Chapter 3

Advances in Microbial Genomics in the Post-Genomics Era

Amjad Ali*; Tanzeela Raza; Hira Sikandar

Atta-ur-Rahman School of Applied Biosciences (ASAB), National University of Sciences & Technology (NUST), H-12, Islamabad, Pakistan 44000.

****Correspondence to: Amjad Ali,** Atta-ur-Rahman School of Applied Biosciences (ASAB), National University of Sciences & Technology (NUST), H-12, Islamabad, Pakistan 44000.*

Email: amjaduni@gmail.com

Abstract

In the pre-genomic era, microbes have been used for hundreds of years due to their applications in products such as bread, beer and wine. The use of these microbes in biotechnology is only possible when scientists know the mystery about this tiny creature. In the post-genomic era, thousands of whole genome sequences along with advanced analysis tools, techniques and technologies have been developed for the exploration of hidden potentials in these microorganisms. In this chapter, we summarize the timeline and advancements in microbial genomics made in the post-genomic era. Microbial evolution through 16S rRNA, bacterial genome sequencing boost by Next-generation and third generation sequencing technologies has also been discussed. Comparative genomics approaches to identify industrial microbes, pathogenic, non-pathogenic, rare and uncultivated microbes have also been described. Pangenome analyses for exploring the genome diversity and plasticity. Finally, reverse vaccinology and subtractive genomics approaches have been discussed in the context of its potentials to identify putative vaccine and drug targets.

Keywords: Post-genomics era; Comparative genomics; Phylogenomics; 16s rRNA; Next-generation sequencing; Pathogenomics; Computational tools; Reverse Vaccinology

1. Introduction

Microbes originated around four billion years ago when the earth was hotter and the environment was anoxic. These old inhabitants of the globe are considered as the foundation of the biosphere in both environmental and evolutionary perspectives. These omnipotent creatures occupy 60% of the earth's biomass. They make their own status by using their high adaptability powers. They are found in extreme environments such as hot springs, marshy places, molten lavas, and Antarctica regions where no other living organism can survive. Moreover, they have huge industrial, medical, forensics and environmental applications. Therefore, after realizing their importance microbiologists tried to explore microbes for their own benefits. However, that was not an easy task. Scientists spend years to perform the morphological and molecular characterization of microbes. Pre-genomic era was difficult because of difficult and costly sequencing techniques. Fortunately, advancements in genomics has now revolutionized every aspect of microbiology. Now after twenty years of first bacterial genome sequencing, it is necessary to find out what we did and what we have to do in this post-genomic era. Pre-genomics era started from the quest of sequence and about finding phylogenetic relationships among microbes and other organisms. The era ended in 1995 when first free-living microbe *Haemophilus influenza* was sequenced by using Whole-genome shotgun sequencing technology. However, the post-genomic era is going to extend over several generations and we will get the fruit of hard work of pre-genomic era in the post genomic era [1].

We have presented a brief history of different events that occurred in last two decades in the chronological order as shown in **Figure 1**. This timeline highlights the progress of sequencing in twenty-two years. From 1995 to 2017, development of advanced sequence technologies such as Next-generation sequencing (NGS) has greatly influenced the microbial genomics. In the past, laborious microbiology and molecular techniques were used for classification and characterization of microbes but now bioinformatics is an alternative to those microbiology and molecular techniques. This approach used to dig out the information about antibiotic resistance, microbial diversity, and to understand microbial communities and their genetic make-up [2].

Due to the advancement of computational approaches, there is huge data in the form of sequences available in different databases like UniProt, NCBI, and GOLD, etc. that is obtained from thousands of environmental microbes, pathogenic bacteria, and other industrially important bacteria. The total number of genome sequences available at NCBI are shown in **Figure 2.**

Now, annotation and analyses of these sequences are quite difficult for microbial bioinformaticians as compared to producing sequence data. They require more advanced and sophisticated data handling pipelines to analyze and interpret genomic or proteomic data. A general way of analyzing data requires commands run on programmes like Ubuntu or Linux operating systems [2]. For quick microbial genome annotation, differently advanced pipelines include RAST, PATRIC, command like software PROKKA, MicroScope etc. are used. Moreover, for metagenomics analysis MG-RAST, EBI metagenomics and Prokaryotic Genome Annotation Pipeline has been developed by NCBI which is capable of analyzing >2000 prokaryotic genome per day [3]. There is only 13-15% of available data of prokaryotes in public databases. There is still a need to discover new environmental microbes to explore more about these tiny creatures' secrets [4]. However, microbes are not easy to culture in the lab because of numerous factors e.g. temperature, fastidious growth, oxygen requirements etc. therefore only less than 1% can be cultured. It was difficult to explore those un-cultured microbes. However, due to advancement in sequencing technology and computational methods, microbial genomes can be obtained directly from environmental samples and sequenced. By using these techniques, we got 8000 genomes that get us closer to the comprehensive genomic representation of the microbial world [5]. There are two categories of post-genomic studies of microbes that include: (a) Direct sequence analyses studies based upon analysis of the genomic sequence information (b) Indirect sequence analysis require only some part of genomic sequence information. Direct sequence analyses enable us to analyze bacteria at the genomic level and help in the determination of small differences like single nucleotide polymorphisms (SNPs) [6].

Timeline of microbial genomics in post-genomic era

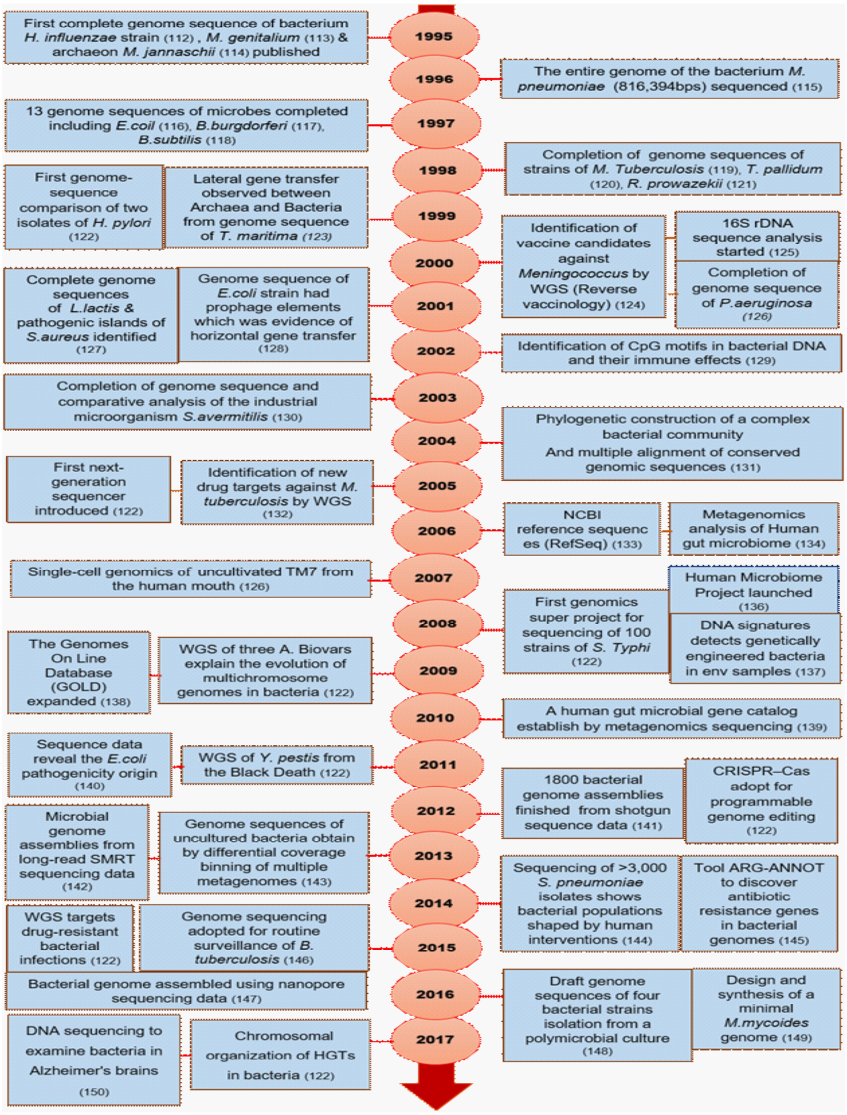
First complete genome sequence of bacterium H. influenzae strain (112) , M. genitalium (113) & archaeon M. jannaschii (114) published
1995
1996
The entire genome of the bacterium M. pneumoniae (816,394bps) sequenced (115)
13 genome sequences of microbes completed including E.coil (116), B.burgdorferi (117), B.subtilis (118)
1997
1998
Completion of genome sequences of strains of M. Tuberculosis (119), T. pallidum (120), R. prowazekii (121)
First genome-sequence comparison of two isolates of H. pylori (122)
Lateral gene transfer observed between Archaea and Bacteria from genome sequence of T. maritima (123)
1999
Identification of vaccine candidates against Meningococcus by WGS (Reverse vaccinology) (124)
16S rDNA sequence analysis started (125)
2000
Completion of genome sequence of P.aeruginosa (126)
Complete genome sequences of L.lactis & pathogenic islands of S.aureus identified (127)
Genome sequence of E.coli strain had prophage elements which was evidence of horizontal gene transfer (128)
2001
2002
Identification of CpG motifs in bacterial DNA and their immune effects (129)
Completion of genome sequence and comparative analysis of the industrial microorganism S.avermitilis (130)
2003
2004
Phylogenetic construction of a complex bacterial community And multiple alignment of conserved genomic sequences (131)
First next-generation sequencer introduced (122)
Identification of new drug targets against M. tuberculosis by WGS (132)
2005
2006
NCBI reference sequenc es (RefSeq) (133)
Metagenomics analysis of Human gut microbiome (134)
Single-cell genomics of uncultivated TM7 from the human mouth (126)
2007
Human Microbiome Project launched (136)
2008
First genomics super project for sequencing of 100 strains of S. Typhi (122)
DNA signatures detects genetically engineered bacteria in env samples (137)
The Genomes On Line Database (GOLD) expanded (138)
WGS of three A. Biovars explain the evolution of multichromosome genomes in bacteria (122)
2009
2010
A human gut microbial gene catalog establish by metagenomics sequencing (139)
Sequence data reveal the E.coli pathogenicity origin (140)
WGS of Y. pestis from the Black Death (122)
2011
1800 bacterial genome assemblies finished from shotgun sequence data (141)
CRISPR–Cas adopt for programmable genome editing (122)
2012
Microbial genome assemblies from long-read SMRT sequencing data (142)
Genome sequences of uncultured bacteria obtain by differential coverage binning of multiple metagenomes (143)
2013
Sequencing of >3,000 S. pneumoniae isolates shows bacterial populations shaped by human interventions (144)
Tool ARG-ANNOT to discover antibiotic resistance genes in bacterial genomes (145)
2014
WGS targets drug-resistant bacterial infections (122)
Genome sequencing adopted for routine surveillance of B. tuberculosis (146)
2015
Bacterial genome assembled using nanopore sequencing data (147)
2016
Draft genome sequences of four bacterial strains isolation from a polymicrobial culture (148)
Design and synthesis of a minimal M.mycoides genome (149)
DNA sequencing to examine bacteria in Alzheimer's brains (150)
Chromosomal organization of HGTs in bacteria (122)
2017

Figure 1: Microbial genomics over the decades: This timeline shows advancements in microbial genome sequencing in chronological order. The concept of the sequencing of microbes started in the nineties (pre- genomic era). In 1995, nonpathogenic *H. Influenza* sequencing by Craig Venter and his team was responsible for the inauguration of post-genomic era. Advanced genome sequencing technologies like Next Generation and Third Generation sequencing boost the microbial DNA sequencing.

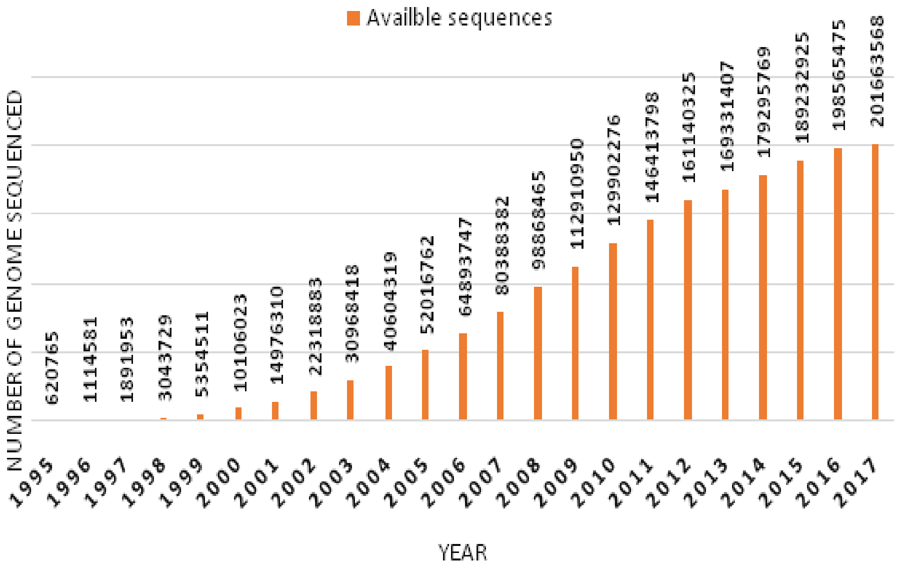

Figure 2*:* Available genome sequences in GenBank (NCBI) increased with the invention of new sequencing technologies (Source: www.ncbi.com).

