

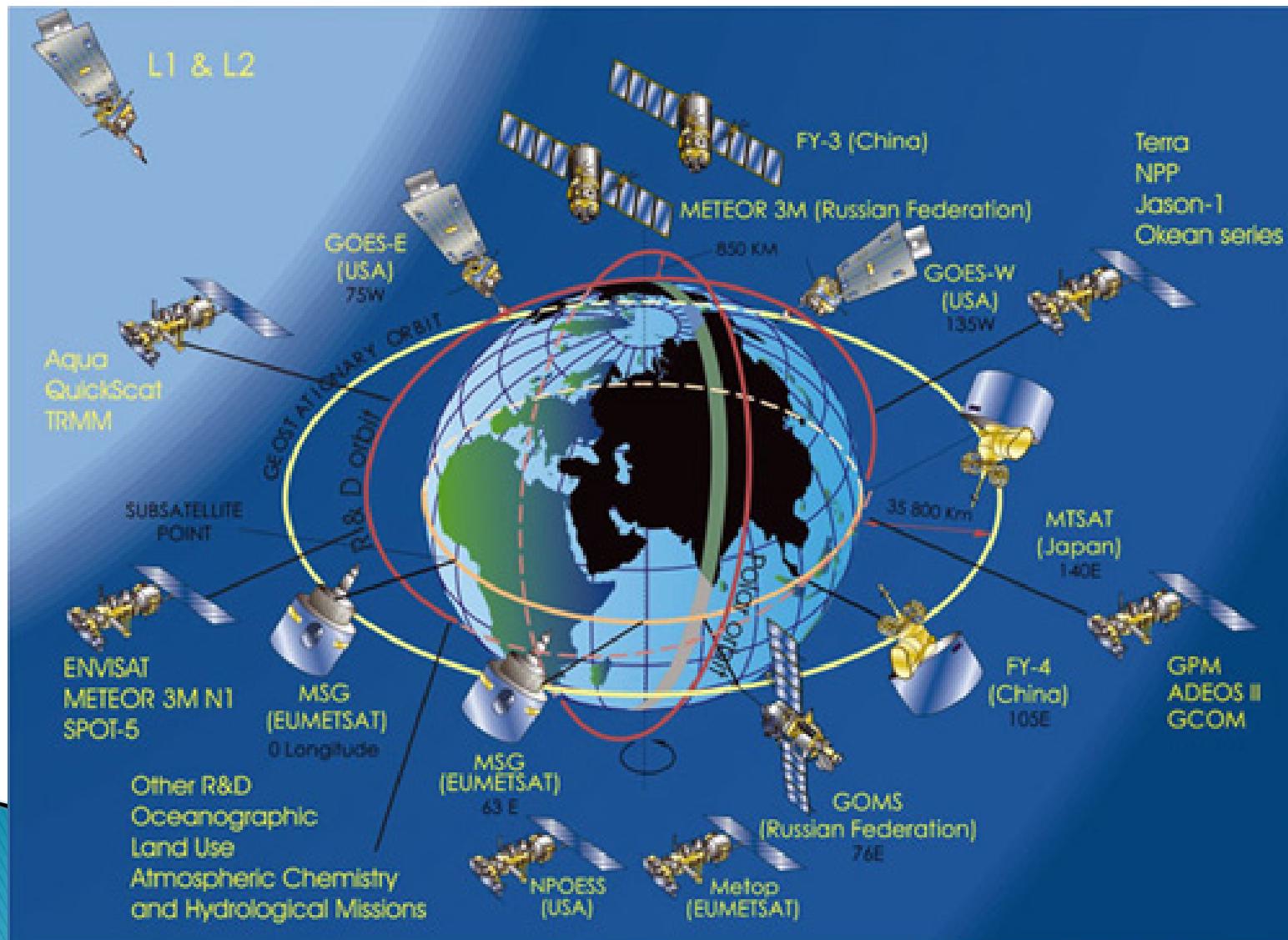
The GMI Global Repository WG

James Ostell

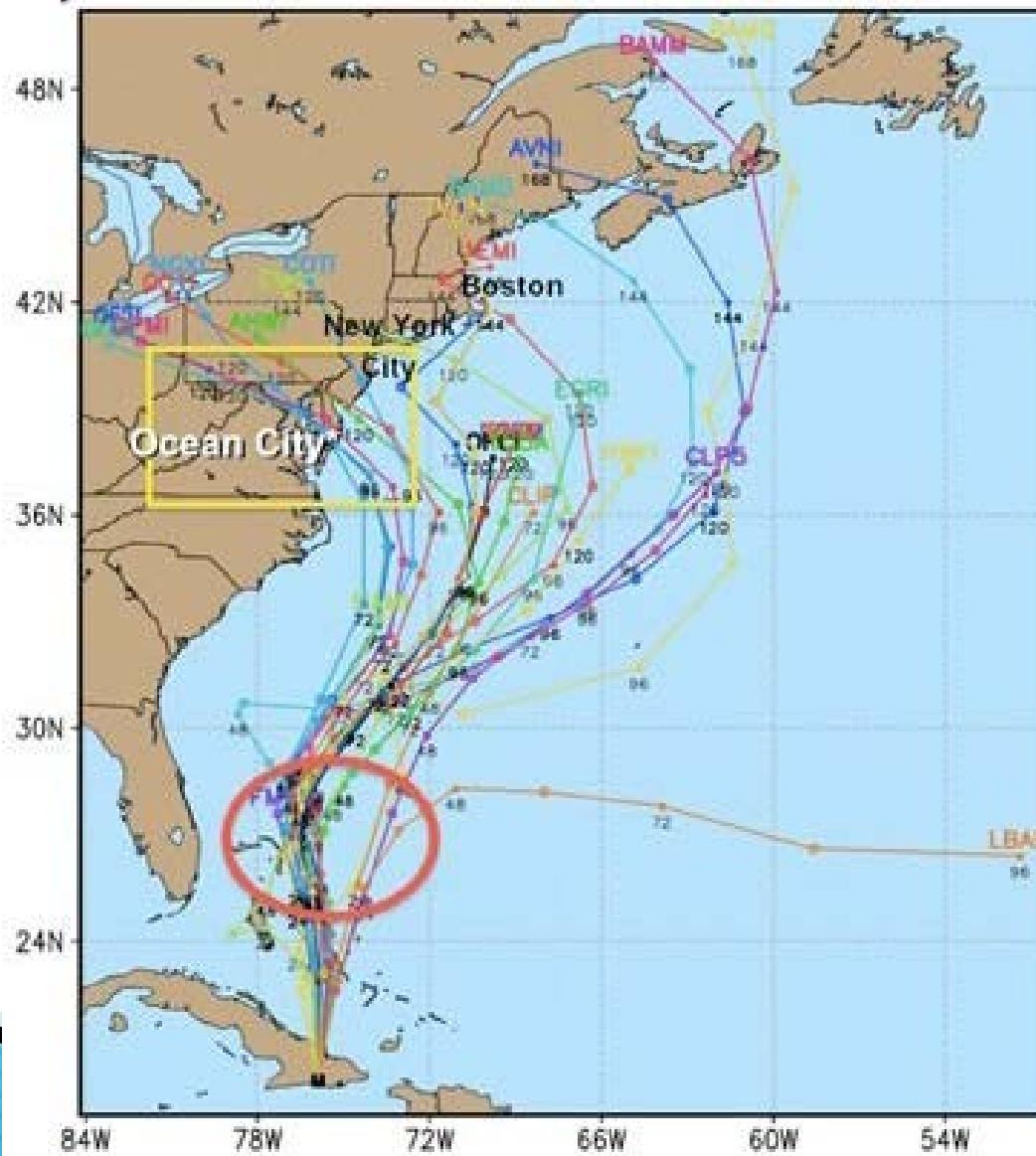
National Center for Biotechnology Information
US National Institutes of Health, Bethesda, MD, 20894–6513



