Institute of Bioinformatics

Bioinformatics and its application in Clinical microbiology

Presented by

Dr. Shruthi Vasanthaiah, M.D

Physician Scientist

Institute of Bioinformatics, Bangalore

Content

- Introduction
- Terminologies
- Databases
- Application in microbial identification
- Drug resistance marker detection
- Application in therapeutic arena
- Application in epidemiology and disease transmission
- Current trends in diagnostics
- Challenges and future directions

Introduction

- Bioinformatics is a discipline developed on the basis of biology, computer science and mathematics
- It effectively acquires and analyzes biological data such as nucleic acid sequences and protein structures, to conduct comprehensive and accurate biological analysis
- The first complete microbial genome sequenced was *Haemophilus influenzae* in 1995
- Apart from genomics, bioinformatics facilitates deeper insights into complex biological data through the integration of transcriptomics, proteomics, and metabolomics

Fleischmann RD, et al., Science, 1995

Terminologies in Bioinformatics

- **Genome**: The complete set of DNA (or RNA in some viruses) within an organism.
- **Transcriptome**: The full range of RNA transcripts produced by the genome under specific conditions.
- **Proteome**: The entire set of proteins expressed by an organism.
- **Metabolome**: The complete set of metabolites present within an organism or biological sample.

Terminologies in Bioinformatics

- **Alignment**: The process of arranging sequences (DNA, RNA, or protein) to identify regions of similarity
- **Multiple Sequence Alignment**: Aligning three or more sequences to identify conserved regions
- **Variant Calling**: Identifying variations (e.g., SNPs, insertions, deletions) between a sample and a reference genome
- **Gene Annotation**: Assigning functions to regions of the genome

Terminologies in Bioinformatics

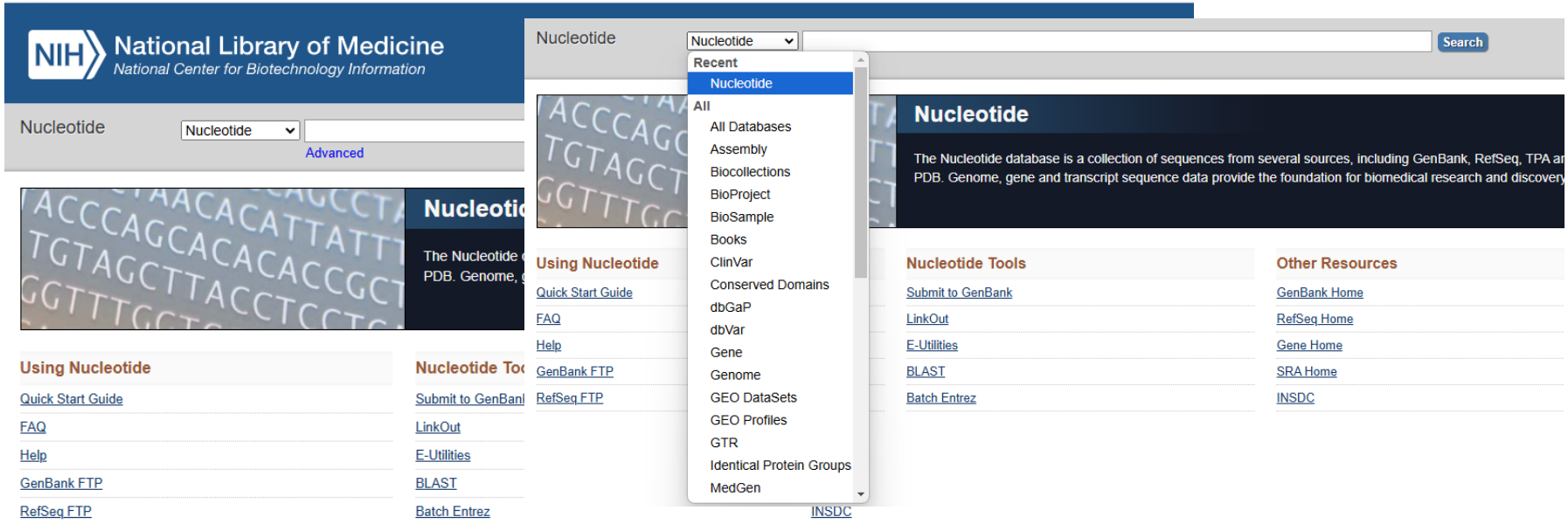
- **Reference Genome**: A representative example of a species genome used for comparison.
- **Transcript Assembly**: Reconstructing RNA sequences from reads obtained through RNA-seq.
- **Functional Annotation**: Associating genomic elements with biological functions, pathways, or phenotypes.
- **Phylogenetics**: The study of evolutionary relationships using sequence data

