International Journal of Zoology and Applied Biosciences **ISSN: 2455-9571**
Volume 10, Issue 1, pp: 15-28, 2025
https://doi.org/10.55126/ijzab.2025.v10.i01.003
http://www.ijzab.com

Review Article

COMPARATIVE GENOMICS: A MODERN-DAY APPROACH FOR UNDERSTANDING THE BLUE PRINT OF LIFE

[1]Gayatri Bhatt, [2]Sushil Bhattarai, [3]Priyanka Negi, [1]Vaibhavi Thapa, [1]Swastika Nandy and [1]*Saranya Joshi

[1]Department of Biotechnology, School of Applied and Life Sciences, Uttaranchal University, Dehradun-248007, Uttarakhand India.
[2]Chemistry & Bioprospecting Division, ICFRE - Forest Research Institute, Dehradun-248006, Uttarakhand, India.
[3]Department of Zoology, School of Allied Sciences, Dev Bhoomi Uttarakhand University, Dehradun, Uttarakhand – 248007

Article History: Received 12th December 2024; Accepted 19th January 2025; Published 27th January 2025

ABSTRACT

Comparative genomics is an inevitable tool for carrying out studies related to evolutionary history of an organism. It unveils the scientific language of nature residing inside the genome of each organism by creating baseline data which is helpful in exploration of studies related to proteomics and drug discovery. The present study elaborates the various aspects associated with comparative genomics by detailing the useful tools available online for this approach. The study also describes a detailed methodology of comparative genomics involving the comparison of genome structure, coding regions and non-coding regions. It also encompasses the technical challenge involved in the approach which is the alignment of the whole genome of organisms. Recommendation has been made for development of tools required for robust comparison of non-coding regions for the identification of regulatory elements involved in controlling the gene expression of the organism which may help in providing significant leads for identification of genes associated with diseases and ailments.

Keywords: Comparative Genomics, Whole Genome Alignment, Phylogenetic studies, Coding, Non-coding regions.

INTRODUCTION

The genes, or the genome as a whole has a foundational role in all the biological processes of an organism. Owing to this fundamental function, studying genomes is essential for advancement of biological studies by revealing the genetic basis of health, evolution and biological functioning (Khan *et al*., 2020) (Charamis *et al*., 2024; Kobras *et al*., 2021). On the other hand, it is claimed that a single organism's genome does not say much of itself. In the development of the phylogenetic process of evolution, genomes and genes must be considered relative, compared and contrasted with other species (or subspecies, or strains) to reveal the important functional elements by highlighting the evolutionary conservation (Clark, 1999). This implies that the study of individual genome sequences will meaningfully provide information relevant to the structural aspect of the genome with limited aspects showing its functional significance (Miller *et al*., 2004). Therefore, in recent years, comparative genomics is a significant approach for interpreting genomic data resulting in its superior comprehension by unveiling patterns and insights that arise only when genomes are compared across various species or populations (WHO, 2024 https://researchrabbitapp.com/home). A vast range of bioinformatics tools and software are used to derive meaningful insights from genomic comparisons and they significantly facilitate and accelerate the analysis, visualization and interpretation of these complex genomic data.

DNA sequences that are common between two species and regulates the expression of genes responsible for similarity in function, shows conservation, whereas the

*Corresponding Author: Dr. Saranya Joshi Department of Biotechnology, School of Applied and Life Sciences, Uttaranchal University, Dehradun, Uttarakhand India, Email: joshisaranya@gmail.com.

sequences that regulate expression of genes that gives rise to dissimilarity between the species show divergence (Hardison, 2003). All eukaryotes, regardless of how distantly they may be related to humans, have a common ancestor and categorization of functions inside and across specialized cells (Bornstein *et al*., 2023; Koonin, 2010). Conserved regions are a direct key to unlock phylogenetic association between species. Comparisons between the genomes of different species, by first studying the chloroplast genome, provided the foundation for applying such genome comparisons for deducing evolutionary relationships as well as functional insights (Palmer, 1985). It was one of the early studies employing near to complete genome data for the purpose of comparison. The analysis of Hox gene clusters among different animal taxa constitutes a major early comparative study, which showed the conservation of such crucial developmental genes thus demonstrating the importance of comparative genomic studies to unravel evolutionary process and its relation to functional conservation (Mulhair & Holland, 2024; Williams & Forey, 2004). Since then, the genomes of several yeasts, two worms, and three mammals- human, mouse, and rat have been sequenced, and compared (Miller *et al*., 2004). Comparative genomics explores these conserved DNA sequences across species which evinces the persistence of these essential biological functions throughout evolutionary events. It was the Human Genome Project's completion in 2003 that transformed comparative genomics.

The human genome project helped scientists advance thei r understanding of genetic sequences and led to the develop ment of different innovative methodologies designed to rapidly analyse large amounts of biological data. It created new sequencing technologies besides providing a reference human genome and allowed them to efficiently collect extensive DNA sequencing data. Such data has led to great comparative analyses at various levels, including cellular and molecular (Smith *et al*., 2012). One of the reasons for large omics datasets production is the aspiration for a better understanding of the molecular basis of our uniqueness, the origin of our species, and life enhancement on earth. Genomic diversity is observed to be significant in both, phylogenetically similar species and more distantly related species. For instance, genetic variation among humans is around 1%, and this accounts for about 1 million discrepancies, including Singl Nucleotide Polymorphism (SNP) and various other modifications. Moreover, around 12% of the human genome is influenced by copy number variations (CNV), which are seen to vary highly among individuals and range up to 5 kilobases in most cases. In addition, complexity is also affected by non-coding sequences as well as gene-gene interactions. This region, has provided genomic variation, including both size and functional complexities (Smith *et al*., 2012). Thus, it demands for whole-genome and transcriptome sequencing to detect all genomic and transcriptomic variations.

Several tools have been developed to aid in comparative genomics in recent years. A broad array of genome comparison tools is currently accessible which can be broadly categorised into pairwise local alignment comparison tools, various global alignment tools (inclusive of pairwise alignment, multisequence alignment and multi-genome alignment), the substring maximum-exact-match tools and lastly the alignment viewing tools (Chain, 2003). The extensive array of tools and technologies used in comparative genomics has enhanced our understanding significantly about the structure, function, and evolutionary processes of genomes. From sequencing and assembly to functional annotation and visualization, these tools facilitate the discovery of complexity in genomic variation and mechanisms of evolution. High-throughput techniques have transformed comparative genomics by enabling extensive, cost-effective, and precise examinations of genetic information across various species that can be completed in short period of time (Kircher & Kelso, 2010). With increased data output, comparative genomics has become approachable and researchers are able to carry out detailed phylogenetic analyses across different species (Lemmon & Lemmon, 2013). Advanced Next Generation Sequencing (NGS) techniques are widely useful in genomics as they can generate millions of sequence data very efficiently in short duration of time and are cost-effective (Sikhakhane *et al*., 2016). NGS and diverse 'omics' technologies (including genomics, proteomics, metabolomics, transcriptomics, and phenomics), present opportunities for full genome annotation (Sikhakhane *et al*., 2016). However, all these technologies have some unique error characteristics and limitations which are to be accounted for in selecting appropriate platforms for specific experiments (Kircher & Kelso, 2010). In the research undertaken in comparative genomics, highly sophisticated computational tools are designed, and strategic utilization of such tools is used for genome-level inspection to arrive at useful biological inferences (Wei *et al*., 2002).

The most crucial and challenging problem that researchers are facing today in the biological science concerns deciphering the functional meaning of a specific genomic sequence. One of the most promising strategies developed so far to face this challenging issue relies on the comparative technique based on the approach of traditional biology, currently modified and developed into the practices of sequence comparison. For effectively confronting challenges in comparative genomics, a multi-layered and structured strategy has been designed and developed for overcoming different problems concerning to databases, computation, and biological understanding (Haubold & Wiehe, 2004; Ptacek, 2005). When studying the genome of a species or a genera, pan genome approach helps in classifying genes unique to the given species as well as the core genome and the accessory genome. This expansion of comparative genomics; the pan genomic approach focuses on the genomic study of the entire set of genes in species or genus as a single data structure for

deeper inference (Carlos Guimaraes *et al*., 2015). On the other hand, the study by Haubold and Wiehe (2004) deals with the thorough analysis of both interspecific and intraspecific genome, which argues that such approaches constructed on a solid biological basis, analyses homology for the identification of conserved regions bearing functional significance in sequences and is essential for the exploration of the genome. The researchers also emphasized the issue of selection in relation to functional significance and explain how one can determine positive selection via patterns of mutation. More importantly, the authors highlighted the need to use various genomic information, including SNP assessments and complete genome-based comparisons, as means to understanding fully the genomic functions and mechanisms of evolution. According to the study, computational methods such as phylogenetic reconstruction and coalescent theory are used to infer evolutionary relationships from genetic data, yet an integrative perspective is necessary for a comprehensive understanding of genomic composition and function (Haubold & Wiehe, 2004) .