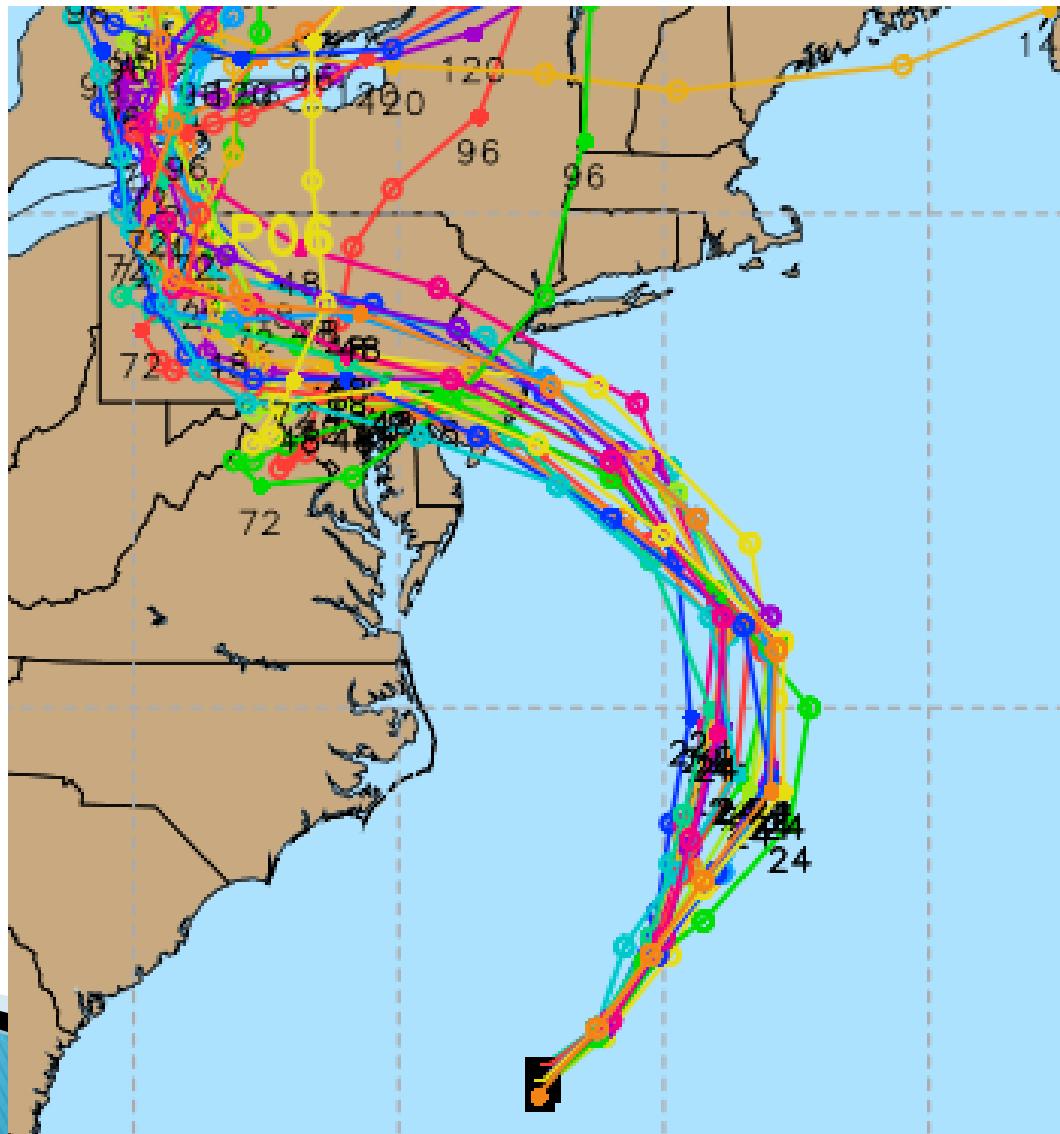
Weather as a Modern Global Infrastructure



Many Models on the Data, Fully Public, On the News



Citizens and Government All See the Power and Limits of Science



DNA Forensics...



Genome sequence is agnostic.
One biological assay could work
on all pathogen species

To be immediately useful all we
need is the genome and a little
metadata.

Functional prediction can be
developed and refined more slowly
from this base.

A Pathogen Genome Is The Fingerprint

LETTUCE

Canada, Chile, Dominican Republic, Mexico, Peru, USA



CUCUMBERS

Canada, Honduras, India, Mexico, Spain, USA



FETA CHEESE

Canada, Denmark, Egypt, Germany, Greece, Israel, Italy, Turkey, UK, USA



VINAIGRETTE

Argentina, Brazil, Canada, Chile, China, France, Germany, Greece, India, Indonesia, Italy, Mexico, Morocco, Peru, Portugal, Spain, Thailand, Tunisia, Turkey, USA, Vietnam



The Well-Traveled Salad. Do You Know Where Your Food Has Been?

As consumers, many of us fail to recognize that even our domestic and local food supplies are part of a global network. The daily activity of consuming food directly links our health as humans to the health of crops and produce, food animals, and the environments in which they are produced.

CROUTONS

Argentina, Australia, Brazil, Canada, China, France, India, Mexico, Netherlands, Poland, Russia, Switzerland, Uruguay, USA, Vietnam



TOMATOES

Canada, Dominican Republic, Holland, Israel, Italy, Mexico, USA



ONIONS

Canada, China, Germany, India, USA



SPROUTS

Argentina, Australia, Bangladesh, Canada, China, Egypt, France, India, Morocco, Nepal, Pakistan, South Africa, Spain, Turkey, USA



MANDARIN ORANGES

Israel, Mexico, Morocco, South Africa, Spain



A "One Health" approach to food safety—bringing together expertise and resources from the clinical, veterinary, wildlife health, and ecology communities—has the potential to reveal the sources, pathways, and factors driving the outbreaks of foodborne illness and possibly prevent them from occurring in the first place.

NOTE: Countries are listed in alphabetical order and not by volume of export.