Important file types in Bioinformatics

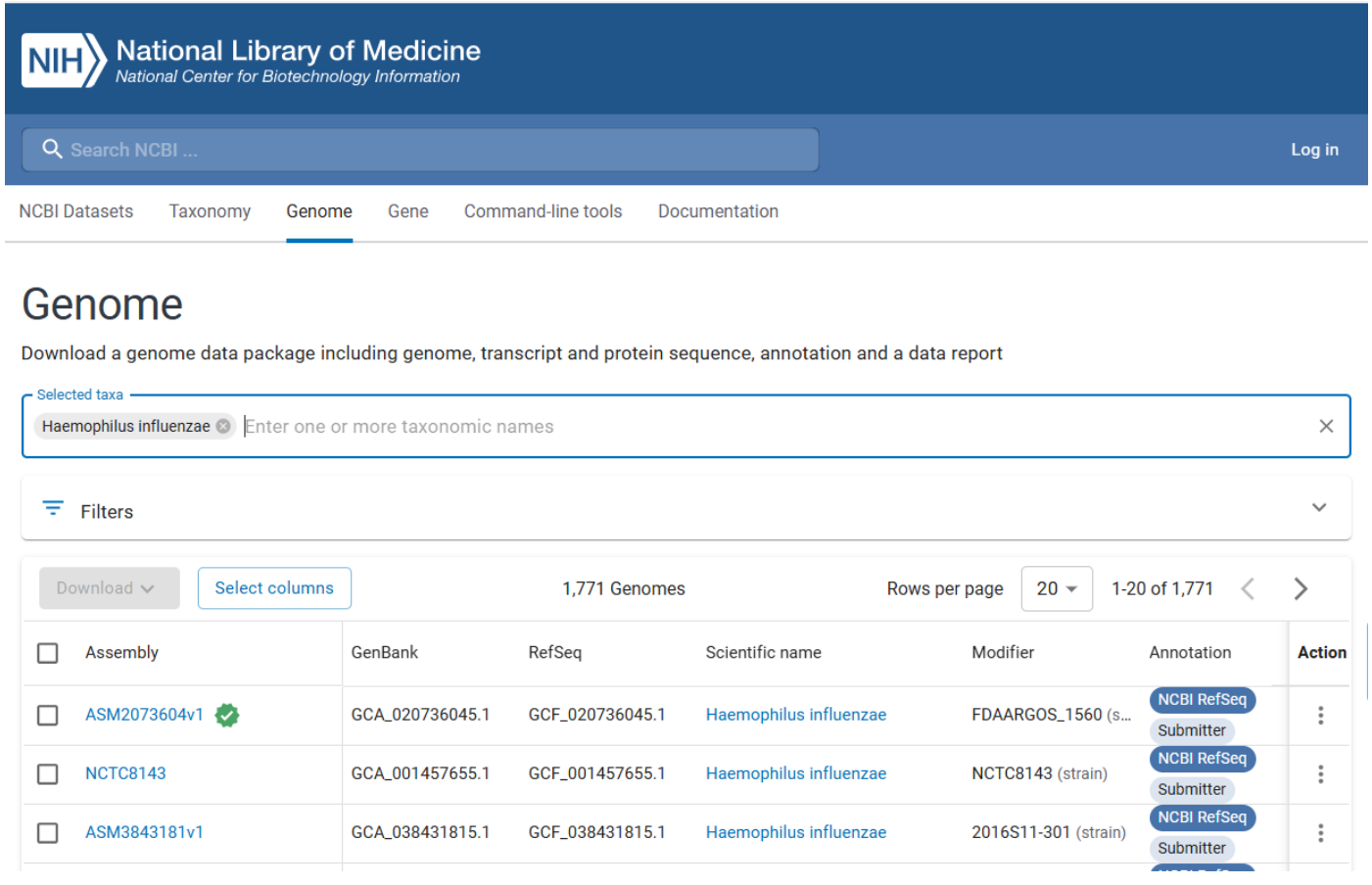
- **BCL(.bcl):** file format is used to store raw basecalling data in binary format. It requires conversion to generate readable **FASTQ** files for downstream
- **FASTA:** File format that stores nucleotide or protein sequences without quality scores, using a header line starting with ">“
- **FASTQ:** File format that stores sequences along with quality scores, with the header starting with "@“.
- **SAM/BAM (.sam, .bam)**: Stores sequence alignment data (SAM is text, BAM is binary)
- **VCF (.vcf)**: Stores variant call data, such as SNPs and indels

