Received: 14 May 2024 | Revised: 18 September 2024 | Accepted: 9 December 2024 | Published online: 24 December 2024

REVIEW

Medinformatics
2024, Vol. 00(00) 1–12
DOI: 10.47852/bonviewMEDIN42023423

BON VIEW PUBLISHING

# Unlocking Nature's Code: The Power of Pan-Genomics in Biological Entities

**Rachana Reghu**[1], **Ramya Latha Karunakaran**[1], **Ranjani Balaji**[1], **Achsha Babu**[1] and **Arun Arumugaperumal**[1,*]

[1]*Department of Biotechnology, Rajalakshmi Engineering College, India*

**Abstract:** Pan-genomics, a holistic approach to genomic analysis, has become a powerful tool with diverse applications in biomedical and environmental sciences. This offers a valuable understanding of genetic variation, adaptation, and evolution highlighting the presence of genes associated with specific traits like disease resistance and the ability to thrive in different environments. The integration of pan-genomics and metagenomics helps understanding the microbial diversity and functional characteristics of microbial communities. In crop improvement, pan-genomics accelerates genetic enhancement through tools like the Rice Pan-genome Genotyping Array, which leverages wild species for improved crop traits. Additionally, pan-genomics is crucial in vaccine development, with efforts focused on creating universal vaccines against Klebsiella and Streptococcus infections. In drug development, it supports the identification of drug targets, proved by research previously focused on *Clostridium botulinum* and *Yersinia pseudotuberculosis*. Furthermore, pan-genomics contributes significantly to public health by revealing unique genomic features in pathogens like *Escherichia coli* and *Mycobacterium tuberculosis*, shedding light on virulence mechanisms and potential therapeutic targets. This exploration underscores the wide-ranging impact and relevance of pan-genomics across multiple scientific domains.

**Keywords:** environment, medicine, microorganisms, pan-genomics, sequencing

## 1. Introduction

Pan-genomics is a transformative approach in genomics that extends beyond the study of individual genomes to encompass the full genetic diversity within a species or group of microorganisms. This approach recognizes that each genome contains unique variations, including single-nucleotide polymorphisms (SNPs) and structural differences. Advancements in sequencing technologies have enabled researchers to efficiently explore these genetic landscapes, providing a comprehensive understanding of a species' genetic makeup, including both shared and unique genes. This broad approach reveals how genetic variations influence traits and offers broad implications across various fields, including medicine, agriculture, and environmental science. The study of microbial genetic diversity has been greatly facilitated by recent developments in sequencing technology. Pan-genomics involves analyzing the collective genes of microbial communities (the pan-genome) and understanding how these genes vary and evolve. Concurrently, the genetic material of these communities can be analyzed directly from their environmental contexts (metagenomes). In pan-genomics, "open" and "closed" pan-genomes describe the genetic diversity within a species. A closed pan-genome has a finite gene pool, with few new genes discovered as more genomes are sequenced, indicating limited genetic diversity in stable environments. In contrast, an open pan-genome features a continuously expanding gene repertoire, reflecting high genetic diversity and frequent gene acquisition through horizontal gene transfer or adaptation to changing environments. Understanding these concepts is essential for studying a species' evolutionary dynamics and adaptability. Pan-genomics provides valuable insights into genetic diversity and evolutionary processes, with significant applications in plant and animal breeding. For instance, in plant improvement, pan-genomics captures the genetic variability essential for understanding species evolution, domestication, and crop enhancement, even though challenges such as high computational costs and data integration persist [1]. A study on Lycopersicon the tomato clad utilized pan-genome analysis to identify structural variants (SVs) linked to important traits such as yield and flavor, aiding in crop breeding by highlighting genetic diversity [2]. Similarly, graph-based pan-genomics addresses missing heritability issues by uncovering structural variations not detected by traditional genome-wide association studies (GWAS). This advancement enhances breeding strategies for traits such as stress tolerance and flavor in maize crops [3]. Furthermore, pan-genome analysis has improved genome editing in maize by revealing chromosomal pattern that hinder recombination, allowing the CRISPR-Cas9 system to target and reinvert these regions to access beneficial genes [4]. Additionally, in the dairy industry pan-genomics has significantly advanced our understanding of *Lactobacillus delbrueckii* subsp. bulgaricus. By sequencing multiple strains of this bacterium, researchers have identified core genes essential for survival and accessory genes responsible for strain-specific traits. Key findings include the loss of certain metabolic pathways as *L. bulgaricus* adapted to milk, as

*__Corresponding author:__ Arun Arumugaperumal, Department of Biotechnology, Rajalakshmi Engineering College, India. Emails: arunbiotechnologist@gmail.com; arun.a@rajalakshmi.edu.in

well as genes crucial for lactose metabolism, lactic acid production, and exopolysaccharide production, all of which affect yogurt fermentation, taste, and texture. Additionally, stress resistance genes have been identified that enhance survival during fermentation and in the gastrointestinal tract. These insights are instrumental in developing *L. bulgaricus* strains with improved fermentation characteristics and product quality [5].

## 2. Combined Aspect of Pan-Genomics and Metagenomics Approaches

The integration of metagenomics and pan-genomics represents a sophisticated approach to understanding the genetic diversity within microbial communities. This technique involves expanding distinct gene clusters from various microbial genomes to encompass diverse ambient microbial communities. One effective strategy is to identify gene clusters from genomes isolated from their environments or from existing databases, infer relationships between pan-genomes, and use metagenomic data to monitor the frequency and abundance of individual genes. For instance, this approach has been applied to identify genomic fragments of Latescibacteria and Bacillus species [6]. An alternative method for reconstructing pan-genomes across populations involves binning co-abundant genes. This reference-free approach allows for the identification of a community's core and variable genes, providing insights into the genetic diversity and population dynamics of environmental microorganisms. A pan-genome study revealed that species exist either as an open or a closed pan-genome, reflecting their genetic diversity and its adaptableness. Species with an open pan-genome, such as Streptococci, Meningococci, Salmonellae, *Helicobacter pylori*, and *Escherichia coli*, typically colonize diverse environments and possess mechanisms for frequent genetic exchange. This genetic flexibility enables them to continually acquire new genes, contributing to an expanding pan-genome. Conversely, species with a closed pan-genome, including *Bacillus anthracis, Mycobacterium tuberculosis (*Mtb), and *Chlamydia trachomatis*, are more genetically stable and inhabit more in isolated environments with limited gene exchange. Their reduced capacity for horizontal gene transfer results in a more conserved and restricted pan-genome [7]. An extreme case of genome stability was observed in *Buchnera aphidicola*, an endosymbiont of aphids, which has undergone no significant genomic rearrangements or gene transfer over the past 50 million years, representing the most stable pan-genome documented to date [8].The concept of the pan-genome was introduced to define the complete set of genes within a bacterial species, including both the core genome (genes shared by all strains) and the accessory genome (genes found in only some strains). This distinction between open and closed pan-genomes has been used to analyze pathogenic bacteria such as *Streptococcus agalactiae, Escherichia coli, Pseudomonas aeruginosa,* Mtb, and *Vibrio cholerae*. Such analyses reveal genes related to virulence, antibiotic resistance, and environmental adaptation, contributing to a deeper understanding of bacterial evolution and aiding in the development of targeted vaccines and treatments, thereby enhancing public health strategies for combating infections and reducing resistance [9]. Table 1 interprets the integrated pan-genomics with metagenomics studies enabling the identification of novel genomic diversity and relationship within microbial communities. The overall pan-genome analysis is demonstrated in Figure 1 [10].

