established guidelines for sequence and meta-data standards how do we go about rolling them out?

a. Create list of journals/organizations to contact (ASM).

b. List of people/labs who are depositing meta-data and sequences and establish collaborations in order to test and develop the system? For each lab, how many draft genomes/metadata will be deposited within the next 6 months?

c. Create a website for publishing metadata standards. (NIH?)

d. Standards for masking location data for submission of sample metadata.

e. Standards for masking patient data for submission of sample metadata.

f. Standards for all steps through to analytical conclusions

g. REPRODUCIBILITY

h. What should be done with data that does not meet the standards? Should the standards be mere guidelines and the level of quality and uncertainty from each step simply be accumulated for final decision making?

2) Policy challenges (political, legal, and global health diplomacy) to sharing genomic data on a global scale:

a. Global Identifier (Digital Immune Surveillance System):
   - What are the threats/opportunities in building 'Global Identifier' capacities at the national resp. International level?
   - Would the Global Identifier stand a chance as preferred method / system related to IHR?

- Which political problems could arise keeping this identification tool fully at open source?

b. Create a list of relevant comparable initiatives in relation to sequence sharing between countries (including human, plant, animal DNA sequences)?

c. Set up partnerships between sequencing capable countries and non-sequencing capable countries.
   - Outline partnerships that already exist (i.e. US military bases in foreign countries that have collections of food-borne illness samples).
   - Propose list of new partnerships
   - Estimate of how many non-USA isolates will be sequenced/submitted in the next 6 months.

d. Which models for financing major global initiatives of this nature could be described/listed?
   - Create list of funding sources (i.e WHO, Gates)
   - Locate meetings/conferences within the next year for networking with large funders
   - Set targets to apply for funding (people, deadlines, etc.)
   - Make list of active members at this meeting who would like to participate in joint grant submission.

e. e. Other incentives to encourage submission, openness, and high quality data nationally and internationally

3) Lessons learned regarding clinical utility and rapid detection

a. Agree on website for distributing/publishing working standards for WGS collection.
   - Host?
   - Editing permissions?

b. Define standards for clinical-grade genome sequencing of pathogens, including quality scores, coverage, calling of clinically-relevant findings per presence/absence of a gene vs SNP or other variant, etc. Can this be done in 6 months? Outline steps towards this goal.
   - List priorities of which parts need to be addressed first.
   - Build a timeline for next 6mo / year.

c. Privacy issues regarding communication of pathogen sequences into public databases. Namely while the pathogen itself is generally not considered "protected health information" what additional components of the clinical record might be allowable to enhance publicly available datasets.
   - HIPAA is fairly clear in what would be allowed or not, but it would help to simplify what contributing institutions need to go through to be able to contribute (similar to DB-GAP, or ClinVar, etc).
   - Publish working list of allowed/not-allowed metadata and format. What will be recommended by this body?
   - Define working group to upload and work with NCBI to build a clinical-grade database.

d. Tie-in of clinical sequencing activities with surveillance efforts via Depts of Health, CDC, food monitoring, etc.
   - Identify partners who will collaborate

    - Set 6-month goals to collect and upload data to NCBI (or partnering public database).
    - Start list-serve, wiki, or google group to help with communication.

e. What resources - US, international, etc. - would help support efforts to leverage pathogen-based sequencing in clinical diagnostics.
    - clinical-grade knowledge base, others. Who will build and compile this database?
    - utility of integrating phenotypic data with genotypes/genomes
    - tools to access/mine the information for developing or updating knowledge bases maintained in laboratories, NCBI and/or by vendors.

f. Plan round-robin blinded validation study. Need specifics.
    - What experiment will be first?
    - Who will participate?
    - Who pays?
    - Plan timeline for highest priority test to accomplish in 6 months.

4) Sources of error and standards

On the path from sample collection to sequencing, assembly/mapping and variant detection and determining isolate relatedness there are a great many possibilities for error to be introduced both natural and unnatural. Do we understand enough about the steps to understand all sources of error and how to minimize them in order to apply confidence levels to the conclusions drawn from the analyses that may take place during outbreak analyses? If not, then what areas need to be addressed? A first step should be to list the steps where error can be introduced and determine steps that already have guidelines, recommendation, or calibration points in order to assign measures of certainty and those which do not and for which standards should be developed.

