Pangenomics in Microbial and Crop Research: Progress, Applications, and Perspectives

Sumit Kumar Aggarwal [1,†], Alla Singh [1,†], Mukesh Choudhary [1,2,3], Aundy Kumar [4], Sujay Rakshit [1,*], Pardeep Kumar [1], Abhishek Bohra [5] and Rajeev K. Varshney [3,6,*]

1   ICAR-Indian Institute of Maize Research, PAU Campus, Ludhiana 141004, India;
     sumit.aggarwal009@gmail.com (S.K.A.); allasingh.panesar@gmail.com (A.S.);
     mukesh.choudhary1@icar.gov.in (M.C.); pardeepkumar656@gmail.com (P.K.)
2   School of Agriculture and Environment, The University of Western Australia, Perth, WA 6009, Australia
3   The UWA Institute of Agriculture, The University of Western Australia, Perth, WA 6009, Australia
4   ICAR-Indian Agricultural Research Institute, New Delhi 110012, India; kumar@iari.res.in
5   ICAR-Indian Institute of Pulses Research (IIPR), Kanpur 208024, India; abhi.omics@gmail.com
6   State Agricultural Biotechnology Centre, Centre for Crop and Food Innovation, Food Futures Institute,
     Murdoch University, Murdoch, WA 6150, Australia
*   Correspondence: s.rakshit@icar.gov.in (S.R.); rajeev.varshney@murdoch.edu.au (R.K.V.)
†   These authors contributed equally to this work.

Abstract: Advances in sequencing technologies and bioinformatics tools have fueled a renewed interest in whole genome sequencing efforts in many organisms. The growing availability of multiple genome sequences has advanced our understanding of the within-species diversity, in the form of a pangenome. Pangenomics has opened new avenues for future research such as allowing dissection of complex molecular mechanisms and increased confidence in genome mapping. To comprehensively capture the genetic diversity for improving plant performance, the pangenome concept is further extended from species to genus level by the inclusion of wild species, constituting a super-pangenome. Characterization of pangenome has implications for both basic and applied research. The concept of pangenome has transformed the way biological questions are addressed. From understanding evolution and adaptation to elucidating host–pathogen interactions, finding novel genes or breeding targets to aid crop improvement to design effective vaccines for human prophylaxis, the increasing availability of the pangenome has revolutionized several aspects of biological research. The future availability of high-resolution pangenomes based on reference-level near-complete genome assemblies would greatly improve our ability to address complex biological problems.

1. Introduction

The initiation of the human genome sequencing project in 1990 served as a breakthrough in biological sciences. It opened the way for many new scientific domains including the study of proteome and metabolome, etc. The genome sequencing efforts have since been extended to many other organisms, including bacteria, fungi, plants, and animals. The breakthroughs in DNA sequencing have led to a considerable reduction in time and cost required for decoding the complete genome sequence [1–3]. This has been reflected in a deluge of genome sequencing datasets that have been deposited in public databases [4]. The increasing availability of a large number of genomes boosted the comparative genomics studies for the estimation of genetic variation among the individuals of a species. Growing realization about the inadequacy of a single reference genome to catalogue entire gene content of a species stimulated interest on sequencing multiple genomes, resulting in the

development of the pangenome concept [5]. Pangenome analysis has the potential to serve as a game-changing approach for covering entire species diversity using advanced sequencing platforms [5,6]. Analysis of genomes of six strains of *Streptococcus agalactiae* laid the foundation of developing a pangenome that has two major components: core genome represented by the fixed genome portion present in all six strains (constituting up to four-fifth proportion in any single individual genome) and the remaining one-fifth "variable" portion corresponded to strain-specific genes, often designated as "dispensable or accessory" [7]. The variable genomic portion can be classified into two distinct parts-unique genes (restricted to one individual only) and dispensable genes (genes common across at least n−1 individuals or absent in some individuals). The core and variable genes signify the essentiality and the diversity of the species, respectively [8]. The development of pangenomes has provided researchers with new tools and approaches to find novel genes and understand how the genome shapes the diversity of an organism. Newly constituted pangenomes throw light on many aspects of basic and applied sciences, including evolution, design of vaccines, and antibacterial.

In the case of crops, wild relatives and diverse germplasm have contributed immensely to domestication and systemic improvement, resulting in the development of modern cultivars. Accounting for the broad genetic diversity contained in wild species in a particular genus extends the pangenome concept from species level to genus level [9]. Super-pangenome, a pangenome of pangenomes, encompasses different wild species in the given genus and hence expands the possibility to harness the maximum genome diversity available in the particular genus. Another remarkable improvement led by the availability of pangenome is the cataloging of large structural variations (SVs) in the genomes [10]. The present article highlights the importance of SVs and discusses the emergence and subsequent maturation of the pangenome concept and its growing applications in diverse fields of biology and plant science.

2. Pangenome: Concept and Types

The term "pangenome" ("pan" in Greek-means "whole") describes the total of essential genes in a complete genome dataset of the given species [5].

The pangenome comprises of three parts: (i) Core genome, formed by genes shared by all genomes and usually involved in essential cellular processes; (ii) accessory or dispensable genome, composed of genes absent in some isolates; and (iii) species-specific or strain-specific genes, genes restricted to the single genome [5,7]. Dispensable and species/strain-specific genes correspond to the variable part of the genome. Thus, pangenome can be constructed by the identification of core and variable genes using the genome of particular individuals or strains of any species (Figure 1). Genes comprising the accessory and species or strain-specific genome are often, but not always involved in the adaptation of an organism to a particular niche. The core genome is highly conserved and involved in basic biological processes like replication, translation, and cellular homeostasis. Dispensable/accessory genome subset of genes emerge by horizontal gene transfer shared between some organisms (but not present in all organisms under study and much more common in prokaryotes compared to Eukaryotes) and hence are associated with specific functions like survival, virulence, or resistance to antibiotics [11]. The accessory genes are under mutational pressure which likely gives rise to new alleles for better adaptation to a particular niche. In contrast, the core genome is under strong selective (or evolutionary) pressure and hence highly conserved. Species-specific/strain-specific genomes are present in a single species that emerged by horizontal gene transfer at the inter-species level, whereas strain-specific genes are only present in one strain and are at the intra-species level and associated with the pathogenicity of a particular strain. Conclusively, core and accessory parts of the genome drive the pangenome diversity.
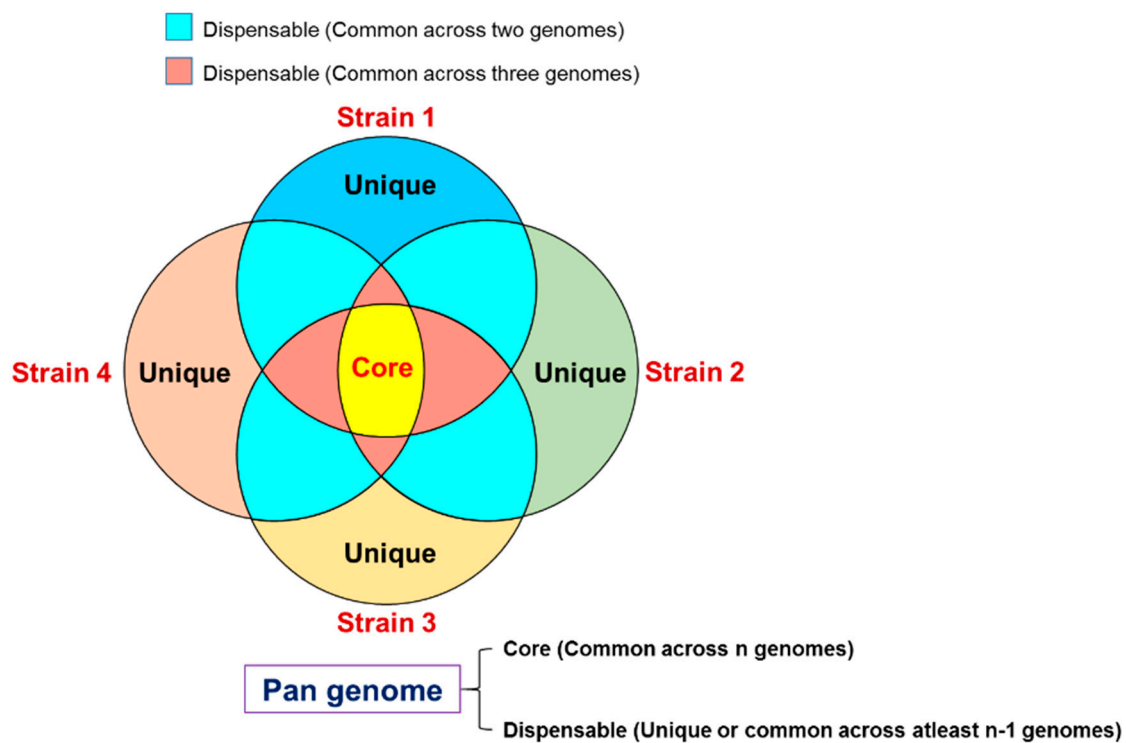
Figure 1. Organization of a pangenome composed of core and dispensable components of the genome.

