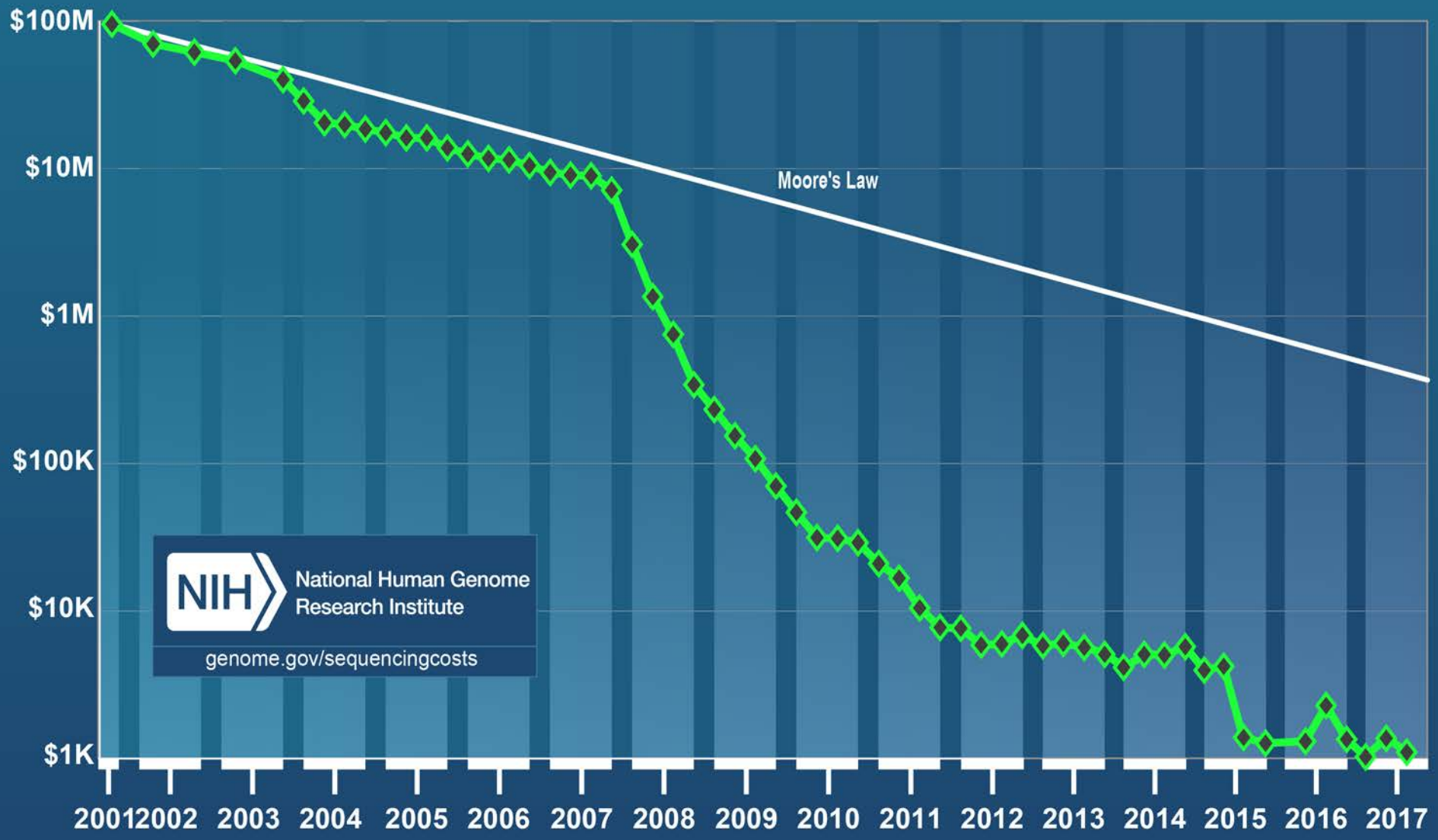
Functional genome annotation enabled by cloud computing

Leonardo Mariño-Ramírez, PhD
NCBI / NLM / NIH

BIOL 7210 A – Computational Genomics
2/20/2018

Cost per Genome



The $1,000 genome is here!

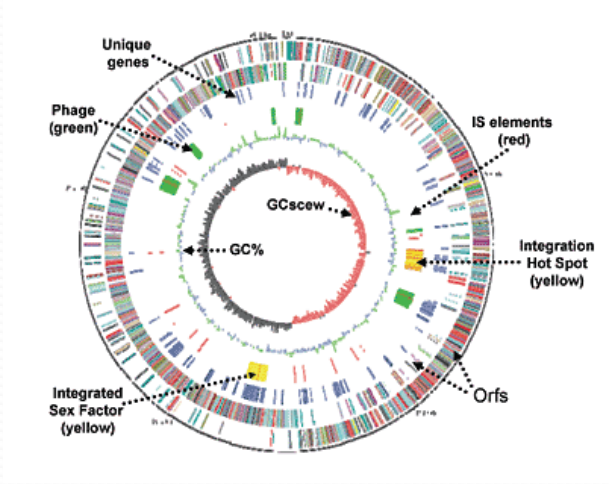
SAN DIEGO--(BUSINESS WIRE)--January 14, 2014--
Illumina, Inc. (NASDAQ:ILMN) today broke the 'sound barrier' of human genomics by enabling the $1,000 genome.

Population power. Extreme throughput. $1,000 human genome.

The HiSeq X Ten is a set of ten ultra-high-throughput sequencers, purpose-built for large-scale human whole-genome sequencing.

http://www.illumina.com/systems/hiseq-x-sequencing-system.ilmn

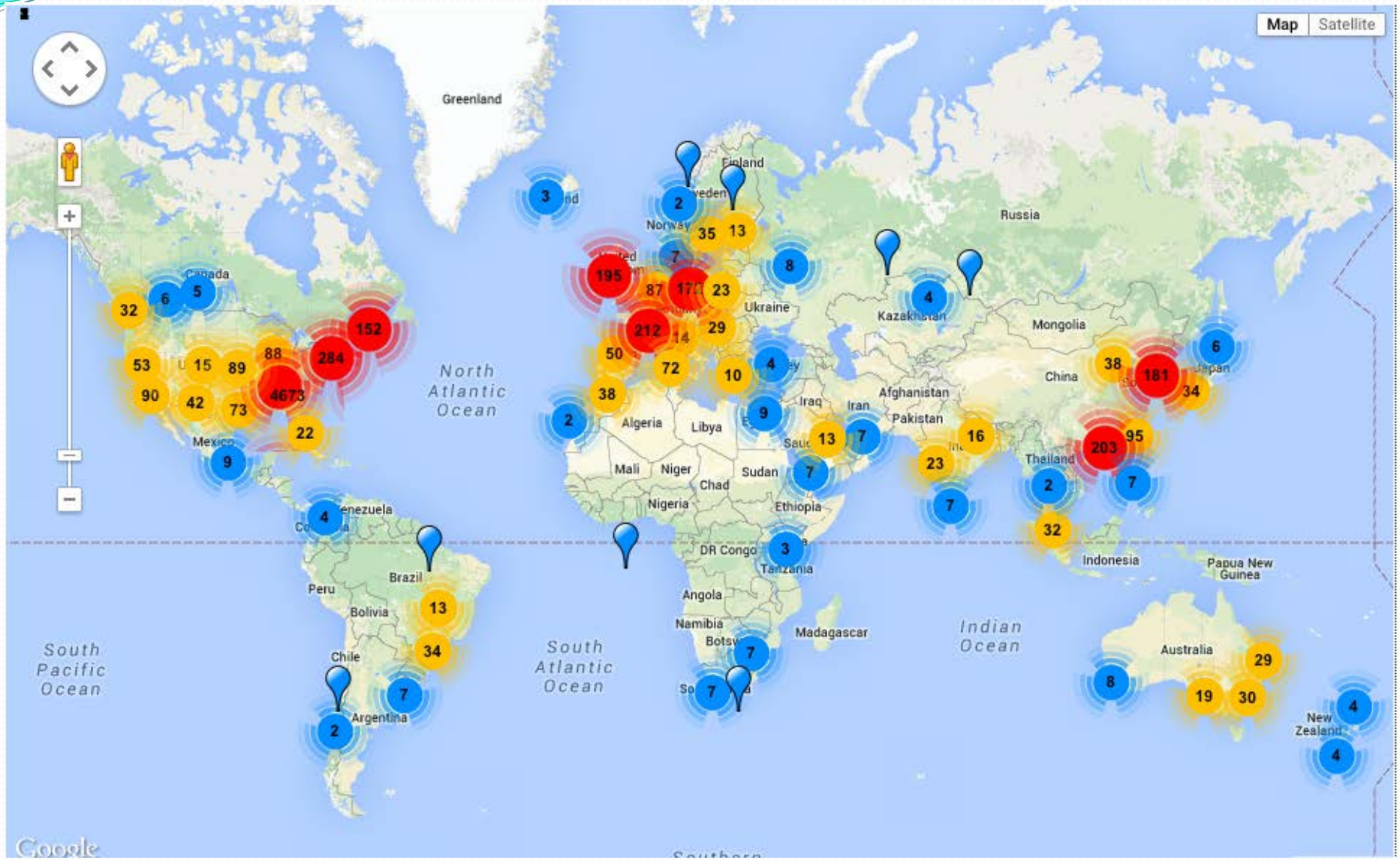
Bioinformatics bottleneck



Bioinformatics challenges

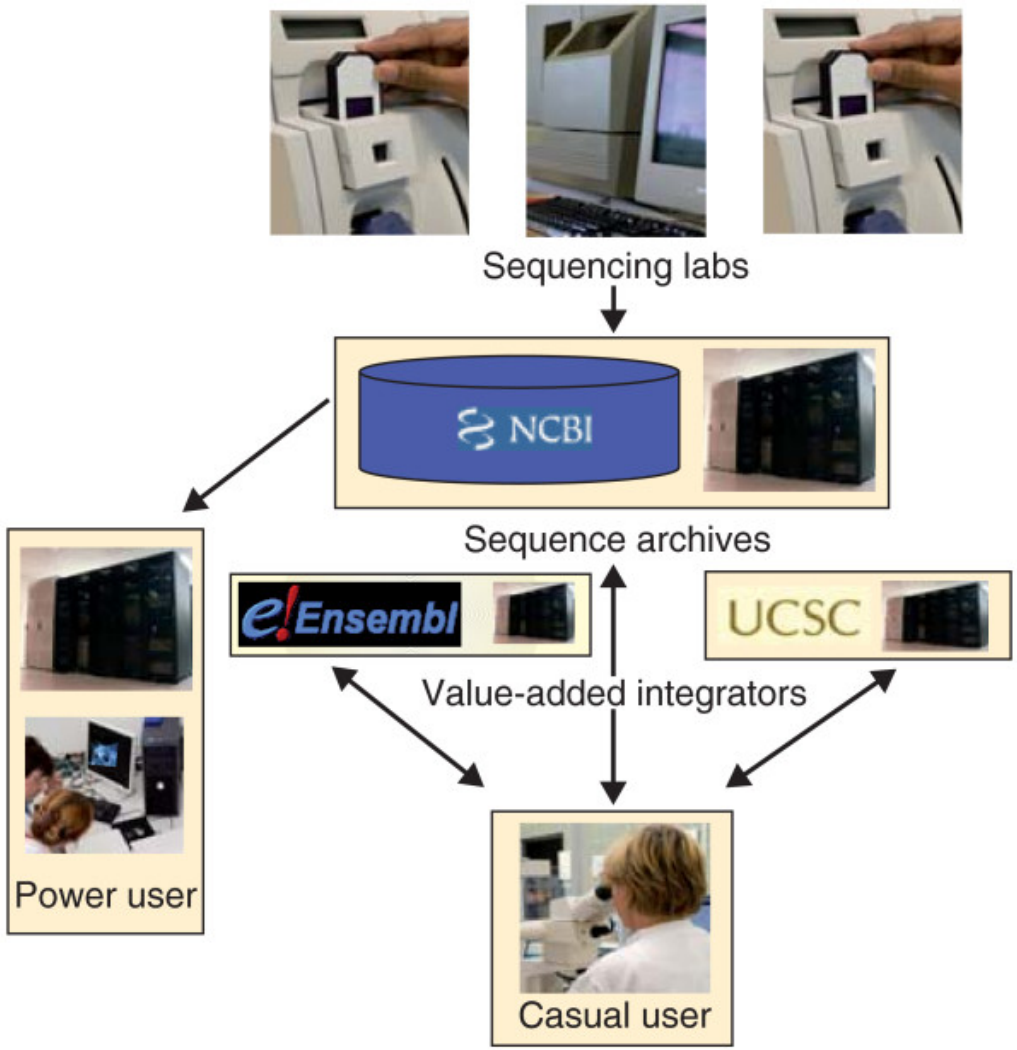
- **Methods:** How do I analyze my data using procedures for various data types?
- **Infrastructure:** Where do I process my data? Large scale compute accessibility, Installing and maintaining software
- **Standards:** How do I ensure my results are useful? Common, shared formats using community developed software and tools