4a) Laboratory - Sources of error and standards for bacterial sequencing

a. SOPs and standard workflows for sequencing.
    a. Form working committee to start development of SOPs for each NGS platform.
    b. Identify website to make SOPs public and to share among collaborators.
    c. Path to update the SOPs? Outline experiments to test how alternative SOPs perform compared to each other.

b. Standards for sample collection – Do the FDA and clinical centers have standardized sample collection protocols, from clinical samples, from environmental sites, industrial sites?
    - How will we ensure the sample was collected properly when all data is electronic? Will there be repositories for the sample itself? cultured? uncultured?
    - Design experiment to show whether collection method matters
    - standards for clinical sample collection and deposition, are the samples made available?
    - what about biothreat agents? are the samples available?

c. Standards for sequencing – this is being worked on by a separate group
    a. List groups that are actively working on standards. Contact members and form joint collaboration. (i.e. the National Institute of Standards and Technology (NIST), Clinical

Laboratory Improvement Amendments (CLIA) has standards for clinical sequencing in humans. Are there others?)

b. Should we adopt another standard or create our own?
c. National vs. international standards?
d. Design standards that are platform agnostic.
e. How will the standard be distributed? Who will pay?
f. Will there be a standardized synthetic spike-in sample to measure the sequence quality and variation (calibration step)?
g. Design plan to generate initial standard. Develop timeline for creation and dissemination of standard with NIST and Eurasian counterparts

4b) Bioinformatics - Sources of error and standards for bacterial identification and clustering

a. Design experiments to test all manner of data analysis. Define priorities for 6 months.
b. Website to distribute/publish bioinformatic standards?
c. Standards for assembly – besides L50 and N50, what are the standard ways to assess assembly quality, ie. in eukaryotes transcript coverage is used. NCBI is using RefSeq proteins for protein coverage and frameshift analysis to assess assembly quality but are there other standard measures to use?
d. Design tests to determine which pipelines are most reproducible. Look for partners who wish to test and develop all aspects of data analysis pipelines.
e. SNP calling
   - Current NCBI and FDA/CFSAN standards,
   - the 1000 Genome Project has guidelines for variant calling
   - Design test to check what variation arises when parameters are modified.
f. What are the gold standard set of finished pathogen genomes (NIST)? What should be done with non-gold standard genomes.
g. Reference Mapping
h. Clustering/Phylogenetic Methods
i. Empirical thresholds for an isolate to isolate “match ” or “non match”. Is there enough known about existing sequences to consider thresholds for matches? If not, then what more needs to be done? More reference sequences? More diversity? More real-time experiments? There are different levels of diversity within different bacterial species. Is it essential to know this diversity beforehand in order to confidently assess sequence similarity? ex. B. anthracis and other recently evolved pathogens show remarkably little diversity, whereas E. coli varies by as much as 20% of genomic content.
j. Validation of SNPs (show that SNPs actually exist)
   - preliminary results from targeted Sanger sequencing?
   - List of current validation studies.
   - the 1000 Genome Project has a set of targeted and validated variants and a method for detection

- Design collaboration to collect SNP data on known genomes for validation and testing. Current work at FDA and NIST 16s projects are both beginning and need volunteers.

5) Computer resources/tools required for a global genome-based disease detection network

a. We need open source tools to perform data QA and submission for NGS of human bacterial pathogens. There are also vendor tools for sale that could be used to assist in data analysis.
   - List of tools available or near-available
   - Define hubs of distribution. Find volunteers who will monitor these portals.
   - List of bioinformatic-center partners (ie, NCBI, EBI, DDBJ . . .)
   - Define and plan where approved database of Gold Standard sequence(s) is kept.
   - Timeline for when software will be made available for distribution, testing.

b. FDA/CFSAN is in discussions with 3rd party bioinformatics software vendors to develop plug-ins for QA/QC desktop sequencing of bacterial isolates. The output of this software should be a data object (reads + metadata) that can be directly deposited to NCBI/SRA (and partners). Implementation will be discussed with NCBI.
   - Timeline for rollout of plugin. Solicit beta testers for software and upload of genomes to NCBI.
   - Working group to test new software.
   - Error checking capability that matches NCBI's internal checks

c. NCBI's pathogen resources
   - Streamlining NCBI submission process
   - Ability to track the submission progress
   - Timeline for public release of submission pipeline
   - Integration with other pathogen resources?
   - List collaborators who are willing make their data public for software testing.
   - What does NCBI need to develop these resources? Timeline for meeting NCBI's list of needs.