An accurate reference genome sequence is an important resource for understanding the biological functions using NGS-based approaches. However, considering the inability of a single individual to represent entire genetic diversity, researchers soon realized the need to look beyond a single reference genome via utilizing the available or generating additional sequence information on multiple genomes. The Computational Pangenomics Consortium [12] notes four types of genomes, *viz.*, complete genome, which consists of all the sequences ever known for an organism; genome of a single individual; functional genome, lacking the disabling mutations known for a genome; and, consensus genome, based on the consensus of available sequence data (Figure 2). The choice of a "reference genome" depends on the objectives and the resource availability. In the post-NGS era, an increase in large-scale genome sequencing projects and the quest to explain hitherto unknown mechanisms have placed more emphasis on a "pangenome" as a new reference, to better understand the genetics of organisms.

Pangenome is classified as "*open*" or "*closed*", depending on the number of new genes added per genome sequenced [5]. If with the addition of a new genome sequence, the number of newly discovered genes keeps on increasing, the pangenome is said to be "open" and warrants further sequencing. On the other hand, if the number of new genes discovered remains the same upon sequencing of new genomes, the pangenome is referred to as "closed". Exemplified by *S. agalactiae* (group B streptococcus; GBS), the addition of sequencing information of new strain led to expanding the pangenome volume by 33 novel genes. For example, an open pangenome was noticed in the case of five strains of *S. pyogenes* exhibiting similar genomic diversity but contributing to the expansion of pangenome by 27 specific genes for the addition of each novel genome [5]. In another study, eight independent *Bacillus anthracis* isolates were sequenced but the pangenome volume expansion via a rise in the number of novel genes halted after the addition of sequence information of only four genomes [5]. Therefore, the *B. anthracis* species is considered to be an example of a "closed" pangenome as only genomic information of four isolates is good enough to represent the entire genomic content of this species. Recent research has reported the development of closed pangenomes in various crop species including rice [4,13–16].
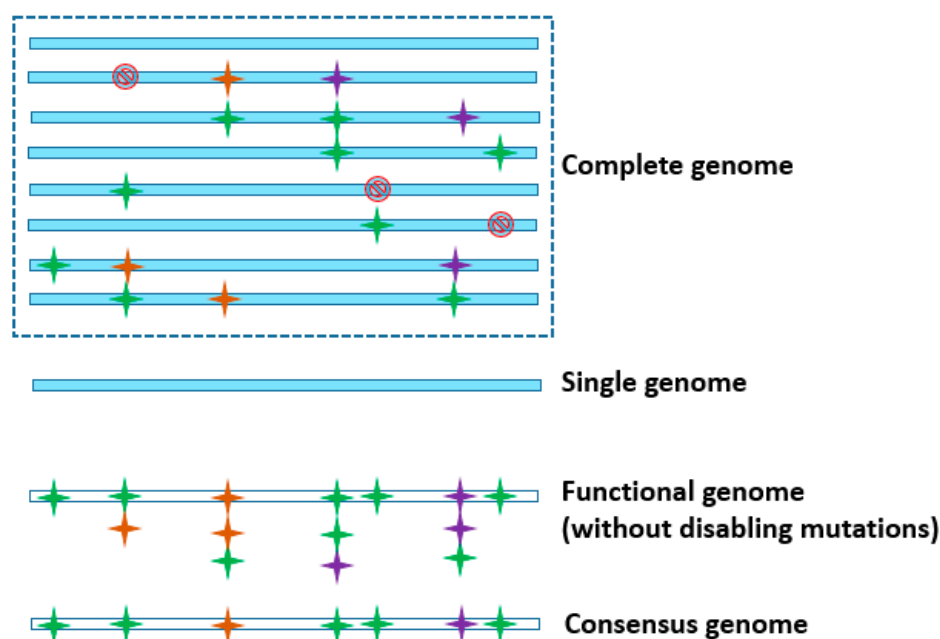
Figure 2. Different forms of a reference genome. The horizontal bars represent the DNA sequence of a genome. ⊘ represents a disabling mutation that disrupts the gene function. ✛, ✛, and ✛ depict various sequence polymorphisms.

3. Importance of Pangenome

The genomic era started over a decade ago, but still, bacterial species have not been explored to a larger extent. The sequencing studies of multiple strains in some species revealed the possibility of finding novel genes with the inclusion of sequencing information of each additional strain. Later, mathematical modeling [17] also supported this fact of discovering novel genes in some species even after the inclusion of hundreds of genomes per species. Therefore, a need was felt to discover a more accurate way to explaining the entire genetic information of bacterial species. Considering that the pangenome of any organism contains the highest amount of genetic information compared to a single genome, changes at the pangenome level may help understand the symptoms and infection in the host [8].

4. Structural Variations Are Crucial for within-Species Diversity

The improvement of any organism depends upon the existing genetic variation. The genetic variation for the agronomic traits among individuals of the same species or different species is caused by the differences in the sequence of nucleotides or bases called sequence variations and large-scale (usually >1 kb) DNA rearrangements referred to as structural variations (SVs). These SVs arise from various mechanisms like recombination, double-strand break repairs, and transposable elements and range from few base pairs to several megabases. The large SVs can be of two types: (i) copy number variation (CNV), defined as the variable number of copies of a particular sequence among different individuals and (ii) the presence-absence variation (PAV) created by the absence of a particular sequence in few individuals which otherwise exists in rest of the individuals [10,18]. Hence, PAV can be considered as an extreme form of CNV where one particular sequence is completely absent in a few individuals. Unlike humans, the abundance of CNVs has been reported from the majority of crop species and hence is considered to assume greater significance for causing variation in trait expressions [19]. One of the key objectives of the pangenome analysis is to capture genome variations caused by the large SVs including PAVs and CNVs. Generally, plant disease-associated defense genes are known to display CNVs [20]. Recent research supports a greater role for PAVs than CNVs in shaping crucial plant

phenotypes [21] (Figure 3a). The role of PAVs in stress response and domestication traits including shattering, photoperiod sensitivity, and male sterility has been evident in major crops such as rice, maize, and sorghum [22–24]. In plant genomes like maize that are characterized by extensive repetitive DNA sequences, the presence of transposable elements (TEs) could explain the abundance of PAVs [19,25]. By contrast, most of the agronomically important traits are governed by CNVs in barley (Figure 3b). Hence, variable distribution of SVs among crops could be the reason for the prominence of PAVs or CNVs in a particular crop. In rice, the SVs are reported to influence gene expression and their distribution among the populations helps understand the domestication process [26].
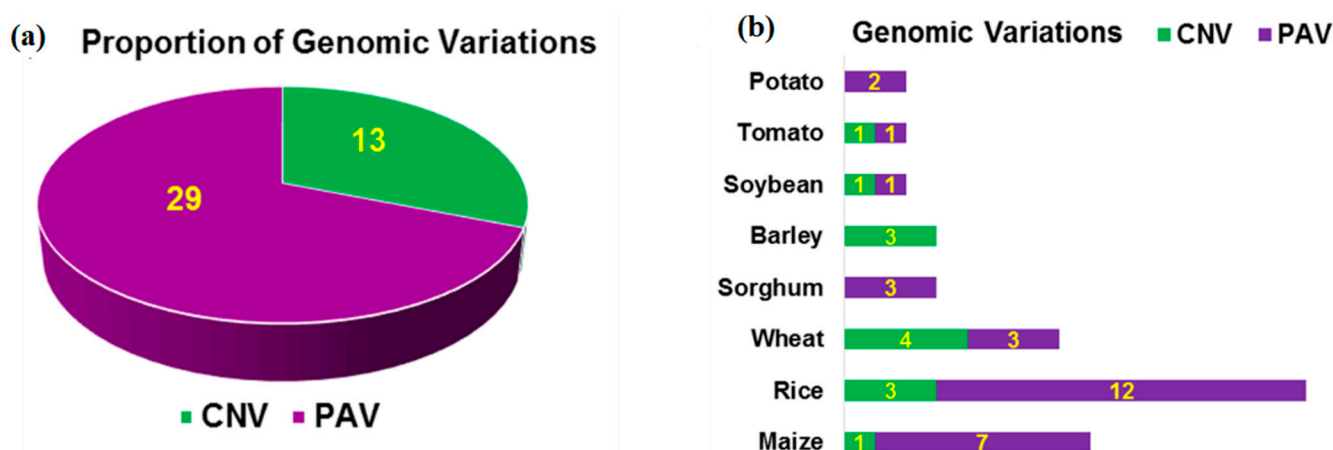

Figure 3. Genomic variation in terms of proportion (**a**) and distribution (**b**) of PAVs and CNVs in the genome of major crops for agriculturally important traits (Interpreted from Tao et al. [21].