**Table 1. Summarizes the pan-genome approaches to microbial communities**

| Microorganism | Study method | Study year | Key features | Drawbacks | References |
|---|---|---|---|---|---|
| *Methanobrevibacter smithii* | By forming syntrophic interactions with other bacterial species, *M. smithii* strains leverage the variability of adhesin-like proteins repertoires to produce changes in their metabolic environments | 2011 | Shows metabolic plasticity through adhesin-like proteins | Limited to interactions with specific microbial species | [11] |
| *Escherichia coli* | To determine the microbial profiles at the strain level in complex *E. coli* communities | 2016 | High resolution of strain-level differences in microbial communities | Complexity of analyzing diverse strain populations | [12] |
| *Bacillus* | There are notable variations in the function and genomic fingerprints of core and strain-specific genes among Bacillus species | 2017 | Highlights strain-specific functional diversity in Bacillus species | Genomic variations make generalization difficult | [13] |
| *Bacteroides* | In a prospective birth cohort of 44 newborns and mothers, bacterial strains among different families were examined. A high-resolution mother-to-child bacterial transfer events with newborn gut colonization were observed | 2018 | Detailed tracking of bacterial transfer between mother and child | Limited scope and requires larger cohorts to confirm findings | [14] |
| *Escherichia coli* | The substantial genetic diversity of *E. coli* isolates and gut microbial populations was shown by pan-genome networks derived from 760 human gut microbiomes and five pathogenic strains of the bacteria | 2018 | Reveals significant genetic diversity in human gut microbiomes | High computational requirements to analyze large datasets | [15] |

(*Continued*)

**Table 1.** (*Continued*)

| Microorganism | Study method | Study year | Key features | Drawbacks | References |
|---|---|---|---|---|---|
| *Prochlorococcus* | Meta pan-genome analysis showed core genes were highly associated with sugar metabolism found in genomic islands of Prochlorococcus species | 2018 | Identified sugar metabolism pathways crucial to core gene functions | Genomic islands are not always fully understood in terms of ecological impact | [16] |
| *Proteobacteria* | Binning co-abundant genes across metagenomic samples especially Proteobacteria revealed gene repertoires of 1661 microbial species are effectively delineated in a matter of few hours, exhibiting comparable specificity and increased sensitivity. | 2019 | Fast and efficient identification of gene repertoires across microbial species | Potential loss of resolution for rare microbial species | [17] |
| *Aeromonas* | In 29 Aeromonas significant degree of genome plasticity with hundreds of gene expansion, and horizontal gene transfers were identified. Numerous virulence factors were found in microbial isolates | 2019 | Highlights high genome plasticity and horizontal gene transfer | Can lead to increased virulence, challenging for clinical applications | [18] |
| *Ruminococcus*, *Alistipes*, *Eubacterium*. | Reassembly of 92,143 genomes from 11,850 human gut microbiome uncovers 1,952 uncultured bacteria. These result in 281% increase in phylogenetic diversity | 2019 | Substantially increases understanding of human gut microbial diversity | High number of uncultured bacteria can complicate functional studies | [19] |
| *Bacteroides* | The study discovered 34,205 genomes with existence of several taxa from under-sampled phyla including Saccharibacteria and Elusimicrobia in gut and oral microbiome | 2019 | Uncovers underrepresented bacterial phyla in human microbiomes | Limited knowledge about the role of newly discovered bacteria | [20] |
| *Haemophilus parainfluenzae*, *Rothia* | The bacteria showed common genomic content from tongue dorsum, supragingival plaque, and buccal mucosa. For the tongue subgroup, an operon encoding oxaloacetate decarboxylase was shown to be indicative | 2020 | Provides insights into shared bacterial functions across oral environments | Focuses on specific subpopulations, potentially missing less abundant bacterial functions | [21] |
| *Lactobacillus gasseri*, *Gardnerella vaginalis* | Although *Gardnerella vaginalis* has hundreds of metagenomes accessible, this significant species has open-meta-genome, whereas *Lactobacillus gasseri* has nearly closed meta-genome network | 2020 | Highlights genomic openness and closure in different bacterial species | Open-meta genomes make it difficult to predict functional traits | [22] |
| *Verminephrobacter eiseniae* | Their symbiotic relationship supports embryogenesis and normal homeostasis thereby focusing on the development and overall fitness of the host healthcare system | 2020 | Key role in maintaining host health and development through symbiosis | Limited application beyond the host-specific relationship | [23] |
| *Pseudomonas parafulva OS-1* | *Pseudomonas parafulva OS-1* encodes glycoside hydrolases and involves in lignocellulose breakdown suggesting the strain possesses biomass degradation potential. | 2023 | Potential for biomass degradation and biofuel production | Limited to specific enzymes involved in lignocellulose breakdown | [24] |
| *Bacillus paralicheniformis* | The bacteria was found to be involved in bacitracin, and bacillibactin synthesis. The strain BP9 also reduces spore germination. | 2023 | Potential for antibiotic and secondary metabolite production | Strain-specific insights, not easily generalized | [25] |
| *Corynebacterium striatum* | The pan-genome revealed 2070-1899 core gene clusters. A total 115 phage species were identified with 44 intact prophage regions. | 2023 | Phage identification provides insights into bacterial defense mechanisms | Phage regions may complicate genomic analysis due to horizontal gene transfer | [26] |

(*Continued*)

**Table 1.** (*Continued*)

| Microorganism | Study method | Study year | Key features | Drawbacks | References |
|---|---|---|---|---|---|
| *Rhodococcus* | A gene from genus Rhodococcus was engaged in the benzoate pathway degradation. | 2024 | Provides functional insights into pollutant degradation | Limited application outside of benzoate pathway studies | [27] |
| *Synechococcus elongatus* | Circular episomal elements and CRISPR-associated protein genes were the unique features that were found | 2024 | Offers potential for CRISPR applications and episomal genetic studies | Episomal elements may have limited functionality depending on environmental conditions | [28] |
| *Xanthomonas fragariae* | The presence of *XopG1, CE8, XopD,* and *GT2* virulence strains plays vital role in early infectious process of Chinese strains | 2024 | Identifies key virulence factors for early infection stages | Limited to specific strains, may not be generalizable to all infections | [29] |
| *Cobetia* | It was determined that the Cobetia *L2A1* strain belongs to an entirely novel species. In addition, *C. pacifica* should be a part of *C. marina*, and *C. litoralis* should be included in the *C. amphilecti* species | 2024 | Novel species identification and taxonomic reclassification | Taxonomic reclassification can lead to confusion in existing literature | [30] |
| *Bacillus* | Probiotic, antibiotic, and secondary metabolite enrichment were observed in *Bacillus licheniformis, Bacillus paralicheniformis,* and *Bacillus spizizenii.* | 2024 | Offers potential for health applications and secondary metabolite production in probiotics | Requires further validation for specific probiotic health benefits | [31] |