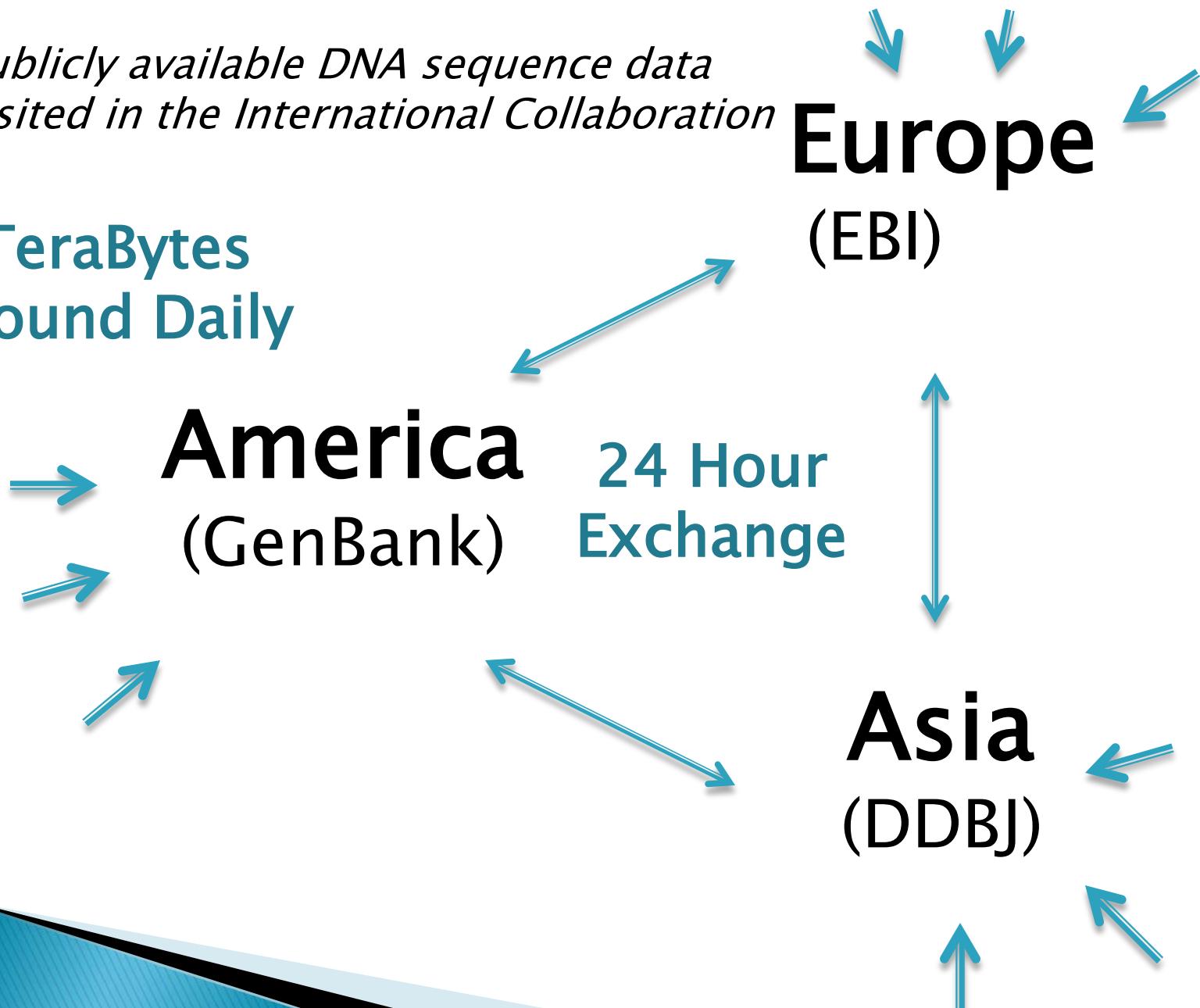


INSTITUTE OF MEDICINE
OF THE NATIONAL ACADEMIES

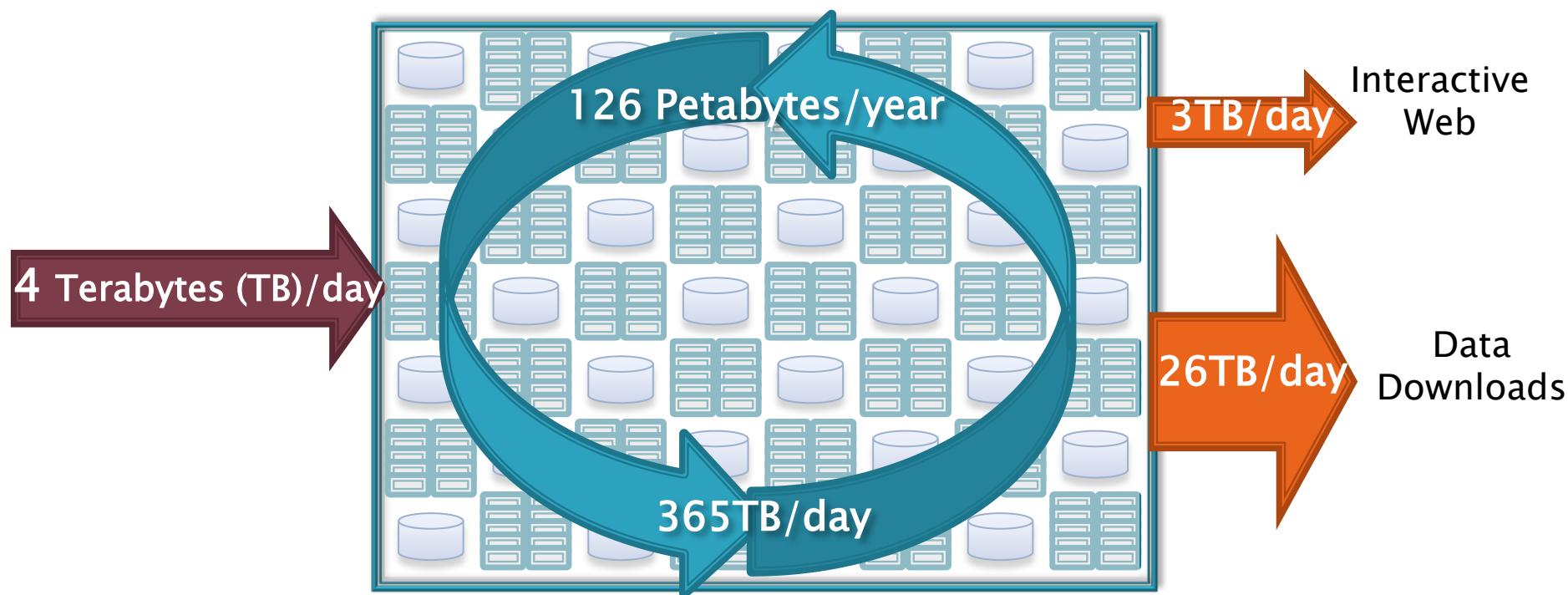
www.iom.edu

*All publicly available DNA sequence data
deposited in the International Collaboration*