5. Pangenome Construction: Basic Approaches and Critical Factors

Pangenomes can be generated by various approaches such as the comparative de novo approach [13,27], an iterative assembly approach [14,16,28,29], and the "map-to-pan" approach [30]. Further, [31] summarized the current approaches for construction pangenome in plants. Figure 4 illustrates the general steps for pangenome construction. It includes the genome assembly of the different strains followed by genome alignment and identification of core and dispensable parts of genomes. The identified genes are then used for functional annotation. The different approaches of pangenome construction rely on this basic procedure with slight modifications. The comparative de novo approach, as exemplified by initial pangenome studies in crops like rice [27], soybean [16], relies on the principle of comparison of annotations of de novo genome assemblies of individuals for identification of core and dispensable genes, whereas the rest two approaches rely on building a pangenome reference sequence. Then, the identified pangenome sequences are annotated. Finally, the genic PAVs are identified via aligning the mapping reads on the pangenome. However, iterative assembly and map-to-pan approaches follow different strategies for the construction of a pangenome sequence as the former uses mapping reads from initial samples to align with whole-genome assembly reference accompanied by reference assembly update by addition of unmapped reads [6]. In contrast, the later approach starts with de novo assembling of individual genomes followed by the use of the reference genome to map low-quality de novo assemblies to construct pangenome [30]. The two approaches have been used in the recent pangenome studies in crop plants based on large-scale genome sequencing of more than 3000 accessions (iterative assembly approach: [4]; map-to-pan approach [32]). The benefits of iterative assembly and "map-to-pan" approaches are low sequencing depth-based identification of genic PAVs via mapping short reads to an annotated genome but with limited applicability to simple genomes with less repetitive gene sequences [33,34].
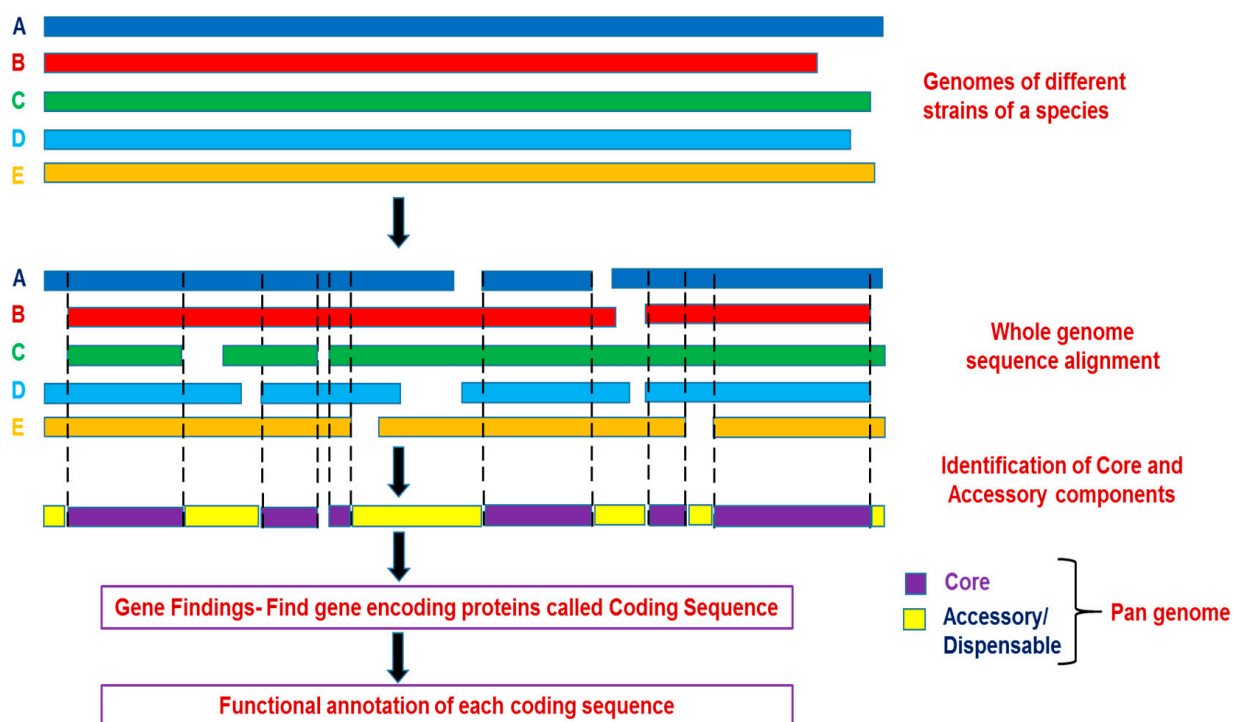
Figure 4. A basic approach for pangenome construction. Genome sequences of different strains represented schematically as A (**blue**), B (**red**), C (**green**), D (**light blue**), and E (**yellow**) are aligned to identify the core and accessory components of the pangenome.

The recent development of long-read sequencing or the third-generation sequencing platforms such as Pacific Biosciences (accessed on 16 March 2022)(PacBio) and Oxford Nanopore Technologies (accessed on 16 March 2022)(Nanopore) is likely to relieve the current drawback of the high-cost associated with the comparative de novo approach, thus greatly enhancing the utility of this approach in pangenome analysis. Alternatively, skim-sequencing can be used for pangenome construction via sequencing multiple varieties and assembling reads that do not align to the reference genome, especially for simpler genomes as it fails to effectively capture SVs in complex genomes [35].

The factors that critically influence the pangenome analysis include the quality of the reference assembly, its annotation quality, orthologous gene detection, selection of appropriate individuals, and suitable pangenome analysis tools or software [6]. The reference genome assembly should be of sufficient quality in terms of its size, completeness, and fragmentation level to facilitate better quality annotation. The long-read sequencing technologies could overcome the problem of fragmented assemblies resulting from the inability of short-read sequencing technologies to resolve the repetitive sequences in complex genomes [36]. The fragmented assemblies cause under-prediction of the total number of genes and also affect the detection of SVs, hence resulting in the poor quality functional annotation. The completeness of genome assemblies can be assessed by using several metrics like Core Eukaryotic Genes Mapping Approach accessed on 16 March 2022) CEGMA) [37] and Benchmarking Universal Single-Copy Orthologs accessed on 16 March 2022 (BUSCO) [38]. Another important factor concerns the selection of candidate individuals for pangenome construction. The candidate individuals should be highly diverse and optimum in number because less diverse and low population size of candidate individuals downgrades the representation quality of pangenome [6]. The optimum number of candidate individuals for a pangenome study can be decided by using the modelling of pangenome expansion and core genome reduction [7].

Mapping and assembling of genes are important issues to consider in the pangenome analysis. Various methods have been reviewed [39]. A pan reference for anchoring additional genes can be created via different approaches. One approach is using the synteny-based co-localization of core genes adjacent to the dispensable genes. Another is anchoring dispensable genes by using genetic marker-based linkage between core and dispensable genes. Alternatively, sequence-similarity-based approaches can be used for anchoring. However, repeat sequences pose a challenge in this case.