d. Reports from analytical pipelines that fit end users needs
   - Prioritize what is needed in a report.
   - What file types will users want back from the pipeline? (i.e. assemblies, annotation files, list of excluded sequences, GenBank files, Popset of variable genes, SNP matrix, etc.)
   - Reports from pathogen pipeline:
   - the following have been requested as outcomes from the NCBI pipeline:
   - 1) assembled genome
   - 1a) coverage plots of reads vs. reference
   - 2) FASTA format of extracted SNPs for end users to do their own analyses
   - 3) phylogenetic tree/clustering in newick format
   - 4) SNPs defining the outbreak lineage (what about other variants?) – with flanking sequence for diagnostic PCR assays

- 5) full annotation of each isolate
- 6) identification of mobile elements
- 5a) phage

- 5b) plasmids
- 7) known virulence/pathogenicity/resistance factors
- What other reports are required? Are there different reports for different scenarios? For clinical surveillance? For environmental surveillance? For outbreak scenarios?

e. Timeline for reporting outcomes, especially during outbreak scenarios? Minutes, hours, days, weeks? What are the bottlenecks in current analyses from sample collection through to variation calling and clustering?

e. Industry and venders: standards and guidelines for developing platform specific tools.
   - What are the standards? Create working group to outline standards. Prioritize standards that do not exist yet.
   - Where will the standards be posted? Define plan for dissemination of report.
   - Who will set and curate the standards?
   - Recruit collaborators to submit data to industry and public databases.
   - Work with NCBI regarding file formats for uploading data.

Metadata

At the October workshop at NCBI, a proposed set of minimal metadata requirements were set forth, discussed, and then implemented in the NCBI Biosample as the pathogen template. The minimal (required) set is:

1) Sample Category
    - 1a) Clinical/Host-associated
        - 1a1) specific_host - Organism name of natural host or disease target
        - 1a2) isolation_source - Describes the physical or environmental location from which the sample was derived.
        - 1a3) host-disease - Name of relevant disease, e.g. Salmonella gastroenteritis. Controlled vocabulary, http://bioportal.bioontology.org/ontologies/1009

    OR

    - 1b) Environmental
        - 1b1) isolation_source- Describes the physical or environmental location from which the sample was derived.
2) sample_name - A unique identifier for the sample. Equivalent to local sample ID/freezer ID
3) strain/isolate – The particular strain/isolate of the sample (should match sample name)
4) organism - Scientific name of the sample
5) collection_date - Equivalent to your "Year of culture"
6) Geographic location
    - 6a geo_loc_name - Geographic location in the form "Country:locality"
    - 6b lat_lon - Latitude and longitude in decimal degrees
7) collected by - Name of the person or organization who collected the sample.
8) additional attributes

The appendix lists detailed information for metadata.

References.

100K Genome Project.
http://www.foodsafetynews.com/2012/07/genome-sequencing-of-100000-foodborne-pathogens-underway/?utm_source=newsletter&utm_medium=email&utm_campaign=120713
http://100kgenome.vetmed.ucdavis.edu/
Genome in a bottle consortium
http://www.nist.gov/mml/biochemical/biomolecular/genome_in_a_bottle_consortium.cfm

NIST has organized the "Genome in a Bottle Consortium" to develop the reference materials, reference data, and reference methods needed to assess performance of human genome sequencing.

Clinical Laboratory Improvements Amendments – assure quality of clinical laboratory testing
In the United States, all clinical laboratory testing performed on humans is regulated by the US Centers for Medicare & Medicaid Services in Baltimore, Maryland, through the Clinical Laboratory Improvement Amendments (CLIA).
http://www.nature.com/nature/journal/v482/n7385/full/482300a.html

Next-generation sequencing data interpretation: enhancing reproducibility and accessibility. Nekrutenko A., and Taylor, J. PMID: 22898652

Routine Use of Microbial Whole Genome Sequencing in Diagnostic and Public Health Microbiology Koser, C., et al., 2012. PMID: 22876174

Performance caomprison of whole-genome sequencing platforms.
Lam, HYK, et al., 2012. Nature Biotechnology. PMID: 22178993

Evaluation of next-generation sequencing software in mapping and assembly. Bao, S. et al. 2011. J. of Human Genetics. PMID: 21525877

Assemblathon 1: A competitive assessment of de novo short read assembly methods. Earl, D. et al., 2011. Genome Res. PMID: 21926179

GAGE: A critical evaluation of genome assemblies and assembly algorithms. Salzberg, S., et al., 2012. Genome Res. PMID: 22147368

A tale of three next generation sequencing platforms: comparison of Ion torrent, pacific biosciences and illumina MiSeq sequencers. Qual, M., et al., 2012. BMC Genomics. PMID: 22827831

Performance comparison of benchtop high-throughput sequencing platforms. Pallen, M. 2012. Nature Biotechnology. PMID: 22522955