High throughput sequencing map



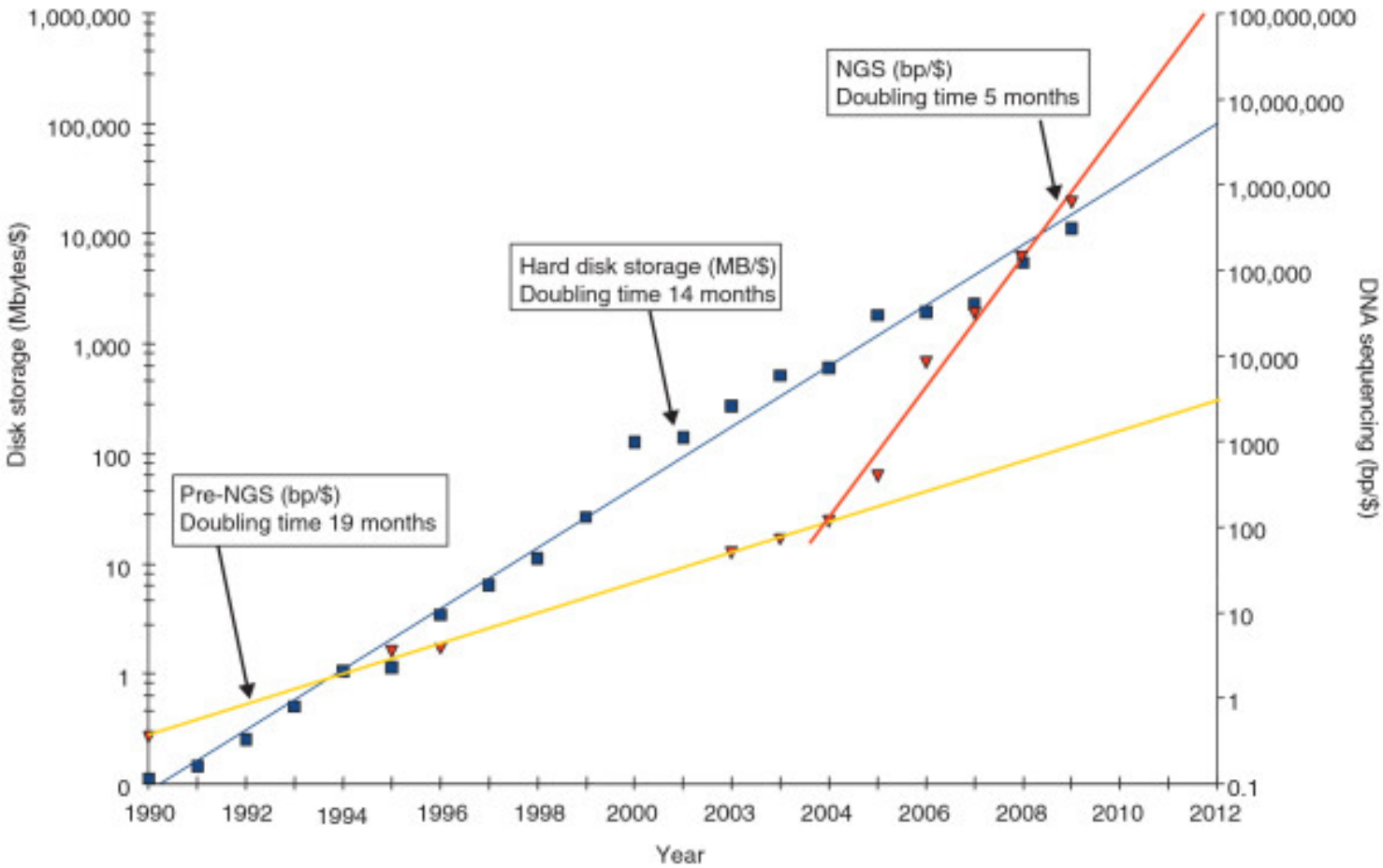
http://omicsmaps.com/

The case for cloud computing in genome informatics



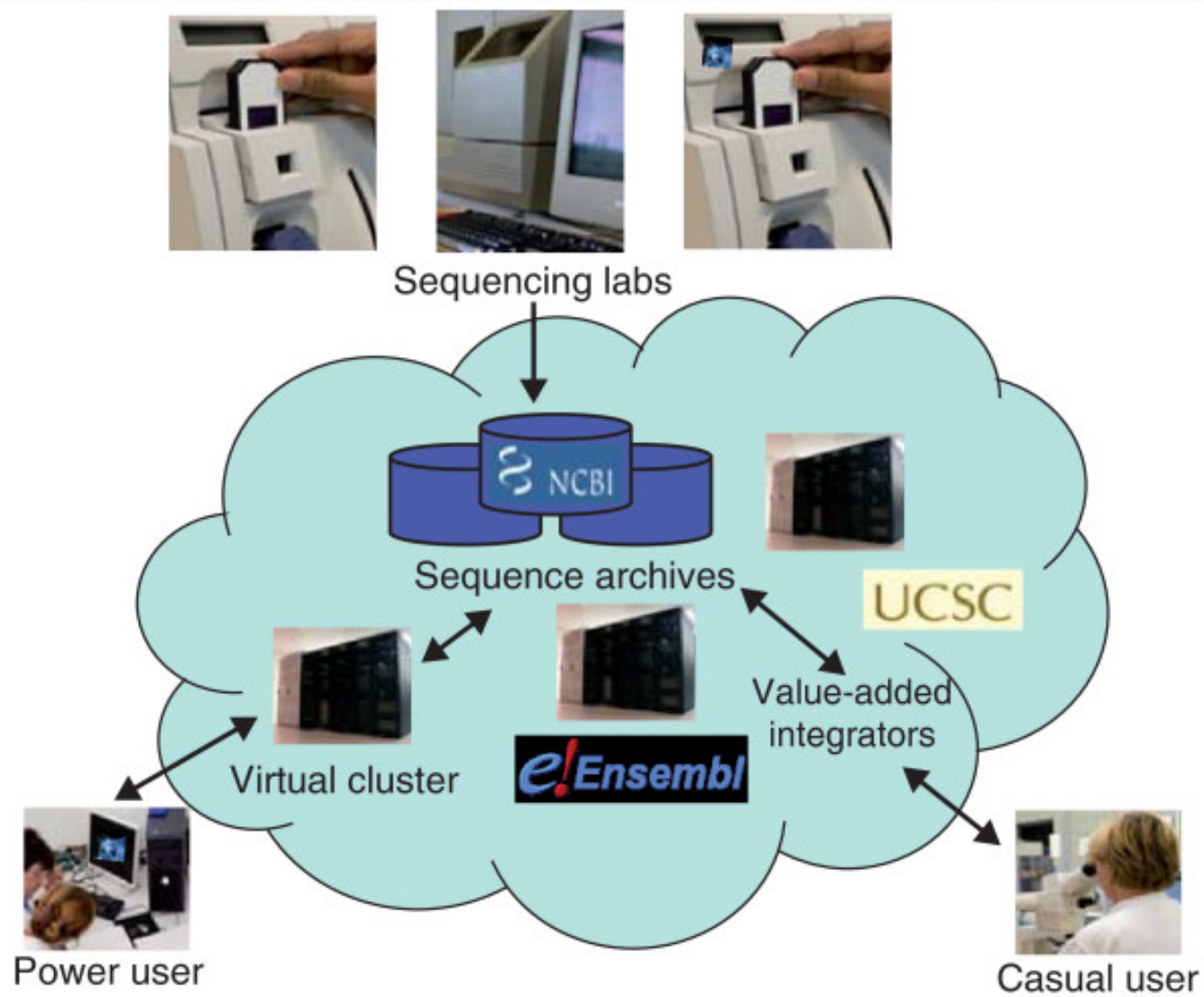
http://genomebiology.com/2010/11/5/207

The case for cloud computing in genome informatics



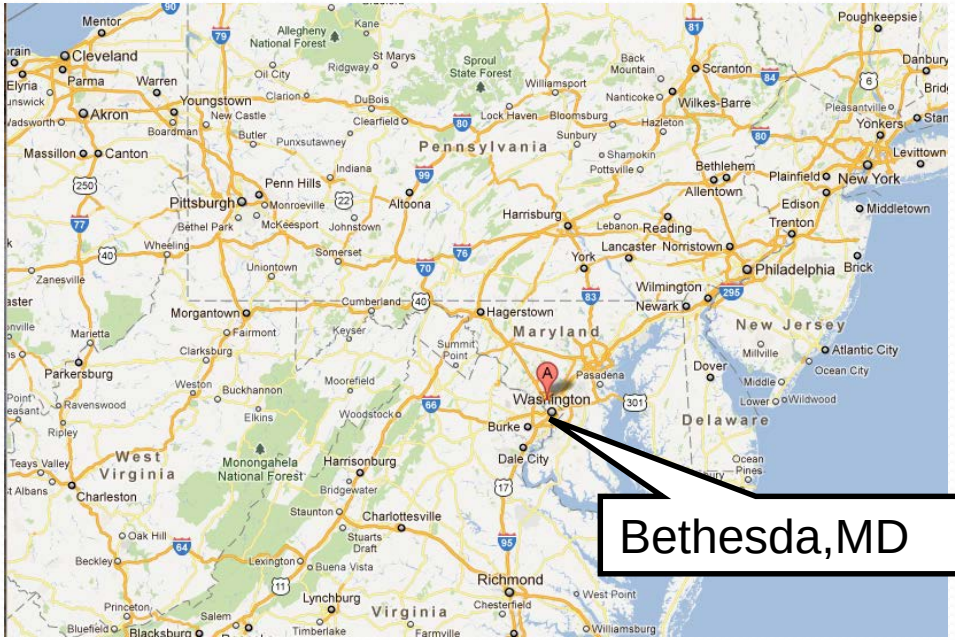
http://genomebiology.com/2010/11/5/207

The case for cloud computing in genome informatics



http://genomebiology.com/2010/11/5/207

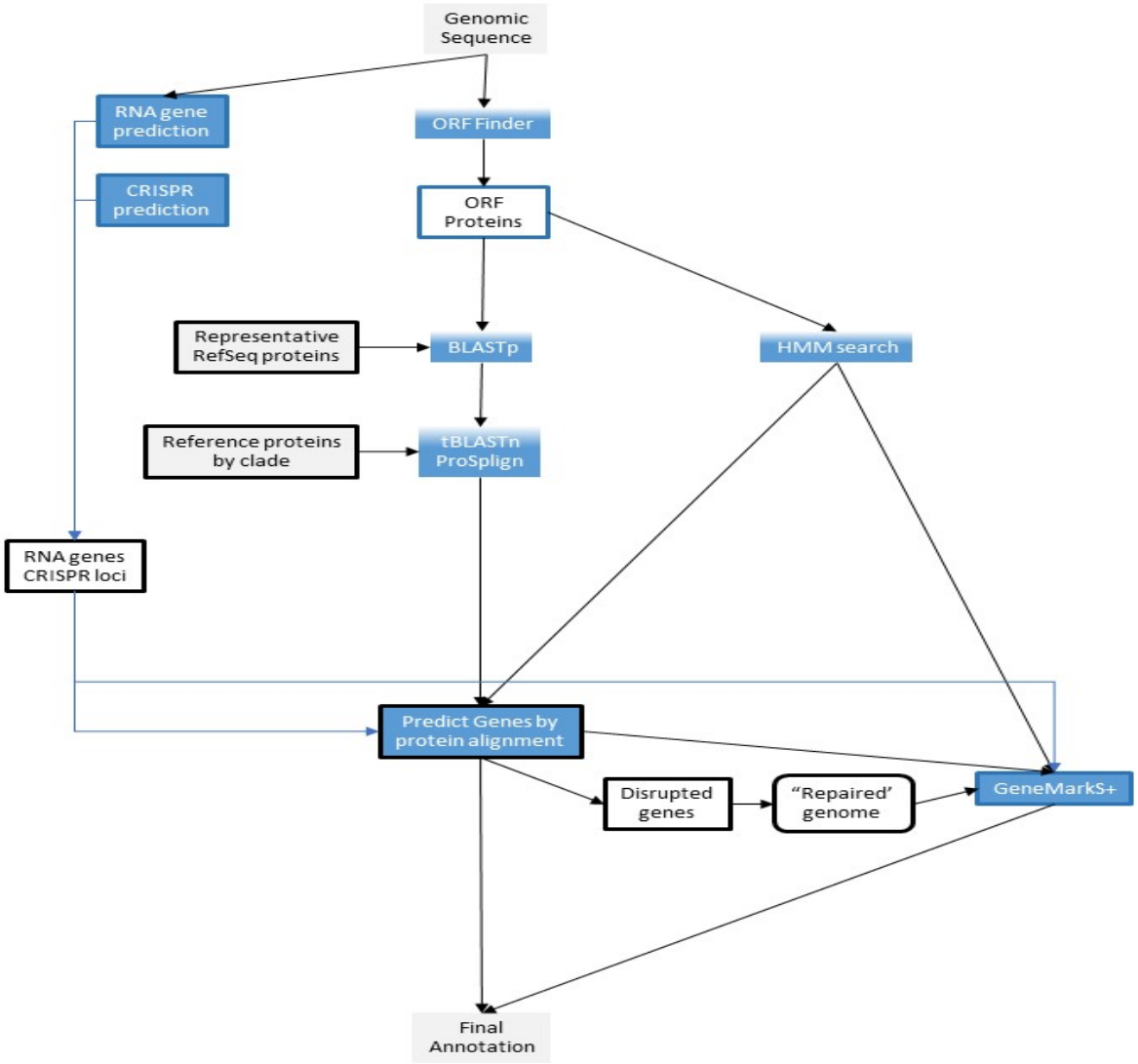
The National Center for Biotechnology Information



Created in 1988 as a part of the National Library of Medicine at NIH

- Establish public databases
- Research in computational biology
- Develop software tools for sequence analysis
- Disseminate biomedical information