Databases used in microbial genomics

- Major databases are NCBI, EMBL and DDBJ(DNA Data Bank of Japan) under International nucleotide sequence database collaboration (INSDC)
- **NCBI GenBank**: is a public database that provides access to sequences and associated data, managed by the National Center for Biotechnology Information (NCBI)



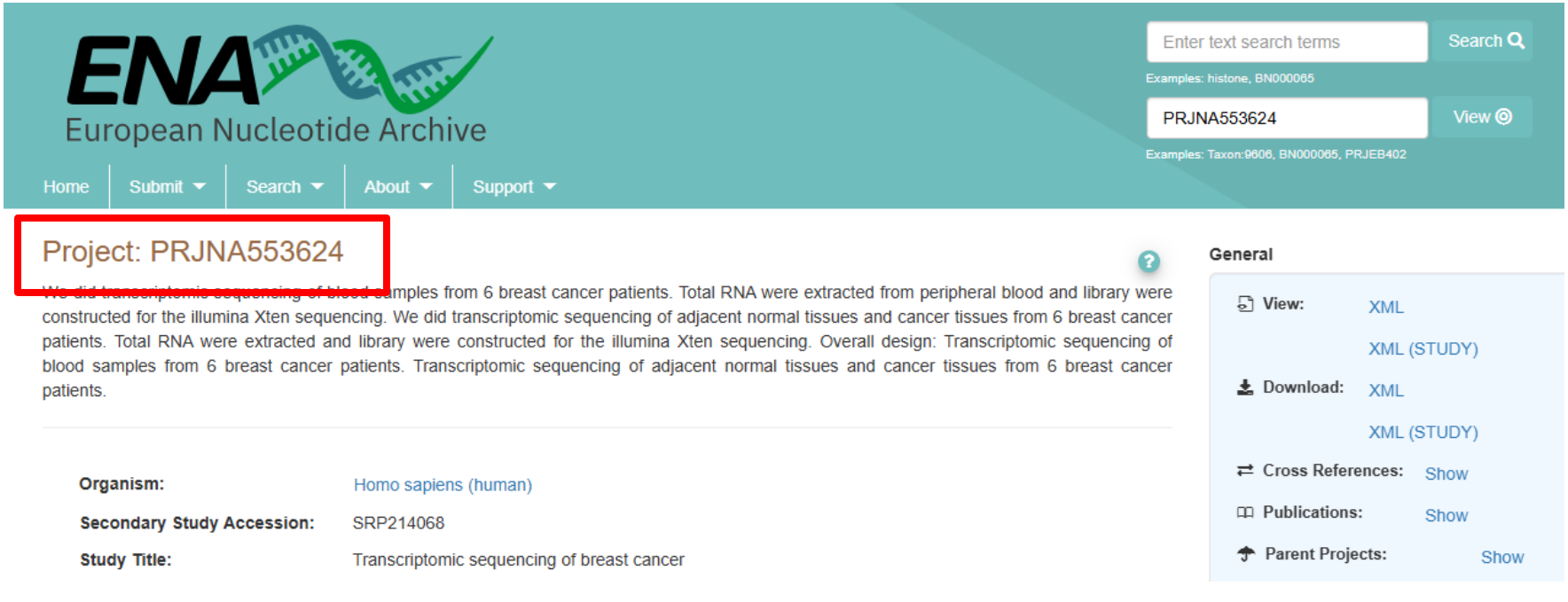
NCBI GenBank



Reference genome is not the genome of a specific individual but rather a consensus or composite of the DNA sequences from multiple individuals of the species.

Databases used in microbial genomics

- **Ensembl**: It is a is a collaborative database project between the European Bioinformatics Institute (EBI) and the Wellcome Trust Sanger Institute



- **PRJ**: Project
- **NA**: Nucleotide Archive

Databases used in other omics technologies

HPRD

General information

URL:	http://www.hprd.org/
Full name:	Human Protein Reference Database
Description:	The Human Protein Reference Database represents a centralized platform to visually depict and integrate information pertaining to domain architecture, post-translational modifications, interaction networks and disease association for each protein in the human proteome.
Year founded:	2003
Last update:	2008-11-06
Version:	v9.0
Accessibility:	Manual: Accessible
Country/Region:	India

HPRD: Human protein reference database was developed as a part of Human Proteome project by Human Proteome Organization(HUPO).

2 of the 25 contributors for this project were Indians!!

UniproT, PRIDE, STRING, Reactome etc.. are other proteome databases

Application in microbial Identification

- **Challenges with Traditional Methods:** Conventional microbial identification techniques, such as culture-based methods, are often time-consuming, labor-intensive, and may fail to identify fastidious, non-culturable, or rare organisms. Additionally in precious sample types (e.g., CSF) as well

- **Importance of accurate identification:** Accurate identification of microorganisms is essential for effective disease management, guiding appropriate antimicrobial therapy (e.g., drug-resistant tuberculosis or DRTB), preventing the spread of outbreaks (e.g., *SARS-CoV-2*), and identifying novel or rare pathogens (e.g., Human *Bocavirus*)

