NEW YORK STATE OF OPPORTUNITY. Department of Health | Wadsworth Center

NCBI National Center for Biotechnology Information

Using the NCBI Pathogen Detection Portal to Aid in Surveillance of Enteric Pathogens

Bill Wolfgang
Bill Klimke
Samantha Wirth

March 8, 2018

william.wolfgang@health.ny.gov
samantha.wirth@health.ny.gov
klimke@ncbi.nlm.nih.gov

Outline

- Why are we hosting this webinar? (Bill W.)
- Introduction to NCBI Pathogen Portal (Bill K.)
- Demonstration of NCBI Pathogen Portal (Samantha)
- Wrap up (Bill W.)
- Questions and discussion

Why are we hosting this webinar?

On a January 12 call with our PN regional labs we asked some questions.

- Do you plan to stop PFGE on *Listeria*?
  - 9 of 10 will stop
- Do you have any concerns with local cluster detection?
  - yes
- Do you plan to use the NCBI Pathogen Browser?
  - 7 of 10

Why we like the NCBI Pathogen Portal

- It is relatively easy and very fast to use
- The outputs are straight forward to understand
- It provides a common analysis tool
- Linked to a huge number of samples
- The portal is evolving

Some things to keep in mind….

- Portal is evolving
- Trees will evolve
- Sometimes samples have very limited metadata

How NYS uses the Pathogen Portal

- Augment our in house pipelines
  - Detect clusters in strains for which we do not have pipelines
  - Detect out of state isolates that are part of in-state clusters
- Quick peek to see if anything matches
- Prescreen requests from our epidemiologists before we forward to CDC

Introduction to NCBI Pathogen Portal (Bill K.)

https://www.ncbi.nlm.nih.gov/pathogens/

Live Demo of NCBI Pathogen Portal (Samantha)

https://www.ncbi.nlm.nih.gov/pathogens/

Wrap up

- The Pathogen Detection Portal at NCBI is a powerful tool for cluster identification.

- Its is a product that is still in development.

- What happens after PulseNet fully transitions to wgMLST - BioNumerics based surveillance?

New York Integrated Food Safety Center of Excellence (NY CoE)

WGS Training in Foodborne Disease Outbreak Investigation

- Over the last year, the CoEs, led by the NY CoE, have provided training on the application of whole genome sequencing (WGS) in foodborne disease outbreak investigation with a focus on the training of epidemiologists.
- CoEs collaborated with CDC to develop four short on-line modules on the basics of WGS, which are available at https://nyfoodsafety.cals.cornell.edu/molecular-epidemiology/modules.
- More in-depth training was provided by a series of four webinars which were recorded and are available at https://nyfoodsafety.cals.cornell.edu/molec epidemiology/webinars.

Integrated Food Safety Centers of Excellence
www.CoEFoodSafetyTools.org

NEW YORK STATE | Department of Health | Wadsworth Center

Comments / Questions / Discussion

NCBI Pathogen Detection Pipeline

William Klimke

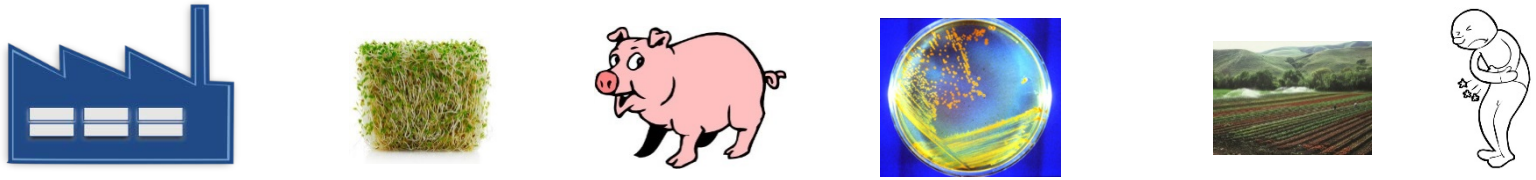
APHL Webinar, March 7th

Analysis goals

1. Are these isolates clonally related?

2. What is the anti-microbial resistance gene repertoire of this isolate

Shared Network pathways and data streams for outbreak detection and investigations



Sampling of clinical/human, food, animal, and environmental bacterial isolates

GenomeTrakr

PulseNet

USDA FSIS labs

State agriculture & food regulatory labs

FSIS & FDA labs food, environmental

State health labs (clinical)