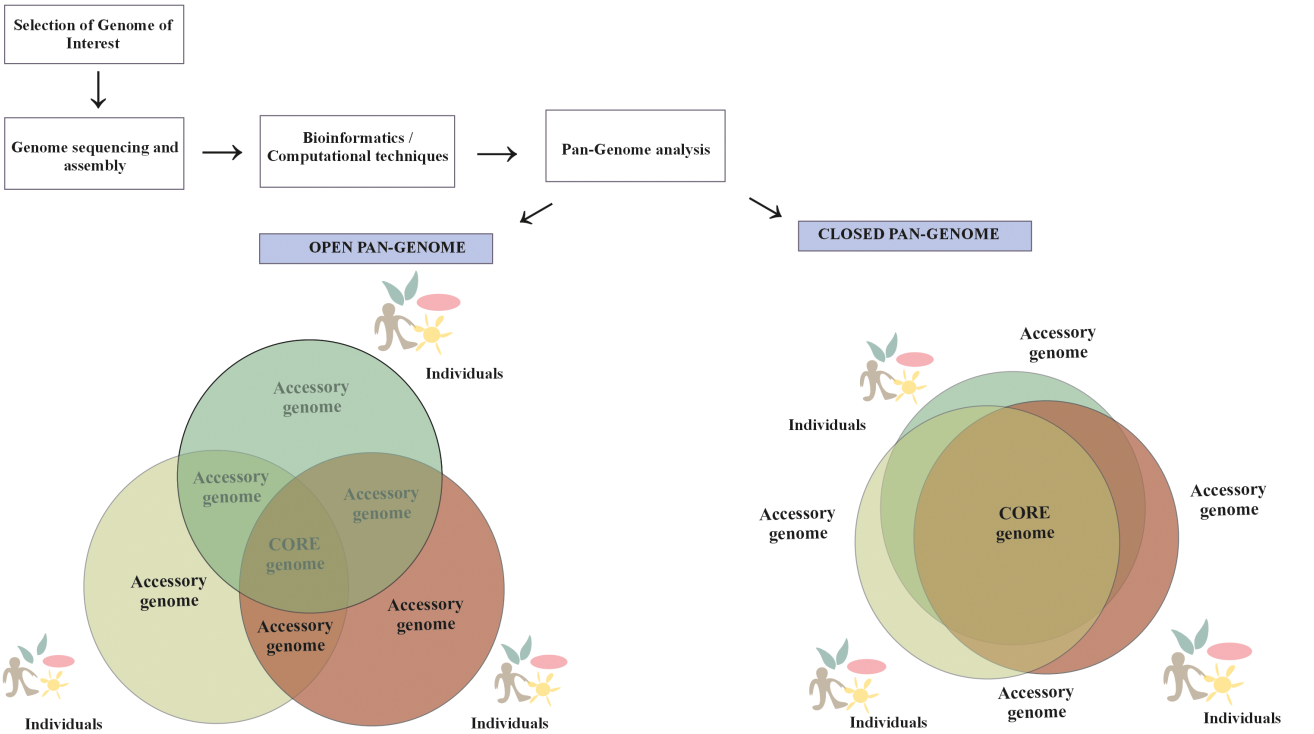

**Figure 1. Overview of pan-genome analysis. The process involves selecting genomes of interest, followed by genome sequencing and assembly. This is then followed by computational methods for pan-genome analysis. The analysis determines whether the pan-genome is open or closed, with individuals represented as follows: human in brown, plants in green, and microbial communities in yellow and red color**

## 3. Human Pan-Genome

In 2010, the first human pan-genome study compared the genomes of an Asian and an African individual to the reference genome, uncovering millions of base pairs of novel sequences specific to populations or individuals [32]. This pioneering work demonstrated the significant genetic diversity that exists beyond the reference genome, highlighting the need for a comprehensive human pan-genome. Subsequent studies have reinforced the importance of developing a human pan-genome. For instance, a 2015 study generated a Danish pan-genome from the genomes of 10 trios, identifying hundreds of thousands of novel single-nucleotide variants and insertions or deletions [33]. In 2017, GenomeDenmark sequenced and assembled the genomes of 50 family trios, establishing a population-specific reference genome and further emphasizing the value of including diverse genomic data [34]. A study expanded the scope by sequencing 910 individuals of African descent and creating an African pan-genome. This research identified 300 megabases of novel sequences unique to African populations, revealing that the African pan-genome contains approximately 10% more DNA than the GRCh38 human reference genome [35]. Another study further advanced the field by analyzing 338 high-quality genome assemblies from genetically diverse human populations. Their analysis uncovered more than 400,000 previously unmapped reads from genomes with approximately 40× coverage, successfully aligning them with a reference sequence using the Human Diversity Reference [36]. This work underscored the progress in capturing and integrating genomic diversity into reference sequences, advancing our understanding of human genetic variation.

## 4. Plant Pan-Genome

It took approximately a decade for scientists to elucidate the pan-genomes of plants following the initial studies of bacterial pan-genomes. This delay was due in part to the high cost of data collection and the assumption that gene presence and absence would exhibit minimal variation among higher organisms like plants, which do not exchange genes as freely as bacteria. The term “pan-genome” was first applied to plants in 2007, when researchers identified small, variable regions in the genomes of rice and maize. However, at that time, the understanding of gene presence or absence was limited due to the lack of detailed genome maps for multiple individuals within the same plant species. With advancements in technology and the reduction in DNA sequencing costs, scientists developed three primary methods to compare whole genomes within plant species. The first method involves assembling and comparing the complete genomes of different individuals within a species, allowing for the identification of variations in gene presence and absence. The second approach involves aligning individual DNA sequences with a reference genome, identifying discrepancies, and adding these variations to a growing database of genetic differences. The most recent technique involves constructing maps that illustrate the diversity of plant genomes, which facilitates the identification of genes present or absent across various individuals. Table 2

**Table 2. Summary of the pan-genome studies conducted in different plant species**

| Plant species | Study year | Genome construction approach | Core genome | Accessory/dispensable genome | References |
|---|---|---|---|---|---|
| *Oryza sativa* | 2014 | De novo assembly | 37,200 | 3162 | [37] |
| *Brassica oleracea* | 2016 | Iterative mapping and assembly | 49,895 | 11,484 | [38] |
| *Brachypodium distachyon* | 2017 | De novo assembly | 13,408 | 17,195 | [39] |
| *Medicago truncatula* | 2017 | De novo assembly | 24,000 | 50,700 | [40] |
| *Oryza sativa* | 2018 | De novo assembly | 12,770 | 9050 | [41] |
| *Oryza sativa* and *Oryza rufipogon* | 2018 | De novo assembly | 26372 | 16,208 | [42] |
| *Sesamum indicum* | 2019 | De novo assembly | 15,409 | 15,890 | [43] |
| *Solanum lycopersicum* L. | 2019 | Presence/absence variation analysis | 29,938 | 1287 | [44] |
| *Helianthus annuus* L. | 2019 | Targeted de novo assembly | 32,917 | 2464 | [45] |
| *Glycine max* | 2020 | Graph-based de novo assembly | 20,623 | 28,679 | [46] |
| *Cucumismelo L.* | 2022 | PAV-GWAS analysis For resistance gene analogs | 603 | 106 | [47] |
| *Delftia tsuruhatensis* | 2022 | Whole-genome sequencing | 4045 | 9856 | [48] |
| Soybean | 2023 | Hi-C sequencing | 708 | 3797 | [49] |
| *Solanum tuberosum L.* | 2023 | De novo sequencing | 23,055 | 1,09,300 | [50] |
| *Setaria* | 2023 | Illumina sequencing | 17,499 | 56,027 | [51] |
| *Lactiplantibacillus plantarum* | 2024 | Roary analysis | 1436 | 15,475 | [52] |
| *J. nigra, Juglans regia, Juglans microcarpa, Castanea mollissima,* and *Carya cathayensis* | 2024 | Illumina, Pacbio, and Hi-C technologies | 11,759 | 13,019 | [53] |
| *Arabidopsis thaliana* | 2024 | Whole-genome sequencing | 19,721 | 7195 | [54] |
| *Vigna unguiculata L.* | 2024 | DNA sequencing and de novo assembly | 26,026 | 35,436 | [55] |

reviews the advancements in pan-genome analysis across diverse plant species, highlighting the progress made in understanding plant genetic diversity and its implications.

## 5. Pan-Genomics in Crop Improvement

### 5.1. Unlocking rice genetic diversity: The role of the rice pan-genome genotyping array (RPGA) in crop improvement and resilience

The RPGA is a tool developed to study the entire genetic variation present in different types of rice plants. This genetic variation contains unique genes that contribute to the diversity seen in rice crops. By studying the genetic variations in different types of rice plants, we can identify specific genes responsible for desirable traits, such as larger grain size, better taste, or resistance to pests [56]. This information is crucial for breeding rice varieties that meet the preferences and needs of farmers and consumers. The RPGA helps to analyze 80,000 specific genetic differences in various types of rice. These differences could be like tiny changes in the DNA code called SNPs or differences in whether certain genes are present or absent. The RPGA addresses a limitation seen in traditional methods that rely on studying only one specific type of rice as a reference. This single-reference approach can introduce bias, as it might not capture the full genetic diversity present in different rice varieties. The RPGA tackles this issue by examining a wide range of rice types, ensuring a more comprehensive understanding of genetic variations. This will help to see a clearer and more complete picture of what makes different types of rice special when it comes to their genetics [57]. This tool will also help scientists map out specific genes connected to important features of rice. The RPGA also helps figure out how different types of rice plants are related, leading to the creation of resilient rice varieties. Pan-genomics also enables the identification of novel genes or gene variants that can improve crop traits. Thus, by incorporating diverse genetic information from pan-genomes, crop breeding programs can develop improved and more resilient varieties with desirable characteristics.