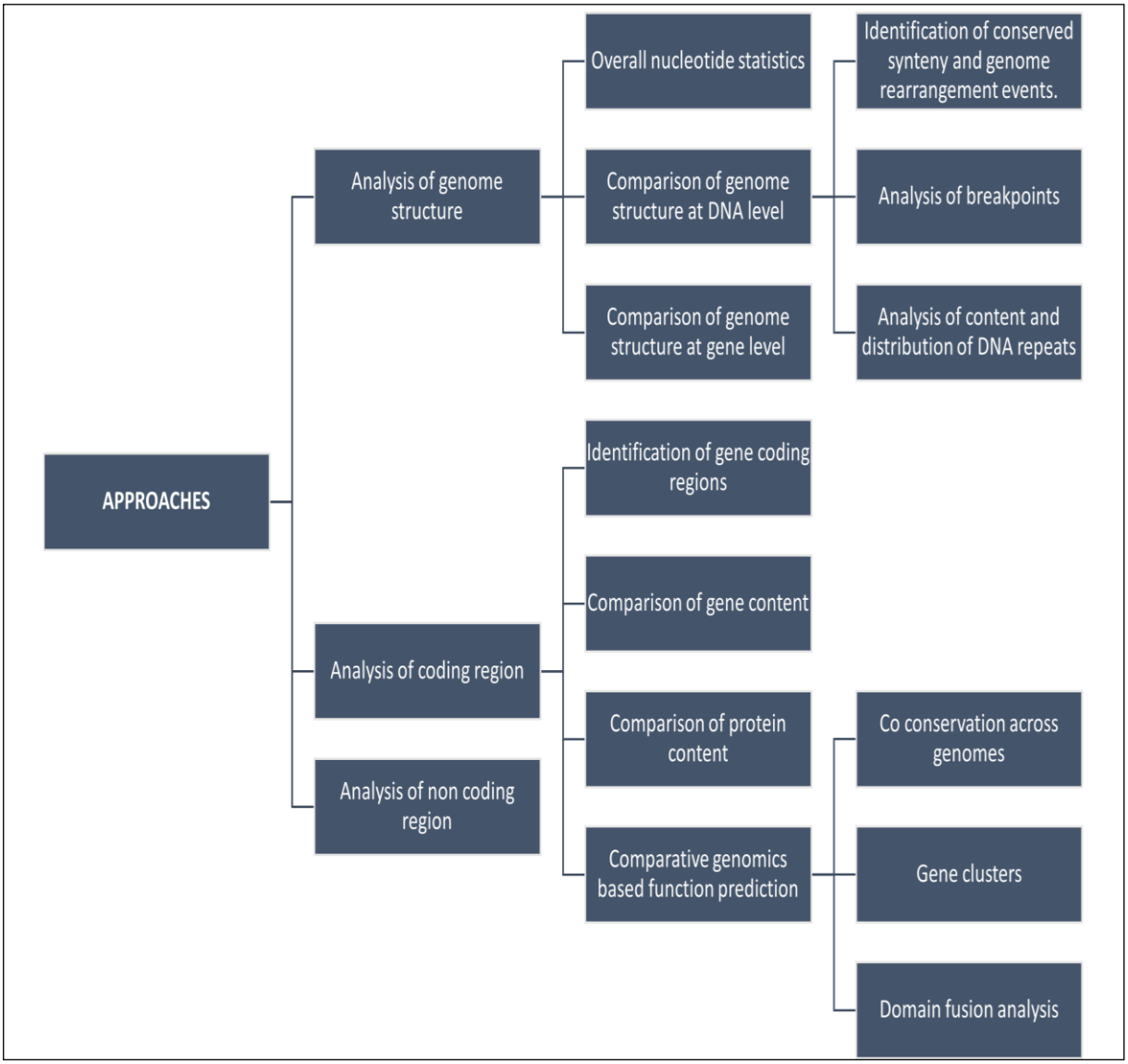

Figure1. Approaches for a comparative genomic study (taken from Wei *et al*., 2002).

Comparative genomics approach in a biological system refers to the analysis and integration of information across different genomes at biological level. The holistic methodology provides a basis through which whole genomic datasets of different organisms are analysed, leading to the findings which cannot be attained by isolated

investigations. It emphasizes understanding relationships and interactions among various macromolecules as a function of time through the help of high-throughput technologies and advanced analytical techniques. It is expected that collaboration between researchers will enhance communication, validate the biological relevance of the results, which in turn may consequently give rise to new discoveries and predictive models in biological research (Lin & Qian, 2007). Despite this, comparative genomics remains largely interdisciplinary, with researchers from various backgrounds exploring genomic phenomena, rather than systematically combining methods and concepts from different disciplines (Sankoff & Nadeau, 2000).

For comparative genomics studies, different approaches can be followed, however the basic of the study as proposed by (Wei *et al*., 2002) is shown at Fig.1. The authors have categorized the comparative genomics analyses into three main areas: genome configuration, coding regions, and non-coding regions. Analysis of the genome structure comprises evaluation of overall nucleotide content for parameters such as size, guanine and cytosine (GC) content, and other aspects including DNA repeats along with changes in structure because of synteny and break points. Coding regions are examined for genes; their composition; their relationships (orthologs and paralogs) which uncover functional similarity or dissimilarity. At last, regulatory elements such as the transcription factor binding sites in the non-coding regions are studied as they impact gene expression responsible for several biological processes (Wei *et al*., 2002). In a study, a genome wide comparison of four complete eukaryote genomes of *Saccharomyces paradoxus*, *Saccharomyces mikatae* and *Saccharomyces bayanus* vs *Saccharomyces cerevisiae* resulted in both identification of full set of conserved genes and regulatory elements (Kellis *et al*., 2003). Altogether, the proposed analyses establishes a comprehensible theoretical framework for studying genetic and functional variation in organisms.

In a nutshell, comparative genomic studies can be proceeded with various approaches individually or in combination to retrieve information regarding the relationships, evolution and functional insights of a species. Hence, it was envisaged that a concise review on comparative genomics should be taken up as it would be very much helpful in deciphering the hidden language residing inside the blueprint of life that is DNA. Further it will help in studies leading to exploration of the genetic material by generating baseline data which is anticipated to be helpful in developing novel strategies to tackle the challenges related to various non-curable diseases and ailments.

MATERIALS AND METHODS

The standard series of methodology and steps followed for the examination of genomes of various species for a comparative genomic study are shown in Figure 2. Outlining the objectives for the comparative genomics study of the target organism is initiated by sequencing the genome which produces high quality data with the help of modern and improved sequencing methods evolved due to the advancement in sequencing technologies and bioinformatic tools (Ali, 2013; Sivashankari & Shanmughavel, 2007). Some of the most commonly used tools employed in Comparative Genomics have been listed at Table 1. Next, structural assessment of the genome with the functional assessment of both the coding and non-coding regions allow for genome annotation (Wei *et al*., 2002). Thorough quality checking of the data is essential to assess its reliability. The identification of conserved and divergent regions is achieved by combining both phylogenetic and pan genomic approaches to interpret obtained data. Evolutionary relationships are established after analysing the genome by applying multiple bioinformatic tools and the utilization of biological databases, aiding researchers in multiple omics integration, contributing to advancements in the field of evolutionary biology, medicine and biotechnology. It must be noted that there are exceptions to general patterns that can still pose challenges, bioinformatics and computational biology have developed methods to identify patterns within large genomic datasets (Commins *et al*., 2009). However, when selecting comparative genomics tools, it's crucial to consider their advantages and disadvantages based on specific applications, such as detecting pathogens or understanding gene function and regulation (Chain, 2003).