The NCBI microbial annotation pipeline

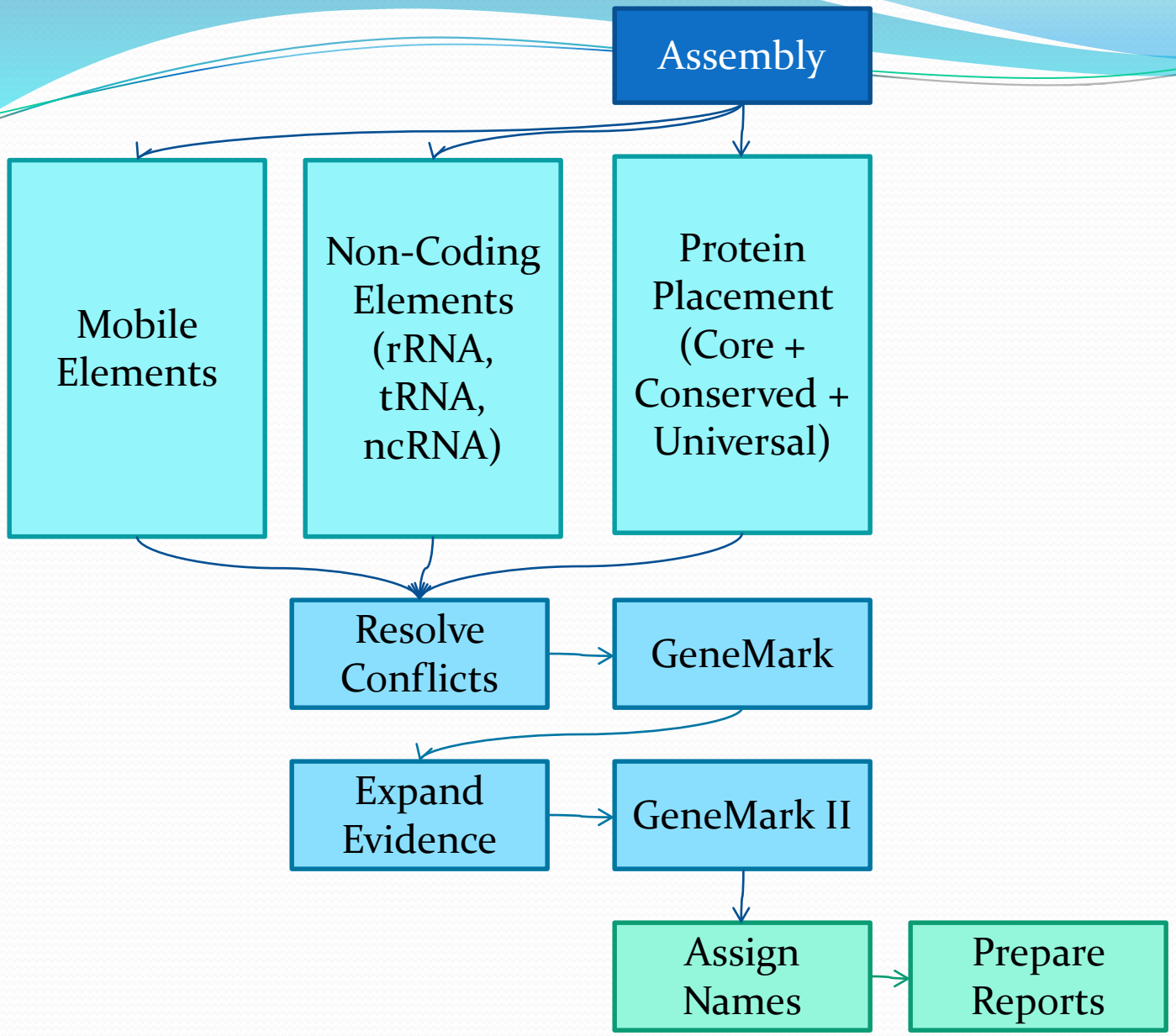


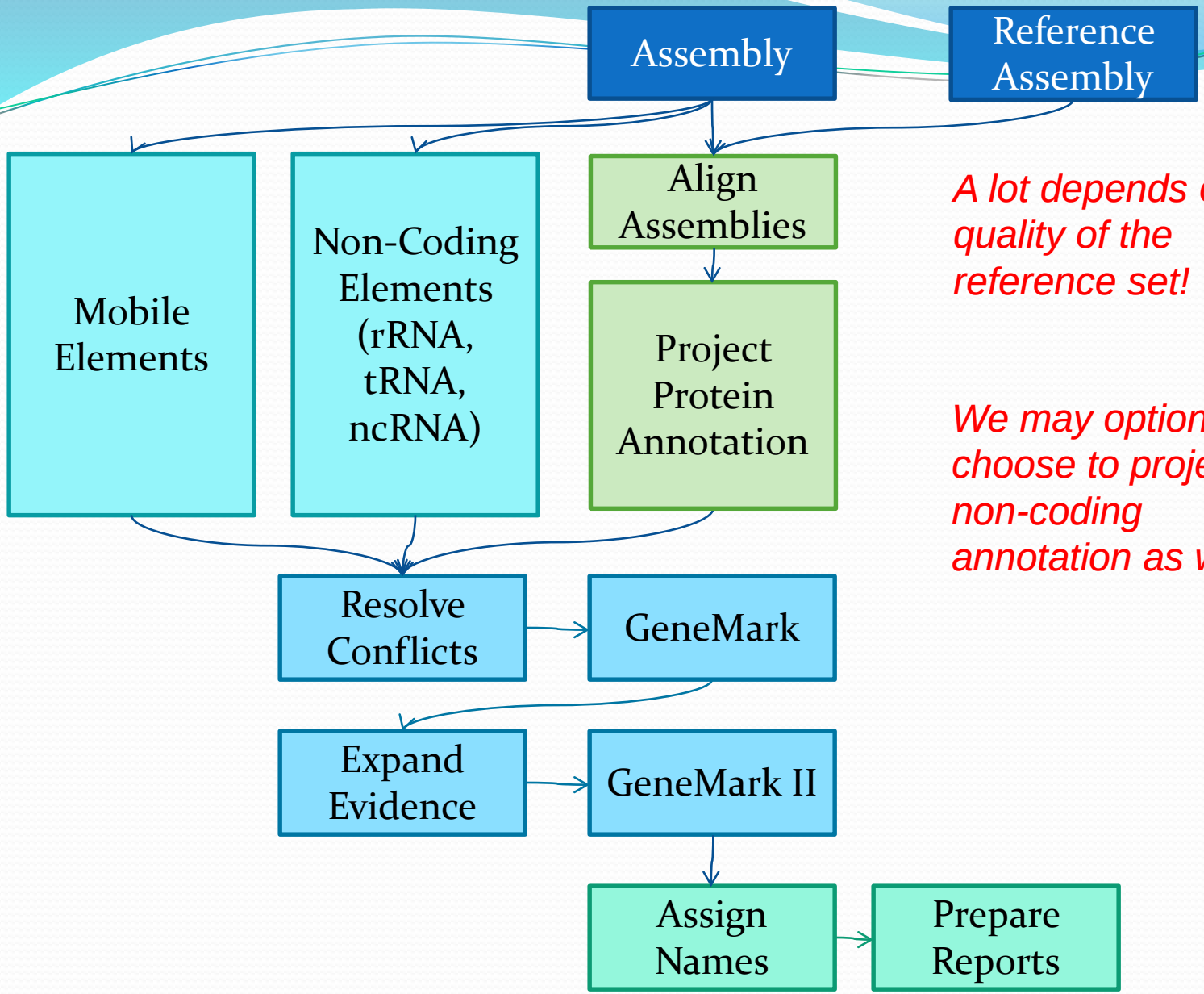
http://www.ncbi.nlm.nih.gov/genome/annotation_prok/process/

Bacterial Annotation

- Released PGAP-2 replacement available for GenBank
- Support large throughput
  - *1,000 assemblies/day*

- Current development tasks:
  - Replace dataflow to create RefSeq assemblies
  - Perform annotation via mapping from “close” assembly


Assembly
Mobile Elements
Non-Coding Elements (rRNA, tRNA, ncRNA)
Protein Placement (Core + Conserved + Universal)
Resolve Conflicts
GeneMark
Expand Evidence
GeneMark II
Assign Names
Prepare Reports



A lot depends on quality of the reference set!

We may optionally choose to project non-coding annotation as well

Pathogen Analysis

- Rapid identification of species and strain
- Rapid assembly and annotation of bacterial short read sequences
- Rapid identification of key characteristics separating “outbreak” strains from background samples

Pathogen Analysis

Health > Pathogen Detection

Pathogen Detection BETA

NCBI Pathogen Detection integrates bacterial pathogen genomic sequences originating in food, environmental sources, and patients. It quickly clusters and identifies related sequences to uncover potential food contamination sources, helping public health scientists investigate foodborne disease outbreaks.

Find isolates now!

Examples:

1. Search for isolates encoding a mobile colistin resistance gene and a KPC beta-lactamase
search: AMR_genotypes:mcr* AND AMR_genotypes:blaKPC*
2. Search for Salmonella isolates from the USA
search: geo_loc_name: USA AND taxgroup_name:"Salmonella enterica"

Explore the Data

Species	New Isolates	Total Isolates
Salmonella enterica	518	104,696
E.coli and Shigella	1	38,269
Listeria monocytogenes	1	16,830
Campylobacter jejuni	2	14,984

See more organisms...

Learn More

About

FAQ

Factsheet

Antimicrobial Resistance

Contributors

Data Resources

Isolates Browser

Antimicrobial resistance reference gene database

Isolates with antibiotic resistant phenotypes

Beta-lactamase resources

Download analysis results (FTP)

Submit

How to submit data

How to submit antibiotic resistance phenotypes

How to submit beta-lactamases

NCBI Submission Portal

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

Pathogen Analysis Timeline

2016

May: NCBI Pathogen Browser released
9th GMI at Food and Agriculture Organization of the United Nations (FAO), Italy

July: "..the Pathogen Detection System ... identified *mcr-1*in the whole genome sequence of an *E. coli* isolate from a Connecticut patient ... this is the fourth isolate from a U.S. patient to contain the *mcr-1*gene."

Aug: CDC reports on pilot project for Listeria
with WGS number of clusters goes up, number of isolates per cluster goes down, outbreaks solved increased, links to food isolates goes up

Nov: investigation into using wgMLST includes
SKESA assembler (other assemblers slow, inaccurate, memory and space reqs)
wgMLST clustering (will improve accuracy and speed)

Dec: addition of antimicrobial resistant genotypes/phenotypes to pathogen browser
NCBI Town Hall - we need to do things differently

Pathogen Analysis Timeline

2017

Feb: maximum compatibility algorithm for reconstruction of recent evolutionary history

May: 10th GMI, Mexico
David Lipman leaves NCBI
development of rapid reports based on SKESA/wgMLST for FDA begins as current SNP pipeline has limitations

July: alpha release of new pathogen browser to FDA
in part to aid navigation as data submissions increases
wgMLST Rapid Reports in production

Sept: first tier1 review for pathogen
prioritize turnaround time to FDA and measurements

Pathogen Analysis 2017

2017 Examples

Large Outbreak, 3-4 months: Papaya

- 251 People from 25 states: Salmonella Thompson (144), Kiambu (54), Anatum (20), Agona (12), Gaminara (7), Urbana (7), Newport & Infantis (4), and Senftenberg (3)
- NCBI Pathogen Detection critical to helping us track and disseminate information about this event and triage product positives (not all positive samples linked to illness)

Warning Letter - Gold Star Smoked Fish Corp.

- https://www.fda.gov/iceci/enforcementactions/warningletters/2017/ucm589689.htm (resident pathogen)

Pathogen Analysis Timeline

2018

Jan: German FDA releases batch of Salmonella genomes
meeting with FDA
new gene categories for reporting, including biocides, metal resistance, virulence
improvements to submission processes, rapid reports

Feb: 2nd meeting with GenFS steering committee to prioritize goals for the next 2 years
new browser released, 2nd tier1 review for pathogen, IEB Seminar
measurement of turnaround time for submissions for Rapid Reports and SNP processing

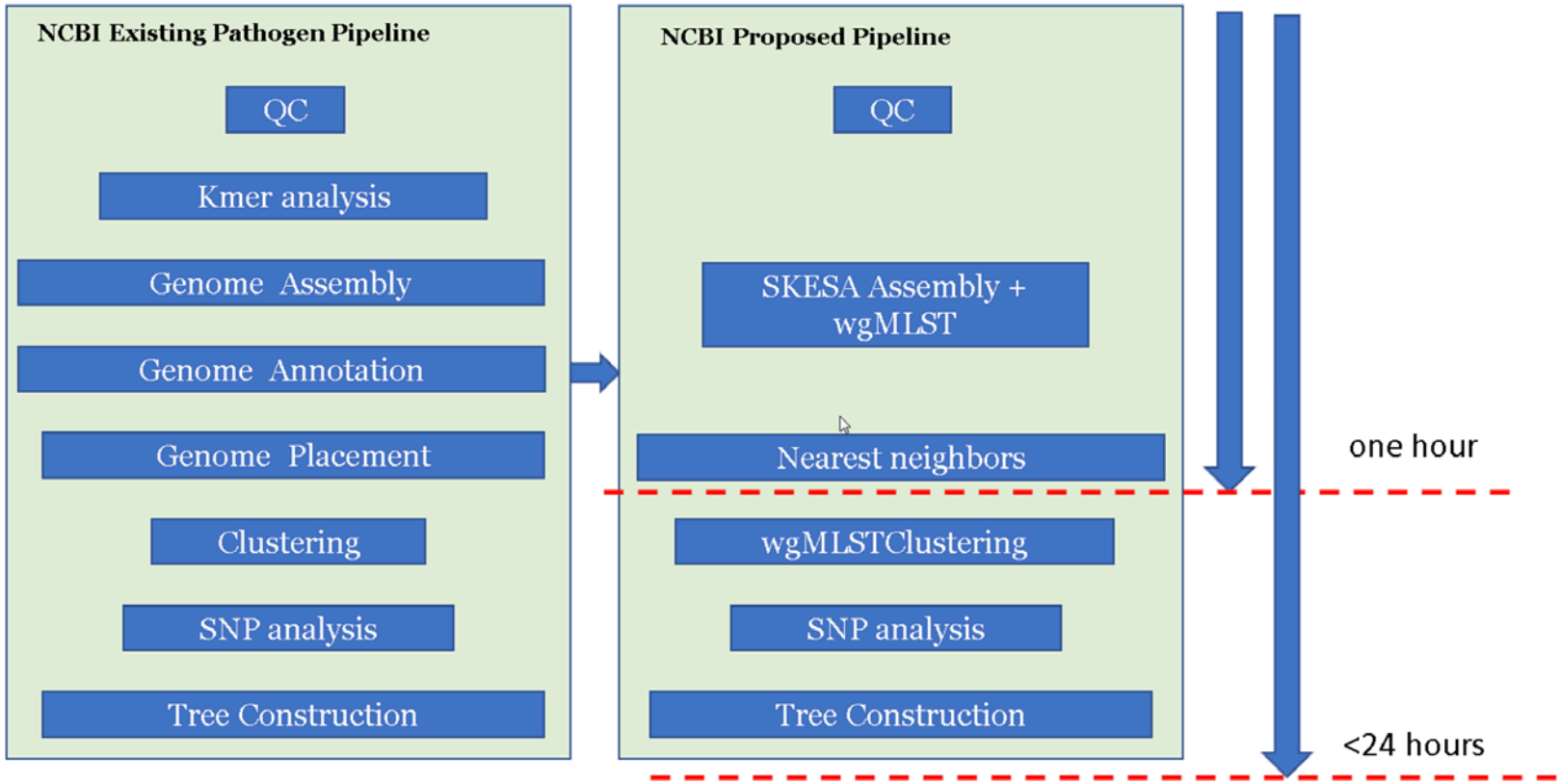
Apr: switch to wgMLST clustering, reduction in turnaround time for Salmonella
turnaround time should improve significantly

May: 11th GMI hosted by WHO

Fall: additional genes/proteins reported into pathogen browser
additional organisms like Staphylococcus added to the system

Dec: expect all 90 000 foodborne pathogens to be sequenced and submitted in real-time
~60 000 are Salmonella

Pathogen Analysis Pipeline

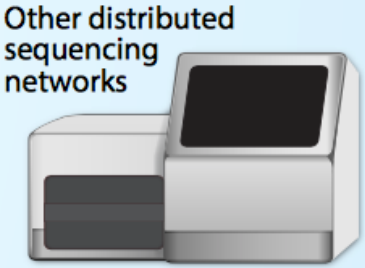


- SKESA de novo assembly + wgMLST will replace and speed up several parts of the existing pipeline
- delivery of nearest neighbors within one hour of data deposition into SRA

Basic Data Flow for Global WGS Public Access Databases

DATA ACQUISITION

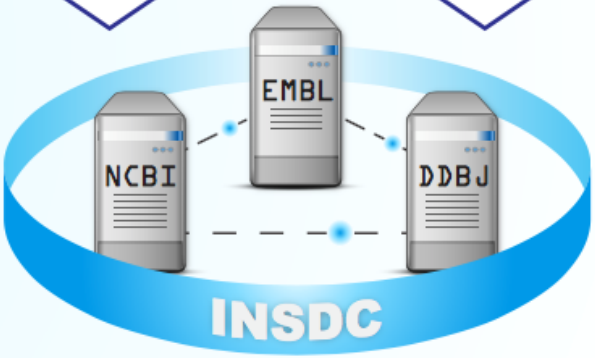
Sequence and upload genomic and geographic data



DATA ASSEMBLY, ANALYSIS, AND STORAGE

International Nucleotide Sequence Database Collaboration (INSDC)
Shared Public Access Databases