### 5.2. Unraveling genetic traits of cultivated rice (*O. sativa*) and wild type (*O. rufipogon*): Implications for domestication, modern breeding, and climate adaptation

Research emphasized the rich genetic diversity in cultivated rice (*O. sativa*) and its wild relative *O. rufipogon* and its implications for domestication, modern breeding, and adaptation to ecological and agronomic conditions. The pan-genome analysis of 66 divergent accessions of *O. sativa* and *O. rufipogon* species revealed 23 million sequence variants including many quantitative trait nucleotides. Both domestication and contemporary breeding have benefited greatly from the enormous genetic diversity of both cultivated and wild rice. This will eventually meet the future demands for food and genetic advancement brought on by climate change. A significant variation among rice accessions would advance the evolutionary and functional aspects of rice. The research also delved into genetic improvement by identifying functionally diverse alleles and presence or absence of variation of coding genes. The authors conducted whole-genome de novo assemblies of 66 rice accessions and identified 16,563,789 SNPs, 5,549,290 small insertions with deletions (indels), and 933,489 SVs [58]. A study determined potential introgression events and identified SNPs with highly differentiated alleles between indica and temperate japonica, implying introgression from indica into tropical japonica [59]. The rich genetic diversity in rice, as revealed by the pan-genome dataset, is anticipated to be useful in GWAS, genetic mapping of complex traits, and functional genomics studies. These studies provide valuable insights into the genetic makeup of rice, its potential for genetic improvement, and its significance in addressing the challenges posed by global climate changes and increasing food demand.

## 6. Pan-Genomics in Vaccine Development

### 6.1. Developing a universal vaccine against dangerous Klebsiella pathogens through the application of pan-genomics

Klebsiella bacteria, often found in hospitals, are a common cause of infections. Among them, *Klebsiella pneumoniae* is especially important, causing various infections like urinary tract infections, pneumonia, and blood infections acquired during hospital stays [60]. These bacteria mainly spread through the gut and the hands of hospital staffs, leading to outbreaks in healthcare settings due to their rapid transmission ability. Additionally, Klebsiella is known for its resistance to many antibiotics, making treatment challenging [61]. These bacteria are on the World Health Organization's critical list for new therapeutics due to their association with high mortality rates and limited treatment options. Employing the principles of pan-genomics, a broad-spectrum vaccine capable of targeting all Klebsiella species was designed. Pan-genomics aids in the comprehensive exploration of the Klebsiella genus, allowing the identification of common genetic features shared among different types of bacteria. These shared elements are then used as a basis for picking out important proteins that could be targeted to create a vaccine that works against multiple species of Klebsiella. Two promising vaccine candidates were identified through an in-depth analysis of the core proteins shared among the bacteria. The two proteins, TonB-dependent siderophore receptor and siderophore enterobactin receptor FepA, play crucial roles in vaccine development. These proteins fulfill essential vaccine parameters, making them promising candidates for an effective vaccine against Klebsiella. Thus, the application of pan-genomics in vaccine development lies in the identification of conserved genes through pan-genome analysis [62]. This approach allows for the creation of broad-spectrum vaccines targeting shared elements in the core genome, effectively addressing the genetic diversity of pathogenic bacterial strains and enhancing vaccine efficacy.

### 6.2. Development and evaluation of recombinant vaccines against *Streptococcus agalactiae* in tilapia: Insights from genomic diversity, adhesin characterization, and cross-protective immunity

A study focused on the development of a recombinant vaccine against *Streptococcus agalactiae* infection in tilapia. The study performed pan-genome analysis and investigated the genetic diversity of *S. agalactiae* precisely in piscine strains [63]. A study found that tilapia from Honduras and marine fish and rays from Queensland showed distinct host-adapted lineages of *S. agalactiae*. Additionally, they characterized the ST-261 adhesins and determined their expression levels in different conditions [64]. The researchers identified the adhesins and found that their expression was affected by culture conditions. They investigated if adhesins were able to elicit cross-protective immune responses to different *S. agalactiae* serotypes in fish models. They found that the recombinant ST-261 adhesins were not as effective as an injectable vaccine against Streptococcus in tilapia. This research highlights the importance of understanding the genomic diversity of *S. agalactiae* and characterizing adhesins for the development of effective vaccines against Streptococcal infections in tilapia. However, the recombinant ST-261 adhesins tested in this study were not found to be protective against infection. Further research is needed to explore other potential vaccine candidates [65].

## 7. Pan-Genomics in Drug Development

### 7.1. Utilizing pan-genome analysis and docking techniques for drug target identification against *Clostridium botulinum*: A computational approach to anti-bacterial drug development

*Clostridium botulinum* a spore-forming bacteria produces strong neurotoxins which can act as bioweapons. A pan-genome analysis was carried out to find possible therapeutic targets of *C. botulinum*. Seven proteins were selected as targets and docked against 105 anti-bacterial compounds. From each protein, the top 10 compounds that had the highest interactions with binding affinity were preferred. Additionally, the study identified compounds that showed high binding energies with multiple proteins, resulting in a total of 39 compounds. In vitro analysis could be carried out for validation of these compounds and may continue for clinical trials. The study highlights the effectiveness of computational approaches in identifying targets for drug development against disease-causing pathogens. While no specific details about individual proteins or compounds are provided, the study emphasizes the potential of using pan-genome analysis and docking techniques to identify promising drug targets [66].

### 7.2. Pan-genomics for screening natural product inhibitors against *Yersinia pseudotuberculosis*

Pan-genome openness is valuable in bioprospecting for enzymes, bioactive compounds, or genes with biotechnological applications. This bacterium can spread globally and produce both sporadic and pandemic infections, exclusively in colder conditions [67]. It possesses a virulent piB plasmid with a single circular chromosome. Virtual screening, a swift method for identifying inhibitors against pathogens, has demonstrated its utility in isolating beneficial compounds from extensive collections of antimicrobial agents across various pathogens in the past [68]. In a study, they used a virtual screening method with special emphasis on phytochemical inhibitors against *Yersinia pseudotuberculosis*, instead of other drugs. This study involved genomic data from 23 genomes to identify essential proteins and drug targets. The multidrug resistance protein mdtC was selected for analysis; however, its function in Yersinia was indefinite. Computational modeling and screening against phytochemical libraries were performed. The results show that the open pan-genome of *Y. pseudotuberculosis* indicates a sympatric lifestyle. There are 85 druggable targets in the core genome of *Y. pseudotuberculosis*. The top-hit phytochemicals did not show any mutagenicity or cytochrome p450 enzyme inhibition over ADMET analysis. By targeting the efflux pump component of the bacterium, which was identified through gene virulence analysis and shown to have no relationship with the human host or gut microbiota, the study provides valuable insights for selective control of pathogens. Moreover, the identification of top-scoring compounds, including Nilotinin D3 and Punicalaginin, which are recognized as generally safe, underscores their potential as effective treatments for the *Y. pseudotuberculosis* infections [69]. These findings highlight the importance of integrating pan-genome analysis with traditional medicine research to identify novel therapeutic targets and develop effective treatments against bacterial pathogens.