Raw Genomic Sequence data

NCBI Submission Portal

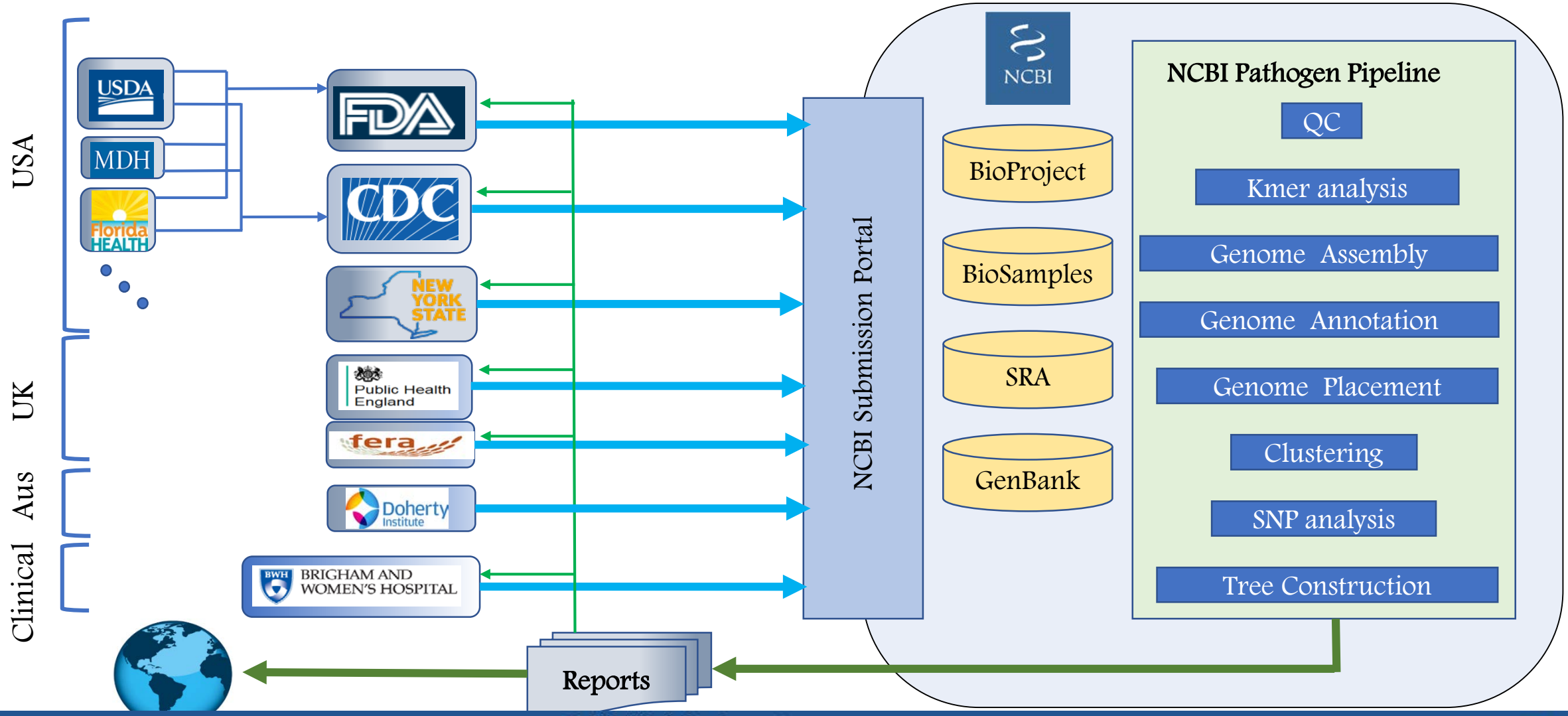
PulseNet National Database

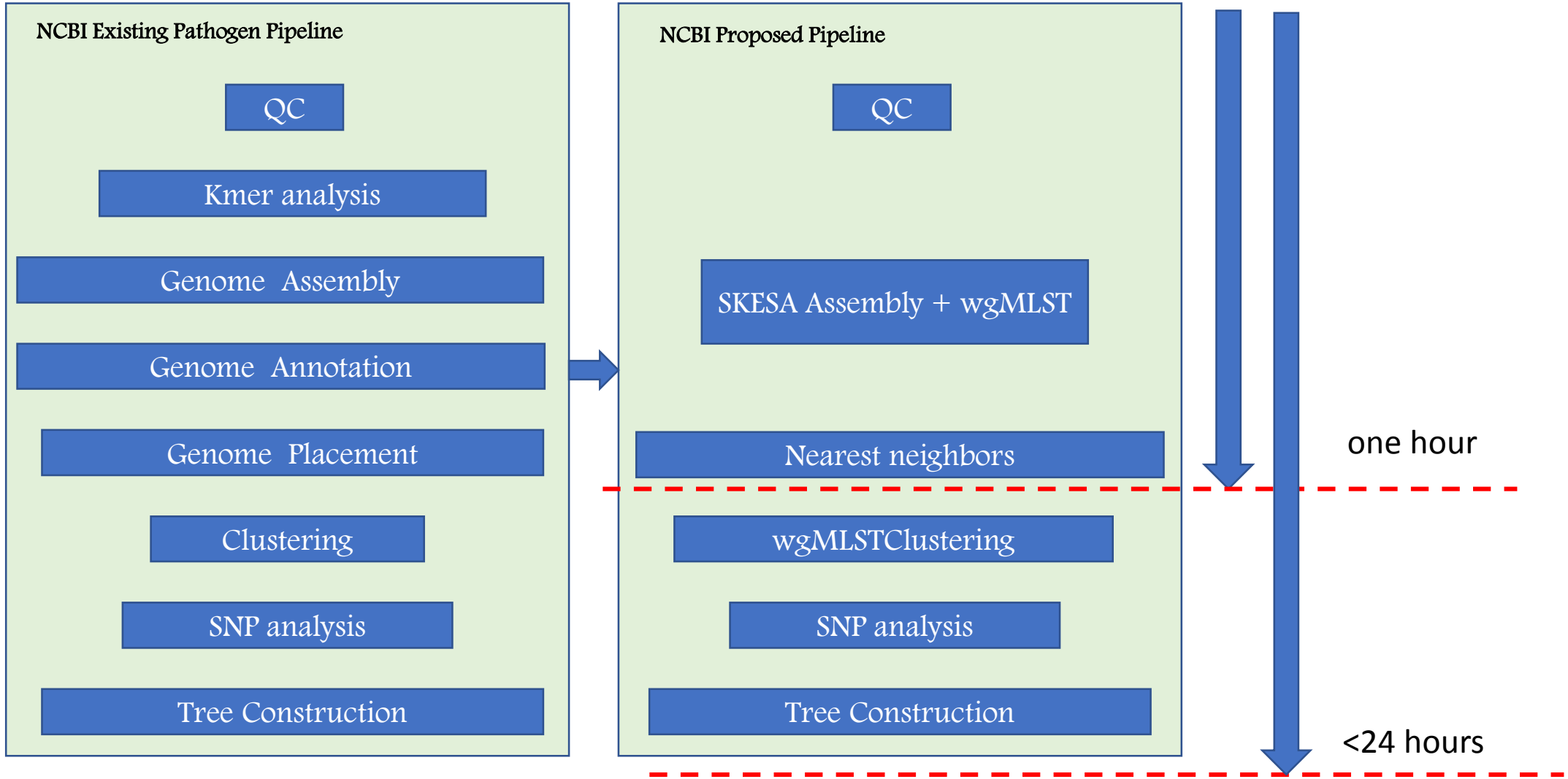
CDC analysis (all raw sequ goes into NCBI

Minimal metadata, everything publically accessible

More extensive metadata, analyzed seq data (tools to translate) data shared among PulseNet labs only

NCBI Pathogen Detection Pipeline Submissions and Analysis





- SKESA de novo assembly + wgMLST will replace and speed up several parts of the existing pipeline
- delivery of nearest neighbors within one hour of data deposition into SRA

Nearest neighbors (rapid reports)

Rapid reports are reported per day for a set of Bioprojects for Salmonella and Listeria

Report nearest 5 neighbors and all neighbors <6 allele differences

Report allele differences, loci in common and SNP accession (if exists)

Put in a tab-delimited file per run

ftp://ftp.ncbi.nlm.nih.gov/pathogen/Results/Listeria/Rapid_reports/
ftp://ftp.ncbi.nlm.nih.gov/pathogen/Results/Salmonella/Rapid_reports/