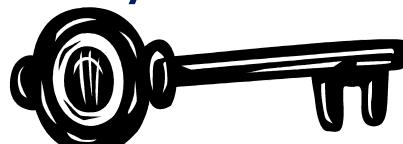
**4 TeraBytes
Inbound Daily**



Daily Data Processing at NCBI



MINIMAL PATHOGEN METADATA (FOODBORNE OUTBREAKS)

sample_name	<u>What</u>	collection_date	<u>When</u>
organism		Geographic location	<u>Where</u>
strain/isolate		6a) geo_loc_name OR 6b) lat_lon	
Category (attribute_package) 1a) Clinical/Host-associated 1a1) specific_host 1a2) isolation_source 1a3) host-disease OR 1b) Environmental/Food/Other 1b1) isolation_source		collected by 	<u>Who</u>

Production Pipelines

NCBI shumwaym: ReadWrite [GPipe Home](#) [GPipe](#)

[GPipe Browser > GPIPE_DEV35 > sci new, GP-3444-day0123-AESNO-subtree-D \(build 14\) > Analysis and SNP-calling for set of pathoge](#)

Tasks

Current Plane: (default)

- [10042: Build SNP Parsimony Tree](#)
- [10032: Calculate Distances](#)
- [10022: Filter SNPs](#)
- [9992: Find Raw SNPs](#)
- [9952: Get Genomic Collection](#)
- [9962: Get Genomic FASTA](#)
- [9972: Get Other Assemblies](#)
- [9982: Master Assembly BlastDb](#)
- [10012: Raw SNP alignments](#)**
- [10002: Sort All Hits By Subject](#)

Planes

(default)
align_463288
align_463308
align_463328
align_463348
align_463628
align_463728
align_463748
align_463768
align_463828
align_463928
align_464008
align_464028
align_464048
align_464068

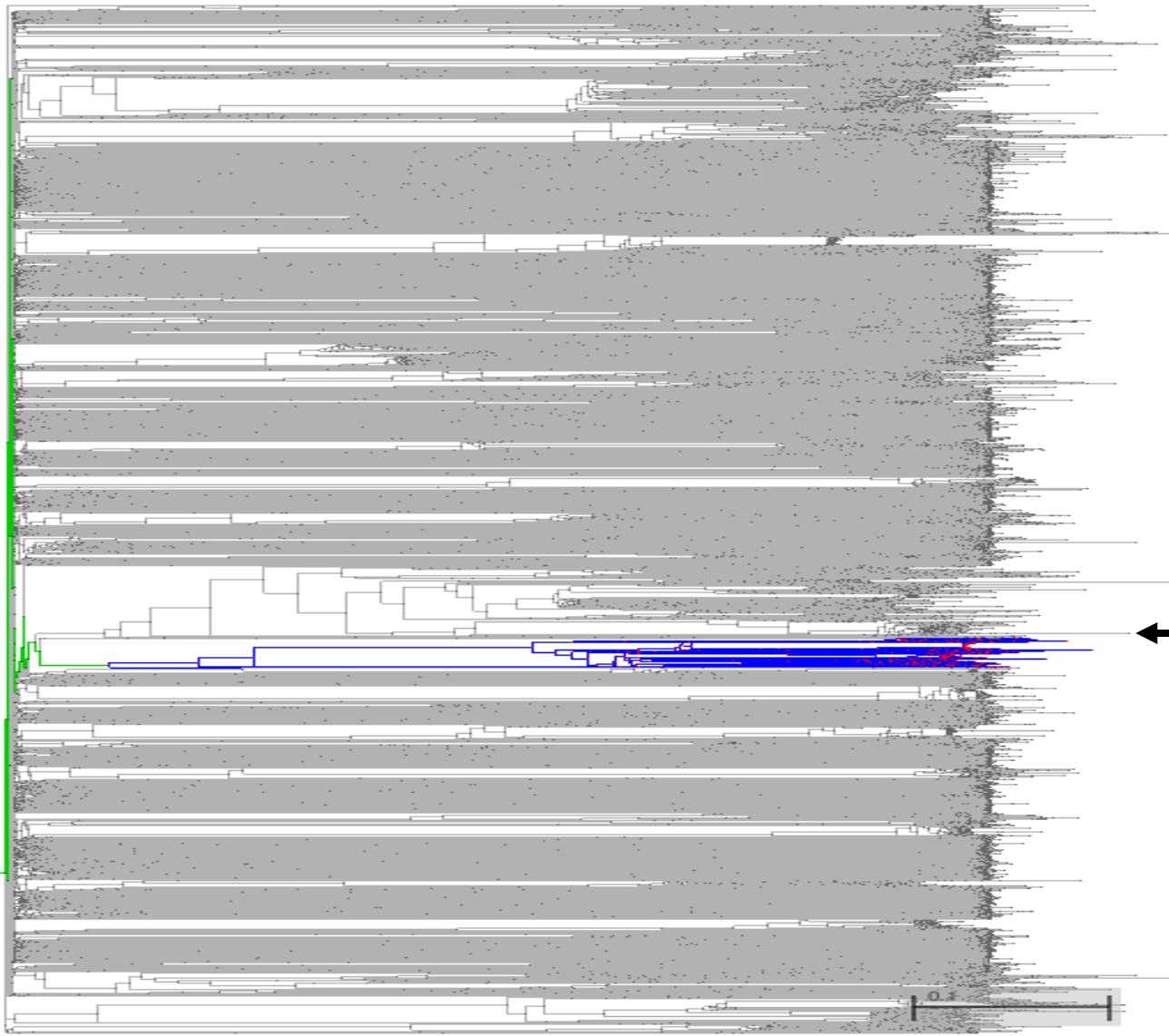
Execution Graph

Show all tasks Show foreign tasks Show dataflow details

```
graph TD; A[Get Other Assemblies] --> B[Find Raw SNPs]; A --> C[Sort All Hits By Subject]; A --> D[Get Genomic FASTA]; E[Get Genomic Collection] --> F[Calculate Distances]; E --> G[Filter SNPs]; E --> H[Build SNP Parsimony Tree]; F --> I[Raw SNP alignments]; G --> I; I --> H; B --> J[Raw SNP alignments]; J --> H; C --> K[Master Assembly BlastDb];
```