## 8. Pan-Genomics in Public Health

### 8.1. Pan-genome analysis of *Escherichia coli* strains and identification of tailocin specific for pandemic ST131

Disease outbreaks related to foodborne contamination and gastrointestinal infections have been long-term associated with *E. coli* and the gut microbiome. In recent times, certain strains of *E. coli*, notably the ST131 variants, have rapidly emerged as widespread and pandemic, exhibiting resistance to multiple drugs [70]. It has been shown that A, B1, B2, C, D, E, F, and G are the major phylogroups of *E.coli* [71]. The study provided valuable insights into the genetic composition and functional features of *E.coli* particularly in B1, B2, and E phylogroups. This comprehensive analysis provides a robust framework for understanding the evolutionary dynamics and biological features of *E. coli* strains, which have significant implications for public health. By systematically analyzing 1324 complete genomes, the study reveals that the softcore genome, comprising approximately 3000 gene families, is invariant in different strains, irrespective of the total number of genomes. The most important and evolutionarily conserved classes of gene families are represented by softcore genome which provides critical insights into the genetic makeup of the species. For example, the unsupervised clustering of *E.coli* phylogroups in the existence or absence of gene families uncovers the distinct bi-lineage nature of phylogroup B1, the variation of secretion and iron acquisition systems in ST11 (E), and the integration of a highly conserved prophage into the ST131 genome (B2). Notably, the study identifies a tailocin specific to ST131 strains, which is hypothesized to play a crucial role in inter-bacterial competition and contribute to the upsurge of the *E. coli* ST131 pandemic [72]. These findings have significant implications for public health applications. Firstly, considering the genetic diversity and functional traits of *E. coli* strains can help in the development of effective diagnostic tools and targeted therapies for combating pathogenic diseases. By identifying specific genetic markers associated with virulence or antibiotic resistance, healthcare professionals can better assess the risk posed by different strains and tailor treatment strategies accordingly. Furthermore, insights into the evolutionary dynamics and population structure of *E. coli* can inform epidemiological surveillance efforts. By tracking the prevalence and distribution of different phylogroups and genetic variants, public health authorities can anticipate emerging threats, monitor the spread of antibiotic resistance, and implement timely interventions to prevent outbreaks and mitigate the impact of infectious diseases caused by *E. coli*. The study underscores the importance of pan-genome analysis in elucidating the genetic basis of microbial diversity and its implications for public health. By providing comprehensive datasets and insights into subtypes, this research helps to develop strategies for disease prevention, surveillance, and control.

### 8.2. Pan-genomics insights into Mycobacterium tuberculosis: Uncovering core genes, super core genes, and strain-specific variations

Since 2014, tuberculosis (TB) has overtaken HIV as the world's most infectious disease. A pan-genome study revealed the genetic composition of Mtb. The study revealed 3,679 core genes as primary genes, secondary genes, and strain-specific genes from Mtb strains. Additionally, there were 1,122 dispensable genes and 964 strain-specific genes, contributing to the individual characteristics of Mtb strains. Furthermore, the study focused on identifying "Super Core Genes" (SCGs) that are crucial in Mtb pathogenicity. 28 SCGs encode antigens, PE/PPE proteins, transposes, and virulence factors, all of which are known to be crucial for Mtb's pathogenicity. The research also unveiled 48 specific single-core genes that differentiate Mtb from *Mycobacterium bovis*, another related bacterium. In short, the pan-genome study offers a significant understanding of the genetic foundation and development of Mtb strains, shedding light on the different gene categories and their roles in Mtb's pathogenicity and diversity [73].

## 9. Future Direction of Pan-Genomics

The future of pan-genomics is poised for groundbreaking advancements, driven by the evolution of sequencing technologies and computational methodologies. As we progress into the era of pan-genomics, the current one-dimensional approach, focused solely on sequence analysis, is transitioning towards a three-dimensional paradigm. This shift involves not only capturing all sequence variations within a species or genus but also encoding their spatial organization and interrelationships. With emerging technologies facilitating three-dimensional pan-genome inference, we anticipate a revolution in how we perceive and analyze genetic data. The pan-genomics will see the convergence of genomic and spatial data, offering a wide view of genetic diversity and its structural implications. This multidimensional approach will open avenues for exploring intricate biological phenomena such as chromatin organization, genome architecture, and functional genomics within a spatial context. Moreover, it will enable us to unravel complex regulatory networks and understand the dynamic nature of genetic interactions within and between genomes [74]. In this trajectory, computational pan-genomics will play a pivotal role in developing novel algorithms and tools tailored for three-dimensional data analysis. The integration of machine learning, artificial intelligence, and spatial genomics techniques will empower researchers to extract meaningful insights from vast and complex pan-genomic datasets. Collaborations across disciplines, including biological science, computer science, and engineering, will drive innovation and propel pan-genomics into a new era of comprehensive genome analysis and interpretation.

## 10. Pan-Genome Challenges and Solutions

Pan-genome studies present several challenges related to the characteristics of Big Data, including the vast volume of data generated and the heterogeneity from various sequencing technologies. The need for rapid processing speeds, ensuring data accuracy and reliability, and leveraging cloud computing for scalable storage and processing further complicate the analysis. Additionally, visualizing complex relationships and homology across multiple genomes poses significant challenges, requiring advanced mathematical and graphical techniques. Key challenges in pan-genomics include homology analysis, software development for data analysis, and the management of large-scale datasets. Addressing these challenges requires collaborative efforts, advancements in computational methodologies, and the development of innovative tools specifically tailored for pan-genome analysis and visualization. Developing advanced algorithms, utilizing cloud computing, and creating innovative visualization tools are crucial steps in overcoming these obstacles and enhancing pan-genome analysis capabilities [74].

## 11. Conclusion

In conclusion, pan-genomics represents a powerful and versatile toolset for genomic analysis, with extensive implications across various fields of study. By integrating with metagenomics, pan-genomics provides invaluable insights into microbial communities and their interactions with the environment. In crop improvement, it facilitates the identification of genetic variations crucial for enhancing plant resilience and productivity. In the realm of public health, pan-genomics enables precise tracking of pathogen evolution and outbreak monitoring, contributing to more effective responses to infectious diseases. Its role in vaccine development promises the creation of tailored solutions against emerging pathogens, while in drug development, it accelerates the identification of novel therapeutic targets and compounds. Embracing pan-genomics fosters a comprehensive understanding of genetic diversity and adaptation, revolutionizing approaches in both biomedical and environmental sciences for the advancement of global health and sustainability.

This review article is structured as follows: the introduction provides a comprehensive overview of pan-genomics, categorizing it into open and closed pan-genome frameworks while highlighting its practical applications. The subsequent section discusses the integration of metagenomics and pan-genomics, elucidating their roles in understanding genetic diversity within microbial communities. The article then progresses thematically, with dedicated sections on human and plant pan-genomics. The plant pan-genome section reviews relevant literature, emphasizing its implications for crop improvement, modern breeding practices, and adaptation to climatic conditions. Additionally, the review addresses the applications of pan-genomics in vaccine and drug development. A section also explores the insights gained from pan-genomics in the context of public health. Finally, a separate section addresses the current challenges faced in the field and offers potential solutions. The conclusion reflects the overarching trends identified throughout the review and outlines directions for future research, emphasizing key areas that require further exploration.

## Acknowledgement

The authors are grateful to the management of Rajalakshmi Engineering College, Chennai, Tamil Nadu, India, for encouraging the authors to carry out this work.

## Ethical Statement

This study does not contain any studies with human or animal subjects performed by any of the authors.

## Conflicts of Interest

The authors declare that they have no conflicts of interest to this work.

## Data Availability Statement

Data sharing is not applicable to this article as no new data were created or analyzed in this study.

## Author Contribution Statement

**Rachana Reghu:** Writing – original draft, Writing – review & editing. **Ramya Latha Karunakaran:** Writing – original draft, Writing – review & editing. **Ranjani Balaji:** Writing – original draft, Writing – review & editing. **Achsha Babu:** Writing – original draft, Writing – review & editing, Visualization, Supervision. **Arun Arumugaperumal:** Conceptualization, Supervision.