6. Software's/Tools for Pangenome Analysis

Software packages and tools are very important to categorize orthologous genes, calculate pangenomic profiles, integrate gene annotations, and construct phylogenies [40]. A detailed description of the various features of different software used in the pangenome analysis has been provided in Table 1. PanSeq (Pangenome Sequence Analysis Program) accessed on 17 September 2021 (Public Health Agency of Canada, Lethbridge, AB, Canada) is an online platform for the identification of core and accessory genomic regions. As a web server, it is platform-independent and makes use of NCBI resources. Similarly, PanFunPro, accessed on 17 September 2021 (Technical University of Denmark, Kongens Lyngby, Denmark) (Pangenome Analysis Based on Functional Profiles) is a tool for pangenome analysis using functional domains from HMM (Hidden Markov Models). GET_HOMOLOGUES is used to perform comparative-genomic analysis of bacterial strains and to build clusters of orthologous groups. ITEP accessed on 17 September 2021 (University of Illinois at Urbana-Champaign, Urbana, IL, USA) (Integrated Toolkit for the Exploration of Microbial Pangenomes) software system has been developed to predict protein families, orthologous genes, functional domains, pangenome (core and variable genes), and metabolic networks for related microbial species. PanGP accessed on 17 September 2021 (Beijing Institute of Genomics, Chinese Academy of Sciences, Beijing, China) (Pangenome Profiles) is a tool for a quick analysis of the bacterial pangenome using a large number of strains. PGAP accessed on 17 September 2021 (Beijing Institute of Genomics, Chinese Academy of Sciences, Beijing, China) (Pangenome Analysis Pipeline) is developed to perform pangenome analysis, genetic variation, evolution, and function analysis of gene clusters. PGAT accessed on 17 September 2021 (University of Washington, Seattle, WA, USA) (Prokaryotic Genome Analysis Tool) is a web tool to compare multiple strains of the same species, to predict genetic differences. Its analyses include pangenome, synteny, identification of genes present or absent in a dataset, comparison of sequence variants in orthologous genes, comparison of genes in metabolic pathways, and improvement of functional annotation. EDGAR accessed 17 September 2021 (Bielefeld University, Bielefeld, Germany) (Efficient Database Framework for Comparative Genome Analyses Using BLAST Score Ratios) is a webtool that performs orthology analysis to calculate pangenome, core-genome, and singletons are computed using BLAST Score Ratio Values (SRV). This method divides the BLAST bit score by the maximum possible bit score, generating the SRV and the cutoff is calculated using a sliding window instead of a fixed SRV threshold of 30, as proposed by [41]. Micropan accessed on 17 September 2021 (Norwegian University of Life Sciences, Norway) package is another tool that helps in pangenome and associated analysis. SplitMem accessed on 17 September 2021 (Stony Brook University, Stony Brook, NY, USA) is a graphical software for producing a compressed colored graph of the pangenome.

Table 1. Software/tools for pangenome analysis.

Software/Tool	Description/Role	URL Link	References
PanSeq	Extract the regions unique in the genome, Identify the SNPs and construct the file for phylogeny programme.	https://lfz.corefacility.ca/panseq/ (accessed on 17 September 2021)	[42]
PanFunPro	Homology detection and pairwise genome analysis in pan/core genome.	https://zenodo.org/record/7583#.YTR36p0zY2w (accessed on 17 September 2021)	[43]
GET_HOMOLOGUES	Clustering proteins and nucleotide sequence into homologous group and analysis of overlapping sets of proteins	http://www.eead.csic.es/compbio/soft/gethoms.php (accessed on 17 September 2021)	[44]
ITEP	It is use for sequence alignment, metabolic, clustering, and protein prediction	https://price.systemsbiology.net/itep (accessed on 17 September 2021)	[45]
PanGP	Use for large-scale bacterial pangenome profile analysis with sampling algorithms.	https://pangp.zhaopage.com/ (accessed on 17 September 2021)	[46]
PGAP	Detection of homologous genes, orthologous genes, SNP, phylogenetic studies, pangenome plotting and functional annotation.	http://pgap.sf.net (accessed on 17 September 2021)	[47]
PGAT	To compare the gene content and sequence across multiple microbial genomes to identify the SNPs.	http://nwrce.org/pgat (accessed on 17 September 2021)	[48]
EDGAR	EDGAR performs homology analyses with a specific cutoff, Venn diagrams and interactive synteny plots.	https://bio.tools/edgar_genomics (accessed on 17 September 2021)	[49]
Micropan	This allows integration of pangenome and additional analyses within a single programming language environment	Package "micropan" in r software (accessed on 17 September 2021)	[50]
SplitMem	A graphic software for pangenome analysis software by de Bruijn graph.	https://sourceforge.net/projects/splitmem/ (accessed on 17 September 2021)	[51]
ClustAGE	Focused on the accessory genomic dimension of pangenome	http://vfsmspineagent.fsm.northwestern.edu/cgi-bin/clustage.cgi (accessed on 17 September 2021)	[52]
DeNoGAP	Help in gene prediction, protein classification and orthology search	https://github.com/DSGlab/DeNoGAP (accessed on 17 September 2021)	[53]
EUPAN	This was first to analyze eukaryotic pangenomes to identify core and accessory gene datasets	http://cgm.sjtu.edu.cn/eupan/index.html (accessed on 17 September 2021)	[30]
Harvest	This is useful for the analysis based on three modules Parsnp (core-genome analysis), Gingr (output visualization), and Harvest Tools (meta-analysis)	https://www.cbcb.umd.edu/software/harvest (accessed on 17 September 2021)	[54]
LS-BSR	Calculates a score ratio per coding sequence within a pangenome dataset using BLAST	https://github.com/jasonsahl/LS-BSR (accessed on 17 September 2021)	[55]
NGSPanPipe	Identify pangenome from short reads and output is compatible with other pangenome analysis tools	https://github.com/Biomedinformatics/NGSPanPipe (accessed on 17 September 2021)	[56]
PanACEA	Identification of genomic regions those are phylogenetically dissimilar.	https://github.com/JCVenterInstitute/PanACEA (accessed on 17 September 2021)	[57]
PanCake	Useful for clustering homologous genes and analyzing core/accessory genome	https://pypi.org/project/pancake/ (accessed on 17 September 2021)	[58]
PanGeT	Pangenome analysis based on comparison at genome and proteome levels.	http://pranag.physics.iisc.ernet.in/PanGeT/ (accessed on 17 September 2021)	[59]
PanGFR-HM	Genomic/functional diversity and phylogenetic on genome-based between human associated microbial genomes	http://www.bioinfo.iicb.res.in/pangfr-hm/ (accessed on 17 September 2021)	[60]
PANINI	For rapid online visualization and analysis of the core and accessory genome evolutionary signal.	http://panini.pathogen.watch (accessed on 17 September 2021) and code at http://gitlab.com/cgps/panini (accessed on 17 September 2021)	[61]
PANNOTATOR	To ensure quality and standards for functional genome annotation among different strains	http://bnet.egr.vcu.edu/iioab/agenote.php (accessed on 17 September 2021)	[62]
PanOCT	PanOCT is a graph-based ortholog clustering tool of closely related prokaryotic genomes.	ftp://ftp.ncbi.nih.gov/blast/executables/release/ (accessed 17 September 2021)	[63]
Pan-Tetris	An interactive and dynamic visual inspection of gene occurrences in a pangenome table.	http://bit.ly/1vVxYZT (accessed on 17 September 2021)	[64]
PanTools	Annotating pangenomes, sequences adding, grouping genes, retrieving genomic regions and querying pangenome	http://www.bif.wur.nl (accessed on 17 September 2021)	[65]
PanViz	It can visualize from range of data formats of pangenomic data and mapping genes from existing pangenome	https://github.com/thomasp85/PanViz (accessed on 17 September 2021)	[66]
PanWeb	It is a graphical interface of pangenome analysis generated from PGAP software.	http://www.computationalbiology.ufpa.br/panweb (accessed on 17 September 2021)	[67]
PanX	This tool identifies orthologous gene clusters in pangenomes, visualization, presence/absence pattern and identify SNPs	https://pangenome.org/ (accessed on 17 September 2021)	[68]

Table 1. *Cont.*

Software/Tool	Description/Role	URL Link	References
PGAdb-Builder	This is used to constructs a pangenome allele database (PGAdb).	http://wgmlstdb.imst.nsysu.edu.tw/ (accessed on 17 September 2021)	[69]
PGAP-X	Genome diversity and visualize genome structure and gene content to understand the evolution.	http://pgapx.ybzhao.com/ (accessed on 17 September 2021)	[22]
Piggy	Detection of highly divergent ("switched") intergenic regions (IGRs) upstream of genes in pangenome	https://github.com/harry-thorpe/piggy (accessed on 17 September 2021)	[70]
Pyseer	This is helpful in genome-wide association studies in the microbes to identify potential genetic variation.	https://github.com/mgalardini/pyseer (accessed on 17 September 2021)	[71]
Seq-seq-pan	For sequential alignment of sequences to build a pangenome data structure and a whole-genome alignment.	https://gitlab.com/rki_bioinformatics (accessed on 17 September 2021)	[72]
Spine and AGEnt	Spine, find core-genome from a group of genomic sequences and AGEnt, find the accessory genome in draft genomic sequences	http://vfsmspineagent.fsm.northwestern.edu/index_age.html (accessed on 17 September 2021)	[73]
BPGA	Pangenome profile analysis, pangenome sequence extraction, exclusive gene family analysis, atypical GC content analysis and species phylogenetic analysis.	http://sourceforge.net/projects/bpgatool/ (accessed on 17 September 2021)	[74]
BGDMdocker	For pangenome analysis, visualization, clustering and genome annotation.	https://www.docker.com/whatisdocker (accessed on 17 September 2021)	[75]
PAN2HGENE	To identify new products, resulting in altering the α value behavior in the pangenome without altering the original genomic sequence.	https://sourceforge.net/projects/pan2hgene-software (accessed on 17 September 2021)	[76]
PATO	Core-genome and accessory genome identification and help to characterize population structure, annotate pathogenic features and create gene sharedness networks.	https://github.com/irycisBioinfo/PATO (accessed on 17 September 2021)	[77]
Panakeia	It analyses synteny and multiple structural patterns of the pangenome, help for biological diversity and evolution studied.	https://github.com/BioSina/Panakeia (accessed on 17 September 2021)	[78]
HUPAN	It is developed for pangenome analysis for humans/mammals	http://cgm.sjtu.edu.cn/hupan/ (17 September 2021) and https://github.com/SJTU-CGM/HUPAN (accessed on 17 September 2021)	[79]