2. Advancements in Sequencing Technologies in Post-Genomics Era

2.1. DNA sequencing

Determining the order of amino acid residues in polynucleotide chains revealed the information about hereditary material and biochemical properties that led to exploration of bacterial communities, their evolution and interaction with each other [7,8]. A milestone of DNA sequencing is shown in **Figure 3**.

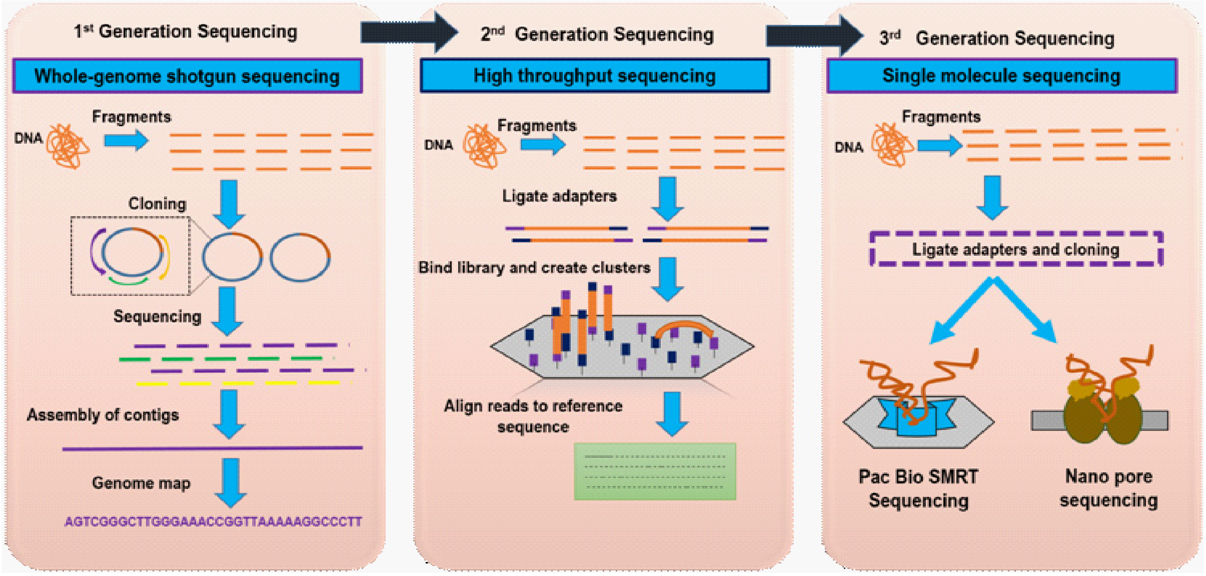
1st Generation Sequencing
Whole-genome shotgun sequencing
DNA
Fragments
Cloning
Sequencing
Assembly of contigs
Genome map
AGTCGGGCTTGGGAAACCGGTTAAAAAGGCCCTT
2nd Generation Sequencing
High throughput sequencing
DNA
Fragments
Ligate adapters
Bind library and create clusters
Align reads to reference sequence
3rd Generation Sequencing
Single molecule sequencing
DNA
Fragments
Ligate adapters and cloning
Pac Bio SMRT Sequencing
Nano pore sequencing

Figure 3: Advancements in microbial genome sequencing technologies in post-genomic era: Whole genome shotgun sequencing requires laborious sample preparation. High throughput sequencing gives high accuracy but short read lengths while single molecule sequencing gives low accuracy with long read lengths.

2.1.1. Whole genome shotgun sequencing

In 1995, Craig venter and co-workers at TIGR, presented the whole genome sequence for *Haemophilus influenza* [9] and *Mycoplasma* [10]. In this method, genomic DNA is subjected to random fragmentation and libraries are produced in *E.coli*. These clones are sequen ced and computationally compared with sequence reads and the matching sequences are assembled [11]. DNA sequencing had some pitfalls; since amplified templates are produced in a single step, certain DNA stretches may skip replication well in *E.coli* [12].

2.1.2. High throughput sequencing or next generation sequencing

Earlier sequencing methods created draft genomes with approximately $50,000 cost. With the advancement in sequencing technology it has reached $1 cost which has revolutionized the microbial genomics [13]. Discovery of restriction enzymes by Hamilton smith and co-workers proved to be a significant event without which Next- generation DNA sequencing would not have been possible. DNA strand to be sequenced are cleaved with RE's to provide specific ends that function as initiating points for sequencing [11]. In 2000's Next-Generation Sequencing was introduced with 100-fold throughput using 454-pyrosequencing approach. Afterward, Illumina and ABI SOLiD were introduced. High-throughput sequencing or Next-generation sequencing can sequence multiple DNA molecules in parallel due to which millions of DNA molecules can be sequenced at a time and at low cost. Next-generation sequencing produces short read length which leads to the taxonomic classification of microbes [14]. The principle behind these technologies is a detection of emission light from the sequenced DNA while Ion torrent was introduced later that detects hydrogen ion [15]. Thus high-throughput sequencing technologies enable us to determine cellular genomics, the transcriptomic signature of various diseases and novel variants responsible for many diseases [16]. HTS provide insights into the genetic and phenotypic diversifications among closely related bacterial infection like *Mycobacterium abscessus, Pseudomonas aeruginosa, Staphylococcus aureus, Mycobacterium tuberculosis* etc [13].

Different commercially available sequencing platforms include; Illumina's platforms, Ion Torrent, 454 and Pacific Biosciences Real Time Sequencer. Illumina platforms have HiSeq2000 and MiSeq that perform an ultra-high-throughput analysis. These machines were tested against 24 host-associated and free-living microbial communities. HiSeq2000 allow large DNA parallel sequencing at low cost, while MiSeq is convenient for smaller projects [17]. Loman NJ and his team compared three benchtop high throughput-sequencing instruments included 454 GS, MiSeq (Illumina) and Ion Torrent PGM [18] by sequencing of *Escherichia coli O104:H4* to know their efficacy. These sequencers can generate bacterial genome sequence data, can identify and characterize bacterial pathogens. They reported that MiSeq had the highest throughput run as compared to Ion Torrent PGM and 454 GS [18]. Another study conducted to characterize *Helicobacter pylori* genome revealed that Illumina Nextera XT sequencing machine produced more accurate multi-locus sequence type in less time and cost as compared to MiSeq and Ion Torrent [19]. In clinical settings, high throughput sequencing technologies are widely used and also used to determine microbial community diversity in food industry during douche-koji making fermentation and in 62 Irish artisanal kinds of cheese

[20,21].

2.1.3. Advanced genomics with single-molecule real-time (SMRT) sequencing

To overcome second generation sequencing problems that included short read length (30-450bases), errors due to short read lengths, and laborious sample preparation methods; a newer system was introduced by Pac Bio’s that is SMRT sequencing after 2007 [15]. Single molecule (SMRT) sequencing is a third- generation sequencing technique, which enables real-time observation of base sequences from individual strands of DNA or RNA [22,23]. Second generation sequencing provides a longer sequenced read length, flexibility, lower cost and higher throughput. In SMRT technology, the polymerase enzyme is affixed at the bottom of Zero-Mode Waveguide (ZMW) nano-holes. Polymerase incorporates fluorescently labeled bases to DNA template and makes immobilized complex at bottom of well. Detectors detect emitted lights as fluorescents base combines with the template [15]. Single-molecule real-time (SMRT) DNA sequencing allows detection of chemical modifications. For example, methylation was detected in *E.coli* [24].

2.1.4. Oxford nanopore sequencing

Nanopores sequencers are also based on single molecule concept but it detect bases without labels, produces long reads, relatively fast and with low GC bias errors. The principle of this technology is tunneling of molecules (polymer) through a pore that separates two sections. This allows identification of specific molecules. Oxford nano-pore has the MinION system that is real-time analyzer of DNA or RNA [25].

Next Generation Sequencing (NGS) Tools

Software mostly commonly use in Next generation sequencing are listed below in **Table 1**.

Table 1: NGS Tools

Sr. NO.	Tool	Function	Web Link	Reference
1.	**mrsFAST**	Map short reads, SNP-aware Mapping,	http://mrsfast.sourceforge.net.	[26]
2.	**ContextMap**	Is RNA sequence mapping algorithm, identification of indels	http://www.bio.ifi.lmu.de/ContextMap	[27]
3.	**SOAPsplice**	Detects splice junction sites from RNA-seq	http://soap.genomics.org.cn/soapsplice.html	[28]
4.	**Bowtie2**	support ultra-fast and memory efficient sequence alignment of local, gapped and paired end modes	https://sourceforge.net/projects/bowtie-bio/files/latest/download?source=files	[29]
5.	**NextGenMap (NGM)**	Read mapping program, memory efficient	http://cibiv.github.io/NextGenMap/	[30]

3. Genome Overview and Browsers

Thousands of genomes are sequenced so far but the follow-up knowledge is still very limited. Structural genomics plays a vital role in understanding the molecular genetics by providing insights into genomic DNA functional stretches [31]. The collection of all genetic material from species is termed as pangenome and could estimate with bioinformatics tools. Data could be visualize and analyze via various online genome browsers [32]. Genome browsers are visualization programs from which researchers can search, retrieve and analyze genomic sequences efficiently and conveniently [33]. Web-based Genome browsers are classified as 'Species-specific genome browser' and 'general genome browsers'. Species-specific genome browsers work on one specific organism while the general genome browsers deal with multiple species. Different genome browsers have different retrieval systems. For example, Ensembl employ BioMart system [34], UCSC system employs table browser [35].

Table 2: List of web-based general microbial genome browsers

Sr. No.	Browser	Description	Web Link
1.	**NCBI**	Provides free access to books of biomedical sciences, microbes	https://www.ncbi.nlm.nih.gov/
2.	**Ensembl**	Genome browser for bacteria, fungi, protists, metazoan, vertebrates, annotate genes, predict regulatory functions and multiple alignment	http://www.ensembl.org/
3.	**Genome Projector**	Hundreds of bacterial genomes with circular and linear maps	http://www.g-language.org/g3/
4.	**UCSC**	Graphical web-based browser, gene annotation and expression, integrates bacterial and archaeal specific tracks	http://archaea.ucsc.edu
5.	**(Integrated Microbial Genome) IMG**	Visualization software tool, Distribute data to public, provide the facility of panning, focus zooming and jump zooming	http://bioviz.org/igb

Table 3: List of web-based microbial species-specific genome browsers

Sr. No.	Browser	Species	Web Link
1.	**Saccharomyces cerevisiae Genome Database(SGD)**	*Saccharomyces cerevisiae*	https://www.yeastgenome.org/
2.	**Paramecium Database (ParameciumDB)**	*Paramecium tetraurelia*	http://paramecium.cgm.cnrs-gif.fr/cgi-bin/gbrowse2/
3.	**DictyBase**	*Dictyosteliumdiscoideum*	http://dictybase.org/db/cgi-bin/ggb/gbrowse/
4.	**CyanoBase**	*Cyanobacteria*	http://genome.kazusa.or.jp/cyanobase
5.	**The Legionella Genome Browser (LGB)**	*Legionella pneumophila*	http://genolist.pasteur.fr/LegioList/
6.	**The Enterobacter Genome Browser**	*Enterobacters*	engene.leibniz-fli.de/
7.	**The Xanthomonas Genome Browser (XGB)**	*Xanthomonas*	xgb.leibniz-fli.de/

3.1. Functionalities and features

High-throughput sequencing and high-performance computing provided with enormous genomic data and web-based genome browsers freely distribute this immense volume of data to researchers. These genome browsers accumulate entire data from different platforms and present it graphically [36]. Images, graphs, cycles, pathways, maps etc are drawn to aggregate the data to present information in less complicated manner to overcome the burden of servers [37].

3.2. Data retrieval and analysis

Data Retrieval and analysis are one of the principle attributes of genome browsers. Different browsers apply different approaches for data retrieval. For example, UCSC present the data in tabular form and ABrowse project apply BioMart system [34].

IGB employ MACS to analyze the results obtained from ChIP-Seq [38]. Genome browsers integrate with other platforms in order to provide better results. Genome browsers provide a platform where researchers collaborate to share their ongoing researches, discoveries and discuss their projects [39].