## References

[1] Li, W., Liu, J., Zhang, H., Liu, Z., Wang, Y., Xing, L., . . . , & Du, H. (2022). Plant pan-genomics: Recent advances, new challenges, and roads ahead. *Journal of Genetics and Genomics*, *49*(9), 833–846. https://doi.org/10.1016/j.jgg.2022.06.004

[2] Li, N., He, Q., Wang, J., Wang, B., Zhao, J., Huang, S., . . . , & Yu, Q. (2023). Super-pangenome analyses highlight genomic diversity and structural variation across wild and cultivated tomato species. *Nature Genetics*, *55*(5), 852–860. https://doi.org/10.1038/s41588-023-01340-y

[3] Hameed, A., Poznanski, P., Nadolska-Orczyk, A., & Orczyk, W. (2022). Graph pangenomes track genetic variants for crop improvement. *International Journal of Molecular Sciences*, *23*(21), 13420. https://doi.org/10.3390/ijms232113420

[4] Tay Fernandez, C. G., Nestor, B. J., Danilevicz, M. F., Marsh, J. I., Petereit, J., Bayer, P. E., . . . , & Edwards, D. (2022). Expanding gene-editing potential in crop improvement with pangenomes. *International Journal of Molecular Sciences*, *23*(4), 2276. https://doi.org/10.3390/ijms23042276

[5] Hao, P., Zheng, H., Yu, Y., Ding, G., Gu, W., Chen, S., . . . , & Zhao, G. (2011). Complete sequencing and pan-genomic analysis of *Lactobacillus delbrueckii* subsp. *bulgaricus* reveal its genetic basis for industrial yogurt production. *PLoS One*, *6*(1), e15964. https://doi.org/10.1371/journal.pone.0015964

[6] Farag, I. F., Youssef, N. H., & Elshahed, M. S. (2017). Global distribution patterns and pangenomic diversity of the candidate phylum "*Latescibacteria*" (WS3). *Applied and Environmental Microbiology*, *83*(10), e00521–17. https://doi.org/10.1128/AEM.00521-17

[7] Medini, D., Donati, C., Tettelin, H., Masignani, V., & Rappuoli, R. (2005). The microbial pan-genome. *Current Opinion in Genetics & Development*, *15*(6), 589–594. https://doi.org/10.1016/j.gde.2005.09.006

[8] Tamas, I., Klasson, L., Canback, B., Naslund, A. K., Eriksson, A. S., Wernegreen, J. J., . . . , & Andersson, S. G. (2002). 50 million years of genomic stasis in endosymbiotic bacteria. *Science*, *296*(5577), 2376–2379. https://doi.org/10.1126/science.1071278

[9] Rouli, L., Merhej, V., Fournier, P. E., & Raoult, D. (2015). The bacterial pangenome as a new tool for analysing pathogenic bacteria. *New Microbes and New Infections*, *7*, 72–85. https://doi.org/10.1016/j.nmni.2015.06.005

[10] Costa, S. S., Guimarães, L. C., Silva, A., Soares, S. C., & Baraúna, R. A. (2020). First steps in the analysis of prokaryotic pan-genomes. *Bioinformatics and Biology Insights*, *14*, 1–9. https://doi.org/10.1177/1177932220938064

[11] Hansen, E. E., Lozupone, C. A., Rey, F. E., Wu, M., Guruge, J. L., Narra, A., . . . , & Gordon, J. I. (2011). Pan-genome of the dominant human gut-associated archaeon, *Methanobrevibacter smithii*, studied in twins. *Proceedings of the National Academy of Sciences*, *108*, 4599–4606. https://doi.org/10.1073/pnas.1000071108

[12] Scholz, M., Ward, D. V., Pasolli, E., Tolio, T., Zolfo, M., Asnicar, F., . . . , & Segata, N. (2016). Strain-level microbial epidemiology and population genomics from shotgun metagenomics. *Nature Methods*, *13*(5), 435–438. https://doi.org/10.1038/nmeth.3802

[13] Kim, Y., Koh, I., Young Lim, M., Chung, W. H., & Rho, M. (2017). Pan-genome analysis of *Bacillus* for microbiome profiling. *Scientific Reports*, *7*(1), 10984. https://doi.org/10.1038/s41598-017-11385-9

[14] Yassour, M., Jason, E., Hogstrom, L. J., Arthur, T. D., Tripathi, S., Siljander, H., . . . , & Xavier, R. J. (2018). Strain-level analysis of mother-to-child bacterial transmission during the first few months of life. *Cell Host & Microbe*, *24*(1), 146–154. https://doi.org/10.1016/j.chom.2018.06.007

[15] Peng, Y., Tang, S., Wang, D., Zhong, H., Jia, H., Cai, X., . . . , & Li, J. (2018). MetaPGN: A pipeline for construction and graphical visualization of annotated pangenome networks. *GigaScience*, *7*(11), giy121. https://doi.org/10.1093/gigascience/giy121

[16] Delmont, T. O., & Eren, A. M. (2018). Linking pangenomes and metagenomes: The *Prochlorococcus* metapangenome. *PeerJ*, *6*, e4320. https://doi.org/10.7717/peerj.4320

[17] Plaza Oñate, F., Le Chatelier, E., Almeida, M., Cervino, A. C., Gauthier, F., Magoulès, F., . . . , & Pichaud, M. (2019). MSPminer: Abundance-based reconstitution of microbial pan-genomes from shotgun metagenomic data. *Bioinformatics*, *35*(9), 1544–1552. https://doi.org/10.1093/bioinformatics/bty830

[18] Zhong, C., Han, M., Yang, P., Chen, C., Yu, H., Wang, L., & Ning, K. (2019). Comprehensive analysis reveals the evolution and pathogenicity of *Aeromonas*, viewed from both single isolated species and microbial communities. *Msystems*, *4*(5), 1110–1128. https://doi.org/10.1128/msystems.00252-19

[19] Almeida, A., Mitchell, A. L., Boland, M., Forster, S. C., Gloor, G. B., Tarkowska, A., . . . , & Finn, R. D. (2019). A new genomic blueprint of the human gut microbiota. *Nature*, *568*(7753), 499–504. https://doi.org/10.1038/s41586-019-0965-1

[20] Pasolli, E., Asnicar, F., Manara, S., Zolfo, M., Karcher, N., Armanini, F., . . . , & Segata, N. (2019). Extensive unexplored human microbiome diversity revealed by over 150,000 genomes from metagenomes spanning age, geography, and lifestyle. *Cell*, *176*(3), 649–662. https://doi.org/10.1016/j.cell.2019.01.001

[21] Utter, D. R., Borisy, G. G., Eren, A. M., Cavanaugh, C. M., & Mark Welch, J. L. (2020). Metapangenomics of the oral microbiome provides insights into habitat adaptation and cultivar diversity. *Genome Biology*, *21*, 1–25. https://doi.org/10.1186/s13059-020-02200-2

[22] Ma, B., France, M., & Ravel, J. (2020). Meta-pangenome: At the crossroad of pangenomics and metagenomics. In H. Tettelin & D. Medini (Eds.), *The pangenome: Diversity, dynamics and evolution of genomes* (pp. 205–218). Springer. https://doi.org/10.1007/978-3-030-38281-0_9

[23] Arumugaperumal, A., Paul, S., Lathakumari, S., Balasubramani, R., & Sivasubramaniam, S. (2020). The draft genome of a new *Verminephrobacter eiseniae* strain: A nephridial symbiont of earthworms. *Annals of Microbiology*, *70*, 1–18. https://doi.org/10.1186/s13213-020-01549-w

[24] Kumari, K., Rawat, V., Shadan, A., Sharma, P. K., Deb, S., & Singh, R. P. (2023). In-depth genome and pan-genome analysis of a metal-resistant bacterium *Pseudomonas parafulva* OS-1. *Frontiers in Microbiology*, *14*, 1140249. https://doi.org/10.3389/fmicb.2023.1140249

[25] Asif, M., Li-Qun, Z., Zeng, Q., Atiq, M., Ahmad, K., Tariq, A., . . . , & Hatamleh, A. A. (2023). Comprehensive genomic analysis of *Bacillus paralicheniformis* strain BP9, pan-genomic and genetic basis of biocontrol mechanism. *Computational and Structural Biotechnology Journal*, *21*, 4647–4662. https://doi.org/10.1016/j.csbj.2023.09.043

[26] Nageeb, W. M., & Hetta, H. F. (2023). Pangenome analysis of *Corynebacterium striatum*: Insights into a neglected multidrug-resistant pathogen. *BMC Microbiology*, *23*(1), 252. https://doi.org/10.1186/s12866-023-02996-6

[27] Benning, S., Pritsch, K., Radl, V., Siani, R., Wang, Z., & Schloter, M. (2024). (Pan)genomic analysis of two *Rhodococcus* isolates and their role in phenolic compound degradation. *Microbiology Spectrum*, *12*(4), e03783–23. https://doi.org/10.1128/spectrum.03783-23