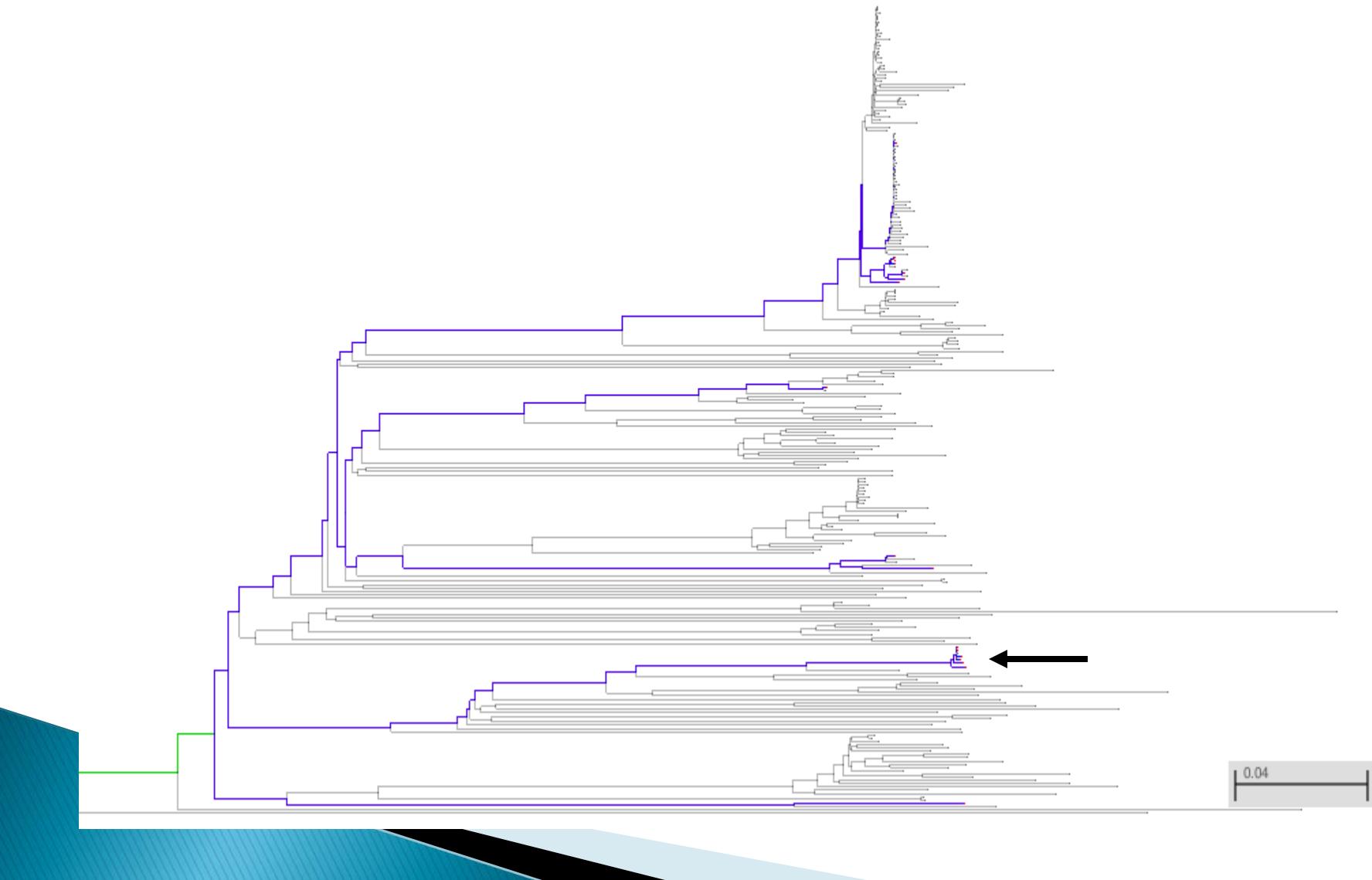
The execution graph illustrates the data flow between various tasks. It starts with three initial tasks: 'Get Other Assemblies', 'Get Genomic Collection', and 'Get Genomic FASTA'. The 'Get Other Assemblies' task feeds into 'Find Raw SNPs' and 'Sort All Hits By Subject', and also provides data to 'Raw SNP alignments'. The 'Get Genomic Collection' task feeds into 'Calculate Distances', 'Filter SNPs', and 'Build SNP Parsimony Tree'. The 'Get Genomic FASTA' task feeds into 'Master Assembly BlastDb'. The 'Find Raw SNPs' task feeds into 'Raw SNP alignments'. The 'Raw SNP alignments' task feeds into 'Build SNP Parsimony Tree'. The 'Sort All Hits By Subject' task feeds into 'Master Assembly BlastDb'. The 'Calculate Distances' task feeds into 'Build SNP Parsimony Tree'. The 'Filter SNPs' task feeds into 'Build SNP Parsimony Tree'.