4. Advanced Computational Tools for Microbial Genomics in Post-Genomic Era

Table 4: Computational tools and their functions

Sr. No	Tool	Function	Web Link	Ref
1.	**BLAST**	Infer evolutionary and functional relationships	http://blast.ncbi.nlm.nih.gov	[40]
2.	**KEGG**	An integrated database resource, provides genomic, chemical and systemic information	http://www.kegg.jp	[41]
3.	**WebACT**	Database provide sequence comparisons between all prokaryotic genomes	webact.org/WebACT/home	[42]
4.	**MUMmer**	Provide ultra-fast alignment of genomes	tar -xvzf MUMmer3.0.tar.gz	[43]
5.	**BASys**	(Bacterial Annotation System)Provides automated bacterial genomic sequencing	http://wishart.biology.ualberta.ca/basys	[44]
6.	**Microbial Genome Viewer (MGV)**	Generate linear and wheel maps for data obtained from annotation and transcriptomic	http://www.cmbi.kun.nl/MGV	[45]
7.	**GeneWiz**	Predict linear or circular genome atlas, by genetic and physical properties of genome, one can make the diagram	http://www.cbs.dtu.dk/services/gwBrowser/	[46]
8.	**GeneMark**	Gene prediction in bacteria, metagenomes, metatranscriptomes, and archaea	http://opal.biology.gatech.edu/GeneMark/	[47]
9.	**(CGV)**	Circular Genome Viewer (CGV) generate static and graphical maps of Circular DNA, providing facilities of zoom in, labeled features and hyperlinks	http://stothard.afns.ualberta.ca/cgview_server/	[48]

10.	**SignalP**	Infer the presence and location of signal peptide cleavage site in nucleotide sequences among different organisms	http://www.cbs.dtu.dk/services/SignalP/	[49]
11.	**Prokka**	Provides genome annotation for bacteria, archaea and viruses	http://www.bioinformatics.net.au/software.prokka.shtml	[50]
12.	**LAST-TRAIN**	Accuracy of sequence alignment improved by inferring better score parameters and re-align	http://last.cbrc.jp/	[51]
13.	**Harvest suite (parsnp, gingr)**	Core genome alignment and visualization tool	HarvestOSX64v1.1.2.tar.gz	[52]
14.	**ClonalFrameML**	Infers recombination in bacterial genome	https://github.com/xavierdidelot/ClonalFrameML	[53]
15.	**POGO-DB**	Provides microbial genomic comparison and visualization tool	http://pogo.ece.drexel.edu	[54]
16.	**JSpeciesWs**	Identifies similarity b/w two genomes, measures average nucleotide identity, analyze correlation indexes of tetra-nucleotide signatures	http://jspecies.ribohost.com/jspeciesws.	[55]
17.	**(SRST2)**	Short Sequence Typing for Bacterial Pathogens (SRST2) detects genes, alleles and MLST from whole genome sequencing data	http://katholt.github.io/srst2/	[56]
18.	**GUBBINS**	Genealogies Unbiased By recomBinations In Nucleotide Sequences Identifies loci containing base substitution and generate phylogenetic tree based on point mutations	Sanger-pathogens.github.io/gubbins/	[57]
19.	**SpeciesFinder**	Predicts the species of a bacterium from complete or partial pre-assembled genomes	http://cge.cbs.dtu.dk/services/SpeciesFinder	[58]
20.	**Velvet**	Genome assembler, for short read sequences, remove errors and generate unique contigs	https://www.ebi.ac.uk/~zerbino/velvet/	[59]
21.	**FgenesB**	Bacterial Operon and gene prediction	http://linux1.softberry.com/	[60]
22.	**SPARTA**	SPARTA (Simple Program for Automated reference-based bacterial RNA-seq Transcriptome Analysis) analyzes differential gene expression, perform quality analysis of the data sets	sparta.readthedocs.org	[61]
23.	**OrthoANI**	OrthoANI(Orthologous Average Nucleotide Identity) measures overall similarity between two genome sequences	http://www.ezbiocloud.net/sw/oat.	[62]
24.	**Oufti**	Quantitative analysis of bacterial count and fluorescent signals	http://www.oufti.org/download/	[63]
25.	**Orione**	Conduct NGS data analysis and annotation by quality control of reads and their trimming	http://orione.crs4.it	[64]
26.	**VacSol**	Scrutinize the whole bacterial pathogen proteome to identify a vaccine candidate proteins	https://sourceforge.net/projects/vacsol/	[65]

5. Microbial Phylogeny and Evolution

Early life on earth was most probably consisted of RNA. According to endosymbiotic theory, archaea was the ancestor and they engulfed mitochondria from gram-negative bacteria or chloroplast from cyanobacteria that lead to the evolution of eukaryotes [66]. Phylogenetic analyses were necessary to explore the microbial diversity, their ecological or niche adaptation, pathogenic potential of unknown microbes, their ability to produce different types of natural products like enzymes etc. The term “Phylogeny” is derived from two Greek words Phylon meaning “clan or race” and genesis meaning “origin”. Therefore, it is the study of the evolutionary history of the organism [67].

Researchers used many approaches for classification of microbes. In 1759, Linnaeus tried to classify all living things and developed the binomial system (Genus species). He divided the world into Animal, Vegetable, and Mineral and put all the microscopic life in one genus i.e. Chaos. In the 1980’s, neo-Darwinian evolutionary theory explained the evolution of plants and animals over the last 560 million years but did not discuss the evolution of microorganisms. Therefore biological scientists from last two decades aimed to build a universal phylogeny [68]. Whittaker in 1969 gave five-kingdom system based on modes of nutrition like photosynthesis, adsorption, and ingestion. The five-kingdom system included Plants, Animals, Fungi, Protists, and Bacteria. However, it did not describe the origin of species. Therefore, microbiologists tried to classified microorganisms on the basis of their morphological, molecular, physiological and metabolic characters. Carl Woese and his coworkers in the 1970s proposed the “Universal tree of life” including Archaea, bacteria, Eucarya (**figure 4**) using 16s rRNA molecular approach for phylogenetic analysis. Phylogenetic analysis increased due to rapid advancements in biology and computational field, which led to the availability of huge genomic data about microbes [69].

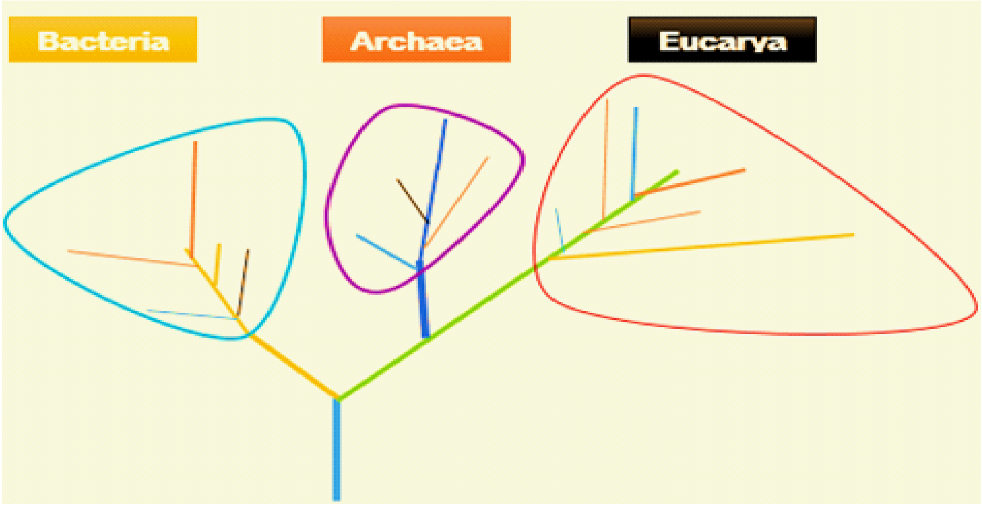

Figure 4: "The Universal Tree of Life" by Carl Woese and co-workers

5.1. Different approaches to construct phylogenetic tree in post-genomic era

The phylogenetic relationship can be determined using morphological (cell size, shape etc.), physiological, molecular (based on genetic material) and comparative genomic approaches. These methods include analyzing the shared gene content, gene order, construction of different phylogenetic trees etc. Due to the limited morphological and physiological characters, along with substantial number of variations among closely related taxa, scientists preferred molecular data. Initially, phylogenetic molecular markers included DNA sequences located on chromosomes and ribosomal RNA gene sequences [69]. Different bacterial genome sequenced after 1995 centered on sequenced data. Based on 16s rRNA sequencing proteobacteria were classified. Proteobacteria are considered as the largest taxonomic group because they comprise 50% of all cultured bacteria. Based on its branching in 16sRNA trees they are divided into five classes; alpha (covers 12% proteobacteria), beta (8%), and gamma (26%), while delta and epsilon covers other 4% [70].

Molecular markers 16s rRNA and rpoB genes (rplB, pyrG, fusA, leuS and rpoB) are compared for Actinobacteria, Bacteroides, Proteobacteria, and Cyanobacteria. Results revealed that rpoB markers were good in detecting minor groups among microbial assemblages [71]. A bulk of sequences allowed scientists to use comparative genomic approaches for phylogenetic study. Ludwig and Schleifer reconstructed the phylogeny of prokaryotes based on comparative sequence analysis of small subunit rRNAs [72]. Phylogenetic relationship of Streptococcus to other species was determined by using comparative genomic approaches. Moreover, these approaches were also used for identification and functional classification of homologous clusters, pan-genome analyses, population structure and virulence factors [73].

5.2. Reasons of evolution of microbes and horizontal gene transfers (HGTs)

Evolution of infectious species can also be determined using 16s rRNA sequences. Derrick and his Co found genus Leptospira pathogenic bacterium with the help of comparative genome analyses. They did pan-genome analyses, 16s rRNA gene sequencing, In-silico DNA-DNA hybridization, metabolic reconstruction and related gene clusters. They reported that Leptospira originated from noninfectious species and adapted different metabolic pathways that became the cause of infection. They also find out a unique signal responsive pathway, gene expressions and chemotaxis systems [74]. Different prokaryotic group's evolution is due to horizontal gene transfer (HGT). In HGT, microorganisms transfer genetic material from one species to other species. Mostly housekeeping genes are involved in HGT. It is an adaptation process and strongly influenced by environment. As earth's environment changed with the passage of time, microorganisms acquired more foreign genes to cope up environmental conditions [75].

5.3. Different phylogenetic molecular markers

Advancement in genomics has led to increasing number of full genomes and gene sequence data resulting in identification of various phylogenetic molecular markers other than 16s rRNA. These include elongation and initiation factors, large subunit rRNA, RNA polymerase, subunits of proton translocation ATPase, DNA gyrase, recA, aminoacyl tRNAsynthetases and so on. Most widely used molecular markers include nuclear ribosomal genes (18S rRNA in eukaryotes and the 16S rRNA in others and large subunit contains the 5S and 23S rRNAs) and more powerful markers in resolving species level phylogenies i.e. mitochondrial genes (cytochrome oxidase I and II (COI/II)), EF-1α, rpoA gene, lux Gene, Nuclear H3, recA, rpoB, rpoC1 etc. These markers can resolve phylogenetic relationship at deep levels of evolution [76]. Secondary structure can also be used for multiple sequence alignment. Le Q and co proposed QuanTest,a fully automated system for protein MSA [77]. However, these markers are more complex. In addition, phylogenetic trees derived from such markers may vary from one another. Therefore, phylogenetic trees of microbes derived from single gene i.e. small subunit rRNA is considered as universal [72].

5.4. Challenges and opportunities for phylogenetic tree reconstruction

Different molecular phylogenetic analysis predicted lateral gene transfer between closely related prokaryotes as well as distantly related prokaryotes. This lateral gene transfer became a hurdle in the understanding of exact evolutionary track of microorganisms. In addition, computing cost involved in the reconstruction of an evolutionary tree. Fortunately, with the advancement in the computational field this hurdle has been overcome. Advancement from16S rRNA genome sequencing to DNA sequencing platform has led to increased number of available sequence data for phylogenetic analysis. Thus, in the post genomic era, a large number of microbial sequences are available in public domains, continuous advancement in high throughput DNA sequencing techniques and the introduction of new phylogenetic inference methods has occurred. These three points provide a challenge and opportunity simultaneously to the researchers to study evolution, ecology, and taxonomy of microbes. One strategy to organize a large set of data in the form of hierarchical distance tree is by using single copy ribosomal protein marker distances. In this tree protein distance measures dissimilarity between the same kinds of markers and measures genomic distance average by ignoring the outlier. As a result, 60,000 organized genomes in a marker distance tree obtained, which result in >6000 species level clade and represented as 7597 taxonomic species. These findings will help the researchers to get pre calculated genomic group [78].