Example: Rapid Report for SRR6110457

#biosample_acc	run_acc	neighbor_biosample_acc	neighbor_run_acc	neighbor_assembly_acc	neighbor_cluster	alleles_different	loci_in_common
SAMN07714220	SRR6110457	SAMN07714122	SRR6109644	NULL	NULL	1	4269
SAMN07714220	SRR6110457	SAMN07714119	SRR6109643	NULL	NULL	2	4270
SAMN07714220	SRR6110457	SAMN06110503	SRR5230172	NULL	**PDS000015357**	2	4267
SAMN07714220	SRR6110457	SAMN06110506	SRR5230329	NULL	**PDS000015357**	3	4268
SAMN07714220	SRR6110457	SAMN06110507	SRR5230169	NULL	**PDS000015357**	6	4244
SAMN07714220	SRR6110457	SAMN07714111	SRR6109637	NULL	NULL	9	4264

rapid reports provide a list of nearest neighbors that aids FDA in deciding on isolate inclusion/exclusion

may provide a 24 hour improvement in turn around time

SNP pipeline

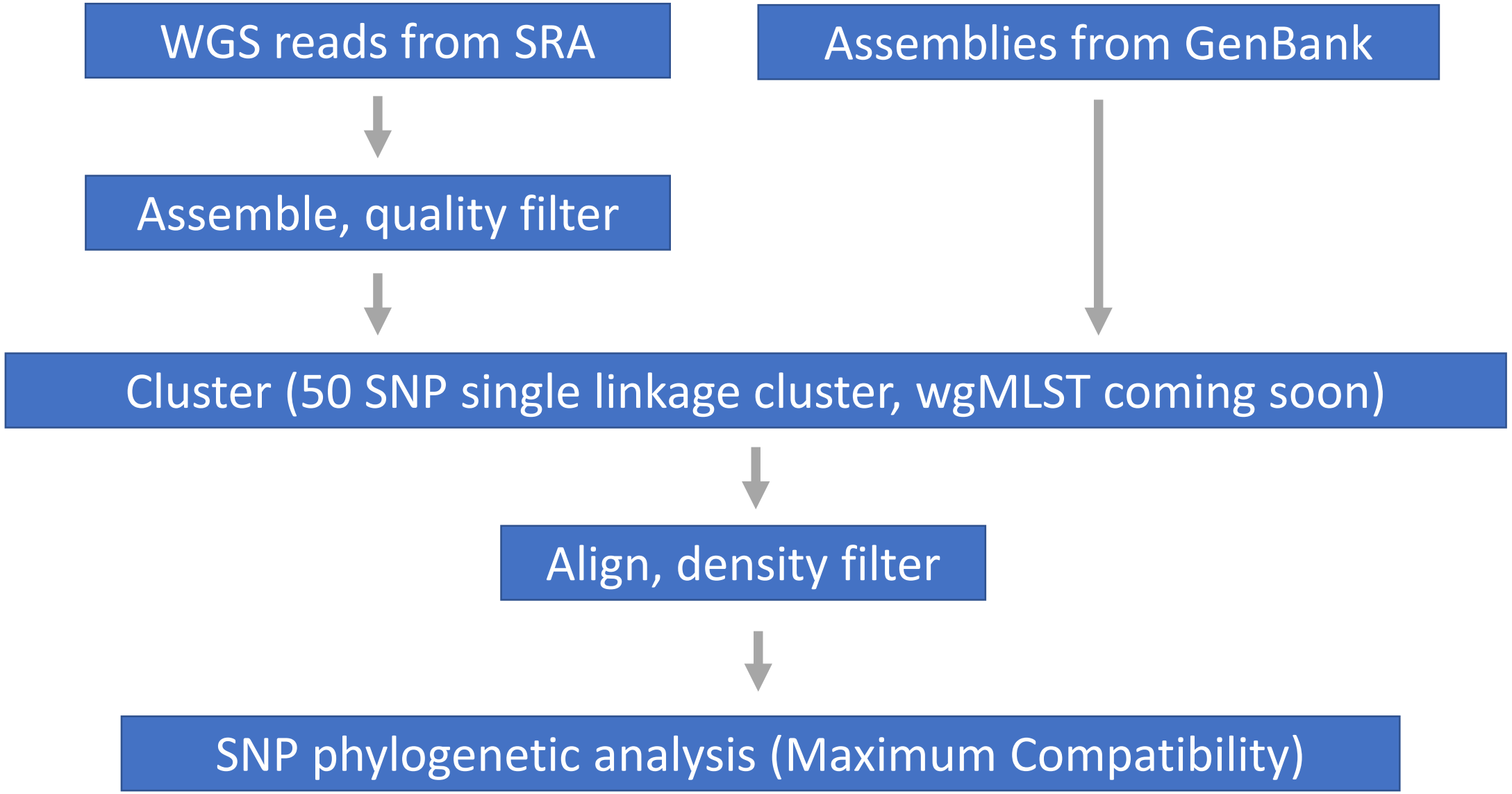
1. **Initial partition of isolates within each species by kmer distances**
2. **Within each partition, blast comparison of all pairs of genomes**
3. **Single linkage clusters with at most 50 SNPs**
4. **Within clusters, SNPs with respect to one reference**
5. **Generate final SNP list and phylogenetic trees**