[28] Jain, V. S., Schubert, M. G., Sarnaik, A. P., Pritam, P., Jaiswal, D., Church, G. M., & Wangikar, P. P. (2024). *De novo* genome assembly and pan-genome analysis of the fast-growing Indian isolates of *Synechococcus elongatus*: Potential chassis for bioproduction. *The Microbe*, *2*, 100048. https://doi.org/10.1016/j.microb.2024.100048

[29] Wei, F., Liang, X., Shi, J. C., Luo, J. N., Qiu, L. J., Li, X. X., . . . , & Feng, J. Y. (2024). Pan-genomic analysis identifies the Chinese strain as a new subspecies of *Xanthomonas fragariae*. *Plant Disease*, *108*(1), 45–49. https://doi.org/10.1094/PDIS-05-23-0933-SC

[30] Nedashkovskaya, O., Balabanova, L., Otstavnykh, N., Zhukova, N., Detkova, E., Seitkalieva, A., . . . , & Isaeva, M. (2024). In-depth genome characterization and pan-genome analysis of strain KMM 296, a producer of highly active alkaline phosphatase; proposal for the reclassification of *Cobetia litoralis* and *Cobetia pacifica* as the later heterotypic synonyms of *Cobetia amphilecti* and *Cobetia marina*, and emended description of the species *Cobetia amphilecti* and *Cobetia marina*. *Biomolecules*, *14*(2), 196. https://doi.org/10.3390/biom14020196

[31] Wang, T., Shi, Y., Zheng, M., & Zheng, J. (2024). Comparative genomics unveils functional diversity, pangenome openness, and underlying biological drivers among *Bacillus subtilis* group. *Microorganisms*, *12*(5), 986. https://doi.org/10.3390/microorganisms12050986

[32] Li, R., Li, Y., Zheng, H., Luo, R., Zhu, H., Li, Q., . . . , & Wang, J. (2010). Building the sequence map of the human pan-genome. *Nature Biotechnology*, *28*(1), 57–63. https://doi.org/10.1038/nbt.1596

[33] Besenbacher, S., Liu, S., Izarzugaza, J. M., Grove, J., Belling, K., Bork-Jensen, J., . . . , & Rasmussen, S. (2015). Novel variation and *de novo* mutation rates in population-wide *de novo* assembled Danish trios. *Nature Communications*, *6*(1), 5969. https://doi.org/10.1038/ncomms6969

[34] Maretty, L., Jensen, J. M., Petersen, B., Sibbesen, J. A., Liu, S., Villesen, P., . . . , & Schierup, M. H. (2017). Sequencing and *de novo* assembly of 150 genomes from Denmark as a population reference. *Nature*, *548*(7665), 87–91. https://doi.org/10.1038/nature23264

[35] Sherman, R. M., Forman, J., Antonescu, V., Puiu, D., Daya, M., Rafaels, N., . . . , & Salzberg, S. L. (2019). Assembly of a pan-genome from deep sequencing of 910 humans of African descent. *Nature Genetics*, *51*(1), 30–35. https://doi.org/10.1038/s41588-018-0273-y

[36] Miga, K. H., & Wang, T. (2021). The need for a human pangenome reference sequence. *Annual Review of Genomics and Human Genetics*, *22*(1), 81–102. https://doi.org/10.1146/annurev-genom-120120-081921

[37] Schatz, M. C., Maron, L. G., Stein, J. C., Wences, A. H., Gurtowski, J., Biggers, E., . . . , & McCombie, W. R. (2014). Whole genome *de novo* assemblies of three divergent strains of rice, *Oryza sativa*, document novel gene space of *aus* and *indica*. *Genome Biology*, *15*, 1–16. https://doi.org/10.1186/s13059-014-0506-z

[38] Golicz, A. A., Bayer, P. E., Barker, G. C., Edger, P. P., Kim, H., Martinez, P. A., . . . , & Edwards, D. (2016). The pangenome of an agronomically important crop plant *Brassica oleracea*. *Nature Communications*, *7*(1), 13390. https://doi.org/10.1038/ncomms13390

[39] Gordon, S. P., Contreras-Moreira, B., Woods, D. P., Des Marais, D. L., Burgess, D., Shu, S., . . . , & Vogel, J. P. (2017). Extensive gene content variation in the *Brachypodium distachyon* pan-genome correlates with population structure. *Nature Communications*, *8*(1), 2184. https://doi.org/10.1038/s41467-017-02292-8

[40] Zhou, P., Silverstein, K. A., Ramaraj, T., Guhlin, J., Denny, R., Liu, J., . . . , & Young, N. D. (2017). Exploring structural variation and gene family architecture with *De Novo* assemblies of 15 *Medicago* genomes. *BMC Genomics*, *18*, 1–14. https://doi.org/10.1186/s12864-017-3654-1

[41] Wang, W., Mauleon, R., Hu, Z., Chebotarov, D., Tai, S., Wu, Z., . . . , & Leung, H. (2018). Genomic variation in 3,010 diverse accessions of Asian cultivated rice. *Nature*, *557*(7703), 43–49. https://doi.org/10.1038/s41586-018-0063-9

[42] Zhao, Q., Feng, Q., Lu, H., Li, Y., Wang, A., Tian, Q., . . . , & Huang, X. (2018). Pan-genome analysis highlights the extent of genomic variation in cultivated and wild rice. *Nature Genetics*, *50*(2), 278–284. https://doi.org/10.1038/s41588-018-0041-z

[43] Yu, J., Golicz, A. A., Lu, K., Dossa, K., Zhang, Y., Chen, J., . . . , & Zhang, X. (2019). Insight into the evolution and functional characteristics of the pan-genome assembly from sesame landraces and modern cultivars. *Plant Biotechnology Journal*, *17*(5), 881–892. https://doi.org/10.1111/pbi.13022

[44] Gao, L., Gonda, I., Sun, H., Ma, Q., Bao, K., Tieman, D. M., . . . , & Fei, Z. (2019). The tomato pan-genome uncovers new genes and a rare allele regulating fruit flavor. *Nature Genetics*, *51*(6), 1044–1051. https://doi.org/10.1038/s41588-019-0410-2

[45] Hübner, S., Bercovich, N., Todesco, M., Mandel, J. R., Odenheimer, J., Ziegler, E., . . . , & Rieseberg, L. H. (2019). Sunflower pan-genome analysis shows that hybridization altered gene content and disease resistance. *Nature Plants*, *5*(1), 54–62. https://doi.org/10.1038/s41477-018-0329-0

[46] Liu, Y., Du, H., Li, P., Shen, Y., Peng, H., Liu, S., . . . , & Tian, Z. (2020). Pan-genome of wild and cultivated soybeans. *Cell*, *182*(1), 162–176. https://doi.org/10.1016/j.cell.2020.05.023

[47] Sun, Y., Wang, J., Li, Y., Jiang, B., Wang, X., Xu, W. H., . . . , & Kong, X. D. (2022). Pan-genome analysis reveals the abundant gene presence/absence variations among different varieties of melon and their influence on traits. *Frontiers in*

*Plant Science*, *13*, 835496. https://doi.org/10.3389/fpls.2022.835496

[48] Yin, Z., Liu, X., Qian, C., Sun, L., Pang, S., Liu, J., . . . , & Xie, Z. (2022). Pan-genome analysis of *Delftia tsuruhatensis* reveals important traits concerning the genetic diversity, pathogenicity, and biotechnological properties of the species. *Microbiology Spectrum*, *10*(2), e02072–21. https://doi.org/10.1128/spectrum.02072-21

[49] Ni, L., Liu, Y., Ma, X., Liu, T., Yang, X., Wang, Z., . . . , & Tian, Z. (2023). Pan-3D genome analysis reveals structural and functional differentiation of soybean genomes. *Genome Biology*, *24*(1), 12. https://doi.org/10.1186/s13059-023-02854-8

[50] Bozan, I., Achakkagari, S. R., Anglin, N. L., Ellis, D., Tai, H. H., & Strömvik, M. V. (2023). Pangenome analyses reveal impact of transposable elements and ploidy on the evolution of potato species. *Proceedings of the National Academy of Sciences*, *120*(31), e2211117120. https://doi.org/10.1073/pnas.2211117120