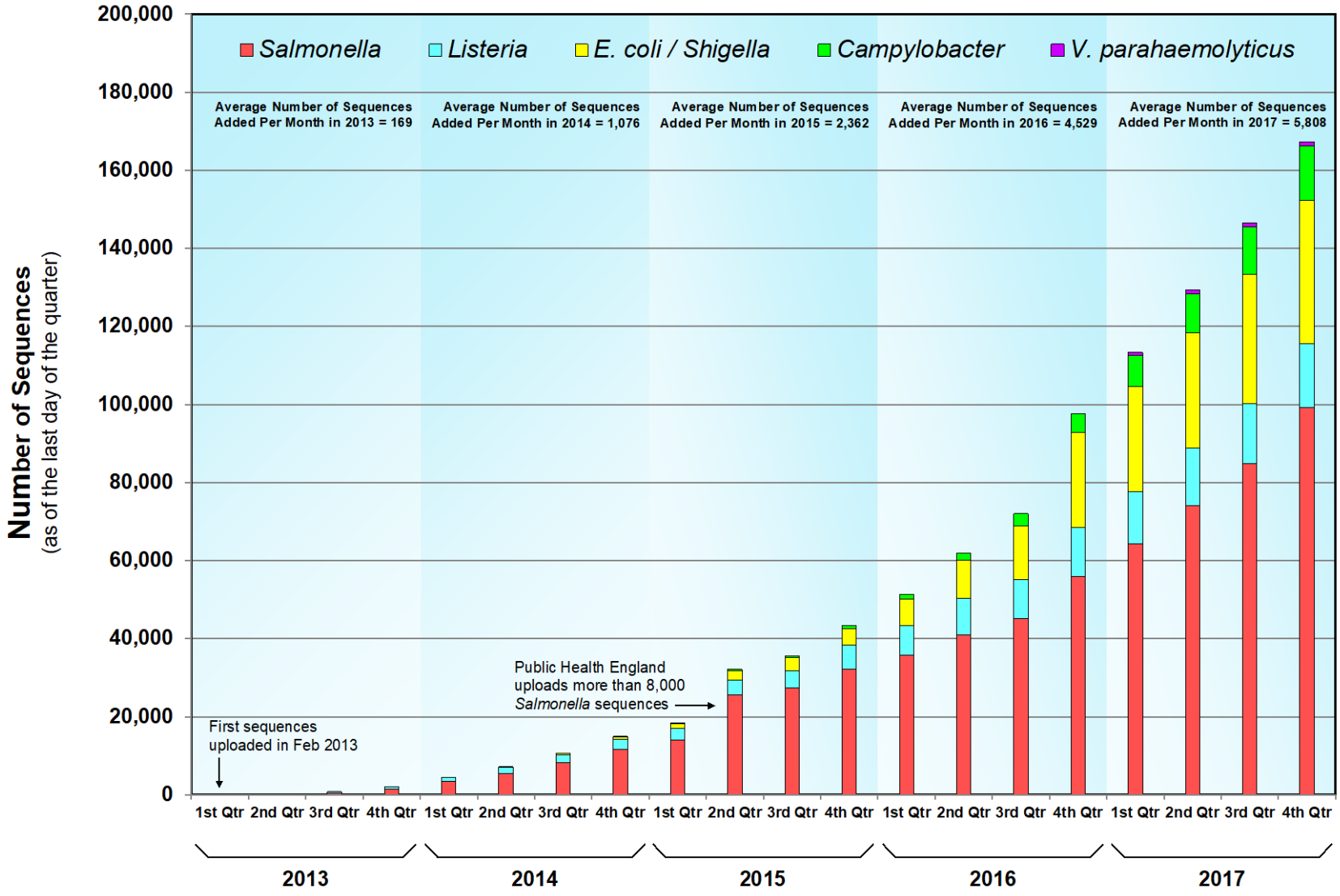
- *NCBI – National Center for Biotechnology Information*
- *EMBL – European Molecular Biology Laboratory*
- *DDBJ – DNA Databank of Japan*



PUBLIC HEALTH APPLICATION AND INTERPRETATION OF DATA

- *Find clinical links*
- *Identify clusters*
- *Conduct traceback*
- *Develop rapid methods*
- *Develop culture independent tests*
- *Develop new analytical software*

Total Number of Sequences in the GenomeTrakr Database



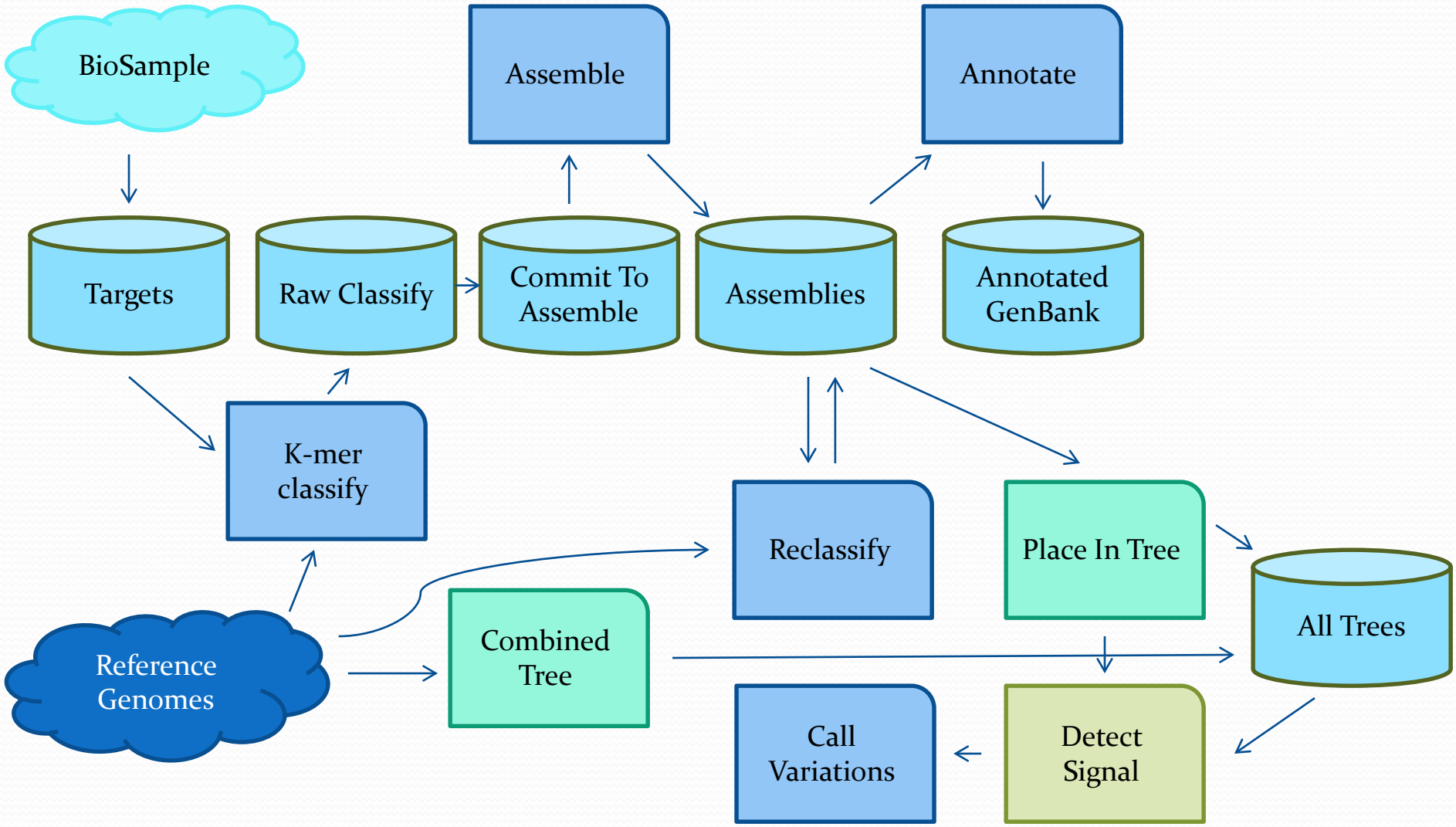
Pathogens of Interest

- Food-borne illness
  - *Salmonella enterica*
  - *Listeria monocytogenes*
  - *Escherichia coli O157* (STEC)
  - *Campylobacter jejuni*

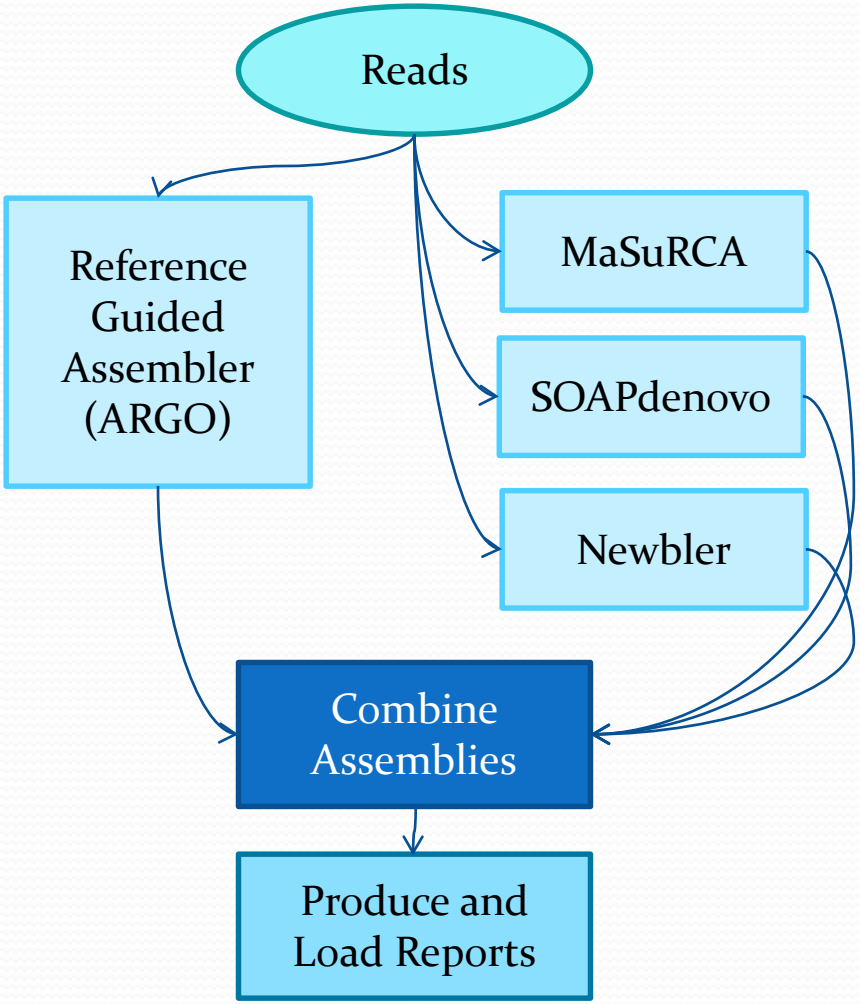
- Resistant / virulent hospital-acquired infections
  - Resistant *Staphylococcus aureus* (MRSA)
  - Resistant *Klebsiella pneumoniae*

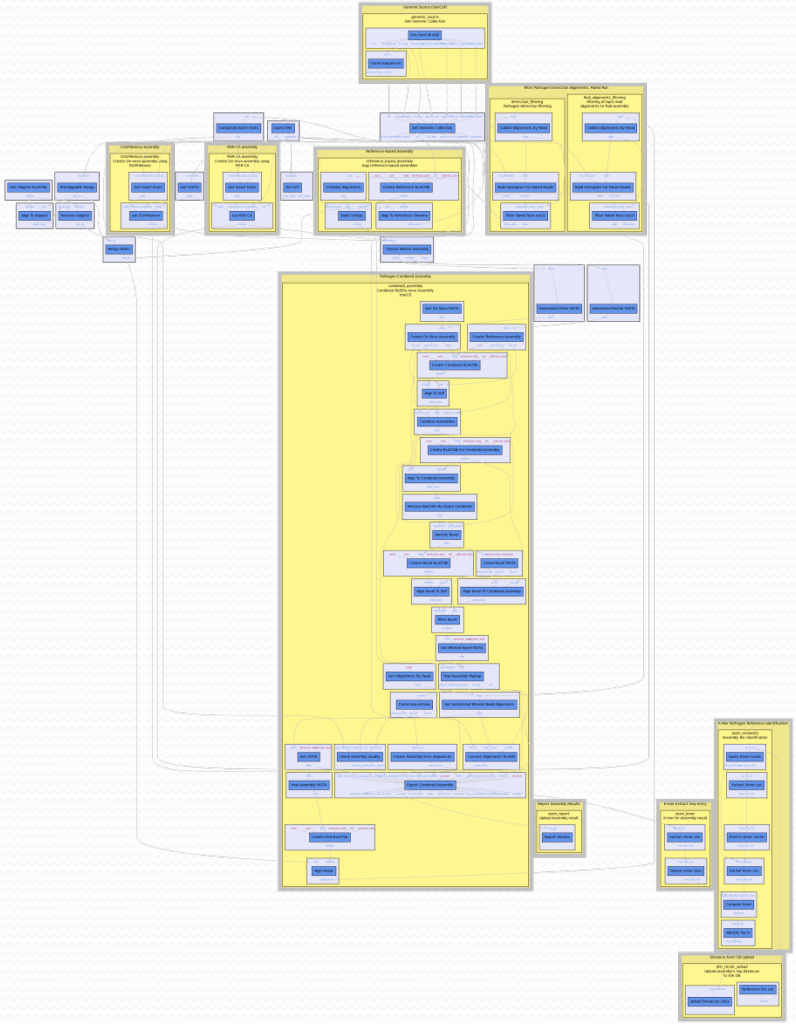
- Difficult to culture
  - *Mycobacterium tuberculosis*

Pathogen Analysis



Assembly Process





Pathogen Future

- Increase throughput
  - Currently, see throughput of 400 samples/day
  - Goal is 1,000 samples/day
- Add automation tasks for signal detection
- Add automated submission and report back to submitter
- Predicting virulence and drug resistance
- Extending to more organisms
- Provide public browsing resources

NCBI BLAST in the Cloud!

Cloud BLAST

BLAST Searches at a Cloud Provider

The NCBI provides a BLAST server image hosted with these cloud vendors: Amazon Web Services (AWS), Google Compute Engine (GCE). This allows users to run stand-alone searches with the BLAST+ applications, submit searches through a subset of the NCBI-BLAST URL API, and perform searches with a simplified webpage. The server image includes a FUSE client that will download BLAST databases during the first search. The server image runs on Ubuntu Linux. This page provides links to the latest server images as well as links to documentation about BLAST in the cloud.