5.5. General steps for phylogenetic tree construction

There are four steps for phylogenetic tree construction of molecular sequences shown in **Figure 5.**

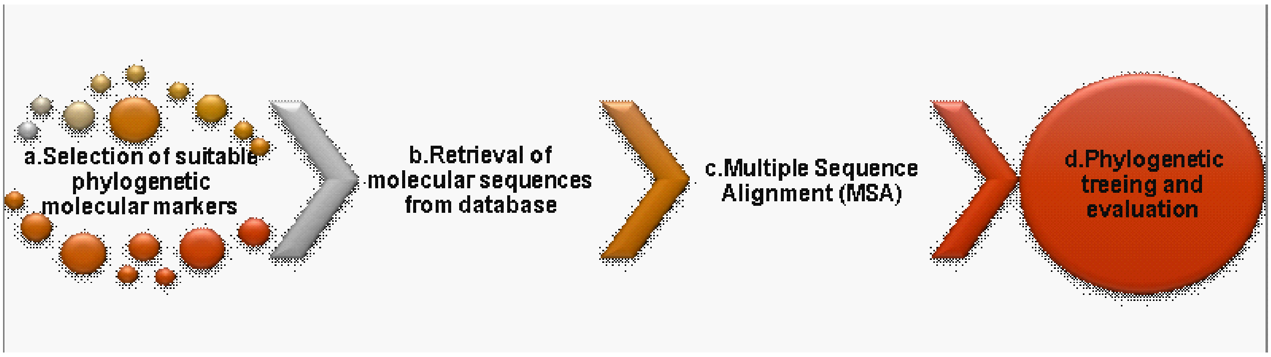
a.Selection of suitable phylogenetic molecular markers
b.Retrieval of molecular sequences from database
c.Multiple Sequence Alignment (MSA)
d.Phylogenetic treeing and evaluation

Figure 5: Four general steps of constructing phylogenetic tree

5.5.1. Selection of suitable phylogenetic markers

The phylogenetic marker is coding or non-coding DNA fragment (locus) used in phylogenetic reconstruction. These phylogenetic markers for microbes include nuclear encoded genes (like 16S rRNA, 5S rRNA, 28S rRNA), mitochondrial (cytochrome oxidase, mitochondrial 12S, cytochrome b, control region) and few chloroplast encoded genes (like rbcL, matK, rpl16) (67). Selection of suitable phylogenetic marker is crucial to study molecular evolution like duplications of genes, mutations, loss or gain of genes, genetic exchange such as recombination events, re-assortment, and horizontal or lateral gene transfer. For an ideal marker it should contain following characteristics:

(a) Single gene should be preferred over multiple genes e.g. use of mitochondrial and nuclear genes.

(b) Marker gene is aligned prior to phylogenetic tree construction; therefore, sequence alignment should be easy and without any ambiguous alignments.

(c) The substitution rate should be optimum to avoid saturation of multiple substitutions.

(d) Primers should be available for amplification of marker genes and universal primers be avoided since they may cause contamination in marker genes.

(e) Markers with too much variation in bases may not represent the true lineage [79].

5.5.2. Retrieval of molecular sequences from database

Molecular data can either be obtained from nucleotide or protein databases. This depends upon chosen organism/s.

5.5.3. Multiple Sequence Alignment (MSA)

Multiple Sequence Alignment (MSA) is for two or more than two molecular sequences. Purpose of MSA is to determine homology and evolutionary relationship between the under study sequences. Different types of alignment homology are obtained after multiple sequence alignment, shown in **Figure 6**.

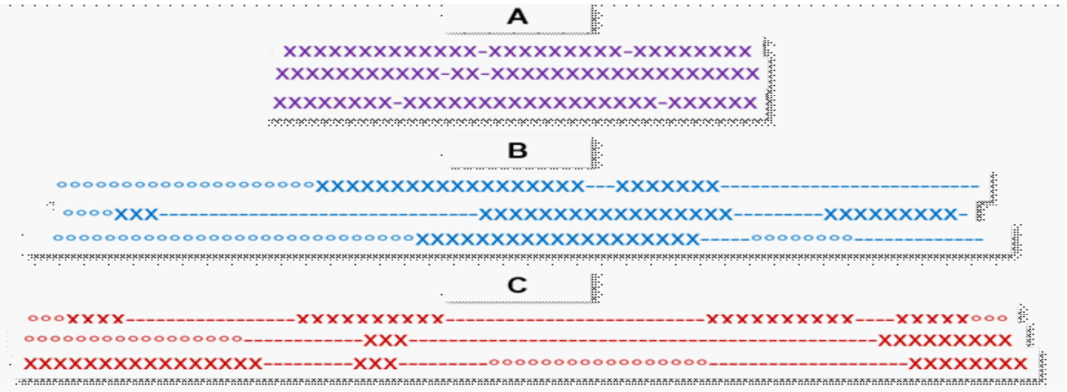

Figure 6: Different types of alignment homology. "x" represents an aligned amino acid residue, and “o” is an unalignable residue, ‘’--‘’ represents a gap. (A) Global sequence alignment (for comparing homologous genes) (B) Local sequence alignment (for finding homologous domains) (C) Long internal gaps.

There are different computer programs for multiple sequence alignment that are listed in **Table 5**.

Table 5: Computational tools for Multiple Sequence Alignment

Sr.No.	Tool	Year	Web link	Ref
1.	**T-Coffee**	2000 & new version in 2008	http://tcoffee.crg.cat/	[80]
2.	**MUSCLE**	2004	https://www.ebi.ac.uk/Tools/msa/muscle/	[81]
3.	**Kalign**	2005	https://www.ebi.ac.uk/Tools/msa/kalign/	[82]
4.	**ClustalW**	2007	http://www.clustal.org/	[83]
5.	**FAMSA**	2016	http://sun.aei.polsl.pl/REFRESH/famsa.	[84]
6.	**MAFFT**	2017	http://mafft.cbrc.jp/alignment/server/large.html	[85]
7.	**HAlign-II**	2017	http://lab.malab.cn/soft/halign/	[86]

5.5.4. Phylogenetic tree construction and evaluation

A phylogenetic tree is a graphical representation of the evolutionary relationships among genes or organisms. Phylogenetic tree is constructed when homologous residues aligned. Different methods or algorithms used to develop phylogenetic tree are distance based method, maximum parsimony, maximum likelihood and Bayesian models. Distance-based method does not use sequences directly while other three methods use sequence information, therefore, known as character-based methods shown in **Figure 7** [67,87].

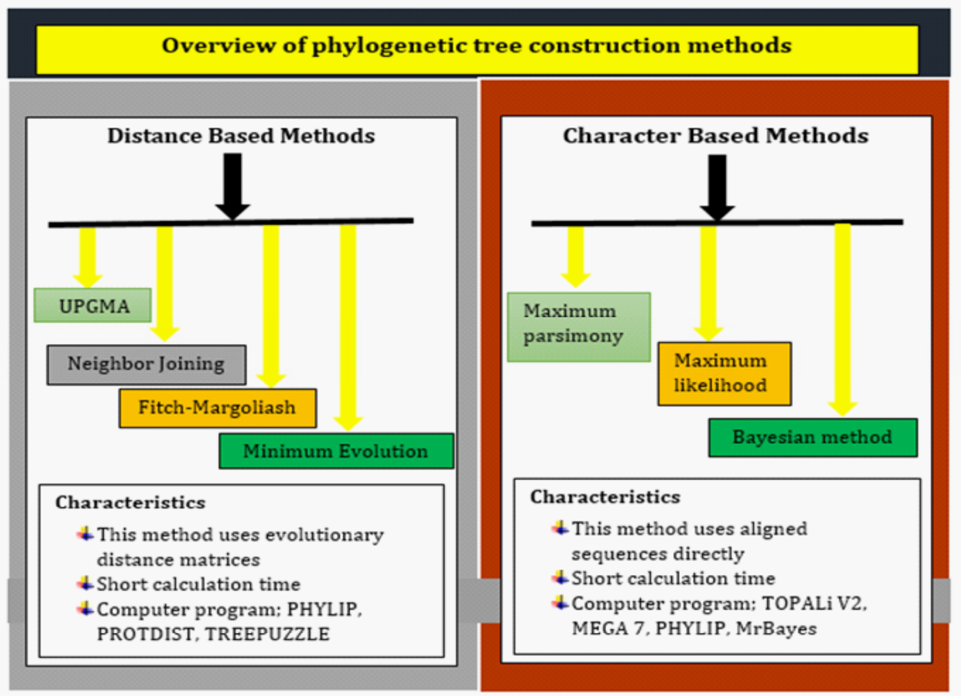
Overview of phylogenetic tree construction methods
Distance Based Methods
UPGMA
Neighbor Joining
Fitch-Margoliash
Minimum Evolution
Characteristics
This method uses evolutionary distance matrices
Short calculation time
Computer program: PHYLIP, PROTDIST, TREEPUZZLE
Character Based Methods
Maximum parsimony
Maximum likelihood
Bayesian method
Characteristics
This method uses aligned sequences directly
Short calculation time
Computer program: TOPALi V2, MEGA 7, PHYLIP, MrBayes

Figure 7: Different Phylogenetic Tree Construction Methods; UPGMA (Unweighted Pair Group Method with Arithmetic) proposed in 1958 by Sokal and Michener, Neighbor-Joining by Saitou and Nei (1987), Maximum parsimony by Henning (1966),maximum likelihood method by Felsenstein (1981).

5.5.5. Phylogenetic tree evaluation

Phylogenetic tree evaluation is necessary for the validity of tree and its shape. The phylogenetic tree represents species phylogeny if species under study are evolved from common ancestor. Branch length in the tree represents evolutionary distance that is tentatively correlated with evolutionary time. Therefore, branch length determines rate of evolution, gene duplication, and speciation events. Moreover, percentage of each external branch is calculated by bootstrapping method. If branch point scores or bootstrapping values is higher (approximately 90% or greater) then it predicts accurate tree. About 500-1000 times bootstrapping is required for reliable results. Other different statistical tests like Jackknifing, Kishino-Hasegawa test, Bayesian analysis and Shimodaira-Hasegawa employed to check the reliability and to confirm which tree is better. The Bayesian analysis is very fast and involves thousands of steps of resampling the results [66]. In an evolutionary tree, there are operational taxonomic units (OTUs) defined as the set of OTUs joined by the same ancestor or parental node [88]. Single 16S rRNA used to differentiate operational taxonomic units (OTUs)(89). How to interpret an evolutionary tree is shown in **Figure 8**.

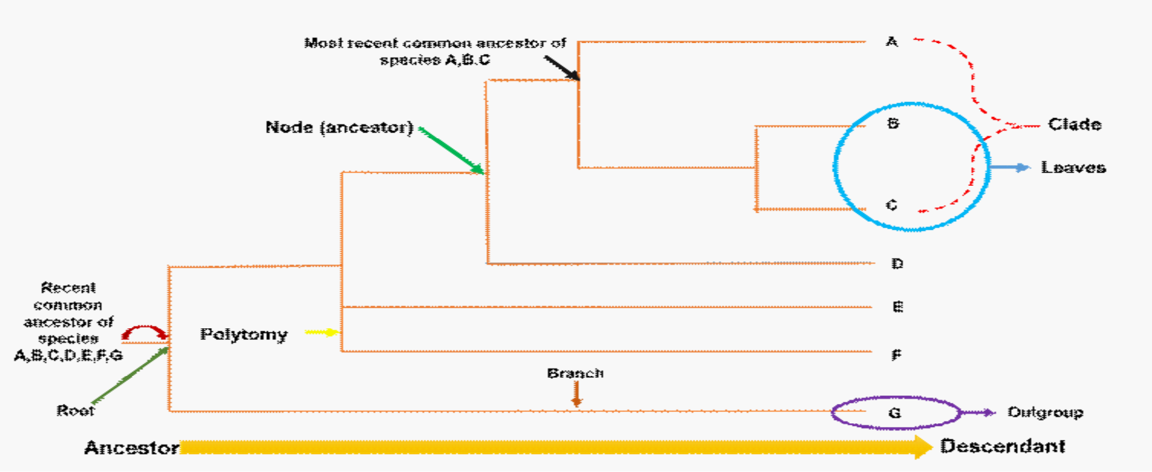
Most recent common ancestor of species A,B,C
Node (ancestor)
A
B
C
D
E
F
G
Clade
Leaves
Recent common ancestor of species A,B,C,D,E,F,G
Polytomy
Branch
Root
Outgroup
Ancestor
Descendant

Figure 8: Interpretation of Evolutionary Tree

6. Comparative Genomics of Microbial Pathogens

Comparative genomics is a holistic approach that compares two or more than two genomes to identify the similarities and differences among the genomes and to study the biology of genomes. Comparative genome analysis can find out the different perspectives of organisms as shown in **Figure 9** [90].

In post-genomic era, comparative genomics has been widely used to distinguish pathogenic and non-pathogenic species; it helped identify virulence factors and genes involved in pathogenicity by sequence analyses [6,91]. More than 1800 bacterial genomes have been sequenced including *Escherichia coli* O157:H7, *Vibrio cholerae, Staphylococcus aureus, Streptococcus pneumoniae, Clostridium difficile* and *Mycobacterium tuberculosis* on which comparative genomics approaches can be applied [92].

Different applications of comparative genomics include gene identification, finding regulatory motifs, in the field of molecular medicine and molecular evolution, selecting model organisms, in clustering of regulatory sites, finding genomic islands, selection of industrially important organism and much more which still need to be explored [93]. These comparative genomes approaches used to differentiate between the multi-drug resistant pathogen *S. maltophilia* and the plant-associated strains *S. maltophilia* R551-3 and *S. rhizophila* DSM14405. *S. maltophilia* contained heat shock proteins and virulence factors that were absent in plant-associated strains [94]. Another disease leptospirosis is a globally widespread zoonotic disease with important health consequences for humans and domesticated animals. This genus *Leptospira* is divided into infectious species for mammals and non-infectious species. Comparative genomics studies revealed that infectious *Leptospira* contained novel virulence modifying proteins, CRISPR-Cas systems and different metabolic pathways like pathogen-specific porphyrin metabolism while non-infectious species did not have these adaptations [74].