Example Surveillance Workflow

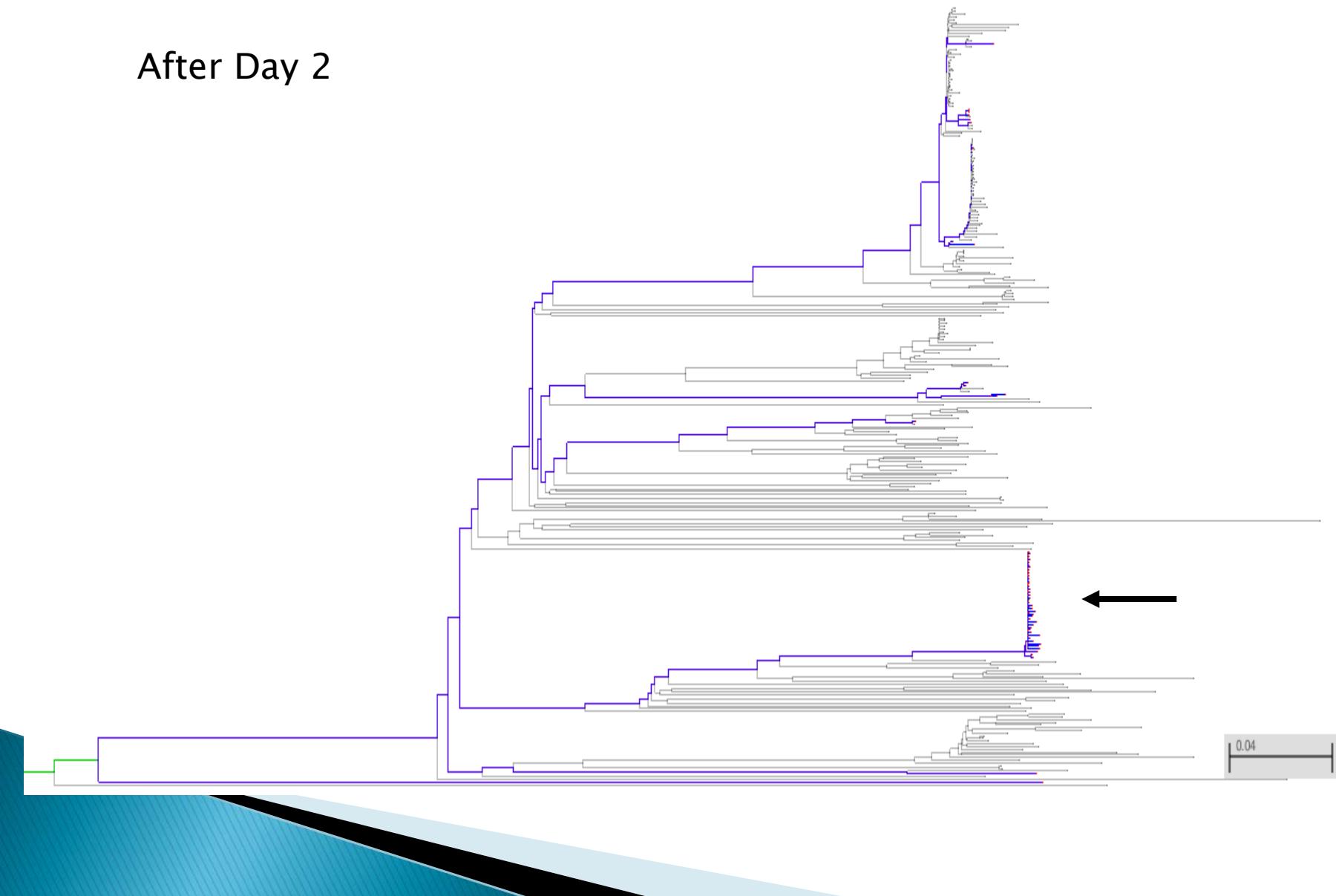


← Existing
Salmonella
clade in
combined
Eubacteria
kmer tree

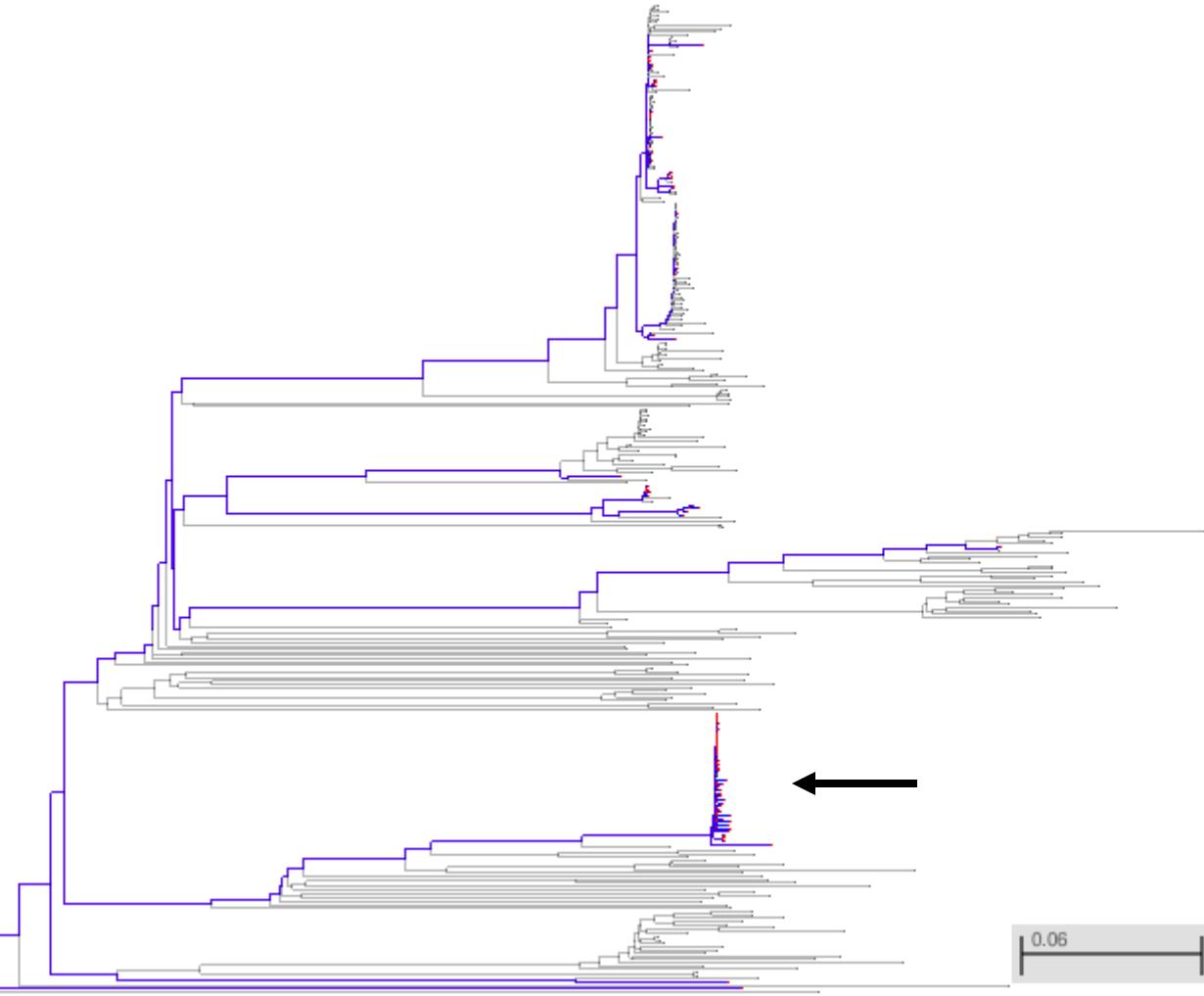
After Day 1



After Day 2



After Day 3



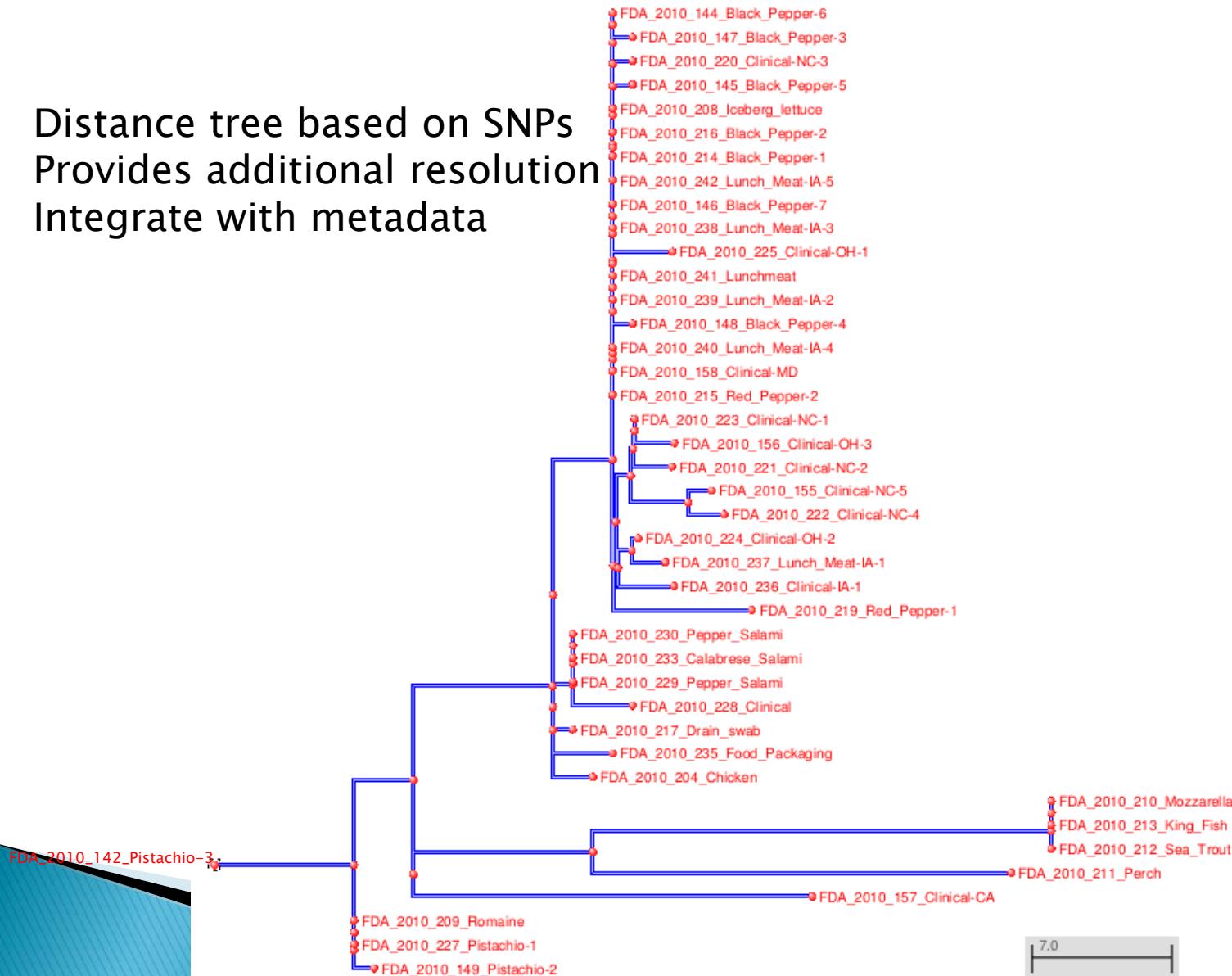