Can an infectious disease genomics project predict and prevent the next pandemic. Gupta, R. 2009. PLoS Biology (Google.org). PMID: 19855828

The role of genomics in the identification, prediction, and prevention of biological threats. Fricke, W.F., et al., 2009. PLoS Biology. PMID: 19855827

Are diagnostic and public health bacteriology ready to become branches of genomic medicine? Pallen, M., and Loman, N. 2011. Genome Medicine. PMID: 21861847

Bacterial epidemiology and biology--lessons from genome sequencing. Parkhill, J., and Wren, B. 2011. Genome Biology. PMID: 22027015

High-throughput sequencing and clinical microbiology: progress, opportunities and challenges. Pallen, M., et al., 2010. Current Opinion in Microbiology. PMID: 20843733

High-throughput bacterial genome sequencing: an embarrassment of choice, a world of opportunity. Loman, N., et al., 2012. Nat Rev Microbiol. PMID: 22864262

APPENDIX

METADATA DEFINITIONS AND PERSPECTIVES

The minimal set (required) for NCBI Biosample is listed here as well as some optional fields. Current implementation is that 'MISSING' is allowed for the required fields for which no data is available (ex: the country and city are known, but exact GPS coordinates are not)

(1) Source of Strain

This category of metadata is also absolutely critical for any sort of investigational insight to be gleaned from a WGS database. In fact, this meta-entry is fundamental to outbreak surveillance and without it, it would be impossible to ascertain which genomes represented clinical submissions versus which represented potential source genomes including food and environmental sources. The challenge will be to find a simple and consistent way to collect a wide array of different "sources" with some degree of specificity.

The current implementation of this is in two basic categories: (pathogen samples are either Clinical/Host-associated OR environmental)

1a) clinical/host-associated

– used to distinguish isolates that came from a living organism, a stool sample from a human in a clinical setting, a leaf or vegetable/fruit from a plant growing on a farm. This simplified distinction versus:

1b) environmental

is to distinguish a sample that comes from an industrial food processing site, a sink in a hotel, a fork in a restaurant, a batch of spices, or even a cut piece of spinach pulled from the shelf of a food market – all of which are no longer living "hosts" and all of which could have been contaminated through various processing steps.

Within both categories there are further fields that help define the sample source.

(1a1) Host

This category is the specific scientific name of the living organism from which the sample was taken in the clinical/host-associated category. Homo sapiens for clinical samples, Solanum lycopersicum for a sample from a tomato plan. These are expected to be scientific names as the organism name will be checked in the NCBI taxonomy database.

(1a2 and 1b1) Isolation source

The exact isolation source is a descriptive field used to further delimit the source. “Stool” or “blood” for a human stool or blood sample in a clinical setting. “Sink” or “salsa” for an enviromental sample from a kitchen sink, or a salsa product in a restaurant. NCBI is looking for dictionaries to validate the submitted text description against, but the number of certified standardized dictionaries to describe food products or ontologies to describe human anatomy are problematic. Nevertheless some do exist, such as the Medical Subject Headings (http://www.ncbi.nlm.nih.gov/mesh), and will be utilized.

1a3) host-disease

This field captures the specific disease afflicting the host from which the sample is. Vocabulary NCBI MeSH http://www.ncbi.nlm.nih.gov/mesh or http://bioportal.bioontology.org/ontologies/1009

(2) Sample/Strain/Isolate name

This field is absolutely required for all NCBI Biosample submssions. The sample is the unique laboratory or field identifier that distinguishes this sample from any other sample and is expected to be unique from a given submitting center for each unique sample (in other words, sample #1 submitted once and then a second time would be a replacement of the record). The isolate or strain name is key as part of a minimal metadata set as it distinguishes the isolate strain from others of the same species. Ideally this would be the same as the sample name, especially when large numbers of samples are collected by the same center for novel strains never before seen. For existing strains such as the classical Escherichia coli K-12 MG1655, then the sample name will not match the strain name. The name ideally will be a unique link back to additional information retained by the submitter. The name is of great importance and if a strain is implicated in a disease cluster, the name will hopefully provide a portal to additional key information that can be transmitted between the submitter and an investigator. These names should also be the names that are attached to the isolate during its lifetime in the WGS database much like these names are now for conventional GenBank submissions. As such, when publications or reports are generated using the adjoining WGS data, the strain/isolate name will be linked from the Biosample record to any sequence data linked to that record.