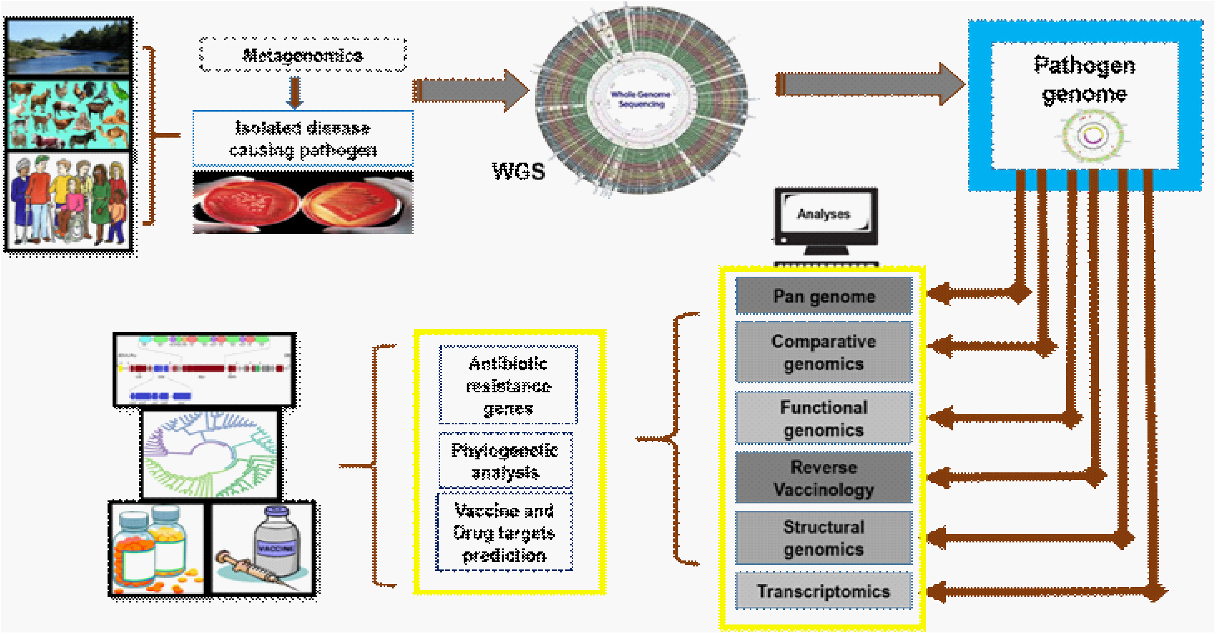

Figure 9: Overview of utilizing computational approaches for analysis of pathogen genome

6.1. Comparative genomic approaches

Comparative genomics considers many approaches for obtaining reliable results. Genome size is an important approach in comparative genomics. Genomic statistics include a number of coding regions, number of chromosomes, GC and AT contents, genome structure, and genome density. For example, genome size of soil-living bacteria has bigger than endosymbiotic bacteria. In addition, while transformation from free-living bacteria to pathogens they gain or lose number of genes. Comparative genomes studies consider these genomic statistics to find out the genomic differences and their reasons. These genomic statistics varies from species to species and even strains to strains [32]. In recent years, increasing number of available genomic information of multiple pathogenic and non-pathogenic bacterial species is also evident that genomic acquisition and reduction have an important role in evolution and pathogenecity. For example, human pathogens *Escherichia coli, Mycobacterium tuberculosis* and *Helicobacter pylori* cause diseases due to genome shifting [95].

Another important approach is finding homologous proteins (including orthologous and paralogous) that remains a challenge for researchers. For this purpose, protein sequences comparison is considered as the powerful tool. This comparison is based upon protein sequences of different species to trace back evolutionary history of many species. Computational tools BLAST, and other clustering tools k-means, affinity propagation, Markov clustering, FORCE, as well as transitivity clustering can be used for finding homologous estimation. In addition, identification of protein-protein interactions plays a vital role in determining biological processes within cells and characterizing those proteins that involved in pathogenicity. Different-proteome-wide common conserved protein-protein interactions (PPIs) for different pathogenic and non-pathogenic bacteria included *C. pseudotuberculosis, C. diphtheriae, C. ulcerans, M. tuberculosis, Y. pestis and E. coli* was determined [32].

6.2. Microbial pathogenomics

Pan-genome analysis of pathogen genome leads to identification of genome plasticity and pathogenic islands. The term pan-genome was first defined in 2005. Pan-genome consists of a core, dispensable and unique genomes. Core genes mostly have housekeeping and essential genes required for growth of bacteria. Dispensable genome carries foreign or modified gees obtained from horizontal genes transfer and these genes could be potential therapeutic targets. Unique genes are novel genes that only confined to particular strains or sometimes in species. These genes increase adaptability to host environment and increase virulence. Therefore comparative pan-genome study is important in studying antibiotic resistance, potential therapeutic targets, epidemiology and phylogenomics. Comparative genome along with pan-genome approach was used to investigate pathogenicity of seven *Campylobacter species*. Pan-genome results revealed 3933 core genome and 1,035 ubiquitous genes [96]. *Streptococcus* genus within phylum *Firmicutes* is among the most significant and diverse zoonotic pathogens. Considerable taxonomic approaches like DNA hybridization, 16S rRNA sequencing did not give the clear evolutionary implications of *Streptococci* species group. Therefore, comparative genomic approaches used to get a clear understanding of evolution of pathogenicity in *Streptococci.* Genome analysis revealed that pan-genome size increases with the addition of newly sequenced strains and core genome size decreases. Population structure analysis and phylogenetic analysis revealed two distinct lineages or clades formed within a species group. Virulence factors also evolved with species evolution [73].

6.3. Genome plasticity

Genome plasticity is the gain or loss of genes and gene rearrangements within specific strains of species for higher adaptability to a new environment. Genome plasticity comprised by several different mechanisms including gene arrangement, inversion, translocation, mutations, plasmid insertions from different organisms, and other insertions like transposons, insertion elements, bacteriophages and genomic islands. Genomic islands are large mobile elements that have cluster or bunch of genes that are directly or indirectly involved in bacterial pathogenicity (**Figure 10**).

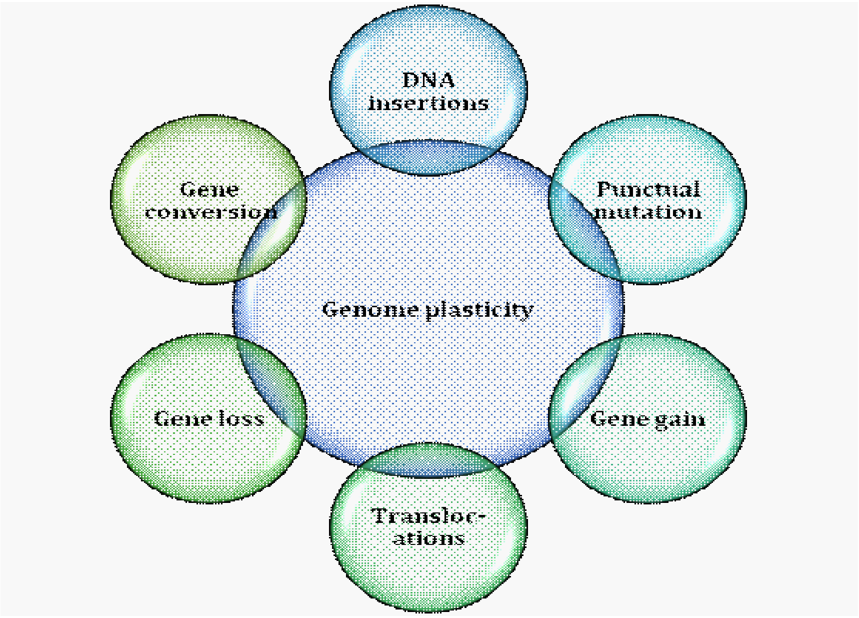

Figure 10: Mechanisms of genome plasticity

Whole genome sequence analysis of *Staphylococcus aureus* revealed mobile genetic elements that carry virulence and antibacterial genes. This horizontal gene transfer of mobile genetic elements mediates the evolution of methicillin resistance *Staph. aureus* [97]. In post-genomic era, researchers explore pathogenicity of microbes by genome comparison. Dao-feng and co predicted the pathogenic potential and international spread of *Staphylococcus argenteus* by genomic comparison analysis. The comparative genomic analysis (based on pan-core genome definition) performed among thirty *S. aureus* genomes, fifteen *Staphylococcus argenteus* and six *S. schweitzeri* genomes. Results revealed that all three species had rare core genome with interspecific recombination. Many virulence genes of *S. aureus, S. argenteus* and *S. schweitzeri* were homologous. Moreover, *S. argenteus* showed ambiguous biogeographical structure that was evidence of its international spread [98].

Pan-genome analysis can use for analysis of minor mutations like single nucleotide polymorphisms (SNPs) that are responsible for any kind of virulence. The pan-genome investigation of two *Mycobacterium tuberculosis* strains helped to identify SNPs, which led to the study of evolution and pathogenesis of these strains. Analysis showed that this species was highly clonal without any lateral gene transfer and these strains lost some genes that were present in other strains [99].

Comparative genomic analyses can be used for finding reasons of bacterial outbreaks in history. In Germany (May-June 2011) an outbreak caused by Shiga-toxin producing *E.coli* O104:H4 that infects more than 3000 people. Scientists tried to find out the reason of this virulence in *E.coli*. After comparative genomic analysis of different strains of pathogenic *E.coli*, they found that it belongs to rare serotype O104:H4. In addition, this strain belonged to enteroaggregative *E.coli* lineage that had acquired Shiga-toxin producing gene and antibiotic resistance gene (i.e. broad-spectrum beta-lactamase gene of CTX-M-15 class). They reported the acquisition of stx2 prophage, gene encoding AAF/III fimbriae which was responsible for alternative adhesion mechanism [100]. *Shigellaflexneri* causes shigellosis that is a leading cause of bacillary dysentery in developing countries, especially in Asia. Infants under five are more susceptible to this disease. Based on O- antigen of outer membrane lipopolysaccharide there are 19 serotypes of *Shigellaflexneri.* Despite its disease causing ability, there was little knowledge about its virulence and genomic structure. Therefore, Pawan Parajuli, Marcin Adamski and Naresh K. Verma, 2017 used hybrid methods of long-read single-molecule real-time (SMRT) and short-read MiSeq (Illumina) sequencing technology to generate a high quality genome sequence of *S. flexneri* serotype 1c for the first time. Results revealed that Y394 chromosome of *S. flexneri* contained mobile genetic elements, IS elements and plasmids. These set of genes was actually responsible for bacterial evolution, diversification, adaptation, pathogen's virulence and antibiotic resistance of bacteria. From the detailed analysis, they also identified novel and highly modified O-antigen structure consisting of three different O-antigen modifying gene clusters that came by horizontal gene transfer from three different bacteriophages. These were the causes of pathogen's virulence and survival in host environment [48]. Pan-genome analysis of *Akkermansia muciniphila* was done for the first time. It is the inhabitant of the intestinal tract and plays a crucial role in human health. Whole genome sequencing and annotation done of 39 isolates. Results revealed the flexible pan-genome consisting of 5644 unique proteins. Comprehensive genomic analysis among human, mouse and pig microbiomes revealed transcontinental distribution of phylogroups of *A. muciniphila* across human gut microbiomes. Qualitative analysis showed its co-relation with anti-diabetic drug usage and body mass index. It also acquired antibiotic resistance genes by lateral gene transfer from symbiotic microbes [101]. Kono N, Tomita M and Arakawa K. Nobuki in 2017, developed the algorithm for reordering of the contigs based on experimental replication profiling (eRP) to facilitate the study of the complete genome sequences, genome rearrangements, and structural variations and to summarize the bacterial genome structure within a draft genome. They also suggested the appropriate timing for genomic sampling i.e. during exponential growth phase of bacteria to obtain information about contig position relative to terminus and replication origins [102].

7. Comparative Genomics for Industrial and Environmental Friendly Microbes

Comparative genomics is also useful for exploration of microbes that are involved in bioremediation and industry. Gang Zhou and his team-mates for the first time gave complete genome sequence of *Citrobacter werkmanii* with genome features and annotation. *Citrobacter werkmanii* BF-6 belongs to family *Enterobacteriaceae*. It has been used for bioremediation of heavy metals because it produced acid type phosphatase enzyme and can accumulate heavy metals due to biofilm formation. *C. werkmanii* BF-6 and *C. werkmanii* NRBC 105721 had closely related evolutionary relationship. They also found different genes involved in biofilm formation. The 12-biofilm producing genes and their location on chromosome BF-6 is illustrating below in **Figure 11** [103].