Genome Data Acquisition

Whole genome sequencing (WGS) has transformed comparative genomics, allowing the investigation of both coding and non-coding regions in the genome for functional elements, ultimately providing a comprehensive genomic information for further use (Benjak *et al*., 2015; Nakagawa & Fujita, 2018; Nobrega & Pennacchio, 2004). Revelations of conserved sequences that are crucial for essential biological functions across various species and identification of key genomic regions associated with complex traits is possible due to Whole Genome Sequencing (WGS) techniques (Morrison *et al*., 2017). Although Whole Genome Sequencing (WGS) as a tool is efficient in the detection of genetic distinctions, studies have demonstrated a variance in outcome that are yielded by combinations of aligners and variant identifying algorithms, where rare and novel variants are particularly susceptible to discrepancies depending on the data analysis framework adopted (Hwang *et al*., 2019). When compared to Whole Exom Sequencing (WES) and targeted gene sequencing, WGS is more sensitive in regions known to have high GC content, as it provides more consistent coverage thus outperforming both (Trudsø *et al*., 2020). However, the extent of sequencing determines the accuracy of variant identification. GATK and SAMtools are toolkits that perform optimally at low coverage, whereas for high coverage, CASAVA is most effective (Cheng *et al*., 2014). Several tools have been developed, which include visualization tools like the ECR Browser, transcription factor analysis tools such as rVista and multiTF, and

alignment tools like zPicture and Mulan and are employed in efficient study of the sequences (Loots & Ovcharenko, 2005). The short fragments produced during sequencing steps are reconstructed to recover the original genome. This process is known as genome assembly and is performed to obtain an organisms actual genome makeup. Tools like PATRIC allow an easy to use experience in genome assembly, its annotation and the comparative analysis of the data (Wattam *et al*., 2018). The structural variant discovery is possible through a framework that allows the identification of specific annotation and orthology, optimization of new genes, and finding isoform by the Comparative Annotation Toolkit (CAT) which was initially meant for comparative gene compilation.

Genome annotation process primarily targets the structural and functional annotation of genome, where identification of gene elements and allotments of its functional role is performed respectively. Collaborative competitions like Assemblathon and Alignathon aim to unite professionals in order to assess and improve both genome assembly and annotation process (Fiddes *et al*., n.d.). To lower the human intervention, making genome annotation a swift process, extensive reliance on computational tools is omnipresent, however, manual reviewing is necessary to maintain precision. It is important for researchers to be knowledgeable about the workings of the annotation processes so that correct reviewing for the reliability of the genomic data can be executed in order to make legitimate inferences (Berriman & Harris, 2004). Although different genome annotation methods may vary in operation and are continuously advancing, standardizing the description of the annotation process is imperative. Despite that, Standard Operating procedures (SOP) lack uniformity in both format and substance. This can be amounted to the lack of a central repository where these protocols can be exchanged and archived (Angiuoli *et al*., 2008).

Data Preparation and Quality Control

Preparation of data for the purpose of comparative genomics is carried out in steps with the assistance from different tools. This is precedented with the preprocessing of the genome in order to avoid issues that may cause errors. As a result, repetitive, missing or redundant sequences are screened out. Even as high throughput sequencing technologies are constantly evolving, strict data quality is always to be maintained through quality control measures to ensure accuracy. Various tools together with robust methodologies are applied by researchers to remove low quality data before comparative genomic analysis. One of the approaches involves use of TagCleaner (web application) for identifying and deleting known and unknown tag sequences from a metagenomic dataset. It is efficient in screening of repetitions and short reads (Schmieder *et al*., 2010). QUAST is another tool that is able to evaluate the quality of the genome by providing statistics and reports for genome assembly comparison, both with or without a genome to refer to. It is worth mentioning that read mapping is not always an efficient way of identifying misassembled contigs; thus, the approaches should be used with utmost care (Gurevich *et al*., 2013; Lehri *et al*., 2017). However, as they enhance the quality of the ensuing genomic and metagenomic data, for superior upstream and downstream analyses, the approaches ensure better conclusions.

Genome Alignment

In comparative genomics, whole genome alignment is a process where DNA sequences are aligned to be compared on the basis of homology and orthology, to detect conserved regions irrespective of different rearrangement and duplication events. The newly sequenced genome is compared against genomes that have studied before in order to understand its characteristics in relation to the pre-existing genomic data. Methods such as progressive alignment, local alignment and reference free alignment are followed for genome mapping (Armstrong *et al*., 2019). Pairwise alignment plays a role in the process of determination of evolutionary relations since it permits more sophisticated form of analysis including orthologous gene prediction and identification of cases of lateral gene transfer. Inference of orthology can be done in two ways: either tree based approach, which involves the construction of gene trees to understand orthology according to speciation events, or graph based approach which highlights orthology based on similarities in sequences (Kapli *et al*., 2020). A number of resources that assist in chordate genome analysis are provided in Ensembl genome browser. These include gene homology, synteny and whole genome alignment (Herrero *et al*., 2016). Whole genome alignment enables the detection of evolution on a large scale by means of estimating the occurrence and the position of structural rearrangements and duplications in addition to detecting the small scale evolution through analysis of substitution and indels on the whole genome level (Dewey, 2019). In a study on measuring the level of agreement between the alignments and their comparison based on coverage and accuracy, the researchers have reported a lack of agreement among the alignments not only in species far from humans but also in mice, a widely researched model organism. Further, Pecan was observed to be most accurate in their study (Chen & Tompa, 2010). Comparative annotation enhances genome annotations across multiple genomes by facilitating the transfer of annotations from a well annotated reference genome to other genomes that have been aligned with it. This allows in function prediction for both coding regions and non-coding regions containing regulatory elements. However, despite such significant advancements in the process of the WGA methodology development, there are numerous challenges that remain unaddressed to date. Arguably, the construction of reliable whole genome multiple sequence alignments remains a need in comparative genomics studies especially when dealing with non-coding regions and distantly related species (Chen & Tompa, 2010). It is suggested that further development of tools is required for exact whole genome alignment and for the identification of

regulatory elements in the non-coding regions involved in controlling the gene expression for providing more accurate results.

Analyses: Structural and Functional

Following genome sequencing and alignment, structural and functional analysis play an vital role in understanding genome by integrating insights from structural data (such as locations of genes, mutations and traits on chromosome) with functional data (like gene expression profiles) after which functional prediction is carried out. The structure and the function of the genome complement each other in extracting complete information of the genome. Toos like rVista and eShadow identify conserved regions in the genome and therefore it is able to relate the architecture of the genome to its function (Loots & Ovcharenko, 2005). Comparative genomics also identifies synteny which reveals gene clusters. Tools like BLASTClust, MCL and OrthoMCL are used to group the sequences into clusters, families and evolutionary relationships which are interpreted using universal protein families and tools like PhyML. Lastly, gene clusters are studied for conservation across species by linking them to metabolic pathways (Alam *et al*., 2007). Therefore, by analysing similarities and differences in genome structure, and studying them with the help of different databases, gene functions and pathways are compared for interpretation. In a nutshell, structural genomics deals in providing a framework of the genome architecture by genome sequencing, mapping and structural analysis. While, functional analyses uncover the roles of genes in biological processes which is achieved through techniques such as gene expression profiling, RNA sequencing and analysing functions through gene knockout methods.