The most recent BLAST+ server image for AWS

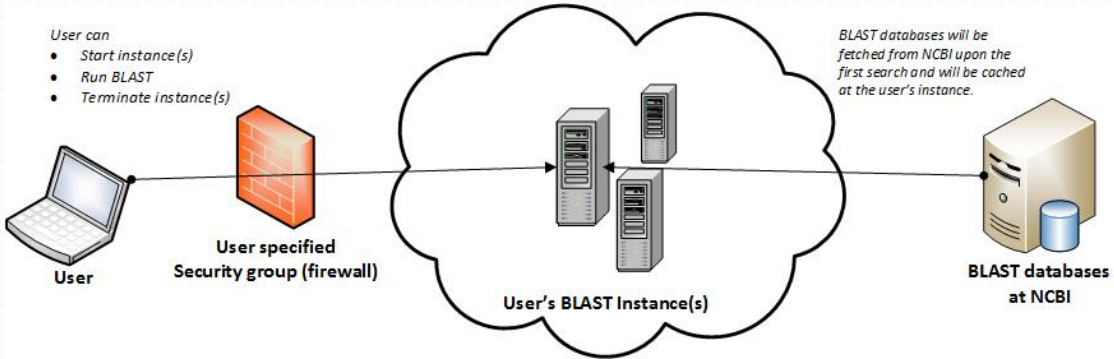
AWS Marketplace: https://aws.amazon.com/marketplace/pp/B00N44P7L6

The most recent BLAST+ server image for GCE

Google Compute Engine: https://googlegenomics.readthedocs.org/en/latest/use_cases/run_familiar_tools/ncbiblast.html

Resources

- NCBI BLAST cloud documentation - How to setup and use the server image, as well as documentation for the simplified URL API.
- BLAST+ user manual - How to run stand-alone BLAST searches.
- Sample PERL code - This script can submit URL API searches to your instance. It can be easily modified for specific tasks.
- YouTube Video recording from the BLAST in the Cloud webinar (July, 2014)
- CloudBlast Poster - Presented at NIH in 2014.



Cloud Options for NCBI BLAST!

- Amazon Web Services
- Google Compute Engine
- Any cloud app with Galaxy project

Select the NCBI CloudBLAST AMI

awsmarketplace

Amazon Web Services Home

Hello, Leonardo Marino. (Sign out)

Your Account | Help | Sell on AWS Marketplace

Shop All Categories ▾ Search AWS Marketplace GO Your Software

NCBI BLAST

Sold by: NCBI

This BLAST AMI is a very exciting development as it allows users to perform sequence similarity searches without restriction they might encounter at a public website and without the work of setting up stand-alone BLAST. The AMI includes a FUSE client that automatically downloads the most popular BLAST databases from the NCBI, and users can still upload their own custom databases. The AMI allows users to run stand-alone searches with the BLAST+ applications, submit searches through a subset of the NCBI-BLAST URL API, and perform searches with a simplified webpage.

Customer Rating	Be the first to review this product
Latest Version	2015-05-18-2.2.31 (Other available versions)
Base Operating System	Linux/Unix, Ubuntu 12.04
Delivery Method	64-bit Amazon Machine Image (AMI) (Learn more)
Support	See details below
AWS Services Required	Amazon EC2, Amazon EBS
Highlights	▪ This AMI is preconfigured with the latest BLAST+ release and has a simplified BLAST web page. ▪ This AMI includes a FUSE client that automatically downloads and caches popular NCBI databases such as nr, nt, swissprot, refseq, and PDB. ▪ This AMI supports a subset of the NCBI BLAST URL API allowing remote submission and formatting of searches.

Continue — You will have an opportunity to review your order before launching or being charged.

Recommended Products

ScaleArc | panzura | Microsoft

ScaleArc for MySQL - Enterprise
ScaleArc, Inc.
Starting from **$0.52/hr** or from **$3,700/yr** for software
Free Trial

Panzura Global NAS Plus Appliance
Panzura
$2.06/hr for software
Free Trial

Microsoft SharePoint Foundation 2010
Amazon Web Services
$0.018 to $9.348/hr incl EC2 charges

Product Description

This BLAST AMI is a very exciting development as it allows users to perform sequence similarity searches without restriction they might encounter at a public website and without the work of setting up stand-alone BLAST. The AMI includes a FUSE client that automatically downloads the most popular BLAST databases from the NCBI, and users can still upload their own custom databases. The AMI allows users to run stand-alone searches with the BLAST+ applications, submit searches through a subset of the NCBI-BLAST URL API, and perform searches with a simplified webpage.

Product Details

Version: 2015-05-18-2.2.31
Available on AWS Marketplace Since: 08/28/2014
Note: Always ensure your operating system is current for your needs.

Resources

BLAST Searches at a Cloud Provider
Documentation
Sample PERL script

Usage Instructions

Please follow the steps called out in the "Blast at AWS document", found under Resources at http://blast.ncbi.nlm.nih.gov/Blast.cgi?

Show more

Support Details

NCBI BLAST
blast-help@ncbi.nlm.nih.gov
Please allow 24 hours

Pricing Details

For region
US East (N. Virginia)

Hourly Fees
Total hourly fees will vary by instance type and EC2 region.

EC2 Instance Type	EC2 Usage	Software	Total
cc2.8xlarge	$2.00/hr	$0.00/hr	$2.00/hr
cr1.8xlarge	$3.50/hr	$0.00/hr	$3.50/hr
g2.8xlarge	$2.60/hr	$0.00/hr	$2.60/hr
m3.medium	$0.07/hr	$0.00/hr	$0.07/hr
m3.large	$0.14/hr	$0.00/hr	$0.14/hr
m3.xlarge	$0.28/hr	$0.00/hr	$0.28/hr
m3.2xlarge	$0.56/hr	$0.00/hr	$0.56/hr
i2.xlarge	$0.853/hr	$0.00/hr	$0.853/hr
i2.2xlarge	$1.705/hr	$0.00/hr	$1.705/hr
i2.4xlarge	$3.41/hr	$0.00/hr	$3.41/hr
i2.8xlarge	$6.82/hr	$0.00/hr	$6.82/hr
c3.large	$0.105/hr	$0.00/hr	$0.105/hr
c3.xlarge	$0.21/hr	$0.00/hr	$0.21/hr
c3.2xlarge	$0.42/hr	$0.00/hr	$0.42/hr
c3.4xlarge	$0.84/hr	$0.00/hr	$0.84/hr
c3.8xlarge	$1.68/hr	$0.00/hr	$1.68/hr
r3.large	$0.175/hr	$0.00/hr	$0.175/hr
r3.xlarge	$0.35/hr	$0.00/hr	$0.35/hr
r3.2xlarge	$0.70/hr	$0.00/hr	$0.70/hr
r3.4xlarge	$1.40/hr	$0.00/hr	$1.40/hr
r3.8xlarge	$2.80/hr	$0.00/hr	$2.80/hr
d2.xlarge	$0.69/hr	$0.00/hr	$0.69/hr
d2.2xlarge	$1.38/hr	$0.00/hr	$1.38/hr
d2.4xlarge	$2.76/hr	$0.00/hr	$2.76/hr
d2.8xlarge	$5.52/hr	$0.00/hr	$5.52/hr

EBS General Purpose (SSD) volumes
$0.10 per GB-month of provisioned storage

Assumes On-Demand EC2 pricing
Learn about instance types

Data Transfer Fees not included
Learn more about Data Transfer Fees

For lower prices you can utilize:
Reserved Instances
Spot Instances
See all Pricing Details

There are no product reviews yet. Be the first to review this product.

Create Your Own Review

Select a Reasonable* Instance

awsmarketplace

Amazon Web Services Home

Hello, Leonardo Marino. (Sign out)

Your Account | Help | Sell on AWS Marketplace

Shop All Categories ▾ Search AWS Marketplace GO Your Software

Version 2015-05-18-2.2.31 of this software is now available. See release notes

Launch on EC2:

NCBI BLAST

1-Click Launch
Review, modify, and launch

Manual Launch
With EC2 Console, APIs or CLI

Click "Launch with 1-Click" to launch this software with the settings below

The default settings are provided by the software seller and AWS Marketplace.

Version
2015-05-18-2.2.31, released 05/20/2015

Region
US East (N. Virginia)

EC2 Instance Type

cc2.8xlarge
cr1.8xlarge
g2.8xlarge
m3.medium
m3.large
m3.xlarge
m3.2xlarge
i2.xlarge
i2.2xlarge

Memory	30 GiB
CPU	26 EC2 Compute Units (8 virtual cores with 3.25 EC2 Compute Units each)
Storage	EBS storage only
Platform	64-bit
Network performance	High
API Name	m3.2xlarge

VPC Settings
Will launch into: subnet-1e506b24

Security Group

A security group acts as a firewall that controls the traffic allowed to reach one or more instances. Learn more about Security Groups.

You can create a new security group based on seller-recommended settings or choose one of your existing groups.

NCBI BLAST-2015-05-18-2-2-31-AutogenByAWSMP-

Description:
This security group was generated by AWS Marketplace and is based on recommended settings for NCBI BLAST version 2015-05-18-2.2.31 provided by NCBI

Connection Method	Protocol	Port Range	Source (IP or Group)
SSH	tcp	22 - 22	0.0.0.0/0
HTTP	tcp	80 - 80	0.0.0.0/0

Warning
Rules with source of 0.0.0.0/0 allows all IP addresses to access your instance. We recommend limiting access to only known IP addresses.

Key Pair
Blast-demo

Price for your selections:

$0.56 / hour
$0.56 m3.2xlarge EC2 Instance usage fees +
$0.00 hourly software fee

$0.10 / GB / month
EBS General Purpose (SSD)

Launch with 1-Click

Cost Estimator

$403.20 / month
m3.2xlarge EC2 Instance usage fees
Assumes 24 hour use over 30 days

Software Charges

$0.00 / month
$0.00 hourly software fees for m3.2xlarge

AWS Infrastructure Charges

$403.20 / month
Cost varies for storage fees
$403.20 hourly EC2 Instance fees for m3.2xlarge
Varied EBS Storage and data transfer fees

Launch the instance

An instance of this software is now deploying on EC2.

- If you would like to check the progress of this deployment, go to the AWS Management Console
- The software will be ready in 2-3 minutes.

Usage Instructions

Please follow the steps called out in the "Blast at AWS document", found under Resources at http://blast.ncbi.nlm.nih.gov/Blast.cgi?CMD=Web&PAGE_TYPE=BlastDocs&DOC_TYPE=CloudBlast

Software Installation Details

Product	NCBI BLAST
Version	2015-05-18-2.2.31, released 05/20/2015
Region	US East (N. Virginia)
EC2 Instance Type	m3.2xlarge
VPC	vpc-cbc7f8ae
Subnet	subnet-1e506b24
Security Group	NCBI BLAST-2015-05-18-2-2-31-AutogenByAWSMP-
Key Pair	Blast-demo

How the Genome has changed?

- More complex genome structures – (chromosomes, organelles, plasmids)
- Genome sequencing – NextGen sequencing
- More complex genome assembly – (chromosomes, scaffolds, contigs)
- Genome-scale projects - (transcriptome, exome, epigenomics, proteomics)
- Multi-isolate genome sequencing - (1001 Arabidopsis, 1000 human genomes)
- Meta-genomes
- Now useful for drug development

New resources at NCBI

BioProject

BioProject | Search

Limits Advanced Help

BioProject

A BioProject is a collection of biological data related to a single initiative, originating from a single organization or from a consortium. A BioProject record provides users a single place to find links to the diverse data types generated for that project.

Using BioProject

- Help
- Submission

Browse BioProject

- By Project attributes
- Download (FTP)

Large Initiatives

- 1000 Genomes
- ENCODE
- HMP

NCBI Resources

- BioSample
- dbGaP
- Genome

External Resources

- Genome projects at DOE
- Genome News Network
- GOLD - Genome On Line Database

New genomic resources at NCBI

Genome

Genome | Search

Limits Advanced Help

Genome

This resource organizes information on genomes including sequences, maps, chromosomes, assemblies, and annotations.

Using Genome

- Help
- Browse by Organism
- Download / FTP
- Submit a genome

Custom resources

- Human Genome
- Microbes
- Organelles
- Plants
- Viruses

Other Resources

- Assembly
- BioProject
- BioSample
- Map Viewer
- Protein Clusters

Genome Tools

- BLAST the Human Genome
- Genomic groups BLAST
- NCBI remap
- Genome Decoration Page

Genome Annotation and Analysis

- Eukaryotic Genome Annotation
- Prokaryotic Genome Annotation
- PASC (Pairwise Sequence Comparison)
- TaxPlot (3-way Genome Comparison)

External Resources

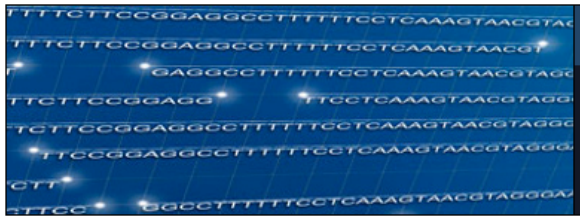
- GOLD - Genomes Online Database
- Ensembl Genome Browser
- Bacteria Genomes at Sanger
- Large-Scale Genome Sequencing (NHGRI)

New resources at NCBI

Assembly

Assembly | Search

Advanced Browse organisms Help



Assembly

Genome assembly organization and additional information.

Using Assembly

- Assembly Help
- Browse by Organism
- NCBI Assembly Data Model
- Assembly Basics

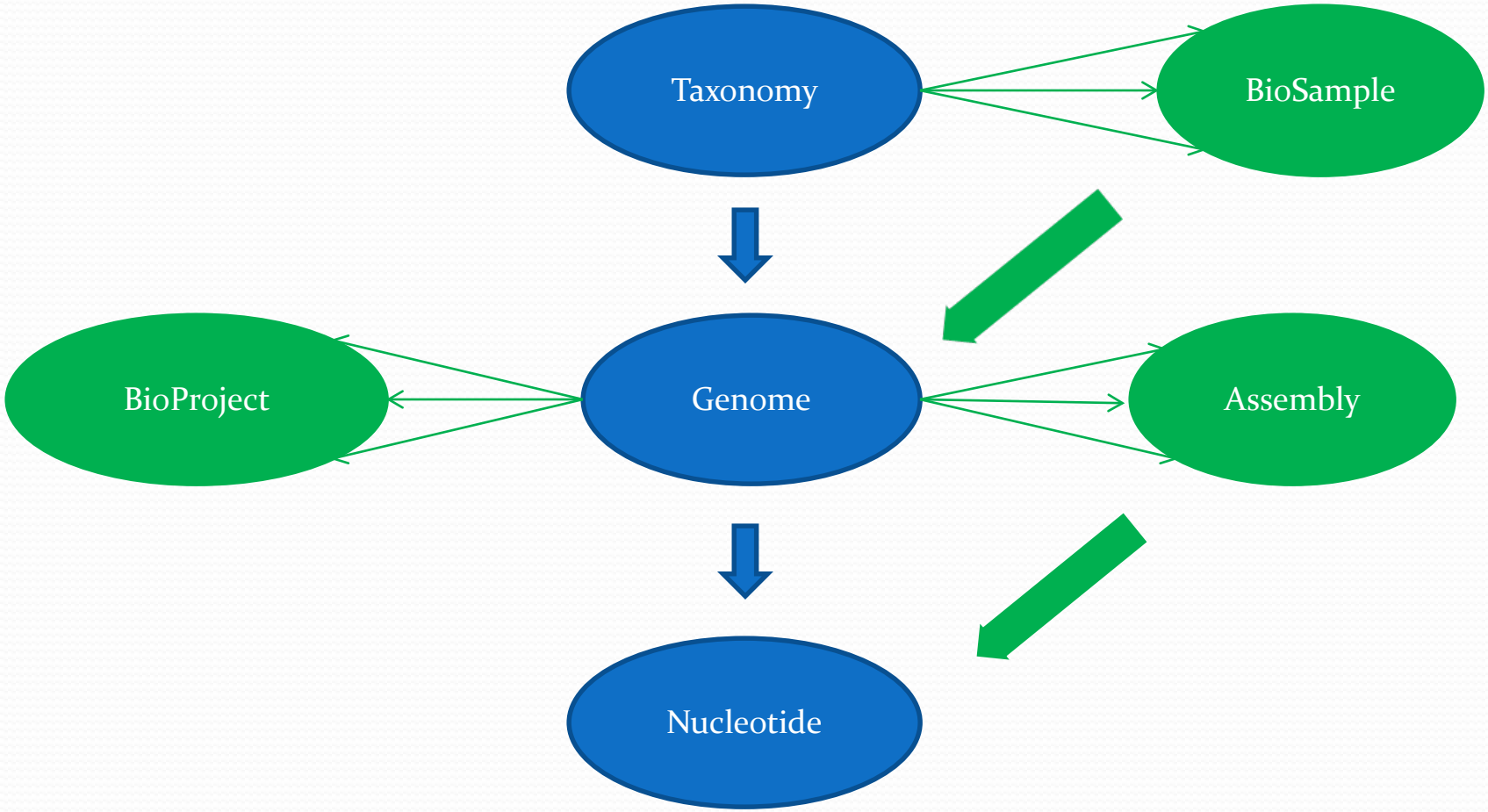
Submitting an Assembly

- Submission Information
- Submission FAQ
- AGP Specifications
- AGP Validation

Related Resources

- Genome
- Genome Reference Consortium

Why do we need new databases?