(3) Organism

The organism is also fundamental and refers directly to the taxonomic identification of the submission (Genus species). Without this essential information concerning the genus/species of the pathogen itself, it may be difficult to track and cluster potential outbreak swarms as outbreaks, by definition, involve, the same pathogen or groups of pathogens. The organism names will either match existing NCBI taxonomy IDs or new ones will be created where necessary (ie. novel organisms can be assigned new taxonomy IDs)

(4) Collection-date (Date of Sampling)

Date remains another fundamental entry into the meta-data fields. Date is as fundamental as source and carries the same weight as geography. In many cases, date is of utmost importance as the date alone can reveal whether an isolate is part of an outbreak cluster or not. Moreover, date can inform the investigator as to whether an outbreak strain has re-emerged or caused illnesses before. Moreover, for the kind of comparative genomic approaches that will be generated from a global WGS database, date allows us to deploy closely related reference and control strains to an analysis without having to be concerned as to their actual involvement in an active outbreak event. Of course, more intuitively, date provides the immediate benefit of being able to define the time course of an outbreak from early onset to the tapering off period of the same event. One cannot overstate the importance of 'date' as an entry into the meta-fields of a WGS submission.

(5) Geographic origin of strain

This information is critical for entry into the NCBI WGS Disease Detection Database. Geographic information linked to matching or closely linked WGS entries will afford a public health investigator the ability to explore potential disease clusters in certain parts of the US, the entire US, and the world. Without this information, a query of the database will not be able to reveal the often essential geographical component that can be important when establishing an outbreak source. This is particularly true for foodborne outbreaks where clusters, identified by matching WGS or other molecular information, comprise clinical strains from the same general geographic locale. This geographical information will be key to traceback to a common food vehicle isolate or to a particular place or environmental source where all of the affected individuals may have visited in common. Conversely, the importance of geographical metadata is also emphasized in cases where seemingly unrelated cases are linked by common WGS-based genome profiles. This would allow investigators to determine the scope of an outbreak and how widespread a pandemic may have become. Also, in the case of food and environmental isolates, accompanying geographic information will be invaluable in providing links back to specific farms or other food production environments. In the case of environmental strains that link back to an outbreak cluster, more careful scrutiny of a region or the ecological features associated with a particular region will be permissible if the proper geographic metadata is present. Finally, data of these nature will be key to aligning outbreak cluster detection with GIS and other "heat-map" style data that can provide even greater insight into the origins of specific outbreaks or emerging pathogens.

In order to be able to consistently collect geographic origin data from around the world, three data fields may be required:

Country
State or sub-region
County (preferred for USA) or city

The implementation capture of geographic location is either GPS coordinates or text consisting of Country:locality

(6) Collected_by

This indicates who collected the sample. Especially important for large centers with many samples and when the sequence data is not submitted by the same person as the sample collector.

(7) Additional characteristics for specific organisms

There are numerous other sample attributes that can aid in distinguishing different samples and providing more information on the sample itself or the host from which the sample was obtained (in the case of clinical/host-associated). Many optional attributes are already implemented in the NCBI Biosample database and include those from the Genome Standards Consortium (http://gensc.org/gc_wiki/index.php/MIGS/MIMS). A few examples are listed but the full list can be found on the Biosample submission page: https://dsubmit.ncbi.nlm.nih.gov/

host_subject_id – a deidentified unique identifier for the host from which the sample was obtained and can be used to group multiple samples from different body sites.

serovar – for those organisms that have serological naming schemes – O157:H7 for E. coli or Typhimurium for Salmonella enterica subsp. enterica

lab_host & passage_history – for viral samples

culture_collection – the specific culture collection identifier if the sample has been deposited at a culture collection (ex: ATCC 12345 or DSM 12345) antibiotic resistance scheme – NCBI is working with clinicians and clinical centers to capture resistance data for each compound

Fields	Description
Antimicrobial	antimicrobial compound (rifampicin, streptomycin)
Interpretation	general interpretation of testing results (resistant, susceptible)
Test Method	general method for testing (disk diffusion, MIC)
Vendor	company producing testing method
Platform	testing platform
Reagent	specific type
Numeric Result	result of the test (inhibition diameter, concentration)
Units	units of the test (mm, micrograms)
Comment	general comment on the test

culturing_laboratory, IRB_protocol, CLIA_certified – for us by clinical centers to capture information on the exact laboratory, the Institutional Review Board protocol, and whether the center asserts that they are CLIA certified or not – NCBI will not verify this information but will provide links to a list of CLIA certified labs as is done in the NCBI Genetic Testing Registry (GTR) - http://www.ncbi.nlm.nih.gov/gtr/