[51] He, Q., Tang, S., Zhi, H., Chen, J., Zhang, J., Liang, H., . . . , & Diao, X. (2023). A graph-based genome and pan-genome variation of the model plant *Setaria*. *Nature Genetics*, *55*(7), 1232–1242. https://doi.org/10.1038/s41588-023-01423-w

[52] Carpi, F. M., Coman, M. M., Silvi, S., Picciolini, M., Verdenelli, M. C., & Napolioni, V. (2022). Comprehensive pan-genome analysis of *Lactiplantibacillus plantarum* complete genomes. *Journal of Applied Microbiology*, *132*(1), 592–604. https://doi.org/10.1111/jam.15199

[53] Zhou, H., Yan, F., Hao, F., Ye, H., Yue, M., Woeste, K., . . . , & Zhang, S. (2023). Pan-genome and transcriptome analyses provide insights into genomic variation and differential gene expression profiles related to disease resistance and fatty acid biosynthesis in eastern black walnut (*Juglans nigra*). *Horticulture Research*, *10*(3), uhad015. https://doi.org/10.1093/hr/uhad015

[54] Lian, Q., Huettel, B., Walkemeier, B., Mayjonade, B., Lopez-Roques, C., Gil, L., . . . , & Mercier, R. (2024). A pan-genome of 69 *Arabidopsis thaliana* accessions reveals a conserved genome structure throughout the global species range. *Nature Genetics*, *56*, 982–991. https://doi.org/10.1038/s41588-024-01715-9

[55] Liang, Q., Muñoz-Amatriaín, M., Shu, S., Lo, S., Wu, X., Carlson, J. W., . . . , & Lonardi, S. (2024). A view of the pan-genome of domesticated Cowpea (*Vigna unguiculata* [L.] Walp.). *The Plant Genome*, *17*(1), e20319. https://doi.org/10.1002/tpg2.20319

[56] Alonge, M., Wang, X., Benoit, M., Soyk, S., Pereira, L., Zhang, L., . . . , & Lippman, Z. B. (2020). Major impacts of widespread structural variation on gene expression and crop improvement in tomato. *Cell*, *182*(1), 145–161. https://doi.org/10.1016/j.cell.2020.05.021

[57] Daware, A., Malik, A., Srivastava, R., Das, D., Ellur, R. K., Singh, A. K., . . . , & Parida, S. K. (2023). Rice pangenome genotyping array: An efficient genotyping solution for pangenome-based accelerated genetic improvement in rice. *The Plant Journal*, *113*(1), 26–46. https://doi.org/10.1111/tpj.16028

[58] Luo, R., Liu, B., Xie, Y., Li, Z., Huang, W., Yuan, J., . . . , & Wang, J. (2012). SOAPdenovo2: An empirically improved memory-efficient short-read *de novo* assembler. *Gigascience*, *1*(1), 2047-217X-1-18. https://doi.org/10.1186/2047-217X-1-18

[59] Li, X. M., Chao, D. Y., Wu, Y., Huang, X., Chen, K., Cui, L. G., . . . , & Lin, H. X. (2015). Natural alleles of a proteasome α2 subunit gene contribute to thermotolerance and adaptation of African rice. *Nature Genetics*, *47*(7), 827–833. https://doi.org/10.1038/ng.3305

[60] Podschun, R., & Ullmann, U. (1998). *Klebsiella spp*. as nosocomial pathogens: Epidemiology, taxonomy, typing methods, and pathogenicity factors. *Clinical Microbiology Reviews*, *11*(4), 589–603. https://doi.org/10.1128/cmr.11.4.589

[61] Chung, P. Y. (2016). The emerging problems of *Klebsiella pneumoniae* infections: Carbapenem resistance and biofilm formation. *FEMS Microbiology Letters*, *363*(20), fnw219. https://doi.org/10.1093/femsle/fnw219

[62] Allemailem, K. S. (2021). A comprehensive computer aided vaccine design approach to propose a multi-epitopes subunit vaccine against genus *Klebsiella* using pan-genomics, reverse vaccinology, and biophysical techniques. *Vaccines*, *9*(10), 1087. https://doi.org/10.3390/vaccines9101087

[63] Delannoy, C. M., Crumlish, M., Fontaine, M. C., Pollock, J., Foster, G., Dagleish, M. P., . . . , & Zadoks, R. N. (2013). Human *Streptococcus agalactiae* strains in aquatic mammals and fish. *BMC Microbiology*, *13*, 1–9. https://doi.org/10.1186/1471-2180-13-41

[64] Delamare Deboutteville, J. (2014). *On the origin of Group B* Streptococcus *from disease outbreaks in wild marine fish in Australia*. PhD Thesis, The University of Queensland.

[65] Kawasaki, M. (2019). *A pan genome reverse vaccinology approach to prevent* Streptococcus agalactiae *infection in farmed tilapia*. PhD Thesis, The University of Queensland.

[66] Riasat, I., Bakhtiar, S. M., Faheem, M., Jaiswal, A. K., Naeem, M., Khan, R., . . . , & Jamal, S. B. (2021). Application of pan genomics towards the druggability of *Clostridium botulinum*. *Applied Nanoscience*, *12*, 3237–3249. https://doi.org/10.1007/s13204-021-02005-0

[67] Hussain, M., Gatherer, D., & Wilson, J. B. (2014). Modelling the structure of full-length Epstein–Barr virus nuclear antigen 1. *Virus Genes*, *49*, 358–372. https://doi.org/10.1007/s11262-014-1101-9

[68] Pyzer-Knapp, E. O., Suh, C., Gómez-Bombarelli, R., Aguilera-Iparraguirre, J., & Aspuru-Guzik, A. (2015). What is high-throughput virtual screening? A perspective from organic materials discovery. *Annual Review of Materials Research*, *45*(1), 195–216. https://doi.org/10.1146/annurev-matsci-070214-020823

[69] Basharat, Z., Jahanzaib, M., Yasmin, A., & Khan, I. A. (2021). Pan-genomics, drug candidate mining and ADMET profiling of natural product inhibitors screened against *Yersinia pseudotuberculosis*. *Genomics*, *113*(1), 238–244. https://doi.org/10.1016/j.ygeno.2020.12.015

[70] Nicolas-Chanoine, M. H., Blanco, J., Leflon-Guibout, V., Demarty, R., Alonso, M. P., Caniça, M. M., . . . , & Johnson, J. R. (2008). Intercontinental emergence of *Escherichia coli* clone O25: H4-ST131 producing CTX-M-15. *Journal of Antimicrobial Chemotherapy*, *61*(2), 273–281. https://doi.org/10.1093/jac/dkm464

[71] Beghain, J., Bridier-Nahmias, A., Le Nagard, H., Denamur, E., & Clermont, O. (2018). ClermonTyping: An easy-to-use and accurate *in silico* method for *Escherichia* genus strain phylotyping. *Microbial Genomics*, *4*(7), e000192. https://doi.org/10.1099/mgen.0.000192

[72] Tantoso, E., Eisenhaber, B., Kirsch, M., Shitov, V., Zhao, Z., & Eisenhaber, F. (2022). To kill or to be killed: Pangenome analysis of *Escherichia coli* strains reveals a tailocin specific for pandemic ST131. *BMC Biology*, *20*(1), 146. https://doi.org/10.1186/s12915-022-01347-7

[73] Yang, T., Zhong, J., Zhang, J., Li, C., Yu, X., Xiao, J., . . . , & Chen, F. (2018). Pan-genomic study of *Mycobacterium tuberculosis* reflecting the primary/secondary genes, generality/individuality, and the interconversion through copy number variations. *Frontiers in Microbiology*, *9*, 1886. https://doi.org/10.3389/fmicb.2018.01886

[74] Barh, D., Soares, S. C., Tiwari, S., & De Car Azevedo, V. A. (2020). *Pan-genomics: Applications, challenges, and future prospects*. USA: Academic Press.

**How to Cite:** Reghu, R., Karunakaran, R. L., Balaji, R., Babu, A., & Arumugaperumal, A. (2024). Unlocking Nature's Code: The Power of Pan-Genomics in Biological Entities. *Medinformatics*. https://doi.org/10.47852/bonviewMEDIN42023423