BioProject, Genome, Assembly

- BioProject is an administrative object (defined by goal, target, funding, collaboration)
- Genome is a biological object defining an organism at molecular level
- Genome assembly is a complex data structure that defines the structure, relative position (scaffold) and chromosome placement of DNA sequences originated from a single sample

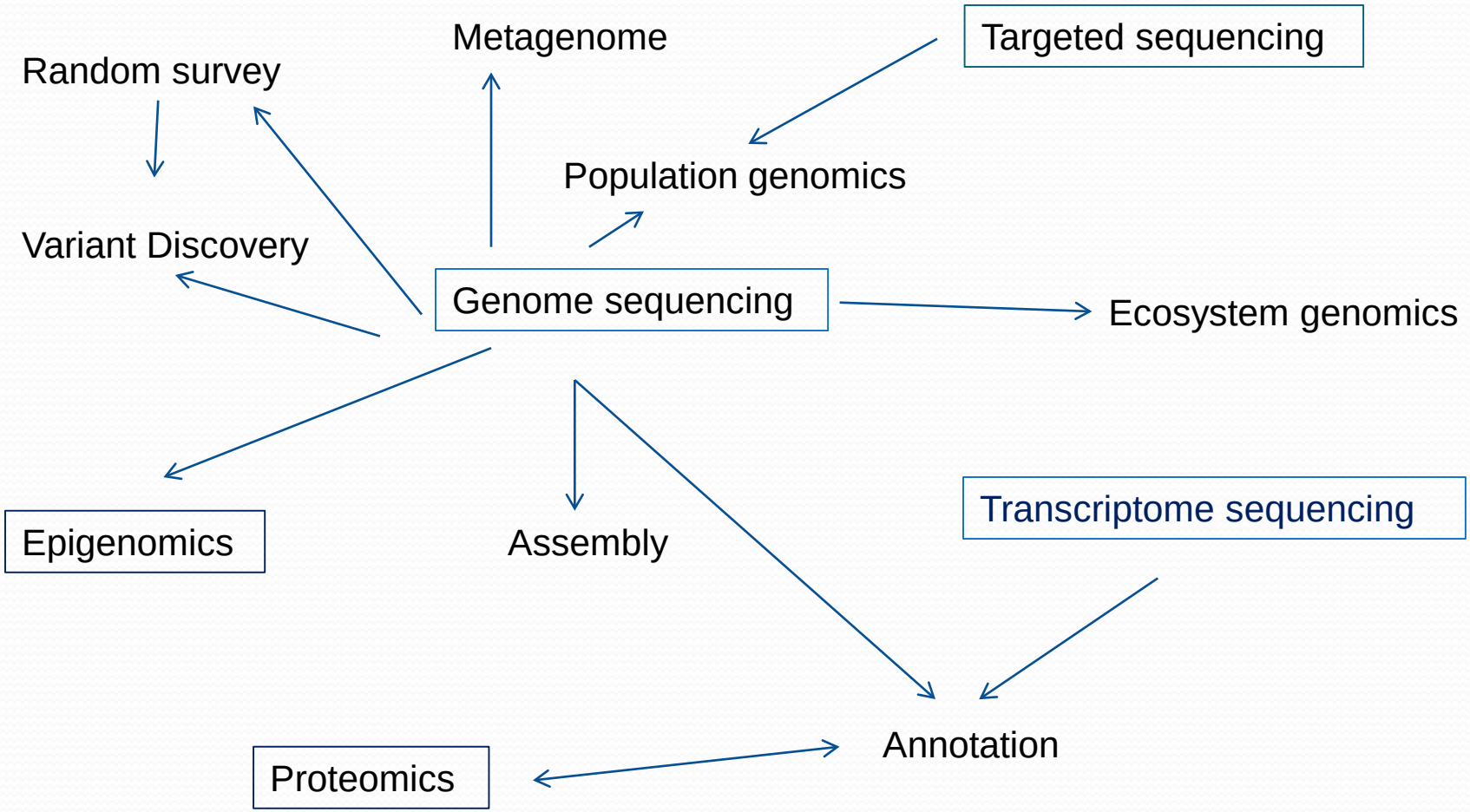
What is a Genome project?

- **Genome project** is a scientific endeavor that ultimately aims to determine the complete genome sequence of an organism and …

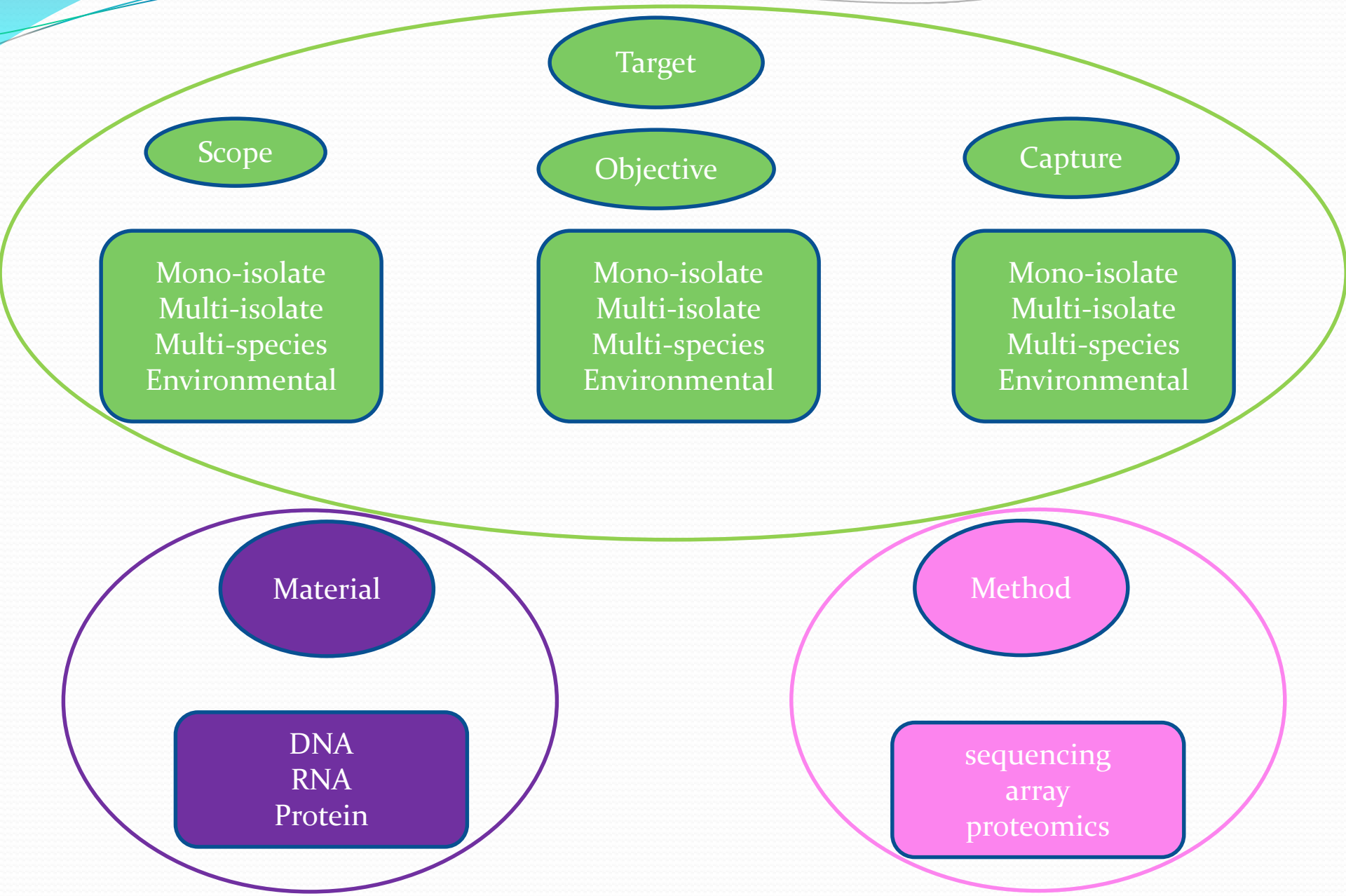
Aims to annotate protein-coding genes and other important genome-encoded features and …

Aims to understand the biology, physiology, and evolution of the organism.

Genome Project -> BioProject



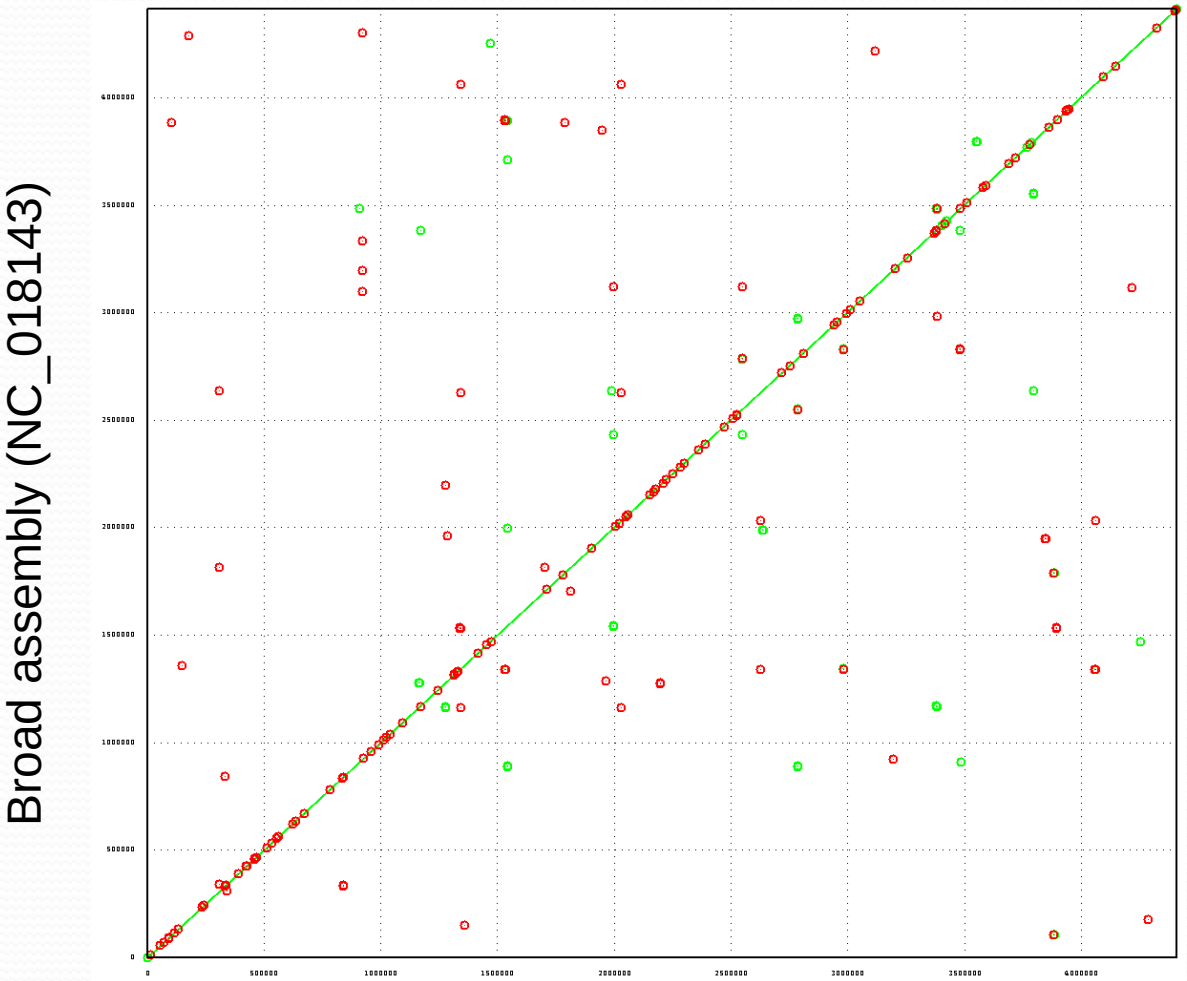
BioProject data model



Why do we need a database of genome assemblies?

- We are in a period of extraordinary growth in genomics data.
- To get the full benefit from all this data, it is important that users can integrate data from different sources. Integration only works, if users know whether or not the different data were reported in the same coordinate system.