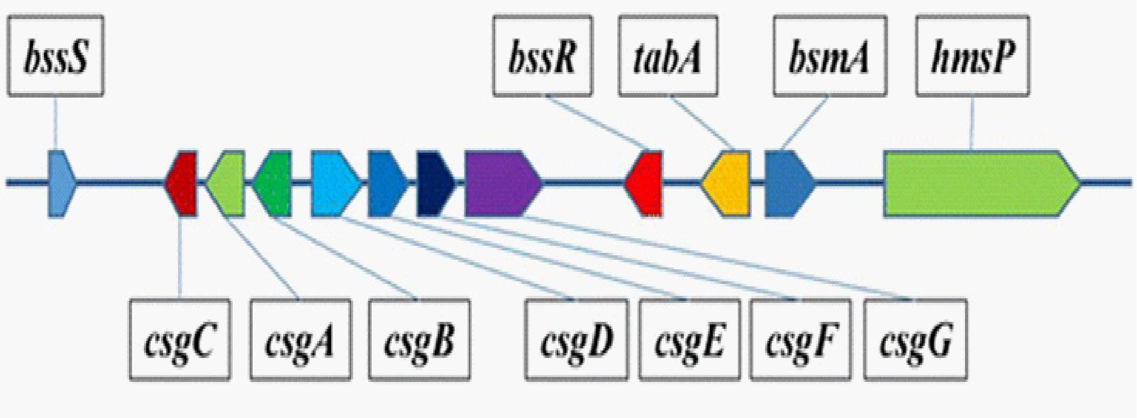

Figure 11: The relative position of biofilm producing genes on chromosome BF-6 by Citrobacterwerkmanii [103].

Industrially important species *Propionibacterium freudenreichii* (member actinobacterial group) genome was completely sequenced by using PacBio RS II sequencing platform. Genomes of 20 strains of P. freudenreichii were compared and results showed. Results showed two conjugative plasmids and three active lysogenic bacteriophages. It also helped in identification of different DNA modifications, which led to the characterization of restriction modification systems; that is CRISPR-Cas systems. The genomic difference observed in specific mucus binding and surface piliation among strains. These characteristics allowed them to grow at unfavorable conditions and help in the formation of biofilm [104].

In post genomic era, computational approaches integrated with "omics" included proteomics, genomics, and metabolomics for selection of drug and vaccine targets. For pathogenic bacteria, comparative and subtractive genomic approaches are being widely used. These identified targeted genes should be non-homologous to host. *Vibrio cholera* is a cholera-causing agent. By using a comparative genomic approach of *Vibrio cholera*, drug target Cholera endotoxin B subunit and membrane proteins like secG, secY, and secE were identified as potential vaccine targets [105].

8. Reverse Vaccinology to Identify Potential Vaccine and Drug Targets for Microbes

Development of vaccines with the help of computational approaches, utilizing genomic data, instead of culturing microbes, is termed as 'reverse vaccinology'. Vaccine development by conventional methods need culturing of pathogenic microbes and all biochemical, microbial and immunological techniques, and all this made it time consuming and laborious. Reverse vaccinology begins with the screening of pathogenic genome, which results in epitope prediction and epitope prediction is said to be the heart of reverse vaccinology [106]. Genomic sequencing discovery had paved the path for predicting the potential antigen candidates from complete genomic data. Predicted candidates are then used in vaccine preparation (**Figure 12**).

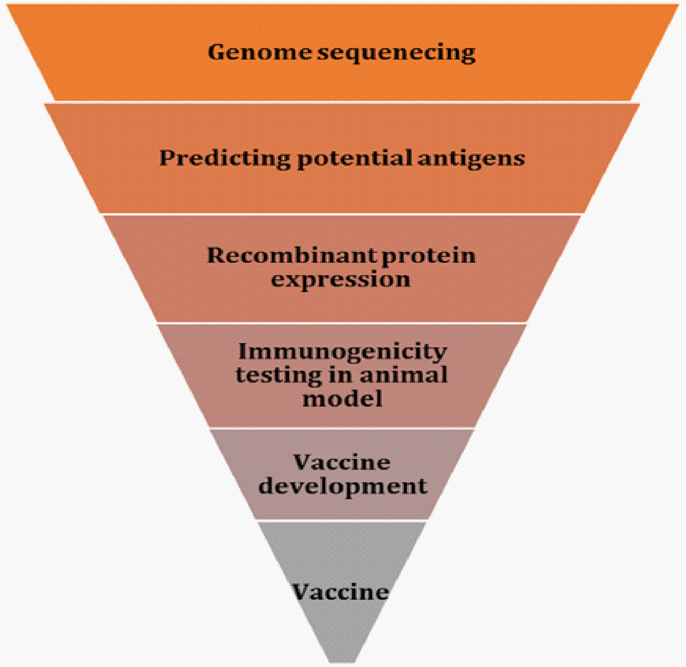
Genome sequenecing
Predicting potential antigens
Recombinant protein expression
Immunogenicity testing in animal model
Vaccine development
Vaccine

Figure 12: Steps involved in vaccine development by reverse vaccinology

Comparative genomics, metabolic pathways analysis, and additional drug prioritizing parameters were used to identify drug and vaccine targets against *Mycoplasma genitalium*, a pathogenic agent responsible for sexually transmitted diseases in human. Total 79 proteins were identified out of which 67 proteins were non-homologous essential proteins that could be potential drug and vaccine targets [107].

Ghosh S and co (2014) also identified drug and vaccine targets in *Staphylococcus aureus* by using comparative genomic approach. They identified 19 proteins as vaccine candidates and 34 proteins as drug targets [107].

Undoubtedly, vaccinologists have successfully eradicated life-threatening diseases. Still, there is a long way to go, to our surprise, there are only ~50 human vaccines out of which only 35-40 are licensed in the US and Europe [108]. The first vaccine developed using reverse vaccinology was against Serogroup *B meningococcus*, by *RinoRappuoli* [109]. They first screened the genome of *B meningococcus*, examined the genome for antigens. Expression of potential candidates was tested in *E. coli* and most potential candidates were applied in vaccine development. After massive efforts, this vaccine was approved safe and potent [110].

Different soft wares are involved in reverse vaccinology a few of them are listed below,

Programs identifying Open Reading Frames

Sr. No.	Software
1.	ORF-FINDER
2.	GLIMMER
3.	GS-FINDER

Programs identifying potential proteins

Sr. No.	Software
1.	ProDom
2.	Pfam
3.	PROSITE

8.1. Applications of reverse vaccinology

Reverse vaccinology (RV) is an efficient and cost-effective as compared to conventional vaccine development approaches. Software for reverse vaccinology includes VacSol, NERVE, VAXIGN, RANKPEP, Vaceed, PGAP. As eukaryotes possess enormous and complicated genome as compared to prokaryotes, therefore RV is more effective towards eukaryotic genome [111].

Bacterial diseases for which licensed vaccines have been developed using 'reverse vaccinology' approach are listed as follows (**Table 6**).

Table 6: Vaccines developed by using "Reverse Vaccinology" approach

Sr. No.	Bacteria	Disease	Vaccine	Trade name
1.	*Neisseria meningitides*	Meningococcal meningitis	Meningococcal Group B Vaccine	BEXSERO
2.	*Bacillus anthracis*	Anthrax	Anthrax Vaccine Adsorbed (AVA)	Biothrax
3.	*Vibrio cholera*	Cholera	Cholera Vaccine Live Oral	Vaxchora
4.	*Corynebacterium diphtheria*	Diphtheria	Diphtheria and Tetanus Toxoid Adsorbed	None
5.	*Yersinia pestis*	Plague	Plague vaccine	None
6.	*Streptococcus pneumoniae*	Pneumococcal	Pneumococcal Vaccine, Polyvalent	Pneumovax 23
7.	*Salmonella enterica*	Typhoid fever	Typhoid Vaccine Live Oral Ty21a	Vivotif

9. Future prospects

This study spells out that microbiology is turning into a data science; potent association of experimental and computational biologists can bring revolution in near future. Considering the present rate of advancements of technology in this discipline, is difficult to predict the future. Nevertheless, we will outline few improvements to be made. Undoubtedly, NGS require small amount of genetic material for analysis, but this is even lesser, for example in case of endangered species. In addition, improvements must be made to produce more and longer sequence reads, reduced sequence errors and algorithms for data analysis, this will surely result in improved transcriptomic and genomic data compilation. Future studies require focusing on genome architecture and regulation as it is link with conservation biology. Cost effective sequencing technique is applied more frequently, generating more sequencing data and hence demands new infrastructures, analysis and data storage approaches and sharing databases. This revolution resulted in enhancements of bringing novel aims and objectives of genetic research in reach of molecular ecologists.

10. References

1. Lengauer T, editor Computational biology at the beginning of the post-genomic era. Informatics; 2001: Springer.

2. Pallen MJ. Microbial bioinformatics 2020. Microbial biotechnology. 2016; 9(5): 681-686.

3. Vallenet D, Calteau A, Cruveiller S, Gachet M, Lajus A, Josso A, et al. MicroScope in 2017: an expanding and evolving integrated resource for community expertise of microbial genomes. Nucleic acids research. 2017; 45(D1): D517-D28.

4. Guizelini D, Raittz RT, Cruz LM, Souza EM, Steffens MB, Pedrosa FO. GFinisher: a new strategy to refine and finish bacterial genome assemblies. Scientific reports. 2016; 6.

5. Parks DH, Rinke C, Chuvochina M, Chaumeil P-A, Woodcroft BJ, Evans PN, et al. Recovery of nearly 8,000 metagenome-assembled genomes substantially expands the tree of life. Nature microbiology. 2017.

6. Raskin DM, Seshadri R, Pukatzki SU, Mekalanos JJ. Bacterial genomics and pathogen evolution. Cell. 2006; 124(4): 703-714.

7. Loman NJ, Pallen MJ. Twenty years of bacterial genome sequencing. 2015; 13: 787.

8. Heather JM, Chain B. The sequence of sequencers: The history of sequencing DNA. Genomics. 2016; 107(1): 1-8.

9. Fleischmann RD, Adams MD, White O, Clayton RA, Kirkness EF, Kerlavage AR, et al. Whole-genome random sequencing and assembly of Haemophilus influenzae Rd. Science (New York, NY). 1995; 269(5223): 496-512.

10. Alimi J-P, Poirot O, Lopez F, Claverie J-M. Reverse transcriptase-polymerase chain reaction validation of 25 "orphan" genes from Escherichia coli K-12 MG1655. Genome research. 2000; 10(7): 959-966.

11. Hutchison Iii CA, Newbold JE, Potter SS, Edgell MH. Maternal inheritance of mammalian mitochondrial DNA. 1974; 251: 536.

12. Hall N. Advanced sequencing technologies and their wider impact in microbiology. 2007; 1518-25 p.

13. McAdam PR, Richardson EJ, Fitzgerald JR. High-throughput sequencing for the study of bacterial pathogen biology. Current opinion in microbiology. 2014; 19: 106-113.

14. Armougom F, Raoult D. Exploring microbial diversity using 16S rRNA high-throughput methods. J Comput Sci Syst Biol. 2009; 2(1): 74-92.

15. Thompson JF, Milos PM. The properties and applications of single-molecule DNA sequencing. Genome biology. 2011; 12(2): 217.

16. Churko JM, Mantalas GL, Snyder MP, Wu JC. Overview of high throughput sequencing technologies to elucidate molecular pathways in cardiovascular diseases. Circulation research. 2013; 112(12): 1613-1623.

17. Caporaso JG, Lauber CL, Walters WA, Berg-Lyons D, Huntley J, Fierer N, et al. Ultra-high-throughput microbial community analysis on the Illumina HiSeq and MiSeq platforms. The ISME journal. 2012; 6(8): 1621-1624.

18. Loman NJ, Misra RV, Dallman TJ, Constantinidou C, Gharbia SE, Wain J, et al. Performance comparison of benchtop high-throughput sequencing platforms. Nature biotechnology. 2012; 30(5): 434-439.

19. Perkins TT, Tay CY, Thirriot F, Marshall B. Choosing a benchtop sequencing machine to characterise Helicobacter pylori genomes. PloS one. 2013; 8(6): e67539.

20. Yang L, Yang H-l, Tu Z-c, Wang X-l. High-Throughput Sequencing of Microbial Community Diversity and Dynamics during Douchi Fermentation. PloS one. 2016; 11(12): e0168166.

21. Quigley L, O'Sullivan O, Beresford TP, Ross RP, Fitzgerald GF, Cotter PD. High-throughput sequencing for detection of subpopulations of bacteria not previously associated with artisanal cheeses. Applied and environmental microbiology. 2012; 78(16): 5717-5723.

22. Eid J, Fehr A, Gray J, Luong K, Lyle J, Otto G, et al. Real-time DNA sequencing from single polymerase molecules. Science. 2009; 323(5910): 133-138.

23. Conlan S, Thomas PJ, Deming C, Park M, Lau AF, Dekker JP, et al. Single-molecule sequencing to track plasmid diversity of hospital-associated carbapenemase-producing Enterobacteriaceae. Science translational medicine. 2014; 6(254): 254ra126-254ra126.

24. Fang G, Munera D, Friedman DI, Mandlik A, Chao MC, Banerjee O, et al. Genome-wide mapping of methylated adenine residues in pathogenic Escherichia coli using single-molecule real-time sequencing. Nature biotechnology. 2012; 30(12): 1232-1239.

25. Buermans H, Den Dunnen J. Next generation sequencing technology: advances and applications. Biochimica et Biophysica Acta (BBA)-Molecular Basis of Disease. 2014; 1842(10): 1932-1941.