Steps in Microbial Strain Identification

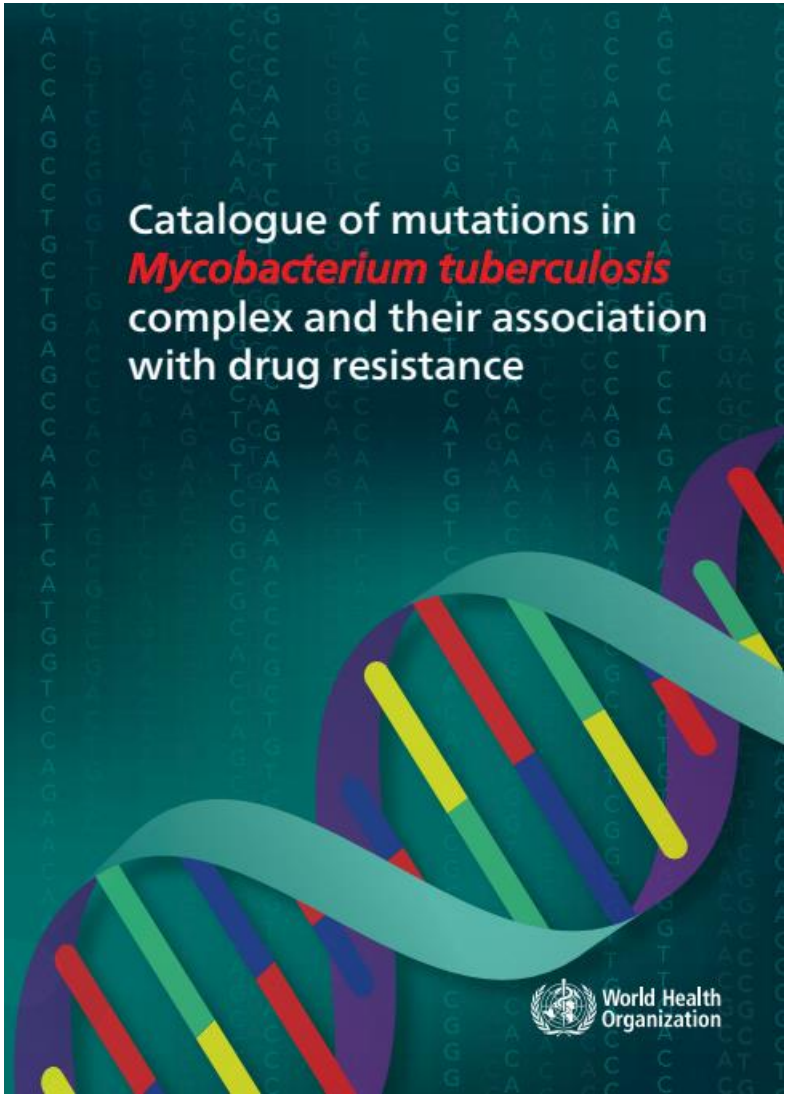
- Sample collection and DNA extraction: Importance of quality samples
- Sequencing types : targeted sequencing, whole genome sequencing
- Sequencing targets: 16srRNA (Bacteria), ITS1 and ITS2 (Fungi), 18srRNA (Parasites) and virus
- Sequencing platforms: Illumina, Oxford Nanopore
- Data analysis and comparison with genomic databases: BLAST search for closest match and Phylogenetic tree construction using software like MEGA, IQ-TREE etc

Other omics based identification

- **Proteomics** : MALDI-TOF (Protein mass fingerprint) and LC-MS/MS (Biomarker discovery) based identification of pathogens
- **Transcriptomics:** Identifying condition specific gene expression. *Example:* Identifying genes expressed during biofilm formation in *Pseudomonas aeruginosa*
- **Metabolomics:** Metabolite profiling to differentiate pathogenic and non-pathogenic *E. coli* strains.
- **Lipidomics:** Biomarker discovery in fungal infections (Mucorales)

Drug resistance marker detection

- Use of sequencing and PCR-based tools for resistance detection
- Clinically need in non-culturable or time taking pathogens
- Example: DeeplexMycoTB, DeeplexMycoLep
- Databases that store drug resistance data (e.g., ResFinder, PATRIC etc)



Application in epidemiology and disease transmission

Genomic epidemiology and tracking disease outbreaks:

- Tracks genetic changes in pathogens during outbreaks to Identifies sources of infection (e.g., SARS-CoV-2 variants, Ebola)
- Use in surveillance programs for emerging infectious diseases