Filtering:

- **Base level**
- **Repeat**
- **Density**

Problematic genomes are eliminated at various points along the way

Analysis pipeline overview



Maximum Compatibility

- Optimality criterion: Minimize homoplasies
- Good for very closely related taxa (e.g., outbreaks)
  - Considers only binary sites
  - Multiple substitutions add to noise
- Fast, exact algorithm (`compat`)
  - Tree with 2,000 isolates and 12,000 variable sites takes 10 seconds on a single core
- Noisy columns in alignment are removed from distance calculations

Cherry, J. L. 2017. A practical exact maximum compatibility algorithm for reconstruction of recent evolutionary history. *BMC Bioinformatics* 18. https://pubmed.gov/82831758

Anti-microbial resistance

1. Curated and released a database of aquired anti-microbial resistance (AMR) genes
2. Created software (AMRFinder) to identify AMR proteins
3. Created database to accept and store antibiograms to associate with sequences

Future directions

1. Point mutations for resistance
2. Additional genes for resistance to sterilizing agents, heat, metals
3. Virulence genes
4. Antigenic genes (serotyping)
5. Mobile elements?

Acknowledgements

Richa Agarwala
Azat Badretdin
Slava Brover
Joshua Cherry
Jinna Choi
Vyacheslav Chetvernin
Robert Cohen
Michael DiCuccio
Boris Fedorov
Mike Feldgarden
Lewis Geer
Dan Haft
Lianyi Han
Avi Kimchi
Michel Kimelman
William Klimke
Alex Kotliarov
Valerii Lashmanov
Aleksandr Morgulis
Eyal Moses
Chris O'Sullivan
Arjun Prasad

Edward Rice
Kirill Rotmistrovskyy
Alejandro A. Schaffer
Stephen Sherry
Sergey Shiryev
Martin Shumway
Oleg Shutov
Alexandre Souvorov
Tatiana Tatusova
Igor Tolstoy
Chunlin Xiao
Leonid Zaslavsky
Alexander Zasypkin
Lukas Wagner
Hlavina Wratko
Eugene Yaschenko

David Lipman
James Ostell
Kim Pruitt

CDC
FDA/CFSAN
GenFS
USDA-FSIS
PHE/FERA
NIHGRI
NIAID
WRAIR
Broad
Wadsworth/MDH
Vendors: PacBio, Illumina, Roche

pd-help@ncbi.nlm.nih.gov

This research was supported by the Intramural Research Program of the NIH, National Library of Medicine.
http://www.ncbi.nlm.nih.gov
National Center for Biotechnology Information – National Library of Medicine – Bethesda MD 20892 USA

Notes from the NCBI Pathogen Detection Portal DEMO (Samantha Wirth)

https://www.ncbi.nlm.nih.gov/pathogens/

ftp://ftp.ncbi.nlm.nih.gov/pathogen/ReadMe.txt

Different ways to browse the portal

1. Explore the data
2. Find isolates now

How to search for one or more isolates

1. Explore the data = new isolates (choose your organism of interest)
2. Find isolates now
    a. Type "cheese" in search bar
        i. "Select an organism group" from drop down (*Listeria*)
        ii. Notice* link in "SNP cluster" field takes you to the SNP cluster that contains the isolate of interest
        iii. NCBI SNP clusters = isolates that are ≤ 50 SNPs apart from each other
    b. EXAMPLE: Sample record has not been created on NCBI yet
        i. PNUSAS036375
        ii. ***"No hits found. Would you like to refine your query?"***
        iii. * most instances → biosample and sequence data has not been uploaded to NCBI yet
        iv. * Be patient and keep checking *
    c. Example: not related to anything on NCBI
        i. Hand type → PNUSAL003624
        ii. Notice* data in "K-mer group" field means isolate has been analyzed by NCBI
        iii. Notice* no link in "SNP cluster" cluster field means no