TB H37Rv Sanger vs. Broad



Mycobacterium genomes at NCBI

Genome | Genome | Search

Genome Information by organism

Mycobacterium (taxid:1763) | Search by organism | Clear

Download Reports from FTP site

Overview [37] | Eukaryotes [0] | Prokaryotes [266] | Viruses [0]

First Previous **Shown: 1 - 37 out of 37 items** Next Last

Download selected records

Organism/Name	Kingdom	Group	SubGroup	Size (Mb)	Chr	Organelles	Plasmids	BioProjects
	All	All	All					
Mycobacterium	Bacteria	Actinobacteria	Actinobacteria	6.26	1	-	2	8
Mycobacterium abscessus	Bacteria	Actinobacteria	Actinobacteria	5.51	1	-	-	46
Mycobacterium africanum	Bacteria	Actinobacteria	Actinobacteria	4.39	1	-	-	1
Mycobacterium avium	Bacteria	Actinobacteria	Actinobacteria	5.48	1	-	1	27
Mycobacterium bovis	Bacteria	Actinobacteria	Actinobacteria	4.37	1	-	-	18
Mycobacterium canettii	Bacteria	Actinobacteria	Actinobacteria	4.48	1	-	-	1
Mycobacterium chelonae	Bacteria	Actinobacteria	Actinobacteria	0	-	-	-	1
Mycobacterium chlorophenolicum	Bacteria	Actinobacteria	Actinobacteria	0	-	-	-	1
Mycobacterium chubuense	Bacteria	Actinobacteria	Actinobacteria	6.34	1	-	2	1
Mycobacterium colombiense	Bacteria	Actinobacteria	Actinobacteria	5.68	-	-	-	1
Mycobacterium fortuitum	Bacteria	Actinobacteria	Actinobacteria	0	-	-	-	1
Mycobacterium gilvum	Bacteria	Actinobacteria	Actinobacteria	5.98	1	-	-	1
Mycobacterium gilvum	Bacteria	Actinobacteria	Actinobacteria	5.78	1	-	-	1
Mycobacterium haemophilum	Bacteria	Actinobacteria	Actinobacteria	0	-	-	-	1
Mycobacterium hassiacum	Bacteria	Actinobacteria	Actinobacteria	0	-	-	-	2
Mycobacterium indicus pranii	Bacteria	Actinobacteria	Actinobacteria	0	-	-	-	1
Mycobacterium intracellulare	Bacteria	Actinobacteria	Actinobacteria	5.5	1	-	-	7
Mycobacterium kansasii	Bacteria	Actinobacteria	Actinobacteria	6.42	1	-	-	1
Mycobacterium leprae	Bacteria	Actinobacteria	Actinobacteria	3.27	1	-	-	3
Mycobacterium mageritense	Bacteria	Actinobacteria	Actinobacteria	6.5	-	-	-	2
Mycobacterium marinum	Bacteria	Actinobacteria	Actinobacteria	6.66	1	-	-	3
Mycobacterium massiliense	Bacteria	Actinobacteria	Actinobacteria	5.2	1	-	-	13
Mycobacterium microti	Bacteria	Actinobacteria	Actinobacteria	0	-	-	-	1
Mycobacterium monacense	Bacteria	Actinobacteria	Actinobacteria	0	-	-	-	1
Mycobacterium mucogenicum	Bacteria	Actinobacteria	Actinobacteria	0	-	-	-	1
Mycobacterium parascrofulaceum	Bacteria	Actinobacteria	Actinobacteria	6.3	-	-	-	1
Mycobacterium phlei	Bacteria	Actinobacteria	Actinobacteria	5.68	-	-	-	2
Mycobacterium pseudoshottsii	Bacteria	Actinobacteria	Actinobacteria	0	-	-	-	1
Mycobacterium rhodesiae	Bacteria	Actinobacteria	Actinobacteria	7.28	1	-	-	2
Mycobacterium smegmatis	Bacteria	Actinobacteria	Actinobacteria	6.99	1	-	-	4
Mycobacterium thermoresistibile	Bacteria	Actinobacteria	Actinobacteria	4.87	-	-	-	1
Mycobacterium tuberculosis	Bacteria	Actinobacteria	Actinobacteria	4.44	1	-	-	104
Mycobacterium tusciae	Bacteria	Actinobacteria	Actinobacteria	7.12	-	-	-	1
Mycobacterium ulcerans	Bacteria	Actinobacteria	Actinobacteria	5.81	1	-	-	2
Mycobacterium vaccae	Bacteria	Actinobacteria	Actinobacteria	0	-	-	-	2
Mycobacterium vanbaalenii	Bacteria	Actinobacteria	Actinobacteria	6.49	1	-	-	1
Mycobacterium xenopi	Bacteria	Actinobacteria	Actinobacteria	4.43	-	-	-	1

Mycobacterium tuberculosis genomes

Genome | Genome | Search

Genome Information by organism

Mycobacterium tuberculosis (taxid:1763) | Search by organism | Clear

Download Reports from FTP site

Overview [1] | Eukaryotes [0] | Prokaryotes [104] | Viruses [0]

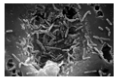
First Previous **Shown: 1 - 100 out of 104 items** Next Last

Download selected records

Organism/Name	BioProject	Group	SubGroup	Size (Mb)	GC%	Chromosomes		Plasmids		WGS	Scaffolds	Gene	Protein	Release Date	Modify Date	Status
		-- All Prokaryotes --	-- All Prokaryotes --			RefSeq	INSDC	RefSeq	INSDC							✓ All Complete No data SRA or Traces Scaffolds or contigs
Mycobacterium tuberculosis CCDC5079	PRJNA161943 PRJNA19585	Actinobacteria	Actinobacteria	4.4	65.60	NC_017523.1	CP001641.1	-	-	-	-	3695	3646	2011/07/05	2012/06/1	
Mycobacterium tuberculosis CCDC5180	PRJNA161941 PRJNA19583	Actinobacteria	Actinobacteria	4.41	65.60	NC_017522.1	CP001642.1	-	-	-	-	3638	3590	2011/07/05	2012/06/1	
Mycobacterium tuberculosis CDC1551	PRJNA57775 PRJNA223	Actinobacteria	Actinobacteria	4.4	65.60	NC_002755.2	AE000516.2	-	-	-	-	4293	4189	2001/10/02	2012/01/19	Complete
Mycobacterium tuberculosis CTRI-2	PRJNA161997 PRJNA43171	Actinobacteria	Actinobacteria	4.4	65.60	NC_017524.1	CP002992.1	-	-	-	-	4001	3944	2011/08/25	2012/06/13	Complete
Mycobacterium tuberculosis F11	PRJNA58417 PRJNA15642	Actinobacteria	Actinobacteria	4.42	65.60	NC_009565.1	CP000717.1	-	-	-	-	3998	3941	2007/06/07	2010/05/12	Complete
Mycobacterium tuberculosis H37Ra	PRJNA58853 PRJNA18883	Actinobacteria	Actinobacteria	4.42	65.60	NC_009525.1	CP000611.1	-	-	-	-	4084	4034	2007/05/31	2012/01/26	Complete
Mycobacterium tuberculosis H37Rv	PRJNA57777 PRJNA224	Actinobacteria	Actinobacteria	4.41	65.60	NC_000962.2	AL123456.2	-	-	-	-	4062	4003	2001/09/07	2012/06/12	Complete
Mycobacterium tuberculosis H37Rv	PRJNA170532 PRJNA37301	Actinobacteria	Actinobacteria	4.41	65.60	NC_018143.1	CP003248.1	-	-	-	-	4170	4111	2012/07/12	2012/07/25	Complete
Mycobacterium tuberculosis KZN 1435	PRJNA59069 PRJNA21055	Actinobacteria	Actinobacteria	4.4	65.60	NC_012943.1	CP001658.1	-	-	-	-	4107	4059	2009/07/09	2011/11/21	Complete
Mycobacterium tuberculosis KZN 4207	PRJNA83619 PRJNA21053	Actinobacteria	Actinobacteria	4.39	65.60	NC_016768.1	CP001662.1	-	-	-	-	4044	3996	2011/04/07	2012/06/18	Complete
Mycobacterium tuberculosis KZN 605	PRJNA54947 PRJNA21057	Actinobacteria	Actinobacteria	4.4	65.60	NC_018078.1	CP001976.1	-	-	-	-	4071	4001	2012/06/19	2012/06/20	Complete
Mycobacterium tuberculosis RGTB327	PRJNA157907 PRJNA73717	Actinobacteria	Actinobacteria	4.38	65.60	NC_017026.1	CP003233.1	-	-	-	-	3739	3691	2012/03/20	2012/08/01	Complete
Mycobacterium tuberculosis RGTB423	PRJNA162179 PRJNA73719	Actinobacteria	Actinobacteria	4.41	65.60	NC_017528.1	CP003234.1	-	-	-	-	3670	3622	2012/03/20	2012/08/08	Complete
Mycobacterium tuberculosis UT205	PRJEA162183 PRJEA74573	Actinobacteria	Actinobacteria	4.42	64.90	NC_016934.1	HE608151.1	-	-	-	-	3814	3796	2012/02/24	2012/08/23	Complete
Mycobacterium tuberculosis '98-R604 INH-RIF-EM'	PRJNA55399 PRJNA30979	Actinobacteria	Actinobacteria	4.29	65.50	-	-	-	-	ABVM01	16	4159	4112	2009/01/13	2010/06/08	Scaffolds or contigs
Mycobacterium tuberculosis 02_1987	PRJNA55097 PRJNA29167	Actinobacteria	Actinobacteria	4.44	65.30	-	-	-	-	ABLM01	22	4126	4081	2008/05/30	2010/06/08	Scaffolds or contigs

Mycobacterium tuberculosis overview

Organism Overview ; Genome Project Report ; Genome Annotation Report



Mycobacterium tuberculosis

Causative agent of tuberculosis

Lineage: Bacteria[3351]; Actinobacteria[547]; Actinobacteria[547]; Actinobacteridae[502]; Actinomycetales[485]; Corynebacterineae[219]; Mycobacteriaceae[38]; Mycobacterium[37]; Mycobacterium tuberculosis complex[5]; Mycobacterium tuberculosis[1]

Mycobacterium. This genus comprises a number of Gram-positive, acid-fast, rod-shaped aerobic bacteria and is the only member of the family *Mycobacteriaceae* within the order *Actinomycetales*. Like other closely related *Actinomycetales*, such as *Nocardia* and *Corynebacterium*, mycobacteria have unusually high More...

Representative

Calculated, Reference genome : Mycobacterium tuberculosis H37Rv

***Mycobacterium tuberculosis* strain H37RV**. This strain has been derived from the original human-lung H37 isolate in 1934, and has been used extensively worldwide in biomedical research. Unlike some clinical isolates, it retains full virulence in animal models of tuberculosis and is susceptible to drugs and receptive to genetic manipulation.

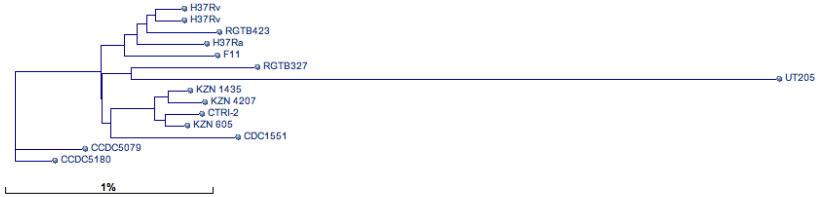
Human Pathogen: yes

Type	Name	RefSeq	INSDC	Size (Mb)	GC%	Protein	rRNA	tRNA	Other RNA	Gene	Pseudogene
Chr	-	NC_000962.2	AL123456.2	4.41	65.6	4,003	3	45	2	4,062	8

Biological Properties

- Morphology
  - Gram : Positive
  - Shape : Bacilli
  - Motility : No
- Environment
  - OxygenReq : Aerobic
  - OptimumTemperature : 37
  - TemperatureRange : Mesophilic
  - Habitat : HostAssociated
- Phenotype
  - Disease : Tuberculosis

Dendrogram (based on genomic BLAST)



Genome Sequencing Projects

◆ Chromosomes [14] ◆ Scaffolds or contigs [46] ◆ SRA or Traces [25] ◆ No data [19]

Organism	BioProject	Assembly	Status	Chrs	Size (Mb)	GC%	Gene	Protein
Mycobacterium tuberculosis H37Rv	PRJNA57777, PRJNA224	ASM19595v1	◆	1	4.41	65.6	4,062	4,003
Mycobacterium tuberculosis F11	PRJNA58417, PRJNA15642	ASM1692v1	◆	1	4.42	65.6	3,998	3,941
Mycobacterium tuberculosis CCDC5079	PRJNA161943, PRJNA19585	ASM27034v1	◆	1	4.4	65.6	3,695	3,646
Mycobacterium tuberculosis CCDC5180	PRJNA161941, PRJNA19583	ASM27036v1	◆	1	4.41	65.6	3,638	3,590
Mycobacterium tuberculosis CDC1551	PRJNA57775, PRJNA223	ASM858v1	◆	1	4.4	65.6	4,293	4,189

See more...

Other BioProjects

Epigenomics	2
Other	4
Transcriptome or Gene expression	80
Variation	5

Tools

1. BLAST Genome

Publications

1. Whole genome shotgun sequencing of one Colombian clinical isolate of Mycobacterium tuberculosis reveals DosR regulon gene deletions. Isaza JP, et al. FEMS Microbiol Lett 2012 May
2. Whole-Genome Sequences of Two Clinical Isolates of Mycobacterium tuberculosis from Kerala, South India. Madhavilatha GK, et al. J Bacteriol 2012 Aug

Mycobacterium tuberculosis genome annotation

Genome | Genome | Search

Limits Advanced Help

Organism Overview ; Genome Project Report ; **Genome Annotation Report**

Mycobacterium tuberculosis

Feature counts are from RefSeq where it is available

Mycobacterium tuberculosis H37Rv — See Protein Details

Sanger Institute

Causative agent of tuberculosis.

Type	Name	RefSeq	INSDC	Size (Mb)	GC%	Protein	rRNA	tRNA	Other RNA	Gene	Pseudogene
Chr	-	NC_000962.2	AL123456.2	4.41	65.6	4,003	3	45	2	4,062	8

Mycobacterium tuberculosis F11 — See Protein Details

Broad Institute

Predominant strain in South African epidemic

Type	Name	RefSeq	INSDC	Size (Mb)	GC%	Protein	rRNA	tRNA	Gene	Pseudogene
Chr	-	NC_009565.1	CP000717.1	4.42	65.6	3,941	3	45	3,998	9

Mycobacterium tuberculosis CCDC5079 — See Protein Details

Beijing Genomics institute and National Institue for Communicable Disease Control and Prevention

Drug-susceptible isolate belonging to the Beijing family.

Type	Name	RefSeq	INSDC	Size (Mb)	GC%	Protein	rRNA	tRNA	Gene	Pseudogene
Chr	-	NC_017523.1	CP001641.1	4.4	65.6	3,646	3	45	3,695	1

Mycobacterium tuberculosis CCDC5180 — See Protein Details

Beijing Genomics institute and National Institue for Communicable Disease Control and Prevetion

Multidrug-resistant clinical isolate.

Type	Name	RefSeq	INSDC	Size (Mb)	GC%	Protein	rRNA	tRNA	Gene
Chr	-	NC_017522.1	CP001642.1	4.41	65.6	3,590	3	45	3,638

Mycobacterium tuberculosis CDC1551 — See Protein Details

TIGR

Causative agent of tuberculosis.