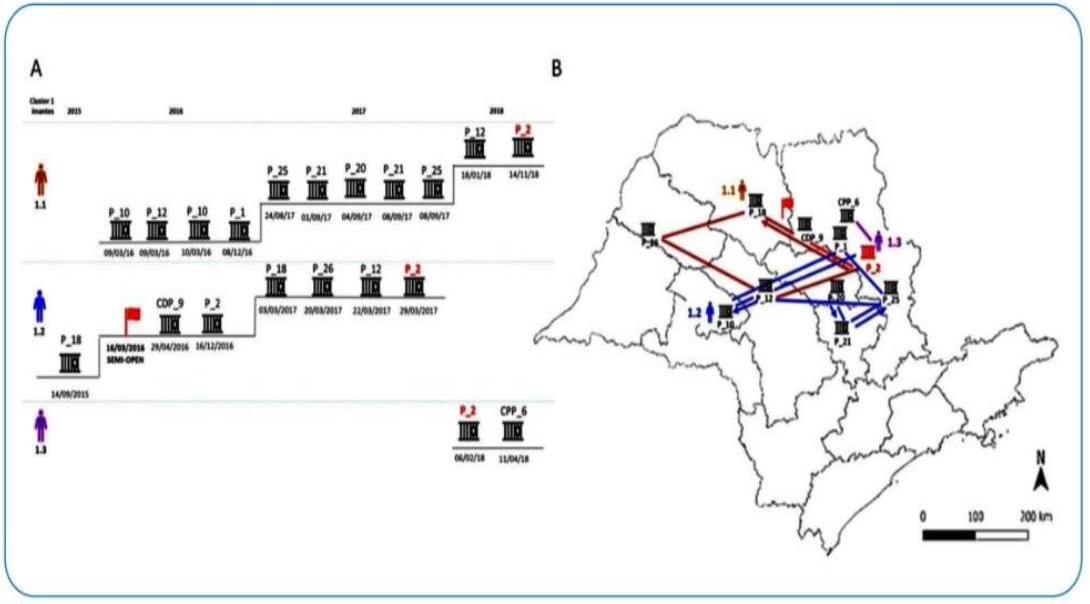
Traditional vs genomic surveillance

Aspect	Traditional Surveillance	Genomic Surveillance
Techniques Used	- Culture-based identification	Whole-genome sequencing (WGS)
	- Phenotypic drug resistance testing	Metagenomics
	- Microscopy	Single nucleotide polymorphism (SNP) analysis
	- Serological assays	Bioinformatics tools and databases
	- PCR for specific markers	
Speed of Results	Relatively slow due to culture growth and lab workflows.	Faster (post-sequencing), as sequencing and bioinformatics analyses can be automated.
Pathogen Identification	Limited to known species; requires culture and phenotypic testing.	Capable of identifying known and novel pathogens, including unculturable or mixed infections.
Resistance Detection	Based on phenotypic tests (e.g., antibiograms).	Identifies resistance genes and mutations, even before phenotypic resistance is observable.
Epidemiological Insights	Relies on epidemiological data to track outbreaks geographically and temporally.	Enables tracking of transmission pathways, outbreak sources, and evolutionary trends using genomic data.
Cost and Infrastructure	- Relatively low upfront cost	Higher upfront cost
	- Requires basic lab facilities	Requires advanced sequencing equipment, bioinformatics expertise, and computational resources
Scalability	Widely accessible in resource-limited settings.	Limited scalability in low-resource settings due to costs and technical requirements.
Use Cases	- Outbreak detection and response	- Tracking pathogen evolution
	- Routine disease reporting	- Identifying new variants or strains
	- Seroprevalence studies	- Antimicrobial resistance monitoring
		- Vaccine effectiveness studies

Application in epidemiology and disease transmission

Transmission dynamics and mutation rates:

- Phylogenetics helps infer how pathogens spread between individuals and population. Identifies superspreading events and transmission pathways.
- Example: Prisons as Reservoir for Community Transmission of Tuberculosis, Brazil - Prisoners with clustered isolates had a high amount of movement between prisons (two to eight moves) during the study period
- Tools: BEAST, Phylopart etc..



Application in therapeutic arena

- **Role in precision therapy:** Personalized medicine tailors treatment based on an individual's genetic factors to improve efficacy and reduces adverse effects
- Identification of Potential Drug Targets and Biomarkers for Targeted Therapy
- Example 1: *NAT2* polymorphisms guided isoniazid dosing in patients with TB. RCT conducted in Japan showed a decrease in DILI from 83% to 0%
- Example 2: Variations in the *CYP2B6* gene can lead to slower metabolism of efavirenz and thus higher drug levels in the bloodstream and an increased risk of side effects

Metagenomics

- **Metagenomics** is the study of genetic material recovered directly from environmental samples, bypassing the need for culturing individual microorganisms.
- Analyzes the collective genome of microorganisms in a sample (bacteria, fungi, viruses, etc.)