Extract cluster, Montevideo serovar

- 2 existing genomes
- 39 new genomes
- Extract sub-tree
- Re-root on outlier



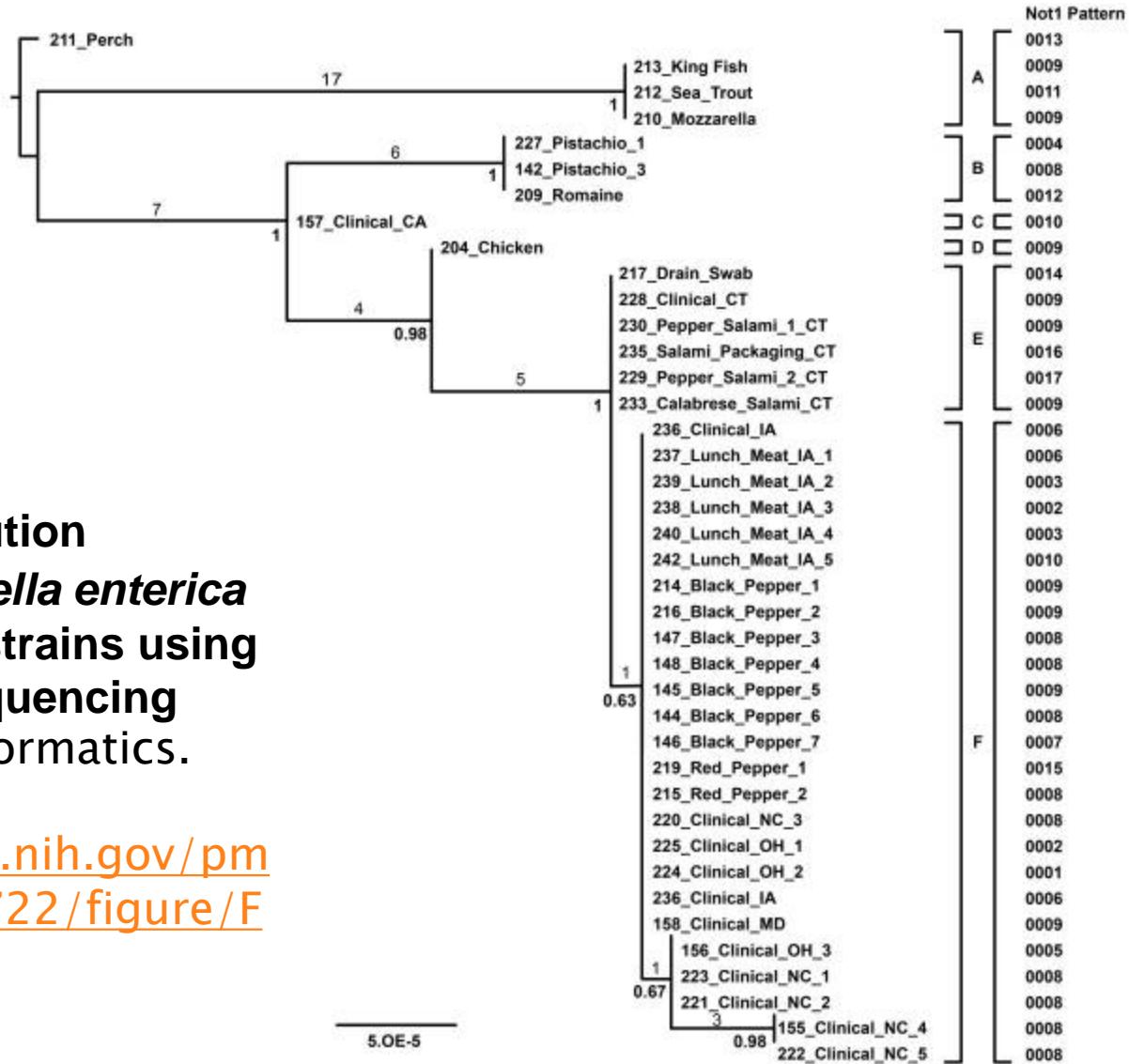
Compute new subtree

- Distance tree based on SNPs
- Provides additional resolution
- Integrate with metadata