---

    d. Example: isolates are related/in SNP cluster
        i. Copy-paste from Excel → WGS IDs in search bar

        ***Listeria monocytogenes* (LM1711-1 // 13-27 alleles)**
        PNUSAL003355
        PNUSAL002413
        PNUSAL002199

    e. Notice link below search bar
        i. PDS# - SNP cluster accessions - are sets of isolates that have been determined by single-linkage clustering (with SNP distance of 50 SNPs)
        ii. 3 NY isolates in same SNP cluster (**PDS000003232.26** →28 total isolates)
    f. Click link – opens a new tab

Layout of new Pathogen Browser: Overview

1. Quickest way to assess your samples;

2. Top Left
    a. Listeria 28 isolates PDG000000001.863 / **PDS000003232.26**
3. Isolates Selected
    a. listed by year (create date)
    b. Shows distance between selected isolates **(11-20 SNPs)**

4. Useful to look at min-same/min-diff → NOW LOOK AT OUR SELECTED SAMPLES

    a. Notice (1) Hightlighted blue, (2) at top of list
    b. Notice all 3 have closely related **min-same**
        i. **Min-same** = min SNP distance to isolate of the "same" type (clinical)
    c. Notice PNUSAL002199 has closely related **min-diff**

i. **Min-diff** = min SNP distance to isolate of the "different" type (environ/other)

5. While we're here……Customize and filter: List of isolates with metadata

a. [] choose columns
   a. Remove by clicking (-) next to items in left panel
      i. Host
      ii. Serovar
   b. Add by clicking (+) next to items in right panel
      i. Collection Date
      ii. Collected by
      iii. Click "OK"

b. Sort by any column header
   a. * note Sample(s) with assembly accession number (GCA#) are the reference genomes used for that SNP cluster

c. \/ **Filters**
   a. Source – if looking for something specific
   b. Often useful to filter by Target Creation
   c. This is small SNP cluster (only 28 isolates), but filtering can help focus your efforts
   d. Notice = can always click "**X**" to remove filters

d. Number isolates per page
   a. Change to "50" --- now can see all isolates that fit the "filtered" criteria

6. NOW LOOK AT THE SNP TREE

a. Notice – our isolates are highlighted in red
b. Remember close min-diff to PNUSAL002199
c. Notice*- the min-diff (env) sample close to PNUSAL002199
d. Can click and drag to move entire tree within window
e. Scroll mouse within tree window to zoom in and zoom out
f. Adjust scale of branch length
g. Adjust vertical node spacing

Different ways to investigate SNP distances and select samples

2. Remember from the table of above that PNUSAL02199 has a closely related sample of a different type
3. Let's find PNUSAL002199 in the SNP tree
4. Create a subtree by clicking on the node (open circle) and selecting "Subtree view" from the pop-up menu
5. Notice there are 6 isolates in this subtree and only one is a different isolation type
   a. Let's select that sample
   b. Notice PNUSAL002199 is 5 SNPs away from this environmental/other isolate

---

1. Let's go back to the main SNP cluster
2. Relatively small cluster, what is SNP range of all isolates in this cluster?
3. Select all isolates in "List of isolates with metadata" window OR by clicking node in tree
   a. Notice = Isolates Selected window = SNP range of 28 selected isolates (0-30 SNPs)

   ---
4. Decide what you want = small cluster = use whole SNP tree = use PNUSAL002199 subtree

Export a tree or subtree (as .pdf)
Download metadata (.csv file)
Share URL

Now I'm going to search for some STEC isolates.
These isolates were collected within the same time frame, same geo area, and produced the same PFGE pattern combination.
Are they closely related?
==========