Type	Name	RefSeq	INSDC	Size (Mb)	GC%	Protein	rRNA	tRNA	Gene	Pseudogene
Chr	-	NC_002755.2	AE000516.2	4.4	65.6	4,189	3	45	4,293	56

Mycobacterium tuberculosis CTRI-2 — See Protein Details

Research Institute for Physical-Chemical Medicine, Moscow, Russia

Mycobacterium tuberculosis CTRI-2 genome sequencing

Type	Name	RefSeq	INSDC	Size (Mb)	GC%	Protein	rRNA	tRNA	Gene	Pseudogene
Chr	-	NC_017524.1	CP002992.1	4.4	65.6	3,944	3	45	4,001	2

Mycobacterium tuberculosis H37Ra — See Protein Details

Chinese National Human Genome Center at Shanghai

An avirulent strain derived from its virulent parent strain H37

Type	Name	RefSeq	INSDC	Size (Mb)	GC%	Protein	rRNA	tRNA	Other RNA	Gene
Chr	-	NC_009525.1	CP000611.1	4.42	65.6	4,034	3	45	2	4,084

Mycobacterium tuberculosis H37Rv — See Protein Details

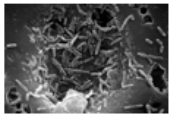
Broad Institute

Mycobacterium tuberculosis H37Rv genome sequencing

Type	Name	RefSeq	INSDC	Size (Mb)	GC%	Protein	rRNA	tRNA	Gene	Pseudogene
Chr	-	NC_018143.1	CP003248.1	4.41	65.6	4,111	3	45	4,170	9

Mycobacterium tuberculosis H37Rv

Organism Overview ; Genome Project Report ; Genome Annotation Report



Mycobacterium tuberculosis H37Rv

Causative agent of tuberculosis.

Lineage: Bacteria[3351]; Actinobacteria[547]; Actinobacteria[547]; Actinobacteridae[502]; Actinomycetales[485]; Corynebacterineae[219]; Mycobacteriaceae[38]; Mycobacterium[37]; Mycobacterium tuberculosis complex[5]; Mycobacterium tuberculosis[1]; Mycobacterium tuberculosis H37Rv[0]

***Mycobacterium tuberculosis* strain H37RV**. This strain has been derived from the original human-lung H37 isolate in 1934, and has been used extensively worldwide in biomedical research. Unlike some clinical isolates, it retains full virulence in animal models of tuberculosis and is susceptible to drugs and receptive to genetic manipulation.

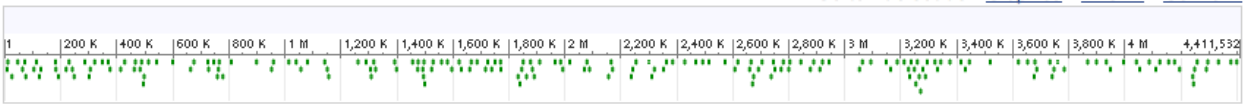
Genome Sequencing Projects

◆ Chromosomes [1] ◆ Scaffolds or contigs [0] ◆ SRA or Traces [0] ◆ No data [0]

Organism	BioProject	Assembly	Status	Chrs	Size (Mb)	GC%	Gene	Protein
Mycobacterium tuberculosis H37Rv	PRJNA57777, PRJNA224	ASM19595v1	◆	1	4.41	65.6	4,062	4,003

Genome Region

Go to nucleotide Graphics FASTA GenBank



Tools

1. GenePlot

Publications

1. Proteomic definition of the cell wall of Mycobacterium tuberculosis. Wolfe LM, et al. J Proteome Res 2010 Nov 5
2. From Corynebacterium glutamicum to Mycobacterium tuberculosis--towards transfers of gene regulatory networks and integrated data analyses with MycoRegNet. Krawczyk J, et al. Nucleic Acids Res 2009 Aug
3. Identification of outer membrane proteins of Mycobacterium tuberculosis. Song H, et al. Tuberculosis (Edinb) 2008 Nov

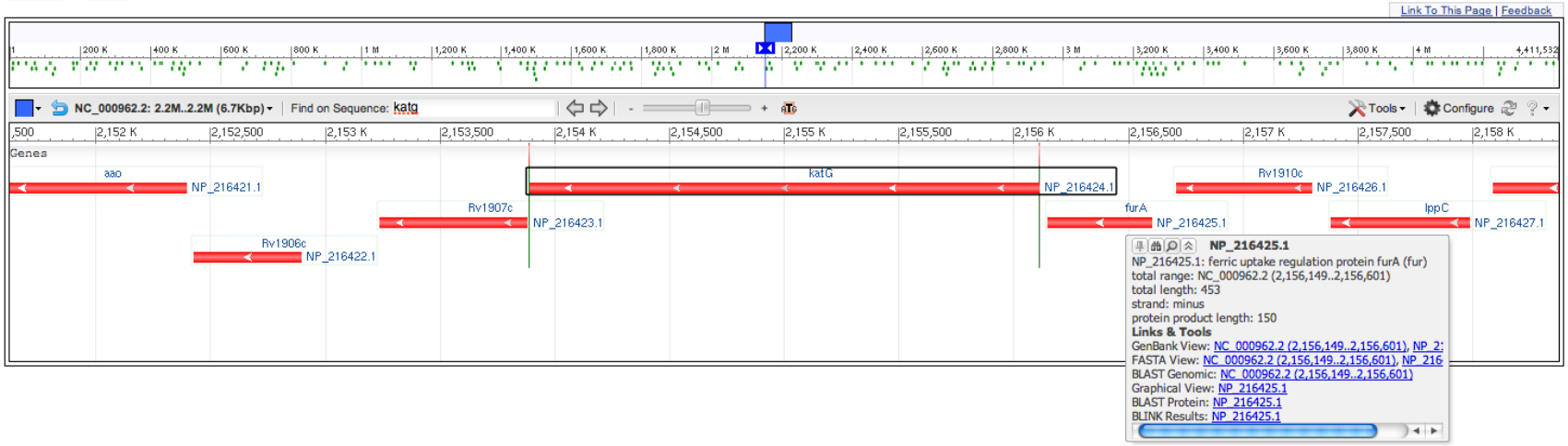
More...

Mycobacterium tuberculosis H37Rv browser

Mycobacterium tuberculosis H37Rv chromosome, complete genome

NCBI Reference Sequence: NC_000962.2

GenBank FASTA



From the Gene record

GeneRIFs: Gene References Into Functions What's a GeneRIF?

1. mechanism of the time-dependent transition from one high spin ferric haem form to another must be more complex than a simple single site oxidation.
2. Findings suggest that OxyS is a negative regulator of katG in mycobacteria.
3. Hotspot mutations within the katG 315 and inhA-15 genes can be used as genetic markers for detection of isoniazid resistance.
4. Downregulation of katG expression is associated with isoniazid resistance.
5. 9 novel KatG mutants with a single-amino-acid substitution were found; all mutants had lower INH oxidase activities than wild type and each had various levels of activity; isolates with mutations with relatively low activity had high-level INH resistance
6. the catalytic properties of the wild-type enzyme to 23 KatG mutants which have been associated with isoniazid resistance in clinical M. tuberculosis isolates were compared.
7. Drug resistance in M tuberculosis is due to mutations in relatively restricted regions of the genome: rpoB for RIF, katG and inhA for INH, embB for ethambutol, pncA for pyrazinamide, and so on.
8. the analysis of region 1 results in an increase in the rate at which the genotypic diagnosis of INH resistance-arising from mutations, deletions or insertions in the katG gene-is reached.
9. An oxyferrous heme/protein-based radical intermediate is catalytically competent in the catalase reaction of Mycobacterium tuberculosis catalase-peroxidase (KatG).
10. Role of the oxyferrous heme intermediate and distal side adduct radical in the catalase activity of Mycobacterium tuberculosis KatG revealed by the W107F mutant

Mycobacterium tuberculosis H37Rv GenePlot

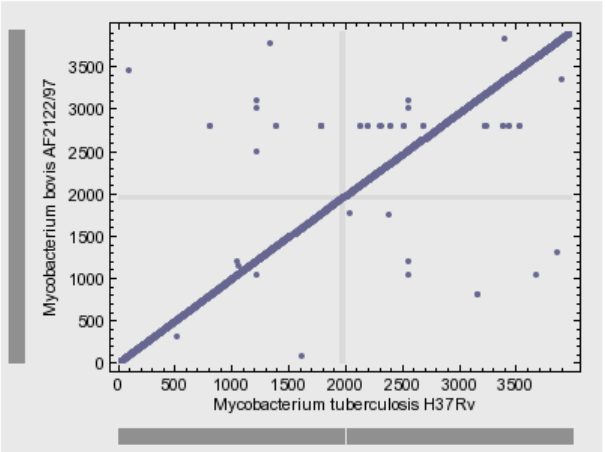
Pairwise genome comparison of protein homologs (symmetrical best hits)

Query organism: Mycobacterium tuberculosis H37Rv

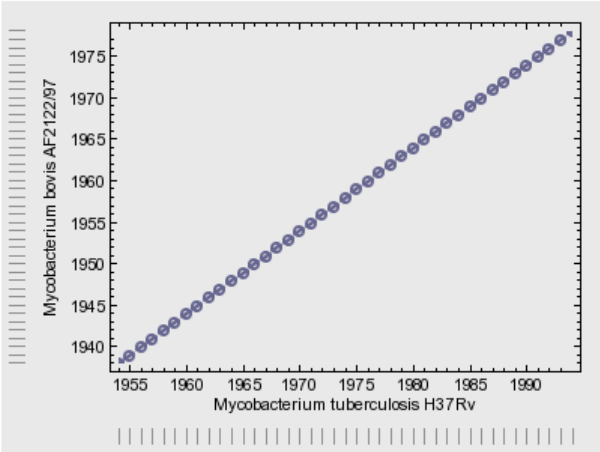
Select two organisms to compare

M. tuberculosis H37Rv versus M. bovis AF2122/97

3920 proteins total



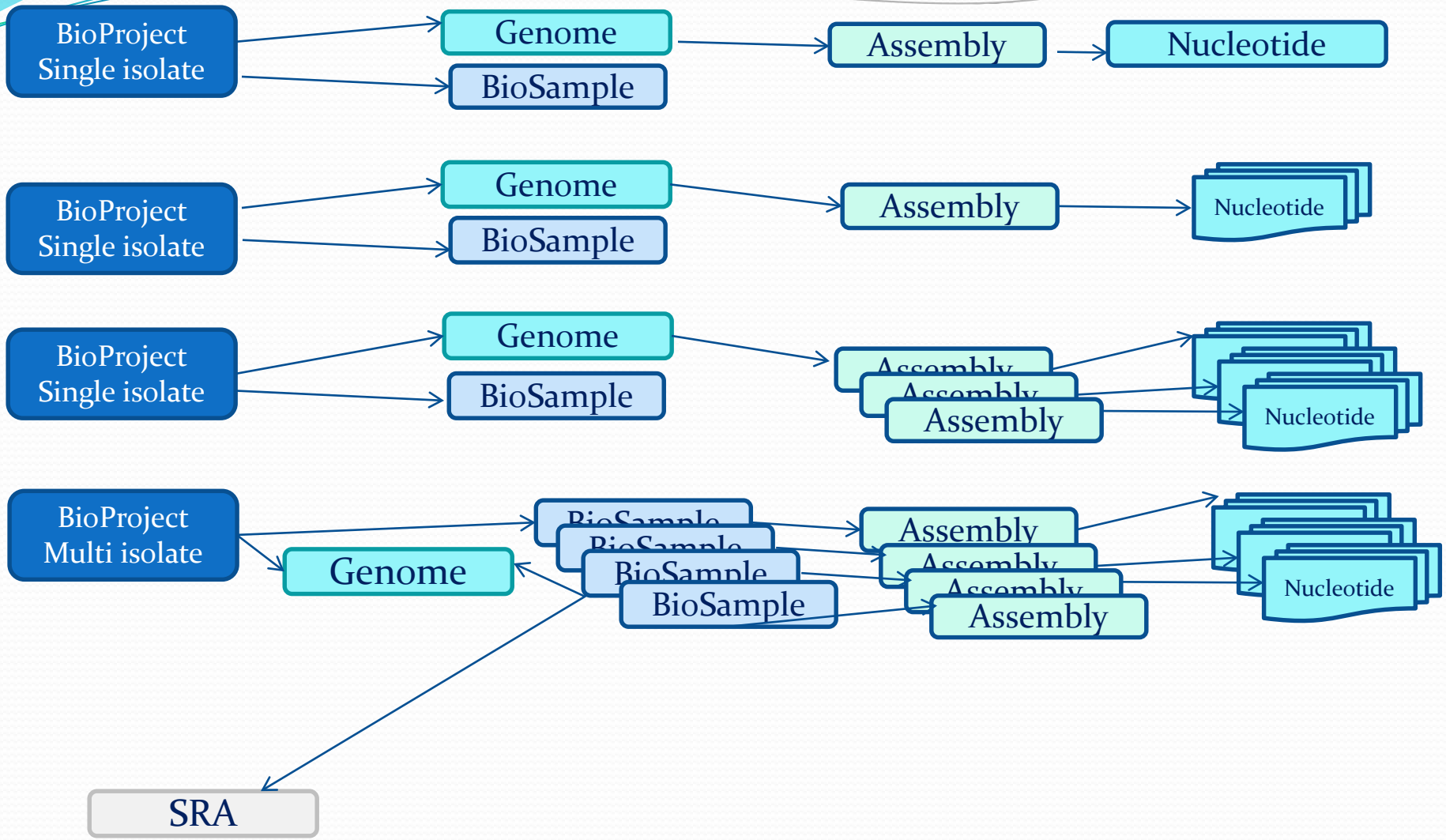
3989



Total number of bets 3872. Save all bets in order on genome.

bl2seq	Locus tags	Protein name
●	Mb1958 - Rv1923	lipase LIPD [Mycobacterium tuberculosis H37Rv]
●	Mb1959c - Rv1924c	hypothetical protein Rv1924c [Mycobacterium tuberculosis H37Rv]
●	Mb1960 - Rv1925	fatty-acid--CoA ligase [Mycobacterium tuberculosis H37Rv]
●	Mb1961c - Rv1926c	hypothetical protein Rv1926c [Mycobacterium tuberculosis H37Rv]
●	Mb1962 - Rv1927	hypothetical protein Rv1927 [Mycobacterium tuberculosis H37Rv]
●	Mb1963c - Rv1928c	short-chain dehydrogenase [Mycobacterium tuberculosis H37Rv]
●	Mb1964c - Rv1929c	hypothetical protein Rv1929c [Mycobacterium tuberculosis H37Rv]
●	Mb1965c - Rv1930c	hypothetical protein Rv1930c [Mycobacterium tuberculosis H37Rv]
●	Mb1966c - Rv1931c	transcriptional regulator [Mycobacterium tuberculosis H37Rv]
●	Mb1967 - Rv1932	thiol peroxidase [Mycobacterium tuberculosis H37Rv]
●	Mb1968c - Rv1933c	acyl-CoA dehydrogenase [Mycobacterium tuberculosis H37Rv]
●	Mb1969c - Rv1934c	acyl-CoA dehydrogenase [Mycobacterium tuberculosis H37Rv]
●	Mb1970c - Rv1935c	enoyl-CoA hydratase [Mycobacterium tuberculosis H37Rv]
●	Mb1971 - Rv1936	monooxygenase [Mycobacterium tuberculosis H37Rv]

BioProject, BioSample, Genome, Assembly, Nucleotide



NCBI genome submission dataflow