Phylogenetic and Evolutionary Studies

Construction of phylogenetic trees based on the similarities identified during sequence alignment is the next step adopted in comparative genomics study which are constructed on the basis of evolutionary relationships. These trees are studied to observe the evolution and the accumulation of mutations over time. The shape of a phylogenetic tree can depict different evolutionary processes, facilitated by robust statistical methods for decoding information merely from its shape (Mooers & Heard, 1997). Reviewing genomes of species or organisms at various phylogenetic distances, which are represented on these trees can resolve many questions. In recent times, due to the improvement in sequencing technologies, generation of genomic data has witnessed increment. That being said, it is very possible for the occurrence of errors when constructing a phylogenetic tree, which can lead researchers to draw false conclusions. Choosing an appropriate model for constructing a tree which mitigates errors is important, otherwise the availability of genomic data stands redundant. Phylogeny is inferred by different methods like the distance based methods (UPGMA & Neighbour Joining), and the character based methods (maximum parsimony, maximum likelihood and the Bayesian inference) (Kapli *et al*., 2020).The Bayesian method counters these uncertainties by giving certain probabilities to each potential tree, the calculations of which are derived from a likelihood function and a prior probability distribution. The Bayesian method takes into consideration the probabilities of all possible trees instead of one which could be incorrect (Huelsenbeck *et al*., 2000). It works on the Markov Chain Monte Carlo (MCMC) algorithm. However, the choice of a specific tree building method depends on the type of study, quality of results needed, the extent of inference expected, and other aspects such as time defined by the researcher. For maintaining accuracy, the tree building method chosen can be followed by bootstrap analysis to check for its reliability.

Data visualization, interpretation and validation

The information of the genome is encoded in its sequence, whose long- and short-range interactions amongst each other actually determines its function. Needless to say, in nature, several biological processes are interconnected which leads to the generation of genomic data that are multifaceted (Nusrat *et al*., 2019). The interpretation of this complex genomic data is aided by visualization tools which can visually represent complex genomic data in formats that are easy to be interpreted by researchers (Siang *et al*., 2024). Specified tools are available to carry out different tasks. Some of them are genome browsers, circular and space filling layout tools and matrix-based platforms. Interpretation of the genomic datasets can be specified by the user in some tools which carries out the analyses accordingly (Nusrat *et al*., 2019). Strudel is one such genomic data visualising application which supports the interactive comparison of large data sets expanding both genetic and physical maps (Bayer *et al*., 2011). Pathline is another interactive visualization tool which represents the phylogenetic relationships and biological pathways in a number of species. It provides linearization of metabolic pathways, resulting in greater accuracy in topological comparison. It also allows the comparison of time series data through its curvemap view (Meyer *et al*., 2010).

Due to the vast size of ever-increasing genomic data, individually determining the expression of genes and thereafter the function of the proteins is impractical. Computational methods are therefore applied to annotate the genome. Comparative genomics allows the prediction of protein function on the principle of 'homology-based function prediction', which involves the identification of homologous proteins with an experimentally characterized function to assign annotation. Methods implying genome context analysis, or assessing properties of proteins such as psychochemical, interaction, localization are alternatives to homology based function prediction (A. Maghawry *et al*., 2015; Gabaldón, 2008). However, protein function prediction using structure based computational methods provide higher accuracy (A. Maghawry *et al*., 2015). Databases like PDB and CATH are integrated along with different tools like CE, CASTp, BioLip, DALI etc to

perform this. Combining the genomic data with other multi-omics data is advantageous in obtaining a much comprehensive interpretation of the function of the genome. This system biological understanding serves real life applications in different fields (Suravajhala *et al*., 2016).

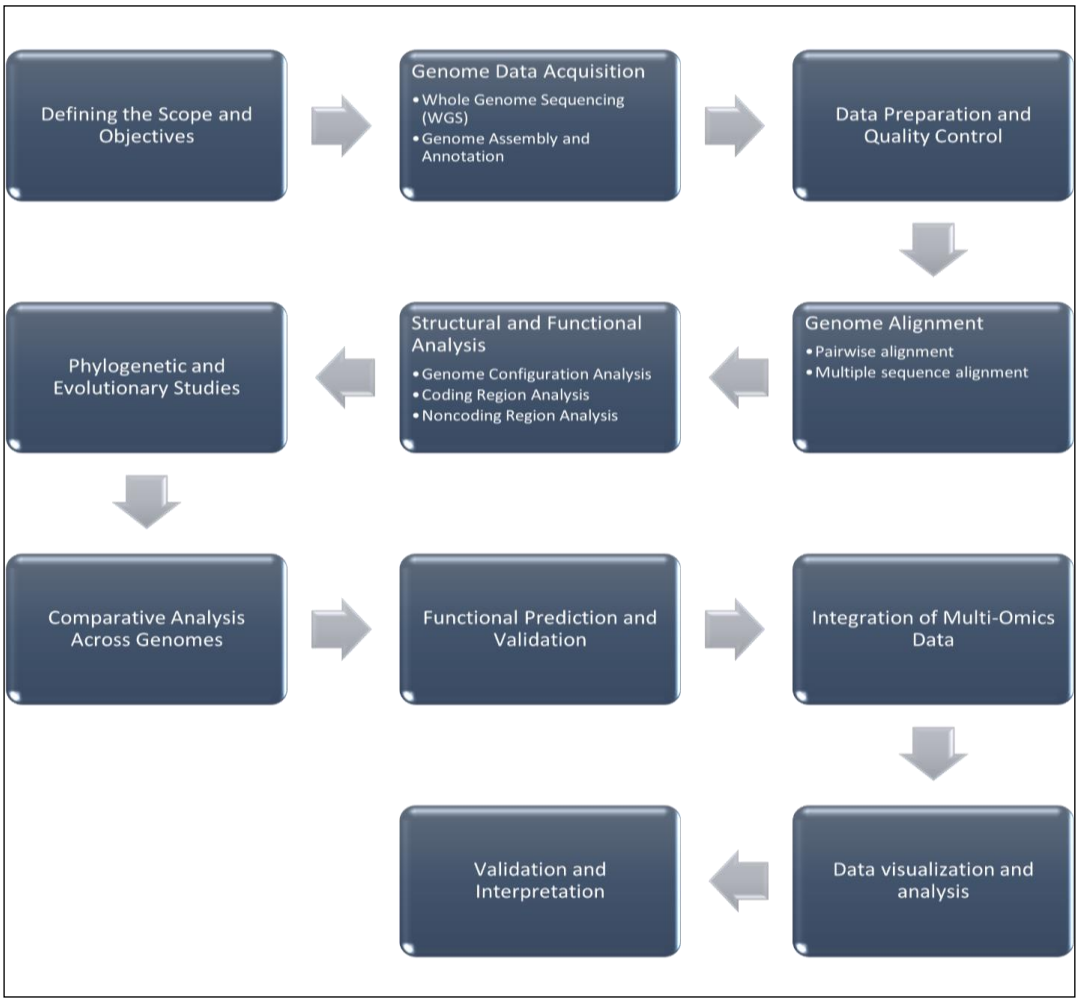

Figure 2. Steps and methodology involved in a comparative genomics study.

Table 1. List of some commonly used tools in Comparative Genomics.

S. no.	Name	URLs	Descriptions	Reference
Whole Genome Alignment				
1.	MUMmer4	https://github.com/mummer4/mummer	It is a quick and simple way to align lengthy DNA sequences. It is capable of handling partial genomes, huge genome assemblies, complete genomes, or a collection of genome reads.	(Marçais *et al*., 2018)
2.	Mauve	http://gel.ahabs.wisc.edu/mauve	It is a software that aligns two or more genomes with occurrences rearrangements within it.	(Darling *et al*., 2010)