1. Find isolates now ***E. coli***
2. Copy/paste WGS IDs in search bar
   Flour cluster (1601MLEXK-1) → PFGE pattern EXKX01.0001/EXKA26.0001
   Search NCBI for these 4 isolates:
   PNUSAE002179
   PNUSAE004513
   PNUSAE003908
   PNUSAE004488 (same PFGE pattern,→ not WGS match)
   *Notice – 3 isolates fall into same SNP cluster
   *Notice – PNUSAE004488, isolate with same PFGE pattern, falls into different SNP cluster
   i. Click on link for 3 isolates
   ii. Quickly Customize and filter
   iii. *Notice 3 NY isolates are highly related (**0-1 SNPs**)
   iv. *Notice highlighted in blue at top of list
   v. *Notice all three have closely related min-same and min-diff samples
   vi. Select node and "Subtree view" for closest isolates
   vii. Select all isolates in subtree by clicking node and view SNP range (**0-16 SNPs**)
   viii. *Notice outliers on tree (branch length) and unselect
   ix. Sort by "isolation type" column→ *Expand # per page, *Notice env samples are all flour

   Export a tree or subtree (as .pdf)
   Download metadata (.csv file)
   Share URL

---

3. What about our 4th isolate?
   a. *Remember it had same PFGE pattern combination and temporal/geographic link, but assigned to a different SNP cluster
   b. Click the link for the SNP cluster
   c. *Notice there are 182 isolates in this SNP cluster
   d. *Notice our isolate is highlighted in blue in the table
   e. It has closely related min-same (sample of same type) but no closely related min-diff
   f. Let's look at SNP tree
   g. *Notice our isolate is highlighted in red
   h. Create subtree and investigate SNP distances to nearby isolates
   i. *Notice there are 4 isolates in this subtree SNP range (4-18 SNPs)

Export a tree or subtree (as .pdf)
Download metadata (.csv file)
Share URL

Download SNP matrix for whole PDG using FTP
==========
Download pairwise SNP distances for whole PDG using FTP
Make note of PDG# and your samples' PDT#s (NCBI's accession number for each pathogen genome)

PDG000000001.863

To download the pairwise SNP distances for the Listeria cluster demonstrated today;

Pathogen Browser → Download analysis results FTP → Listeria → PDG000000001.863 → SNP trees→ PDS000003232.26.tar.gz →PDS000003232.26.dnapars_input.dnapars.gz

Here is the explanation from ftp://ftp.ncbi.nlm.nih.gov/pathogen/ReadMe.txt ;

```
c. SNP_trees

The directory now contains gzipped tarballs of each SNP cluster instead of
subdirectories.
The files are named as PDS# accession.version.tar.gz for each cluster. Clusters can
consist of two or more closely related isolates.

Inside each tarball are files that correspond to three things:

1. Phylogenetic trees. Currently these files are maximum compatibility trees generated
with compat on the input SNP matrix.
    - *.newick_tree.newick - Newick / New Hampshire formatted maximum compatibility tree.
    - *.biotree.asn - ASN.1 formatted maximum compatibility tree, includes metadata and
can be loaded into Genome Workbench.
    - *.snp_tree.pdf - PDF image of the above tree
2. SNP matrix used to generate trees
    - *.dnapars_input.dnapars.gz - phylip formatted SNP matrix generated from the NCBI
Pathogen Detection Pipeline
3. Variant calls
    - *.variation.vcf.gz - Variant Call Format (VCF) file for this SNP cluster compressed
with bgzip
    - *.variation.vcf.gz.csi - tabix generated index for the .vcf.gz file

NOTE: the .variation.vcf.gz file contains the SNP call output of the NCBI Pathogen
Detection pipeline. Positions marked with filter PASS and I are included in the
.dnapars.gz file; see the header in that file for additional information. The .vcf.gz
file can be unzipped with standard tools for gzipped files, but is formatted and indexed
for use with bcftools. See http://www.htslib.org/ for more information on the bgzip
format.
```

Listeria monocytogenes cluster
WGS ID
PNUSAL003355
PNUSAL002413
PNUSAL002199

Listeria singleton
PNUSAL003624

S
WGS ID
PNUSAE002179
PNUSAE004513
PNUSAE003908
PNUSAE004488
PNUSAE001868
PNUSAE004999

STEC cluster
PFGE Pattern
EXKX01.0001/EXKA26.0001
EXKX01.0001/EXKA26.0001
EXKX01.0001/EXKA26.0001
EXKX01.0001/EXKA26.0001
EXKX01.0001/EXKA26.0005
EXKX01.0011/EXKA26.0001