7. Applications of Pangenomics in Biological Research

The concept of pangenome started with bacterial species and was extended to other organisms, including crops later on with diverse applications. Pangenomics has facilitated applied research directly in some cases by identifying industrially relevant microbial resources and fostering the design of vaccines. It has helped in the identification of novel genes for agriculturally important traits in different crops. In the following section, the role of pangenomics in advancing the basic biological research eventually leading to real-life applications has been discussed.

7.1. Finding Novel Genes

The advent of massively parallel sequencing at relatively cheaper costs has facilitated the large-scale generation of genome sequence information. However, computational algorithms are required to derive meaningful inferences from these huge datasets. The urgency of robust algorithms is greater for pangenome studies, as it does not discard any data, rather attempts to map the DNA sequences obtained to relevant genomic locations in already sequenced strains. Pangenomic comparison often relies on the relationship of homology between newly generated DNA sequences and those already available in the repositories. The genes arising from a speciation event, are termed orthologs whereas paralogs result originate from DNA duplication events. Bosi et al. [80] have reviewed the concept of homology, particularly orthology, for data mining of pangenomic sequences. The authors describe Bidirectional Best Hits (BBH) as a simple and fast approach to identifying orthologous genes. This approach relies on the assumption that orthologous genes are more similar to each other than they are to any other sequences in the genome. Databases like Clusters of Orthologous Groups of proteins (COGs) (accessed on 16 March 2022) and Kyoto

Encyclopedia of Genes and Genomes (KEGG) (accessed on 16 March 2022) are used to define orthology relationships and categorize pangenome into the core- and accessory-genomes. Pangenomics adds information to public sequence repositories. Novel genes, elucidated by pangenomic analysis have high potential in biotechnology applications. Othoum et al. [81] describe the use of pan-genomic analysis to mine genetic regions capable of biosynthetic capabilities in *Virgibacillus* strains. The novel biosynthetic capabilities carry industrial importance, especially as pharmaceuticals. The study revealed the involvement of genes encoding for protein classes like non-ribosomal peptide synthetases (NRPS), polyketide synthetases (PKS), ribosomally synthesized and post-translationally modified peptides (RiPPs), etc., in anti-tumor, antimicrobial and immunosuppressive properties. Analysis of nine *Virgibacillus* strains showed that most genes encoding for NRPS are present in genomic islands, predicted to have been transferred by lateral DNA flow. The authors deciphered two strains, *V. dokdonensis Bac330* and *Virgibacillus Bac332* to be important and containing more modular genes as compared to other species. The two strains being isolated from Red Sea mangrove mud may have attained a higher proportion of biosynthetic capabilities as a result of the potential environmental stress encountered. Hence, apart from elucidating novel genomic islands, gene findings also led to the identification of potential industrially important strains. In another study, ref. [82] demonstrated the potential of two more Red Sea strains *B. paralicheniformis* (*Bac48*) and *B. halosaccharovorans* (*Bac94*), which are capable of secreting twice as much protein as the model strain *B. subtilis* 168. The strain *Bac94* was shown to be enriched with genes associated with the Tat and Sec protein secretion system, hence making it a promising system for recombinant protein expression work.

7.2. Revealing Niche-Specific Fitness

The genes in the accessory genome are often linked to traits that influence an organism's ability to migrate to a new niche. For example, unlike autochthonous organisms that colonize the intestine permanently, some lactobacilli are not capable of permanently residing in the intestine. These bacteria then reside in the gut for a shorter time as commensals. The organisms in a commensal relationship neither harm nor benefit each other. *Lactobacillus rhamnosus* is a good example of commensalism. It is used as a pro-biotic and has great potential in functional foods. Implications of surface-exposed proteins for niche-specific fitness were evident in *L. rhamnosus* based on a pangenomic analysis using genomes of 13 strains isolated from various origins [83]. An operon spaCBA that encodes SpaCBA-called pili has been implicated to be essential for niche adaptability of *L. rhamnosus* strains. The pilli enables the microbe to have a mucoadhesive phenotype. This phenotype is a rare and uncommon occurrence in *L. rhamnosus*. The above finding also explains why some strains can adapt to particular niches better than others.

Pangenomics analysis can elucidate genes that impart niche specificity. McInerney et al. [84] have argued that extensive pangenomes in prokaryotes are the result of adaptive evolution, which contributes to the fitness of an organism. By linking organism's lifestyle with the proportion of the core genome in the pangenome, the authors presented a perspective that most of the accessory genes appear to confer capabilities that are advantageous for the fitness of the organism. However, ref. [85] contested the above notion. The authors argued that accessory genes could have deleterious effects. As such, the accessory genome is not composed of genes that only confer a fitness advantage. For accessory genes with deleterious effects, there is a selection to lower the uptake. Such genes, even when taken up, are consequently lost from the genome. Earlier, ref. [86] argued that gene loss events could lead to an underestimation of the core genome in some cases. Together, the above studies indicate that niche fitness is not the only function of the accessory genome. Livingstone et al. [87] studied the pangenome of *Corallococcus*, which is an abundant genus of predatory soil myxobacteria. Its accessory genome was found to encode for proteins that are involved in predatory defense mechanisms or the generation of secondary metabolites. This also makes the genus a promising candidate for novel bioactive compounds with antimicrobial properties like corallopyronin, corallorazine, and

coralmycin. Pangenome serves a broad role, including host-pathogen interactions and predation in microbes, speciation, and contributing to domestication and heterosis in the case of plants.

7.3. Evolution, Domestication and Breeding History

Crop domestication commenced around 10,000 years ago in the Fertile Crescent. Attempts to modify wild crops according to human needs have led to marked changes in the crucial plant phenotypes, referred to as domestication traits. Evidence suggests that crop domestication has been associated with trade-offs that reduce the fitness of the crops due to the accumulation of deleterious genetic variations. Availability of the sequence information on multiple genomes provides an enormous opportunity to refine the crop domestication and breeding history. For instance, a large-scale analysis of the genome-wide diversity patterns and domestication-associated loci in rice suggested the first domestication of *Oryza sativa* ssp. *japonica* from the *O. rufipogon* whereas the *O. sativa* ssp. *indica* resulted from the cross involving *japonica* and local wild rice [88]. Zhang and colleagues [89] studied 10 species of poplar to understand their evolutionary history. The authors found substantial DNA variations between the species and reported that the major differences among the poplar species were attributed to R genes for disease resistance with loss-of-function mutations, and the genes for self-incompatibility. Due to the comprehensive coverage of the genome of a particular species, pangenomics is a promising tool for phylogenetic analysis to understand evolutionary dynamics, as exemplified by a recent pangenome analysis to understand eggplant domestication [90]. For studies on evolution, a key question has been to evaluate the number of genomes to sequence for consideration in the analysis of evolution. Bacteria tend to have open pangenomes due to higher gene flow between them. Several pangenomes in different crops have been constructed, and were enlisted in recent studies. [77,91–94] reviewed the relation between pangenome, machine learning, and genomic selection in plants.