3.	Mugsy	http://mugsy.sf.net.	multiple alignments of whole genomes without requiring a reference genome.	(Angiuoli & Salzberg, 2011)
4.	MUM&Co	https://github.com/SAMtoBAM/MUMandCo	MUM&Co is a computational tool used for whole genome alignment with a focus on identifying maximal unique matches (MUMs).	(O'Donnell & Fischer, 2020)
5.	Smash++	https://github.com/smortezah/smashpp/tree/master/experiment/dataset	The tool is designed for fast whole-genome alignment, detecting unique subsequences, efficiently handling both closely and distantly related genomes, and incorporating advanced indexing techniques for efficient memory usage.	(Hosseini *et al*., 2020)
Genome Assembly				
1.	SPAdes	https://github.com/ablab/spades	SPAdes allows for the assembly and analysis of sequenced data. It is mainly tailored for handling Illumina sequencing data.	(Bankevich *et al*., 2012)
2.	Velvet	https://github.com/dzerbino/velvet	It is mainly applicable for assembly of de novo genome, particularly arising from Illumina technology.	(Zerbino & Birney, 2008)
3.	CANU	https://github.com/marbl/canu	Evolved from Celera Assembler, this tool performs de novo genome assembly for long read data	(Koren *et al*., 2017)
4.	ABySS	http://www.bcgsc.ca/platform/bioinfo/software/abyss	It helps in efficiently assembling large-scale data from sequencing projects. It is designed for assembling large genomes using a distributed computing approach.	(Simpson *et al*., 2009)
5.	MEGAHIT v1.0	https://hku-bal.github.io/megabox	It is a highly efficient and scalable metagenome assembler, suitable for large and complex sequencing datasets, short reads, and whole genome assembly of single organisms.	(D. Li *et al*., 2016)
Genome Annotation Tools				
1.	PROKKA	http://vicbioinformatics.com/	It is a software tool which annotates prokaryotic genome and can generate files that can be viewed in genome browsers for further evaluation.	(Seemann, 2014)
2.	AUGUSTUS	http://augustus.gobics.de	It is a software which can predict genes in eukaryotes and some prokaryotes. Its functions on the Generalized Hidden Markov Model. It can predict multiple splice variants and is the primary ab initio gene finder to do so. The software also offers motif searching for user-defined regular expressions.	(Stanke *et al*., 2006)
3.	RAST	http://rast.nmpdr.org/	It is an automatic annotation server for microbial genomes, built on the SEED system. RAST consistently produces annotations comparable to human annotators and extends them to as many protein-encoding genes as possible.	(Overbeek *et al*., 2014)
4.	InterProScan	https://github.com/ebi-pf-team/interproscan	It is used for automatic annotation of protein sequences and genome analysis, providing reliable characterisation of sequences for functional annotation.	(Biswas, 2002)
5.	BLAST2GO	http://www.blast2go.de	It is a tool used for Functional annotation of genomes based on GO terms and BLAST results. It integrates similarity	(Conesa *et al*., 2005)

			searches, statistical analysis along with visualization which are run on acyclic graphs, making it suitable for functional genomics research.	
Alignment (Pairwise and Multiple Sequence)				
1	BLAST	https://ftp.ncbi.nlm.nih.gov/blast/db/	It is a widely used tool for comparison of nucleotide bases or protein sequences to user defined databases, offering fast and efficient pairwise alignments.	(Altschul *et al*., 1990)
2.	ClustalW (Pairwise)	http://www.bii.a-star.edu.sg/software/clustalw-mpi/	It is a tool that can be used for pairwise alignments. It functions on progressive alignment algorithm.	(K.-B. Li, 2003)
3.	Clustal Omega	https://www.ebi.ac.uk/jdispatcher/msa/clustalo	It is a multiple sequence alignment (MSA) program. It is an upgrade of the previous Clustal programs.	(Sievers & Higgins, 2014)
4.	MUSCLE	https://www.ebi.ac.uk/jdispatcher/msa/muscle?stype=protein	It is a MSA tool known for its accuracy and speed, often recommended for aligning large sequence datasets.	(Edgar, 2004)
5.	MAFFT	https://mafft.cbrc.jp/alignment/software/	It is a fast and highly accurate tool for aligning multiple sequences using several algorithms to handle large numbers of sequences.	(Katoh & Standley, 2013)
Phylogenetic Analysis				
1.	MEGA	https://www.megasoftware.net/	It is a tool for constructing phylogenetic trees using methods like Neighbor-Joining, Maximum Likelihood, and Maximum Parsimony.	(Kumar *et al*., 1994)
2.	RAxML	https://github.com/amkozlov/raxml-ng	It is a high-performance tool for Maximum Likelihood-based phylogenetic inference. It handles large datasets efficiently and supports a variety of substitution models.	(Rokas, 2011)
3.	PhyML	http://atgc.lirmm.fr/phyml	It is a web interface to that implements a fast and accurate prediction of phylogeny of organism via maximum parsimony method.	(Guindon *et al*., 2005)
4.	BEAST2	https://www.beast2.org/	It is a cross-platform program. It reconstructs phylogenies and tests evolutionary hypotheses without a single tree topology, using Markov chain Monte Carlo. Utilized for Bayesian phylogenetic analysis.	(Bouckaert *et al*., 2014)
5.	IQ-TREE 2.2.0	http://www.iqtree.org/	It is a phylogenetic inference software that has been enhanced with new features which supports DNA, protein, codon sequences, binary and morphological data, and supports partitioned and mixed models.	(Minh *et al*., 2020)
Data visualization tools				
1.	VISTA	https://genome.lbl.gov/vista/index.shtml	Curve based genome browser that supports the interactive comparison of genomic data, allowing users to identify conserved regions and evaluate genomic data across species.	(Frazer *et al*., 2004)
2.	GenoFig	https://forgemia.inra.fr/public-pgba/genofig/-/tree/main	Application that facilities the visualization of prokaryotic genomic data	(Branger & Leclercq, 2024)
3.	OrthoVenn 3	https://orthovenn3.bioinfotoolkits.net/	Platform containing Venn diagrams to visualize, identify and annotate	(Sun *et al*., 2023)

			orthologous clusters.	
4.	BactoGeNIE	https://www.evl.uic.edu/research/2038	Large scale comparative genome analysis and visualization tool for bacterial genome.	(Aurisano *et al*., 2015)
5.	PhylomeDB	https://phylomedb.org/	Public database which allows the visualization of phylogenetic trees along with the evolutionary history of genes	(Fuentes *et al*., 2022)
Data integration tools				
1.	MicrobesOnline	http://www.microbesonline.org/	Website based tool serving as a portal of comparative and function genomic integration for prokaryotes.	(Alm *et al*., 2005)
2.	JCVI	https://www.jcvi.org/research/software-tools	Versatile Python based library containing a range of tools integrating genome assembly, annotation and comparative analysis across species	(Tang *et al*., 2024)
3.	MOSAIC	https://arxiv.org/abs/1309.2319	Integrates methodologically diverse algorithms which improves detection of orthologs.	(Maher & Hernandez, n.d.)
4.	mixOmics	https://mixomics.org/	Its is a package which supports the interlinking and exploration of different types of 'omics' data sets.	(Rohart *et al*., 2017)
5.	Galaxy	https://galaxyproject.org/	Platform for data integration and visualization across various stages of genomic analysis.	(Goecks *et al*., 2010)

For successful interpretation of genomic study, both pre and post genomic data analysis validation is crucial. Pre analysis validation of genomic data is achieved in quality assessment of genomic assemblies. The post analysis data validation crosschecks the functional predictions and the inference of the genomic study across species by examining the elements across known conserved regions. Statistical methods such as the bootstrapping methods are also employed to validate derived conclusions. Sometimes validation encounters misinterpretation of genomic studies.

CONCLUSION

In this technological era, where tremendous amount of genome sequence data is generated daily and is made available on the public databases, it has become easier for researchers to use the obtained data for comparative genomics approach. Further, the approach is also applied to newly sequenced genome for its exploration in various fields such as discovery of phylogenetic history, proteomics, and drug discovery. Hence, this handy approach should not be skipped to understand the language of genomic data which is an important tool for reading the blue print of life residing in the genome of every organism. To overcome the problem associated with the alignment of the whole genome of organisms, it is suggested that further development of tools is required for robust comparison of non-coding regions for the identification of regulatory elements involved in controlling the gene expression of the organism. It is anticipated that it may help in providing significant leads for identification of virulent genes related to diseases and ailments.

ACKNOWLEDGMENT

The authors are thankful to their respective departments for constant motivation and moral support in undertaking the present review work.

CONFLICT OF INTERESTS

The authors declare no conflict of interest

ETHICS APPROVAL

Not applicable

REFERENCES

A. Maghawry, H., G. M. Mostafa, M., H. Abdul-Aziz, M., & F. Gharib, T. (2015). Structural Protein Function Prediction- A Comprehensive Review. *International Journal of Modern Education and Computer Science*, 7(10), 49–57. https://doi.org/10.5815/ijmecs.2015.10.07.

Alam, I., Cornell, M., Soanes, D. M., Hedeler, C., Wong, H. M., Rattray, M., Hubbard, S. J., Talbot, N. J., Oliver, S. G., & Paton, N. W. (2007). A Methodology for Comparative Functional Genomics. *Journal of Integrative Bioinformatics*, 4(3), 112–122. https://doi.org/10.1515/jib-2007-69.