26. Hach F, Hormozdiari F, Alkan C, Hormozdiari F, Birol I, Eichler EE, et al. mrsFAST: a cache-oblivious algorithm for short-read mapping. Nature methods. 2010; 7(8): 576-577.

27. Baruzzo G, Hayer KE, Kim EJ, Di Camillo B, FitzGerald GA, Grant GR. Simulation-based comprehensive benchmarking of RNA-seq aligners. Nature methods. 2017; 14(2): 135-139.

28. Huang S, Zhang J, Li R, Zhang W, He Z, Lam T-W, et al. SOAPsplice: genome-wide ab initio detection of splice junctions from RNA-Seq data. Frontiers in genetics. 2011; 2.

29. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nature methods. 2012; 9(4): 357-359.

30. Sedlazeck FJ, Rescheneder P, Von Haeseler A. NextGenMap: fast and accurate read mapping in highly polymorphic genomes. Bioinformatics. 2013; 29(21): 2790-2791.

31. Baumbach J, Tauch A, Rahmann S. Towards the integrated analysis, visualization and reconstruction of microbial gene regulatory networks. Briefings in Bioinformatics. 2009; 10(1): 75-83.

32. Ali A, Soares S, Barbosa E, Santos A, Barh D, Bakhtiar S. Microbial comparative genomics: an overview of tools and insights into the genus Corynebacterium. J Bacteriol Parasitol. 2013; 4(167): 2.

33. Liu X, Wu J, Wang J, Liu X, Zhao S, Li Z, et al. WebLab: a data-centric, knowledge-sharing bioinformatic platform. Nucleic Acids Research. 2009; 37(Web Server issue): W33-W9.

34. Smedley D, Haider S, Ballester B, Holland R, London D, Thorisson G, et al. BioMart – biological queries made easy. BMC Genomics. 2009; 10(1): 22.

35. Karolchik D, Hinrichs AS, Furey TS, Roskin KM, Sugnet CW, Haussler D, et al. The UCSC Table Browser data retrieval tool. Nucleic Acids Res. 2004; 32(Database issue): D493-D496.

36. Wang J, Kong L, Gao G, Luo J. A brief introduction to web-based genome browsers. Brief Bioinform. 2013; 14(2): 131-143.

37. Karolchik D, Baertsch R, Diekhans M, Furey TS, Hinrichs A, Lu YT, et al. The UCSC Genome Browser Database. Nucleic Acids Res. 2003; 31(1): 51-54.

38. Zhang Y, Liu T, Meyer CA, Eeckhoute J, Johnson DS, Bernstein BE, et al. Model-based analysis of ChIP-Seq (MACS). Genome biology. 2008; 9(9): R137.

39. Nielsen CB, Cantor M, Dubchak I, Gordon D, Wang T. Visualizing genomes: techniques and challenges. Nature methods. 2010; 7(3 Suppl): S5-s15.

40. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. Journal of molecular biology. 1990; 215(3): 403-410.

41. Kanehisa M, Goto S. KEGG: kyoto encyclopedia of genes and genomes. Nucleic acids research. 2000; 28(1): 27-30.

42. Abbott JC, Aanensen DM, Bentley SD. WebACT: an online genome comparison suite. Comparative Genomics. 2008: 57-74.

43. Kurtz S, Phillippy A, Delcher AL, Smoot M, Shumway M, Antonescu C, et al. Versatile and open software for comparing large genomes. Genome biology. 2004; 5(2): R12.

44. Chiang S, Burch T, Van Domselaar G, Dick K, Radziwon A, Brusnyk C, et al. The interaction between thymine DNA glycosylase and nuclear receptor coactivator 3 is required for the transcriptional activation of nuclear hormone receptors. Molecular and cellular biochemistry. 2010; 333(1-2): 221.

45. Kerkhoven R, Van Enckevort FH, Boekhorst J, Molenaar D, Siezen RJ. Visualization for genomics: the microbial genome viewer. Bioinformatics. 2004; 20(11): 1812-1814.

46. Hallin PF, Stærfeldt H-H, Rotenberg E, Binnewies TT, Benham CJ, Ussery DW. GeneWiz browser: an interactive tool for visualizing sequenced chromosomes. Standards in genomic sciences. 2009; 1(2): 204.

47. Vallenet D, Engelen S, Mornico D, Cruveiller S, Fleury L, Lajus A, et al. MicroScope: a platform for microbial genome annotation and comparative genomics. Database. 2009; 2009: bap021.

48. Parajuli P, Adamski M, Verma NK. Bacteriophages are the major drivers of Shigella flexneri serotype 1c genome plasticity: a complete genome analysis. BMC genomics. 2017; 18(1): 722.

49. Nielsen H. Predicting Secretory Proteins with SignalP. Protein Function Prediction: Methods and Protocols. 2017: 59-73.

50. Seemann T. Prokka: rapid prokaryotic genome annotation. Bioinformatics. 2014; 30(14): 2068-2069.

51. Hamada M, Ono Y, Asai K, Frith MC. Training alignment parameters for arbitrary sequencers with LAST-TRAIN. Bioinformatics. 2017; 33(6): 926-928.

52. Treangen TJ, Ondov BD, Koren S, Phillippy AM. The Harvest suite for rapid core-genome alignment and visualization of thousands of intraspecific microbial genomes. Genome biology. 2014; 15(11): 524.

53. Didelot X, Wilson DJ. ClonalFrameML: efficient inference of recombination in whole bacterial genomes. PLoS computational biology. 2015; 11(2): e1004041.

54. Lan Y, Morrison JC, Hershberg R, Rosen GL. POGO-DB—a database of pairwise-comparisons of genomes and conserved orthologous genes. Nucleic acids research. 2013; 42(D1): D625-D32.

55. Yarza P, Richter M, Peplies J, Euzeby J, Amann R, Schleifer K-H, et al. The All-Species Living Tree project: a 16S rRNA-based phylogenetic tree of all sequenced type strains. Systematic and applied microbiology. 2008; 31(4): 241-250.

56. Inouye M, Dashnow H, Raven L-A, Schultz MB, Pope BJ, Tomita T, et al. SRST2: Rapid genomic surveillance for public health and hospital microbiology labs. Genome medicine. 2014; 6(11): 90.

57. Croucher NJ, Page AJ, Connor TR, Delaney AJ, Keane JA, Bentley SD, et al. Rapid phylogenetic analysis of large samples of recombinant bacterial whole genome sequences using Gubbins. Nucleic acids research. 2014; 43(3): e15-e.

58. Larsen MV, Cosentino S, Lukjancenko O, Saputra D, Rasmussen S, Hasman H, et al. Benchmarking of methods for genomic taxonomy. Journal of clinical microbiology. 2014; 52(5): 1529-1539.

59. Zerbino DR, Birney E. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. Genome research. 2008; 18(5): 821-829.

60. Hyatt D, Chen G-L, LoCascio PF, Land ML, Larimer FW, Hauser LJ. Prodigal: prokaryotic gene recognition and translation initiation site identification. BMC bioinformatics. 2010; 11(1): 119.

61. Johnson BK, Scholz MB, Teal TK, Abramovitch RB. SPARTA: Simple Program for Automated reference-based bacterial RNA-seq Transcriptome Analysis. BMC bioinformatics. 2016; 17(1): 66.

62. Lee I, Kim YO, Park S-C, Chun J. OrthoANI: an improved algorithm and software for calculating average nucleotide identity. International journal of systematic and evolutionary microbiology. 2016; 66(2): 1100-1113.

63. Paintdakhi A, Parry B, Campos M, Irnov I, Elf J, Surovtsev I, et al. Oufti: an integrated software package for high-accuracy, high-throughput quantitative microscopy analysis. Molecular microbiology. 2016; 99(4): 767-77.

64. Cuccuru G, Orsini M, Pinna A, Sbardellati A, Soranzo N, Travaglione A, et al. Orione, a web-based framework for NGS analysis in microbiology. Bioinformatics. 2014; 30(13): 1928-1929.

65. Rizwan M, Naz A, Ahmad J, Naz K, Obaid A, Parveen T, et al. VacSol: a high throughput in silico pipeline to predict potential therapeutic targets in prokaryotic pathogens using subtractive reverse vaccinology. BMC bioinformatics. 2017; 18(1): 106.

66. Brenchley PJ, Brenchley P, Harper D. Palaeoecology: Ecosystems, environments and evolution: CRC Press; 1998.

67. Patwardhan A, Ray S, Roy A. Molecular markers in phylogenetic studies-A review. Journal of Phylogenetics & Evolutionary Biology. 2014; 2014.

68. Sapp J. The prokaryote-eukaryote dichotomy: meanings and mythology. Microbiology and molecular biology reviews. 2005; 69(2): 292-305.

69. Plyusnin A, Elliott RM. Bunyaviridae: molecular and cellular biology: Horizon Scientific Press; 2011.

70. Gao B, Gupta RS. Microbial systematics in the post-genomics era. Antonie van Leeuwenhoek. 2012; 101(1): 45-54.

71. Roux S, Enault F, le Bronner G, Debroas D. Comparison of 16S rRNA and protein-coding genes as molecular markers for assessing microbial diversity (B acteria and A rchaea) in ecosystems. FEMS microbiology ecology. 2011; 78(3): 617-628.

72. Ludwig W, Schleifer K. Bacterial phylogeny based on 16S and 23S rRNA sequence analysis. FEMS microbiology reviews. 1994; 15(2-3): 155-173.

73. Gao X-Y, Zhi X-Y, Li H-W, Klenk H-P, Li W-J. Comparative genomics of the bacterial genus Streptococcus illuminates evolutionary implications of species groups. PloS one. 2014; 9(6): e101229.

74. Fouts DE, Matthias MA, Adhikarla H, Adler B, Amorim-Santos L, Berg DE, et al. What makes a bacterial species pathogenic?: Comparative genomic analysis of the genus Leptospira. PLoS neglected tropical diseases. 2016; 10(2): e0004403.

75. Jain R, Rivera MC, Lake JA. Horizontal gene transfer among genomes: the complexity hypothesis. Proceedings of the National Academy of Sciences. 1999; 96(7): 3801-3806.

76. Horiike T, Hamada K, Kanaya S, Shinozawa T. Origin of eukaryotic cell nuclei by symbiosis of Archaea in Bacteria is revealed by homology-hit analysis. Nature Cell Biology. 2001; 3(2): 210-214.

77. Le Q, Sievers F, Higgins DG. Protein multiple sequence alignment benchmarking through secondary structure prediction. Bioinformatics. 2017; 33(9): 1331-1337.

78. Field D, Garrity G, Gray T, Morrison N, Selengut J, Sterk P, et al. The minimum information about a genome sequence (MIGS) specification. Nature biotechnology. 2008; 26(5): 541-547.

79. Chun J, Hong S. Methods and programs for calculation of phylogenetic relationships from molecular sequences. Molecular phylogeny of microorganisms Caister Academic Press, Norfolk. 2010: 23-39.

80. Notredame C, Higgins DG, Heringa J. T-Coffee: A novel method for fast and accurate multiple sequence alignment. Journal of molecular biology. 2000; 302(1): 205-217.

81. Edgar RC. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic acids research. 2004; 32(5): 1792-1797.

82. Lassmann T, Sonnhammer EL. Kalign–an accurate and fast multiple sequence alignment algorithm. BMC bioinformatics. 2005; 6(1): 298.

83. Larkin MA, Blackshields G, Brown N, Chenna R, McGettigan PA, McWilliam H, et al. Clustal W and Clustal X version 2.0. bioinformatics. 2007; 23(21): 2947-2948.

84. Deorowicz S, Debudaj-Grabysz A, Gudyś A. FAMSA: Fast and accurate multiple sequence alignment of huge protein families. Scientific reports. 2016; 6.

85. Katoh K, Rozewicki J, Yamada KD. MAFFT online service: multiple sequence alignment, interactive sequence choice and visualization. Briefings in Bioinformatics. 2017: bbx108.

86. Wan S, Zou Q. HAlign-II: efficient ultra-large multiple sequence alignment and phylogenetic tree reconstruction with distributed and parallel computing. arXiv preprint arXiv: 170400878. 2017.

87. Horiike T. AN INTRODUCTION TO MOLECULAR PHYLOGENETIC ANALYSIS. Reviews in Agricultural Science. 2016; 4: 36-45.

88. Sicheritz-Pontén T, Andersson SG. A phylogenomic approach to microbial evolution. Nucleic acids research. 2001; 29(2): 545-552.

89. Case RJ, Boucher Y, Dahllöf I, Holmström C, Doolittle WF, Kjelleberg S. Use of 16S rRNA and rpoB genes as molecular markers for microbial ecology studies. Applied and environmental microbiology. 2007; 73(1): 278-288.

90. Wei L, Liu Y, Dubchak I, Shon J, Park J. Comparative genomics approaches to study organism similarities and differences. Journal of biomedical informatics. 2002; 35(2): 142-50.

91. Prentice MB. Bacterial comparative genomics. Genome biology. 2004; 5(8): 338.

92. Donkor ES. Sequencing of bacterial genomes: principles and insights into pathogenesis and development of antibiotics. Genes. 2013; 4(4): 556-572.