7.4. Elucidating Host-Pathogen Interactions

Pangenomics has been used to understand the genes coding for the pathogenicity repertoire of pathogens and how they interact with host systems. Hu et al. [95] describe that comparative genomics has been used to understand strain-to-strain variation and estimate differences between pathogens and their near neighbor, non-pathogenic organisms. The interaction between host and pathogen is ever-evolving, with both adapting ways to ensure their survival in the antagonistic interaction. An open pangenome, where a species can acquire new genes, which could, among other factors, be due to its particular lifestyle, carries the potential to influence host–pathogen interactions in novel ways. DNA transfers and inherent genomic diversity can both lead to an increase in the repertoire of genes responsible for pathogenesis. Casa-Esperón et al. [96] have reviewed the role of horizontal DNA transfer in the evolution of host–pathogen interactions. Perna et al. [97] compared non-pathogenic *Escherichia coli* strain K12 with pathogenic *E. coli* O157: H7. The authors concluded that the phage-mediated horizontal flow of DNA was responsible for the pathogenicity of *E. coli* O157: H7. Pangenomic analysis [98] has concluded that considerable genomic diversity exists between *E. coli* species, besides phage-mediated transfers. Analysis of 17 *E. coli* strains revealed that in *E. coli*, while the core genome consists of approximately 2200 genes, the accessory or dispensable genome consists of about 13,000 genes. Thus, at the time of the study, *E. coli* dispensable genome represented a staggering proportion (~83%) of its pangenome. Hence, pangenomics analysis is essential to understand the different ways by which an organism, especially with an open pangenome, can interact with its host. Badet et al. [99] studied the pangenome of the fungal pathogen of wheat, *Zymoseptoria tritici*, taking 19 samples from six continents. Major chromosomal rearrangements that include presence/absence variation were observed in the fungal strains. The authors reported that the dispensable genome contains pathogenesis-related genes, which encode proteins responsible for plant tissue degradation and manipulation

of host functions. Plissonneau et al. [100] also reported that in *Z. tritici*, the dispensable genome largely accounts for its adaptive evolution. A similar study identified pangenome for host–pathogen in *Pantoeastewartia* subsp. *indologenes* (Psi) and foxtail millet and pearl millet [101]. This way, pangenomics add a new dimension to the study of host–pathogen interactions by moving beyond the historical events of lateral DNA transfer and along with the former, focusing also on the pan-genetic complement for understanding essential genes related to pathogenicity.

7.5. Explaining Heterosis

Large SVs influence many phenomena including metabolism, flowering, nutrient use efficiency, and stress response [27,102–104]. It was earlier hypothesized that CNVs and PAVs may not result in large phenotypic differences as many genes in plants are organized in the form of a gene family [105]. Hence, there is "partial redundancy of the function", whereby loss-of-function or altered function resulting from CNVs/PAVs in one gene would be partially offset and compensated by other genes of the family.

Given this understanding, gene function was conceptualized in the form of a "functional block", whereby each gene product contributes a certain function to the concerned phenotype. The authors explained that although gene function loss in one family may not result in much difference, loss or alteration of function of some genes in many gene families can lead to decreased vigor. In a hybrid, this effect will be partially nullified, explaining the "hybrid vigor". Pangenomics can play an important role in unraveling gene members and families contributing to heterosis, according to the proposed model (Figure 5). Thus, it is clear that based on the model proposed [105], a new gene and variant finding is essential to explaining and utilizing heterosis for crop improvement. Single reference genomes cannot be used for novel gene discovery. Zhao et al. [13] utilized divergent species of rice *O. sativa* and relative *O. rufipogon* to map the rice pangenome. Based on an analysis of 1529 rice accessions, the divergent 57 accessions along with nine popular cultivars were sequenced to assemble the rice pangenome. Extensive PAVs were found in rice accessions based on the assembled pangenome, which is a useful resource for further studies. Hirsch et al. [15] analyzed 503 maize inbred lines for understanding developmental transitions from juvenile to vegetation and then to reproduction. The authors found that 16.4% of representative transcript assemblies were observed in all lines, while 82.7% expressed in a subset of lines. This shows the limitation of using a single genome for transcript mapping and reveals the importance of pangenomics for molecular characterization of heterosis phenomenon.

7.6. Facilitating Taxonomic Identification

Pangenomics is a useful tool to identify a species as well as gain a detailed understanding of its lifestyle habit and habitat. Species identification is important for various reasons, including diagnostics. Rouli et al. [106] describe an interesting case whereby the distinct morphological and biochemical features of *E. coli* and *Shigella* species led to their differential categorization. However, despite a myriad of differences, the mechanism of pathogenicity in *Shigella* and *E. coli*, particularly the enterohaemorrhagic invasive *E. coli* EIEC is identical as both enter epithelial cells to cause local inflammation leading to ulceration of the colon. The authors argue that *Shigella* and *E. coli* should be grouped together and that their distinction and individualization were due to medical diagnosis. Indeed, based on pangenomics studies, and based on cluster analysis using Clusters of Orthologous Groups of proteins (COGs) and Kyoto Encyclopedia of Genes and Genomes (KEGG), *Shigella* was found to be distributed among the different *E. coli* clusters. The principal component analysis revealed two clusters, both of which contain a mix of *E. coli* and *Shigella* species. It has also been observed that the lifestyle of a micro-organism also influences the type of genome that it contains. Microbes can exist in two states of lifestyle: allopatry (living alone in an environmental niche) and sympatry (living in a large community in an environmental niche). The allopatric microbes tend to have closed pangenomes, while those in sympatric lifestyles, have open pangenomes. Sympatric microbes gain genes to survive in diverse

niches, while allopatric microbes face gene loss. This indicates the complementation of gene gain and loss events for the pangenome. Hence, the nature of pangenome can also indicate the lifestyle habit of an organism. The pangenome nature reveals intricate details of a microbe's interaction with the environment. It is intriguing to note that *B. anthracis* is a soil bacterium but still contains a closed pangenome. This is because it stays dormant in the form of a spore with minimal interactions with outside. On the other hand, *Legionella pneumonia* stays intracellularly in amoeba, but is in a metabolically active state and thus possesses an open pangenome.
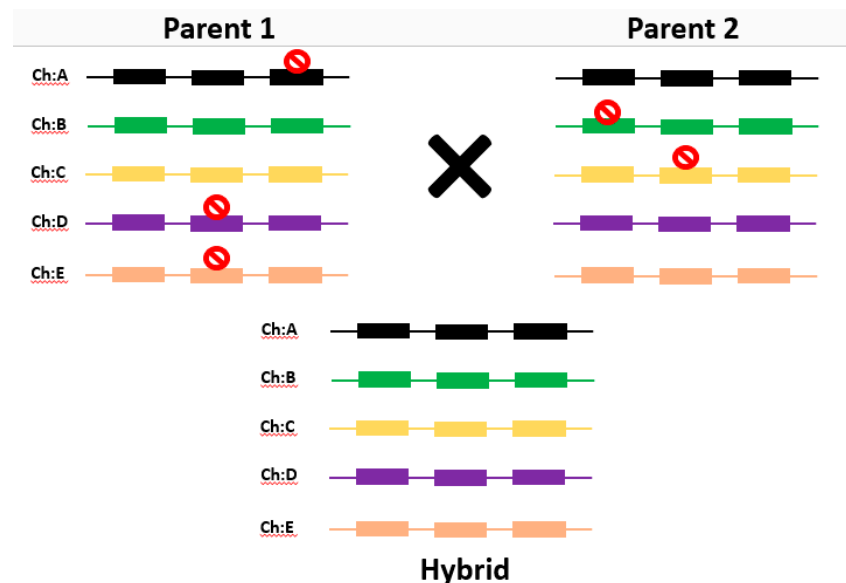

Figure 5. A model of heterosis proposed by Swanson-Wagner et al. [105]. Bars represent genes. Three genes are considered in each hypothetical gene family, situated on different chromosomes. 🚫 represents "functional block" leading to null or altered protein function. In a real scenario, accumulation of a similar effect with many gene families leads to reduced vigor in inbreeds and heterosis in hybrid. Pangenomics can help to unravel heterosis in a phenotypic trait by discovering new gene variants.

7.7. Strengthening Proteogenomics

In terms of functional annotation, a pangenome may also be looked upon in the form of pan-metabolome (complement of all metabolic reactions in a species), pan-regulon (collection of co-expressed genes), resistome (repertoire of all genes encoding for proteins that confer resistance to other organisms), etc. Pangenomes further improve our understanding of microbial species and can be utilized in proteogenomics-based identification of microbial flora in diverse biological samples. In this approach, the sample to be identified is taken, proteins are isolated and digested to result in a mix of peptides, which is specific for a particular species. This is referred to as Peptide Mass Fingerprint. Mass spectrometry has been used in the identification of microbial strains by analyzing the peptide fingerprint patterns of the sample proteins to proteomics databases. With the addition of new strains in the repositories, data mining becomes an issue as computational search becomes more and more demanding. Among the various techniques to accomplish peptide fingerprint matching, de Souza et al. [107] utilized computational algorithms to reduce the redundancy of protein databases of related bacterial species. This was denoted as MSMSpdbb (Multi-Strain Mass Spectrometry Prokaryotic DataBase Builder) (The Gade Institute, University of Bergen, Haukeland University Hospital, Bergen, Norway) approach. Given the increasingly more pangenomic data being generated, one way to allow robust microbial identification is to create customized databases where peptides from homologous proteins are not present in all the related bacterial strains.