Ali, A. (2013). Microbial Comparative Genomics: An Overview of Tools and Insights Into The Genus Corynebacterium. *Journal of Bacteriology & Parasitology*, 04(02). https://doi.org/10.4172/2155-9597.1000167.

Alm, E. J., Huang, K. H., Price, M. N., Koche, R. P., Keller, K., Dubchak, I. L., & Arkin, A. P. (2005). The MicrobesOnline Web site for comparative genomics. *Genome Research*, 15(7), 1015–1022. https://doi.org/10.1101/gr.3844805.

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., & Lipman, D. J. (1990). Basic local alignment search tool. *Journal of Molecular Biology*, 215(3), 403–410. https://doi.org/10.1016/S0022-2836(05)80360-2.

Angiuoli, S. V., Gussman, A., Klimke, W., Cochrane, G., Field, D., Garrity, G. M., Kodira, C. D., Kyrpides, N., Madupu, R., Markowitz, V., Tatusova, T., Thomson, N., & White, O. (2008). Toward an Online Repository of Standard Operating Procedures (SOPs) for (Meta)genomic Annotation. OMICS: *A Journal of Integrative Biology*, 12(2), 137–141. https://doi.org/10.1089/omi.2008.0017.

Angiuoli, S. V., & Salzberg, S. L. (2011). Mugsy: Fast multiple alignment of closely related whole genomes. Bioinformatics, 27(3), 334–342. https://doi.org/10.1093/bioinformatics/btq665.

Armstrong, J., Fiddes, I. T., Diekhans, M., & Paten, B. (2019). Whole-Genome Alignment and Comparative Annotation. *Annual Review of Animal Biosciences*, 7(1), 41–64. https://doi.org/10.1146/annurev-animal-020518-115005.

Aurisano, J., Reda, K., Johnson, A., Marai, E. G., & Leigh, J. (2015). BactoGeNIE: A large-scale comparative genome visualization for big displays. BMC *Bioinformatics*, 16(S11), S6. https://doi.org/10.1186/1471-2105-16-S11-S6.

Bankevich, A., Nurk, S., Antipov, D., Gurevich, A. A., Dvorkin, M., Kulikov, A. S., Lesin, V. M., Nikolenko, S. I., Pham, S., Prjibelski, A. D., Pyshkin, A. V., Sirotkin, A. V., Vyahhi, N., Tesler, G., Alekseyev, M. A., & Pevzner, P. A. (2012). SPAdes: A New Genome Assembly Algorithm and Its Applications to Single-Cell Sequencing. *Journal of Computational Biology*, 19(5), 455–477. https://doi.org/10.1089/cmb.2012.0021.

Bayer, M., Milne, I., Stephen, G., Shaw, P., Cardle, L., Wright, F., & Marshall, D. (2011). Comparative visualization of genetic and physical maps with Strudel. *Bioinformatics*, 27(9), 1307–1308. https://doi.org/10.1093/bioinformatics/btr111.

Benjak, A., Sala, C., & Hartkoorn, R. C. (2015). Whole-Genome Sequencing for Comparative Genomics and De Novo Genome Assembly. In T. Parish & D. M. Roberts (Eds.), Mycobacteria Protocols (Vol. 1285, pp. 1–16). Springer New York. https://doi.org/10.1007/978-1-4939-2450-9_1.

Berriman, M., & Harris, M. (2004). Annotation of Parasite Genomes. In S. E. Melville, Parasite Genomics Protocols (Vol. 270, pp. 017–044). Humana Press. https://doi.org/10.1385/1-59259-793-9:017.

Biswas, M. (2002). Applications of InterPro in protein annotation and genome analysis. Briefings in Bioinformatics, 3(3), 285–295. https://doi.org/10.1093/bib/3.3.285.

Bornstein, K., Gryan, G., Chang, E. S., Marchler-Bauer, A., & Schneider, V. A. (2023). The NIH Comparative Genomics Resource: Addressing the promises and challenges of comparative genomics on human health. BMC *Genomics*, 24(1), 575. https://doi.org/10.1186/s12864-023-09643-4.

Bouckaert, R., Heled, J., Kühnert, D., Vaughan, T., Wu, C.-H., Xie, D., Suchard, M. A., Rambaut, A., & Drummond, A. J. (2014). BEAST 2: A Software Platform for Bayesian Evolutionary Analysis. PLoS Computational Biology, 10(4), e1003537. https://doi.org/10.1371/journal.pcbi.1003537.

Branger, M., & Leclercq, S. O. (2024). GenoFig: A user-friendly application for the visualization and comparison of genomic regions. Bioinformatics, 40(6), btae372. https://doi.org/10.1093/bioinformatics/btae372.

Carlos Guimaraes, L., Benevides De Jesus, L., Vinicius Canario Viana, M., Silva, A., Thiago Juca Ramos, R., De Castro Soares, S., & Azevedo, V. (2015). Inside the Pan-genome—Methods and Software Overview. Current Genomics, 16(4), 245–252. https://doi.org/10.2174/1389202916666150423002311.

Chain, P. (2003). An applications-focused review of comparative genomics tools: Capabilities, limitations and future challenges. *Briefings in Bioinformatics*, 4(2), 105–123. https://doi.org/10.1093/bib/4.2.105.

Charamis, J., Balaska, S., Ioannidis, P., Dvořák, V., Mavridis, K., McDowell, M. A., Pavlidis, P., Feyereisen, R., Volf, P., & Vontas, J. (2024). Comparative Genomics Uncovers the Evolutionary Dynamics of Detoxification and Insecticide Target Genes Across 11 Phlebotomine Sand Flies. *Genome Biology and Evolution*, 16(9), evae186. https://doi.org/10.1093/gbe/evae186.

Chen, X., & Tompa, M. (2010). Comparative assessment of methods for aligning multiple genome sequences. Nature Biotechnology, 28(6), 567–572. https://doi.org/10.1038/nbt.1637.

Cheng, A. Y., Teo, Y.-Y., & Ong, R. T.-H. (2014). Assessing single nucleotide variant detection and genotype calling on whole-genome sequenced individuals. Bioinformatics, 30(12), 1707–1713. https://doi.org/10.1093/bioinformatics/btu067.

Clark, M. S. (1999). Comparative genomics: The key to understanding the human genome project. BioEssays, 21(2), 121–130. https://doi.org/10.1002/(SICI)1521-1878(199902)21:2<121::AID-BIES6>3.0.CO;2-O.

Commins, J., Toft, C., & Fares, M. A. (2009). Computational Biology Methods and Their Application to the Comparative Genomics of Endocellular Symbiotic Bacteria of Insects. Biological Procedures Online, 11(1), 52. https://doi.org/10.1007/s12575-009-9004-1.

Conesa, A., Götz, S., García-Gómez, J. M., Terol, J., Talón, M., & Robles, M. (2005). Blast2GO: A universal tool for annotation, visualization and analysis in functional genomics research. Bioinformatics, 21(18), 3674–3676. https://doi.org/10.1093/bioinformatics/bti610.

Darling, A. E., Mau, B., & Perna, N. T. (2010). progressive Mauve: Multiple Genome Alignment with Gene Gain,

Loss and Rearrangement. PLoS ONE, 5(6), e11147. https://doi.org/10.1371/journal.pone.0011147.

Dewey, C. N. (2019). Whole-Genome Alignment. In M. Anisimova (Ed.), Evolutionary Genomics (Vol. 1910, pp. 121–147). Springer New York. https://doi.org/10.1007/978-1-4939-9074-0_4.

Edgar, R. C. (2004). MUSCLE: Multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Research, 32(5), 1792–1797. https://doi.org/10.1093/nar/gkh340.

Fiddes, I. T., Armstrong, J., Diekhans, M., Nachtweide, S., Underwood, J. G., Gordon, D., Earl, D., Keane, T., Eichler, E. E., Haussler, D., Stanke, M., & Paten, B. (n.d.). Comparative Annotation Toolkit (CAT)—Simultaneous clade and personal genome annotation.

Frazer, K. A., Pachter, L., Poliakov, A., Rubin, E. M., & Dubchak, I. (2004). VISTA: Computational tools for comparative genomics. Nucleic Acids Research, 32(Web Server), W273–W279. https://doi.org/10.1093/nar/gkh458.