93. Sivashankari S, Shanmughavel P. Comparative genomics-A perspective. Bioinformation. 2007; 1(9): 376.

94. Alavi P, Starcher MR, Thallinger GG, Zachow C, Müller H, Berg G. Stenotrophomonas comparative genomics reveals genes and functions that differentiate beneficial and pathogenic bacteria. BMC genomics. 2014; 15(1): 482.

95. Ahmed N, Dobrindt U, Hacker J, Hasnain SE. Genomic fluidity and pathogenic bacteria: applications in diagnostics, epidemiology and intervention. Nature reviews microbiology. 2008; 6(5): 387-394.

96. Méric G, Yahara K, Mageiros L, Pascoe B, Maiden MC, Jolley KA, et al. A reference pan-genome approach to comparative bacterial genomics: identification of novel epidemiological markers in pathogenic Campylobacter. PloS one. 2014; 9(3): e92798.

97. Jamrozy D, Coll F, Mather AE, Harris SR, Harrison EM, MacGowan A, et al. Evolution of mobile genetic element composition in an epidemic methicillin-resistant Staphylococcus aureus: temporal changes correlated with frequent loss and gain events. BMC genomics. 2017; 18(1): 684.

98. Zhang D-F, Zhi X-Y, Zhang J, Paoli GC, Cui Y, Shi C, et al. Preliminary comparative genomics revealed pathogenic potential and international spread of Staphylococcus argenteus. BMC Genomics. 2017; 18(1): 808.

99. Alland D, Whittam TS, Murray MB, Cave MD, Hazbon MH, Dix K, et al. Modeling bacterial evolution with comparative-genome-based marker systems: application to Mycobacterium tuberculosis evolution and pathogenesis. Journal of bacteriology. 2003; 185(11): 3392-3399.

100. Rohde H, Qin J, Cui Y, Li D, Loman NJ, Hentschke M, et al. Open-source genomic analysis of Shiga-toxin–producing E. coli O104: H4. New England Journal of Medicine. 2011; 365(8): 718-724.

101. Guo X, Li S, Zhang J, Wu F, Li X, Wu D, et al. Genome sequencing of 39 Akkermansia muciniphila isolates reveals its population structure, genomic and functional diverisity, and global distribution in mammalian gut microbiotas. BMC genomics. 2017; 18(1): 800.

102. Kono N, Tomita M, Arakawa K. eRP arrangement: a strategy for assembled genomic contig rearrangement based on replication profiling in bacteria. BMC genomics. 2017; 18(1): 784.

103. Zhou G, Peng H, Wang Y-s, Huang X-m, Xie X-b, Shi Q-s. Complete genome sequence of Citrobacter werkmanii strain BF-6 isolated from industrial putrefaction. BMC genomics. 2017; 18(1): 765.

104. Deptula P, Laine PK, Roberts RJ, Smolander O-P, Vihinen H, Piironen V, et al. De novo assembly of genomes from long sequence reads reveals uncharted territories of Propionibacterium freudenreichii. BMC genomics. 2017; 18(1): 790.

105. Chawley P, Samal HB, Prava J, Suar M, Mahapatra RK. Comparative genomics study for identification of drug and vaccine targets in Vibrio cholerae: MurA ligase as a case study. Genomics. 2014; 103(1): 83-93.

106. Kanampalliwar AM, Soni R, Girdhar A, Tiwari A. Web Based Tools and Databases for Epitope Prediction and Analysis: A Contextual Review. International Journal of Computational Bioinformatics and In Silico Modeling. 2013; 2: 180-185.

107. Butt AM, Tahir S, Nasrullah I, Idrees M, Lu J, Tong Y. Mycoplasma genitalium: a comparative genomics study of metabolic pathways for the identification of drug and vaccine targets. Infection, Genetics and Evolution. 2012; 12(1): 53-62.

108. Barrett AD. Vaccinology in the twenty-first century. npj Vaccines. 2016; 1: 16009.

109. Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, et al. The sequence of the human genome. science. 2001; 291(5507): 1304-1351.

110. Sette A, Rappuoli R. Reverse Vaccinology: Developing Vaccines in the Era of Genomics. Immunity. 2010; 33(4): 530-541.

111. Davies MN, Flower DR. Harnessing bioinformatics to discover new vaccines. Drug discovery today. 2007; 12(9): 389-395.

112. Fraser CM, Gocayne JD, White O, Adams MD, Clayton RA, Fleischmann RD, et al. The minimal gene complement of Mycoplasma genitalium. science. 1995; 270(5235): 397-404.

113. Fleischmann RD, Adams MD, White O, Clayton RA, Kirkness EF, Kerlavage AR, et al. Whole-genome random sequencing and assembly of Haemophilus influenzae Rd. science. 1995; 269(5223): 496-512.

114. Garrett RA. Genomes: Methanococcus jannaschii and the golden fleece. Current Biology. 1996; 6(11): 1377-1380.

115. Himmelreich R, Hilbert H, Plagens H, Pirkl E, Li B-C, Herrmann R. Complete sequence analysis of the genome of the bacterium Mycoplasma pneumoniae. Nucleic acids research. 1996; 24(22): 4420-4449.

116. Blattner FR, Plunkett G, Bloch CA, Perna NT, Burland V, Riley M, et al. The complete genome sequence of Escherichia coli K-12. science. 1997; 277(5331): 1453-1462.

117. Fraser CM, Casjens S, Huang WM, Sutton GG, Clayton R, Lathigra R, et al. Genomic sequence of a Lyme disease spirochaete, Borrelia burgdorferi. Nature. 1997; 390(6660): 580-586.

118. Kunst F, Ogasawara N, Moszer I, Albertini A, Alloni G, Azevedo V, et al. The complete genome sequence of the gram-positive bacterium Bacillus subtilis. Nature. 1997; 390(6657): 249-256.

119. Cole S, Brosch R, Parkhill J, Garnier T, Churcher C, Harris D, et al. Deciphering the biology of Mycobacterium tuberculosis from the complete genome sequence. Nature. 1998; 393(6685): 537-544.

120. Fraser CM, Norris SJ, Weinstock GM, White O, Sutton GG, Dodson R, et al. Complete genome sequence of Treponema pallidum, the syphilis spirochete. Science. 1998; 281(5375): 375-388.

121. Bentley SD, Parkhill J. Comparative genomic structure of prokaryotes. Annu Rev Genet. 2004; 38: 771-791.

122. Loman NJ, Pallen MJ. Twenty years of bacterial genome sequencing. Nature Reviews Microbiology. 2015; 13(12): 787-795.

123. Nelson KE, Clayton RA, Gill SR, Gwinn ML, Dodson RJ, Haft DH, et al. Evidence for lateral gene transfer between Archaea and bacteria from genome sequence of Thermotoga maritima. Nature. 1999; 399(6734): 323-329.

124. Drancourt M, Bollet C, Carlioz A, Martelin R, Gayral J-P, Raoult D. 16S ribosomal DNA sequence analysis of a large collection of environmental and clinical unidentifiable bacterial isolates. Journal of clinical microbiology. 2000; 38(10): 3623-3630.

125. Stover C, Pham X, Erwin A, Mizoguchi S, Warrener P, Hickey M, et al. Complete genome sequence of Pseudomonas aeruginosa PAO1, an opportunistic pathogen. Nature. 2000; 406(6799): 959-964.

126. Bolotin A, Wincker P, Mauger S, Jaillon O, Malarme K, Weissenbach J, et al. The complete genome sequence of the lactic acid bacterium Lactococcus lactis ssp. lactis IL1403. Genome research. 2001; 11(5): 731-753.

127. Kuroda M, Ohta T, Uchiyama I, Baba T, Yuzawa H, Kobayashi I, et al. Whole genome sequencing of meticillin-resistant Staphylococcus aureus. The Lancet. 2001; 357(9264): 1225-1240.

128. Hayashi T, Makino K, Ohnishi M, Kurokawa K, Ishii K, Yokoyama K, et al. Complete genome sequence of enterohemorrhagic Eschelichia coli O157: H7 and genomic comparison with a laboratory strain K-12. DNA research. 2001; 8(1): 11-22.

129. Krieg AM. CpG motifs in bacterial DNA and their immune effects. Annual review of immunology. 2002;20(1):709-60.

130. Ikeda H, Ishikawa J, Hanamoto A, Shinose M, Kikuchi H, Shiba T, et al. Complete genome sequence and comparative analysis of the industrial microorganism Streptomyces avermitilis. Nature biotechnology. 2003; 21(5): 526-531.

131. Acinas SG, Klepac-Ceraj V, Hunt DE, Pharino C, Ceraj I, Distel DL, et al. Fine-scale phylogenetic architecture of a complex bacterial community. Nature. 2004; 430(6999): 551-554.

132. Andries K, Verhasselt P, Guillemont J, Göhlmann HW, Neefs J-M, Winkler H, et al. A diarylquinoline drug active on the ATP synthase of Mycobacterium tuberculosis. Science. 2005; 307(5707): 223-227.

133. Pruitt KD, Tatusova T, Maglott DR. NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. Nucleic acids research. 2006; 35(suppl_1): D61-D5.

134. Margulies M, Egholm M, Altman WE, Attiya S, Bader JS, Bemben LA, et al. Genome sequencing in microfabricated high-density picolitre reactors. Nature. 2005; 437(7057): 376-380.

135. Marcy Y, Ouverney C, Bik EM, Lösekann T, Ivanova N, Martin HG, et al. Dissecting biological "dark matter" with single-cell genetic analysis of rare and uncultivated TM7 microbes from the human mouth. Proceedings of the National Academy of Sciences. 2007; 104(29): 11889-11894.

136. Simpson KT, Thomas JG. Oral Microbiome: Contributions to Local and Systemic Infections. Current Oral Health Reports. 2016; 3(1): 45-55.

137. Allen JE, Gardner SN, Slezak TR. DNA signatures for detecting genetic engineering in bacteria. Genome biology. 2008; 9(3): R56.

138. Liolios K, Chen I-MA, Mavromatis K, Tavernarakis N, Hugenholtz P, Markowitz VM, et al. The Genomes On Line Database (GOLD) in 2009: status of genomic and metagenomic projects and their associated metadata. Nucleic acids research. 2009; 38(suppl_1): D346-D54.

139. Qin J, Li R, Raes J, Arumugam M, Burgdorf KS, Manichanh C, et al. A human gut microbial gene catalogue established by metagenomic sequencing. nature. 2010; 464(7285): 59-65.

140. Loo VG, Bourgault A-M, Poirier L, Lamothe F, Michaud S, Turgeon N, et al. Host and pathogen factors for Clostridium difficile infection and colonization. New England Journal of Medicine. 2011; 365(18): 1693-1703.

141. Ribeiro FJ, Przybylski D, Yin S, Sharpe T, Gnerre S, Abouelleil A, et al. Finished bacterial genomes from shotgun sequence data. Genome research. 2012; 22(11): 2270-2277.

142. Larsen MV, Cosentino S, Rasmussen S, Friis C, Hasman H, Marvig RL, et al. Multilocus sequence typing of total genome sequenced bacteria. Journal of clinical microbiology. 2012: JCM. 06094-11.

143. Albertsen M, Hugenholtz P, Skarshewski A, Nielsen KL, Tyson GW, Nielsen PH. Genome sequences of rare, uncultured bacteria obtained by differential coverage binning of multiple metagenomes. Nature biotechnology. 2013; 31(6): 533-538.

144. Chewapreecha C, Harris SR, Croucher NJ, Turner C, Marttinen P, Cheng L, et al. Dense genomic sampling identifies highways of pneumococcal recombination. Nature genetics. 2014; 46(3): 305-309.

145. Gupta SK, Padmanabhan BR, Diene SM, Lopez-Rojas R, Kempf M, Landraud L, et al. ARG-ANNOT, a new bioinformatic tool to discover antibiotic resistance genes in bacterial genomes. Antimicrobial agents and chemotherapy. 2014; 58(1): 212-220.

146. Trewby H, Wright D, Breadon EL, Lycett SJ, Mallon TR, McCormick C, et al. Use of bacterial whole-genome sequencing to investigate local persistence and spread in bovine tuberculosis. Epidemics. 2016; 14: 26-35.

147. Rosana ARR, Orata FD, Xu Y, Simkus DN, Bramucci AR, Boucher Y, et al. Draft genome sequences of seven bacterial strains isolated from a polymicrobial culture of coccolith-bearing (C-type) Emiliania huxleyi M217. Genome announcements. 2016; 4(4): e00673-16.

148. Hutchison CA, Chuang R-Y, Noskov VN, Assad-Garcia N, Deerinck TJ, Ellisman MH, et al. Design and synthesis

of a minimal bacterial genome. Science. 2016; 351(6280): aad6253.

149. Hanisch U-K, Kettenmann H. Microglia: active sensor and versatile effector cells in the normal and pathologic brain. Nature neuroscience. 2007; 10(11): 1387-1394.

150. Oliveira PH, Touchon M, Cury J, Rocha EP. The chromosomal organization of horizontal gene transfer in bacteria. Nature communications. 2017; 8: 841.