Techniques:

- 16S rRNA Sequencing: Common for bacterial identification
- Shotgun Sequencing: Provides a broader view of microbial genomes and functional pathways

Applications and Benefits of metagenomics

- **Microbial Diversity:** Identifies and characterizes microbial communities in various environments (soil, water, human gut)
- **Environmental Monitoring:** Detects pollutants or harmful microbes in ecosystems
- **Human Health:** Assesses gut microbiome for personalized medicine and disease diagnostics
- **Culture-Independent:** Can study microbes that cannot be cultured in labs
- **Comprehensive:** Captures a wide variety of microorganisms in a single sample
- **Dynamic:** Provides real-time insight into microbial communities responses to environmental changes

Other applications

- Bioinformatics in vaccine development and drug discovery helps to identify potential targets, optimize drug design, and predict vaccine efficacy.
- **Synthetic Biology:** Bioinformatics aids in designing engineered microbes for various applications, such as biofuel production, waste treatment, and pharmaceutical production
- **Infectious diseases model:** to predict new outbreaks and identify potential pathways in vitro before testing new drug substitutes, enabling faster and more efficient drug development and response strategies.

Current trends in diagnostics

- Artificial intelligence and machine learning in diagnosis, Automated antibiogram preparation
- Point-of-care diagnostics and portable technologies
- Example of handheld sequencers and use in field with Cloud based software like EPI2ME
- Directly from clinical sample sequencing using RNA Baits

EPI2ME


EPI2ME Desktop Agent
EPI2ME Agent Edit View Help
EPI2ME
PROFILE [SHRUTHIVA
NEW ANALYSIS
23 hours ago
Fastq Custom Alignment
v2023.06.14-1866743
449439
Thu Nov 30 2023 11:12:54 AM
Fastq Custom Alignment
v2023.06.14-1866743
436984
Thu Nov 30 2023 11:12:20 AM
Fastq Custom Alignment
v2023.06.14-1866743
436983
EPI2ME
FOR THE INTERNET OF LIVING THINGS
ANYTHING
ANYONE
ANYWHERE

Challenges and future directions

- Data privacy and ethical concerns.
- Challenges in implementing genomics in resource-limited settings owing to the cost, scalability and limited computational resources
- **Data Quality and Standardization:** Variations in sequencing protocols and platforms lead to inconsistent data quality, complicating comparison and integration. Standardized methods are needed for reproducibility and large-scale analyses
- **Integration of Multi-Omics Data:** Combining genomic, transcriptomic, and proteomic data for a comprehensive understanding of microbial systems is challenging

References

1. The Applications of Bioinformatics in Microbial Technology - CD Genomics. file:///C:/Users/shrut/Downloads/Applications%20of%20Bioinformatics%20in%20Microbial%20Technology%20-%20CD%20Genomics.pdf
2. Saeb ATM. Current Bioinformatics resources in combating infectious diseases. Bioinformation. 2018 Jan 31;14(1):31-35
3. Hiraoka S, Yang C, Iwasaki W. Metagenomics and bioinformatics in microbial ecology: current status and beyond Microbes and environments, 2016, 31(3): 204-212
4. Anselmo LMP, Gallo JF, Pinhata JMW, Peronni KC, et al New insights on tuberculosis transmission dynamics and drug susceptibility profiles among the prison population in Southern Brazil based on whole-genome sequencing. Rev Soc Bras Med Trop. 2023 Feb 20;56:e0181. PMID: 36820651

Additional study material

- Phylogenetic analysis using MEGA software

https://www.youtube.com/watch?v=eysZVTwRjc0

- How to download and install MEGA software

https://www.youtube.com/watch?v=jbwAJru32_E

- How to construct a Maximum Likelihood ML tree using an example dataset in MEGA?

https://www.youtube.com/watch?v=cg12Q5IIJBg