Fuentes, D., Molina, M., Chorostecki, U., Capella-Gutiérrez, S., Marcet-Houben, M., & Gabaldón, T. (2022). PhylomeDB V5: An expanding repository for genome-wide catalogues of annotated gene phylogenies. Nucleic Acids Research, 50(D1), D1062–D1068. https://doi.org/10.1093/nar/gkab966.

Gabaldón, T. (2008). Comparative Genomics-Based Prediction of Protein Function. In M. Starkey & R. Elaswarapu (Eds.), Genomics Protocols (Vol. 439, pp. 387–401). Humana Press. https://doi.org/10.1007/978-1-59745-188-8_26.

Goecks, J., Nekrutenko, A., Taylor, J., & Galaxy Team. (2010). Galaxy: A comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. Genome Biology, 11(8), R86. https://doi.org/10.1186/gb-2010-11-8-r86.

Guindon, S., Lethiec, F., Duroux, P., & Gascuel, O. (2005). PHYML Online A web server for fast maximum likelihood-based phylogenetic inference. Nucleic Acids Research, 33(Web Server), W557–W559. https://doi.org/10.1093/nar/gki352.

Gurevich, A., Saveliev, V., Vyahhi, N., & Tesler, G. (2013). QUAST: Quality assessment tool for genome assemblies. Bioinformatics, 29(8), 1072–1075. https://doi.org/10.1093/bioinformatics/btt086.

Hardison, R. C. (2003). Comparative Genomics. PLoS Biology, 1(2), e58. https://doi.org/10.1371/journal.pbio.0000058

Haubold, B., & Wiehe, T. (2004). Comparative genomics: Methods and applications. Naturwissenschaften, 91(9). https://doi.org/10.1007/s00114-004-0542-8

Herrero, J., Muffato, M., Beal, K., Fitzgerald, S., Gordon, L., Pignatelli, M., Vilella, A. J., Searle, S. M. J., Amode, R., Brent, S., Spooner, W., Kulesha, E., Yates, A., & Flicek, P. (n.d.). Ensembl comparative genomics resources.

Hosseini, M., Pratas, D., Morgenstern, B., & Pinho, A. J. (2020). Smash++: An alignment-free and memory-efficient tool to find genomic rearrangements. GigaScience, 9(5), giaa048. https://doi.org/10.1093/gigascience/giaa048

Huelsenbeck, J. P., Rannala, B., & Masly, J. P. (2000). Accommodating Phylogenetic Uncertainty in Evolutionary Studies. Science, 288(5475), 2349–2350. https://doi.org/10.1126/science.288.5475.2349

Hwang, K.-B., Lee, I.-H., Li, H., Won, D.-G., Hernandez-Ferrer, C., Negron, J. A., & Kong, S. W. (2019). Comparative analysis of whole-genome sequencing pipelines to minimize false negative findings. Scientific Reports, 9(1), 3219. https://doi.org/10.1038/s41598-019-39108-2

Kapli, P., Yang, Z., & Telford, M. J. (2020). Phylogenetic tree building in the genomic age. Nature Reviews Genetics, 21(7), 428–444. https://doi.org/10.1038/s41576-020-0233-0

Katoh, K., & Standley, D. M. (2013). MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability. Molecular Biology and Evolution, 30(4), 772–780. https://doi.org/10.1093/molbev/mst010

Kellis, M., Patterson, N., Endrizzi, M., Birren, B., & Lander, E. S. (2003). Sequencing and comparison of yeast species to identify genes and regulatory elements. Nature, 423(6937), 241–254. https://doi.org/10.1038/nature01644

Khan, M. I., Khan, Z. A., Baig, M. H., Ahmad, I., Farouk, A.-E., Song, Y. G., & Dong, J.-J. (2020). Comparative genome analysis of novel coronavirus (SARS-CoV-2) from different geographical locations and the effect of mutations on major target proteins: An in-silico insight. PLOS ONE, 15(9), e0238344. https://doi.org/10.1371/journal.pone.0238344

Kircher, M., & Kelso, J. (2010). High-throughput DNA sequencing – concepts and limitations. BioEssays, 32(6), 524–536. https://doi.org/10.1002/bies.200900181

Kobras, C. M., Fenton, A. K., & Sheppard, S. K. (2021). Next-generation microbiology: From comparative genomics to gene function. Genome Biology, 22(1), 123. https://doi.org/10.1186/s13059-021-02344-9

Koonin, E. V. (2010). The origin and early evolution of eukaryotes in the light of phylogenomics. Genome Biology, 11(5), 209. https://doi.org/10.1186/gb-2010-11-5-209

Koren, S., Walenz, B. P., Berlin, K., Miller, J. R., Bergman, N. H., & Phillippy, A. M. (2017). Canu: Scalable and accurate long-read assembly via adaptive k -mer weighting and repeat separation. Genome Research, 27(5), 722–736. https://doi.org/10.1101/gr.215087.116

Kumar, S., Tamura, K., & Nei, M. (1994). MEGA: Molecular Evolutionary Genetics Analysis software for microcomputers. Bioinformatics, 10(2), 189–191. https://doi.org/10.1093/bioinformatics/10.2.189

Lehri, B., Seddon, A. M., & Karlyshev, A. V. (2017). The hidden perils of read mapping as a quality assessment tool in genome sequencing. Scientific Reports, 7(1), 43149. https://doi.org/10.1038/srep43149

Lemmon, E. M., & Lemmon, A. R. (2013). High-Throughput Genomic Data in Systematics and Phylogenetics. Annual Review of Ecology, Evolution, and Systematics, 44(1), 99–121. https://doi.org/10.1146/annurev-ecolsys-110512-135822

Li, D., Luo, R., Liu, C.-M., Leung, C.-M., Ting, H.-F., Sadakane, K., Yamashita, H., & Lam, T.-W. (2016). MEGAHIT v1.0: A fast and scalable metagenome assembler driven by advanced methodologies and community practices. Methods, 102, 3–11. https://doi.org/10.1016/j.ymeth.2016.02.020.

Li, K.-B. (2003). ClustalW-MPI: ClustalW analysis using distributed and parallel computing. Bioinformatics, 19(12), 1585–1586. https://doi.org/10.1093/bioinformatics/btg192.

Lin, J., & Qian, J. (2007). Systems biology approach to integrative comparative genomics. Expert Review of Proteomics, 4(1), 107–119. https://doi.org/10.1586/14789450.4.1.107.

Loots, G. G., & Ovcharenko, I. (2005). Dcode.org anthology of comparative genomic tools. Nucleic Acids Research, 33(Web Server), W56–W64. https://doi.org/10.1093/nar/gki355

Maher, M. C., & Hernandez, R. D. (n.d.). A MOSAIC of methods: Improving ortholog detection through the integration of algorithmic diversity.

Marçais, G., Delcher, A. L., Phillippy, A. M., Coston, R., Salzberg, S. L., & Zimin, A. (2018). MUMmer4: A fast and versatile genome alignment system. PLOS Computational Biology, 14(1), e1005944. https://doi.org/10.1371/journal.pcbi.1005944.

Meyer, M., Wong, B., Styczynski, M., Munzner, T., & Pfister, H. (2010). Pathline: A Tool For Comparative Functional Genomics. Computer Graphics Forum, 29(3), 1043–1052. https://doi.org/10.1111/j.1467-8659.2009.01710.x.

Miller, W., Makova, K. D., Nekrutenko, A., & Hardison, R. C. (2004). COMPARATIVE GENOMICS. Annual Review of Genomics and Human Genetics, 5(Volume 5, 2004), 15–56. https://doi.org/10.1146/annurev.genom.5.061903.180057.

Minh, B. Q., Schmidt, H. A., Chernomor, O., Schrempf, D., Woodhams, M. D., Von Haeseler, A., & Lanfear, R. (2020). IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era. Molecular Biology and Evolution, 37(5), 1530–1534. https://doi.org/10.1093/molbev/msaa015.

Mooers, A. O., & Heard, S. B. (1997). Inferring Evolutionary Process from Phylogenetic Tree Shape. The Quarterly Review of Biology, 72(1), 31–54. https://doi.org/10.1086/419657.