Pangenomics keeps on adding new genomic information of new species. The information about Open Reading Frames (ORF) in the new genes, new features of homologous genes like differences in translational starting site (TSS) can be supplemented to existing protein databases to allow robust identification of biological samples. Caputo et al. [108] reported that while pangenomes identify novel strains like *Akkermansia muciniphila*, *Microvirga massiliensis*, etc., analysis of data like a discontinuity in the core/pangenome ratio can also indicate the presence of novel species. de Souza et al. [107] reported that concatenating protein sequences obtained from the pangenomics analysis of multiple organisms contained in public repositories leads to a thoroughly covered microbial sequence database, for sample identification. Thus, pangenomics contributes to microbial identification in multiple ways, including finding new strains, indicating the presence of other strains, and complementing protein sequence databases in public repositories.

7.8. Advancing Reverse Vaccinology

Vaccine development has witnessed paradigm shifts in the genomics and pangenomics era. Conventional vaccinology requires cultivable microorganisms, purification of components responsible for immunogenicity, immunogenicity testing in animal models, and the development of vaccines. However, there are disadvantages, as vaccines cannot be fabricated for non-culturable pathogens and the correlation of animal models with human subjects may not be high in certain cases. Moreover, only abundantly expressed antigens are generally tested. With the availability of pan-genomic sequences, virtually all antigens can be tested. The development of vaccines using genomic sequences has been referred to as reverse vaccinology, as its operating procedure is essentially the reverse of the steps taken in conventional vaccine development. There is no issue of the non-culturability of a pathogen for which genomic sequence has been made available. The proteins involved in host–pathogen interaction can be utilized for the prioritization of targets for vaccine development. The vast repertoire of applicable gene products, as revealed by pangenomics, have immense potential to develop specific vaccines for various pathogens and subtypes. Naz et al. [109] demonstrated the use of pangenomic data for vaccine development. The authors have designed a pipeline to scan the entire genomic complement of a pathogen to design effective vaccines against it. This approach termed as Pangenome-Reverse Vaccinology is a cost-effective technique to overcome the limitations associated with conventional vaccine development, by employing pangenomic DNA sequences. Dalsass et al. [110] reviewed the open-source platforms for bacterial vaccine antigen discovery. The shortcomings of the current prediction pipelines were highlighted. There is a need to expand the curation of protein datasets by incorporating negative results and inclusion of high-throughput secondary structure prediction methods like Circular Dichroism spectroscopy. This would enhance the prediction power for better translation in wet laboratory results.

8. Conclusions and Future Perspectives

Pangenomics has augmented both basic and applied research. It has contributed to a variety of interesting areas such as identification of industrially relevant microbial resources, vaccine designing, refining evolution and taxonomic identification, proteogenomics, deeper knowledge about host–pathogen interaction, and genetic makeup of important agronomic phenotypes. There remains an immense scope in pangenomics for understanding complex biological phenomena. Further refinements in core and accessory genome characterization leverage understanding of crop adaptation. It will also aid the discovery of new variations through novel haplotypes and functional molecular markers which will facilitate trait introgression or genomics-assisted breeding. This is expected to result in better utilization of heterosis for breeding improvement. Further, the development of portable sequencing technologies like Oxford Nanopore and integration of more robust and open-source high throughput visualization tools, along with efficient storage and retrieval of huge pangenomics data, would lead to progressive deployment of pangenomics for addressing different scientific issues in a cost-effect way by researchers.

References

1. Bohra, A.; Jha, U.C.; Godwin, I.; Varshney, R.K. Genomic Interventions for Sustainable Agriculture. *Plant Biotechnol. J.* **2020**, *18*, 2388–2405. [CrossRef] [PubMed]
2. Heather, J.M.; Chain, B. The Sequence of Sequencers: The History of Sequencing DNA. *Genomics* **2016**, *107*, 1–8. [CrossRef] [PubMed]
3. Varshney, R.K.; Nayak, S.N.; May, G.D.; Jackson, S.A. Next-Generation Sequencing Technologies and Their Implications for Crop Genetics and Breeding. *Trends Biotechnol.* **2009**, *27*, 522–530. [CrossRef]
4. Varshney, R.K.; Bohra, A.; Yu, J.; Graner, A.; Zhang, Q.; Sorrells, M.E. Designing Future Crops: Genomics-Assisted Breeding Comes of Age. *Trends Plant Sci.* **2021**, *26*, 631–649. [CrossRef] [PubMed]
5. Tettelin, H.; Masignani, V.; Cieslewicz, M.J.; Donati, C.; Medini, D.; Ward, N.L.; Angiuoli, S.V.; Crabtree, J.; Jones, A.L.; Durkin, A.S.; et al. Genome Analysis of Multiple Pathogenic Isolates of *Streptococcus Agalactiae*: Implications for the Microbial "pangenome". *Proc. Natl. Acad. Sci. USA* **2005**, *102*, 13950–13955. [CrossRef] [PubMed]
6. Golicz, A.A.; Batley, J.; Edwards, D. Towards Plant Pangenomics. *Plant Biotechnol. J.* **2016**, *14*, 1099–1105. [CrossRef] [PubMed]
7. Vernikos, G.; Medini, D.; Riley, D.R.; Tettelin, H. Ten Years of Pangenome Analyses. *Curr. Opin. Microbiol.* **2015**, *23*, 148–154. [CrossRef] [PubMed]
8. Medini, D.; Donati, C.; Tettelin, H.; Masignani, V.; Rappuoli, R. The microbial pan-genome. *Curr. Opin. Genet. Dev.* **2005**, *15*, 589–594. [CrossRef]
9. Khan, A.W.; Garg, V.; Roorkiwal, M.; Golicz, A.A.; Edwards, D.; Varshney, R.K. Super-Pangenome by Integrating the Wild Side of a Species for Accelerated Crop Improvement. *Trends Plant Sci.* **2020**, *25*, 148–158. [CrossRef] [PubMed]
10. Saxena, R.K.; Edwards, D.; Varshney, R.K. Structural Variations in Plant Genomes. *Brief. Funct. Genom.* **2014**, *13*, 296–307. [CrossRef] [PubMed]
11. Carlos Guimaraes, L.; Benevides de Jesus, L.; Vinicius Canario Viana, M.; Silva, A.; Thiago Juca Ramos, R.; de Castro Soares, S.; Azevedo, V. Inside the Pangenome—Methods and Software Overview. *Curr. Genom.* **2015**, *16*, 245–252. [CrossRef] [PubMed]
12. Marschall, T.; Marz, M.; Abeel, T.; Dijkstra, L.; Dutilh, B.E.; Ghaffaari, A.; Kersey, P.; Kloosterman, W.P.; Mäkinen, V.; Novak, A.M.; et al. Computational Pangenomics: Status, Promises and Challenges. *Brief. Bioinform.* **2018**, *19*, 118–135. [CrossRef]
13. Zhao, Q.; Feng, Q.; Lu, H. Pangenome Analysis Highlights the Extent of Genomic Variation in Cultivated and Wild Rice. *Nat. Genet.* **2018**, *50*, 278–284. [CrossRef] [PubMed]
14. Golicz, A.A.; Bayer, P.E.; Barker, G.C.; Edger, P.P.; Kim, H.; Martinez, P.A.; Chan, C.K.K.; Severn-Ellis, A.; McCombie, W.R.; Parkin, I.A.; et al. The Pangenome of an Agronomically Important Crop Plant Brassica Oleracea. *Nat. Commun.* **2016**, *7*, 1–8. [CrossRef]
15. Hirsch, C.N.; Foerster, J.M.; Johnson, J.M.; Sekhon, R.S.; Muttoni, G.; Vaillancourt, B.; Peñagaricano, F.; Lindquist, E.; Pedraza, M.A.; Barry, K. Insights into the Maize Pangenome and Pantranscriptome. *Plant Cell* **2014**, *26*, 121–135. [CrossRef] [PubMed]
16. Li, Y.H.; Zhou, G.; Ma, J.; Jiang, W.; Jin, L.G.; Zhang, Z.; Guo, Y.; Zhang, J.; Sui, Y.; Zheng, L.; et al. De Novo Assembly of Soybean Wild Relatives for Pangenome Analysis of Diversity and Agronomic Traits. *Nat. Biotechnol.* **2014**, *32*, 1045–1052. [CrossRef] [PubMed]
17. Vernikos, G.S. A Review of Pangenome Tools and Recent Studies. In *The Pangenome: Diversity, Dynamics and Evolution of Genomes*; The Pangenome; Springer: Cham, Switzerland, 2020.
18. Baker, M. Structural Variation: The Genome's Hidden Architecture. *Nat. Methods* **2012**, *9*, 133–137. [CrossRef] [PubMed]
19. Springer, N.M.; Ying, K.; Fu, Y.; Ji, T.; Yeh, C.T.; Jia, Y.; Wu, W.; Richmond, T.; Kitzman, J.; Rosenbaum, H.; et al. Maize Inbreds Exhibit High Levels of Copy Number Variation (CNV) and Presence/Absence Variation (PAV) in Genome Content. *PLoS Genet.* **2009**, *5*, e1000734. [CrossRef] [PubMed]
20. Ellis, J.; Dodds, P.; Pryor, T. Structure, Function and Evolution of Plant Disease Resistance Genes. *Curr. Opin. Plant Biol.* **2000**, *3*, 278–284. [CrossRef]