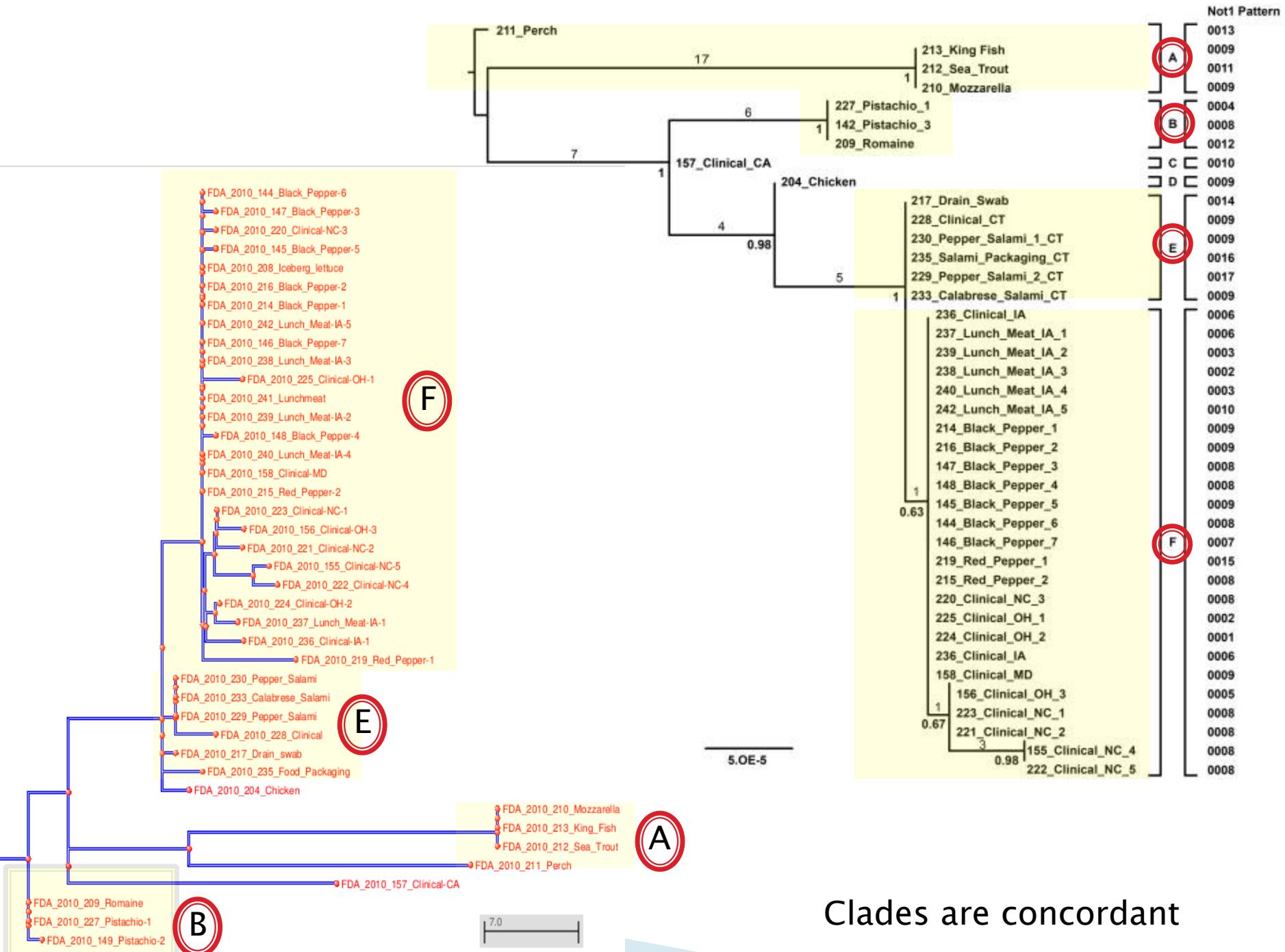
Compare with published dataset



Luo et al. High resolution clustering of *Salmonella enterica* serovar Montevideo strains using a next-generation sequencing approach BMC Bioinformatics.

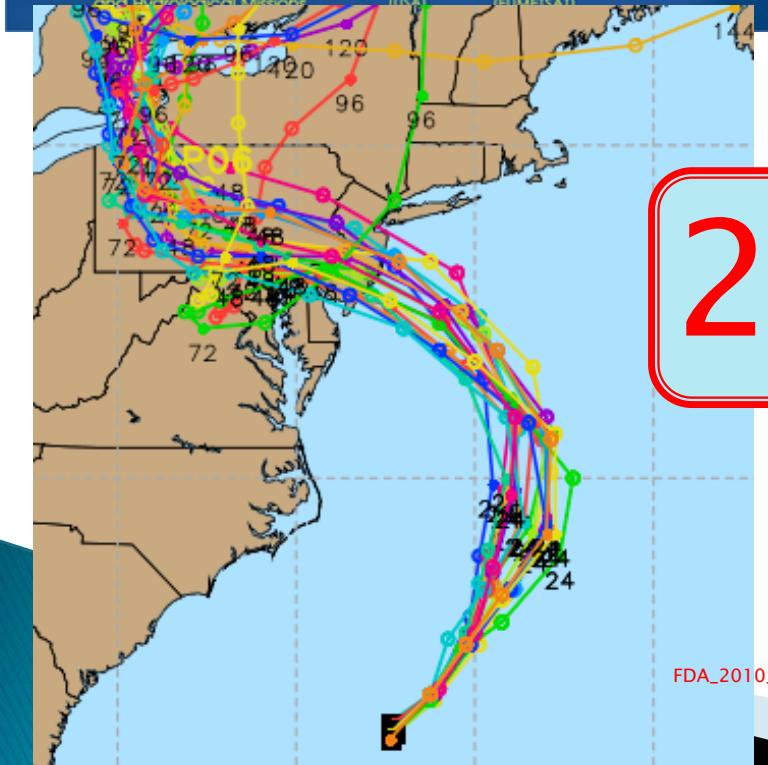
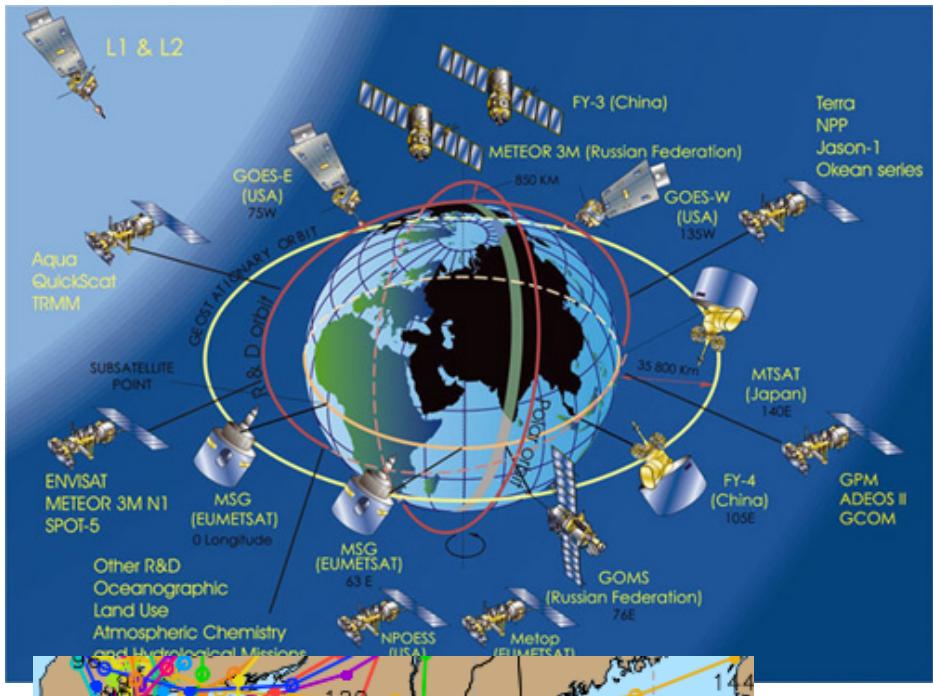
2012; 13: 32.

<http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3368722/figure/F3/>



Pipeline Performance Today for 42 Salmonella Genomes

- ▶ Place Reads in Taxonomic Tree to Clade Level by Kmers
 - ▶ Assemble Reads by Both de novo and Reference Guided Methods and Integrate
 - ▶ Annotate All CDS and Structural RNAs
 - ▶ Call SNPs Against Closely Related Genomes and Compute a New SubTree
-
- ▶ Wall Clock Time = 4 hours



FDA_2010_142_Pistachio-3

America (GenBank)

Europe (EBI)

Asia (DDBJ)