Morrison, A. C., Huang, Z., Yu, B., Metcalf, G., Liu, X., Ballantyne, C., Coresh, J., Yu, F., Muzny, D., Feofanova, E., Rustagi, N., Gibbs, R., & Boerwinkle, E. (2017). Practical Approaches for Whole-Genome Sequence Analysis of Heart- and Blood-Related Traits. The American Journal of Human Genetics, 100(2), 205–215. https://doi.org/10.1016/j.ajhg.2016.12.009.

Mulhair, P. O., & Holland, P. W. H. (2024). Evolution of the insect Hox gene cluster: Comparative analysis across 243 species. Seminars in Cell & Developmental Biology, 152–153, 4–15. https://doi.org/10.1016/j.semcdb.2022.11.010.

Nakagawa, H., & Fujita, M. (2018). Whole genome sequencing analysis for cancer genomics and precision medicine. Cancer Science, 109(3), 513–522. https://doi.org/10.1111/cas.13505.

Nobrega, M. A., & Pennacchio, L. A. (2004). Comparative genomic analysis as a tool for biological discovery. The Journal of Physiology, 554(1), 31–39. https://doi.org/10.1113/jphysiol.2003.050948.

Nusrat, S., Harbig, T., & Gehlenborg, N. (2019). Tasks, Techniques, and Tools for Genomic Data Visualization. Computer Graphics Forum, 38(3), 781–805. https://doi.org/10.1111/cgf.13727.

O'Donnell, S., & Fischer, G. (2020). MUM&Co: Accurate detection of all SV types through whole-genome alignment. Bioinformatics, 36(10), 3242–3243. https://doi.org/10.1093/bioinformatics/btaa115.

Overbeek, R., Olson, R., Pusch, G. D., Olsen, G. J., Davis, J. J., Disz, T., Edwards, R. A., Gerdes, S., Parrello, B., Shukla, M., Vonstein, V., Wattam, A. R., Xia, F., & Stevens, R. (2014). The SEED and the Rapid Annotation of microbial genomes using Subsystems Technology (RAST). Nucleic Acids Research, 42(D1), D206–D214. https://doi.org/10.1093/nar/gkt1226.

Palmer, J. D. (1985). Comparative organization of chloroplast genomes. Annual Review of Genetics, 19, 325–354. https://doi.org/10.1146/annurev.ge.19.120185.001545.

Ptacek, T. (2005). A tiered approach to comparative genomics. Briefings in Functional Genomics and Proteomics, 4(2), 178–185. https://doi.org/10.1093/bfgp/4.2.178.

Rohart, F., Gautier, B., Singh, A., & Lê Cao, K.-A. (2017). mixOmics: An R package for 'omics feature selection and multiple data integration. PLoS Computational Biology, 13(11), e1005752. https://doi.org/10.1371/journal.pcbi.1005752.

Rokas, A. (2011). Phylogenetic Analysis of Protein Sequence Data Using the Randomized Axelerated Maximum Likelihood (RAXML) Program. Current Protocols in Molecular Biology, 96(1). https://doi.org/10.1002/0471142727.mb1911s96.

Sankoff, D., & Nadeau, J. H. (2000). Comparative Genomics. In D. Sankoff & J. H. Nadeau (Eds.), Comparative Genomics (Vol. 1, pp. 3–7). Springer Netherlands. https://doi.org/10.1007/978-94-011-4309-7_1.

Schmieder, R., Lim, Y. W., Rohwer, F., & Edwards, R. (2010). TagCleaner: Identification and removal of tag sequences from genomic and metagenomic datasets. BMC Bioinformatics, 11(1), 341. https://doi.org/10.1186/1471-2105-11-341.

Seemann, T. (2014). Prokka: Rapid prokaryotic genome annotation. Bioinformatics, 30(14), 2068–2069. https://doi.org/10.1093/bioinformatics/btu153.

Siang, C. V., Mohamed, F., Salleh, F. M., Iglesias, A., & Abdul-Rahman, A. (2024). A Systematic Literature Review of Comparative Visualization Designs and Techniques for Comparative Genomics Analysis. Open Science Framework. https://doi.org/10.31219/osf.io/k9zjg.

Sievers, F., & Higgins, D. G. (2014). Clustal Omega. Current Protocols in Bioinformatics, 48(1). https://doi.org/10.1002/0471250953.bi0313s48.

Sikhakhane, T. N., Figlan, S., Mwadzingeni, L., Ortiz, R., & Tsilo, T. J. (2016). Integration of Next-generation Sequencing Technologies with Comparative Genomics in Cereals. In I. Y. Abdurakhmonov (Ed.), Plant Genomics. InTech. https://doi.org/10.5772/61763.

Simpson, J. T., Wong, K., Jackman, S. D., Schein, J. E., Jones, S. J. M., & Birol, İ. (2009). ABySS: A parallel assembler for short read sequence data. Genome Research, 19(6), 1117–1123. https://doi.org/10.1101/gr.089532.108.

Sivashankari, S., & Shanmughavel, P. (2007). Comparative genomics—A perspective.

Smith, C. L., Oh, H., & Stamenović, D. (2012). Comparative Genomics Approaches and Technologies. In R. A. Meyers (Ed.), Encyclopedia of Analytical Chemistry (1st ed.). Wiley. https://doi.org/10.1002/9780470027318.a1406.pub2.

Stanke, M., Keller, O., Gunduz, I., Hayes, A., Waack, S., & Morgenstern, B. (2006). AUGUSTUS: Ab initio prediction of alternative transcripts. Nucleic Acids Research, 34(Web Server), W435–W439. https://doi.org/10.1093/nar/gkl200.

Sun, J., Lu, F., Luo, Y., Bie, L., Xu, L., & Wang, Y. (2023). OrthoVenn3: An integrated platform for exploring and visualizing orthologous data across genomes. Nucleic Acids Research, 51(W1), W397–W403. https://doi.org/10.1093/nar/gkad313.

Suravajhala, P., Kogelman, L. J. A., & Kadarmideen, H. N. (2016). Multi-omic data integration and analysis using systems genomics approaches: Methods and applications in animal production, health and welfare. Genetics Selection Evolution, 48(1), 38. https://doi.org/10.1186/s12711-016-0217-x.

Tang, H., Krishnakumar, V., Zeng, X., Xu, Z., Taranto, A., Lomas, J. S., Zhang, Y., Huang, Y., Wang, Y., Yim, W. C., Zhang, J., & Zhang, X. (2024). JCVI: A versatile toolkit for comparative genomics analysis. iMeta, 3(4), e211. https://doi.org/10.1002/imt2.211.

Trudsø, L. C., Andersen, J. D., Jacobsen, S. B., Christiansen, S. L., Congost-Teixidor, C., Kampmann, M.-L., & Morling, N. (2020). A comparative study of single nucleotide variant detection performance using three massively parallel sequencing methods. PLOS ONE, 15(9), e0239850. https://doi.org/10.1371/journal.pone.0239850.

Wattam, A. R., Brettin, T., Davis, J. J., Gerdes, S., Kenyon, R., Machi, D., Mao, C., Olson, R., Overbeek, R., Pusch, G. D., Shukla, M. P., Stevens, R., Vonstein, V., Warren, A., Xia, F., & Yoo, H. (2018). Assembly, Annotation, and Comparative Genomics in PATRIC, the All Bacterial Bioinformatics Resource Center. In J. C. Setubal, J. Stoye, & P. F. Stadler (Eds.), Comparative Genomics (Vol. 1704, pp. 79–101). Springer New York. https://doi.org/10.1007/978-1-4939-7463-4_4.

Wei, L., Liu, Y., Dubchak, I., Shon, J., & Park, J. (2002). Comparative genomics approaches to study organism similarities and differences. *Journal of Biomedical Informatics*, *35*(2), 142–150. https://doi.org/10.1016./S1532-0464(02)00506-3.

Williams, D. M., & Forey, P. L. (2004). *Milestones in Systematics*. CRC Press.

Zerbino, D. R., & Birney, E. (2008). Velvet: Algorithms for de novo short read assembly using de Bruijn graphs. *Genome Research*, *18*(5), 821–829. https://doi.org/10.1101/gr.074492.107.