21. Tao, Y.; Zhao, X.; Mace, E.; Henry, R.; Jordan, D. Exploring and Exploiting Pangenomics for Crop Improvement. *Mol. Plant* **2019**, *12*, 156–169. [CrossRef] [PubMed]

22. Ashikawa, I.; Hayashi, N.; Yamane, H.; Kanamori, H.; Wu, J.; Matsumoto, T.; Ono, K.; Yano, M. Two Adjacent Nucleotide-Binding Site–Leucine-Rich Repeat Class Genes Are Required to Confer Pikm-Specific Rice Blast Resistance. *Genetics* **2008**, *180*, 2267–2276. [CrossRef] [PubMed]

23. Lin, Z.; Li, X.; Shannon, L.M.; Yeh, C.T.; Wang, M.L.; Bai, G.; Peng, Z.; Li, J.; Trick, H.N.; Clemente, T.E.; et al. Parallel domestication of the Shattering. *Nat. Genet.* **2012**, *44*, 720–724. [CrossRef] [PubMed]

24. Yang, Q.; Li, Z.; Li, W.; Ku, L.; Wang, C.; Ye, J.; Li, K.; Yang, N.; Li, Y.; Zhong, T.; et al. CACTA-like Transposable Element in ZmCCT Attenuated Photoperiod Sensitivity and Accelerated the Post Domestication Spread of Maize. *Proc. Natl. Acad. Sci. USA* **2013**, *110*, 16969–16974. [CrossRef] [PubMed]

25. Fu, H.; Dooner, H.K. Intraspecific Violation of Genetic Colinearity and Its Implications in Maize. *Proc. Natl. Acad. Sci. USA* **2002**, *99*, 7578–9573. [CrossRef] [PubMed]

26. Qin, P.; Lu, H.; Du, H.; Wang, H.; Chen, W.; Chen, Z.; He, Q.; Ou, S.; Zhang, H.; Li, X.; et al. Pangenome Analysis of 33 Genetically Diverse Rice Accessions Reveals Hidden Genomic Variations. *Cell* **2021**, *184*, 3542–3558.e16. [CrossRef]

27. Schatz, M.C.; Maron, L.G.; Stein, J.C.; Wences, A.H.; Gurtowski, J.; Biggers, E.; Lee, H.; Kramer, M.; Antoniou, E.; Ghiban, E.; et al. Whole Genome de Novo Assemblies of Three Divergent Strains of Rice, *Oryza Sativa*, Document Novel Gene Space of Aus and Indica. *Genome Biol.* **2014**, *15*, 506. [PubMed]

28. Montenegro, J.D.; Golicz, A.A.; Bayer, P.E.; Hurgobin, B.; Lee, H.; Chan, C.K.K.; Visendi, P.; Lai, K.; Doležel, J.; Batley, J.; et al. The Pangenome of Hexaploid Bread Wheat. *Plant J.* **2017**, *90*, 1007–1013. [CrossRef]

29. Hurgobin, B.; Golicz, A.A.; Bayer, P.E.; Chan, C.K.K.; Tirnaz, S.; Dolatabadian, A.; Schiessl, S.V.; Samans, B.; Montenegro, J.D.; Parkin, I.A.; et al. Homoeologous Exchange Is a Major Cause of Gene Presence/Absence Variation in the Amphidiploid Brassica Napus. *Plant Biotechnol. J.* **2018**, *16*, 1265–1274. [CrossRef] [PubMed]

30. Hu, Z.; Sun, C.; Lu, K.C.; Chu, X.; Zhao, Y.; Lu, J.; Shi, J.; Wei, C. EUPAN Enables Pangenome Studies of a Large Number of Eukaryotic Genomes. *Bioinformatics* **2017**, *33*, 2408–2409. [CrossRef] [PubMed]

31. Jayakodi, M.; Schreiber, M.; Stein, N.; Mascher, M. Building Pangenome Infrastructures for Crop Plants and Their Use in Association Genetics. *DNA Res.* **2021**, *28*, dsaa030. [CrossRef]

32. Wang, W.; Mauleon, R.; Hu, Z. Genomic Variation in 3010 Diverse Accessions of Asian Cultivated Rice. *Nature* **2018**, *557*, 43–49. [CrossRef] [PubMed]

33. Gan, X.; Stegle, O.; Behr, J.; Steffen, J.G.; Drewe, P.; Hildebrand, K.L.; Lyngsoe, R.; Schultheiss, S.J.; Osborne, E.J.; Sreedharan, V.T.; et al. Multiple Reference Genomes and Transcriptomes for Arabidopsis Thaliana. *Nature* **2011**, *477*, 419–423. [CrossRef] [PubMed]

34. Zapata, L.; Ding, J.; Willing, E.M.; Hartwig, B.; Bezdan, D.; Jiao, W.B.; Patel, V.; James, G.V.; Koornneef, M.; Ossowski, S.; et al. Chromosome-Level Assembly of Arabidopsis Thaliana Ler Reveals the Extent of Translocation and Inversion Polymorphisms. *Proc. Natl. Acad. Sci. USA* **2016**, *113*, E4052–E4060. [CrossRef] [PubMed]

35. Adams, K.L.; Wendel, J.F. Polyploidy and Genome Evolution in Plants. *Curr. Opin. Plant Biol.* **2005**, *8*, 135–141. [CrossRef] [PubMed]

36. Cao, M.D.; Nguyen, S.H.; Ganesamoorthy, D.; Elliott, A.G.; Cooper, M.A.; Coin, L.J. Scaffolding and Completing Genome Assemblies in Real-Time with Nanopore Sequencing. *Nat. Commun.* **2017**, *8*, 14515. [CrossRef] [PubMed]

37. Parra, G.; Bradnam, K.; Korf, I. CEGMA: A Pipeline to Accurately Annotate Core Genes in Eukaryotic Genomes. *Bioinformatics* **2007**, *23*, 1061–1067. [CrossRef] [PubMed]

38. Simão, F.A.; Waterhouse, R.M.; Ioannidis, P.; Kriventseva, E.V.; Zdobnov, E.M. BUSCO: Assessing Genome Assembly and Annotation Completeness with Single-Copy Orthologs. *Bioinformatics* **2015**, *31*, 3210–3212. [CrossRef]

39. Tranchant-Dubreuil, C.; Rouard, M.; Sabot, F. Plant Pangenome: Impacts on Phenotypes and Evolution. *Annu. Rev. Plant Biol.* **2018**, *15*, 453–478.

40. Xiao, J.; Zhang, Z.; Wu, J.; Yu, J. A Brief Review of Software Tools for Pangenomics. *Genom. Proteom. Bioinform.* **2015**, *13*, 73–76. [CrossRef] [PubMed]

41. Lerat, E.; Daubin, V.; Moran, N.A. From Gene Trees to Organismal Phylogeny in Prokaryotes: The Case of the Gammaproteo Bacteria. *PLoS Biol.* **2003**, *1*, 101–109. [CrossRef] [PubMed]

42. Laing, C.; Buchanan, C.; Taboada, E.N.; Zhang, Y.; Kropinski, A.; Villegas, A.; Thomas, J.E.; Gannon, V.P. Pangenome Sequence Analysis Using Panseq: An Online Tool for the Rapid Analysis of Core and Accessory Genomic Regions. *BMC Bioinform.* **2010**, *11*, 461. [CrossRef] [PubMed]

43. Lukjancenko, O.; Thomsen, M.C.; Voldby, L.M.; Ussery, D.W. PanFunPro: Pangenome Analysis Based on FUNctionalPROfiles. *F1000Research* **2013**, *2*, 265. [CrossRef]

44. Contreras-Moreira, B.; Vinuesa, P. GET_HOMOLOGUES, a Versatile Software Package for Scalable and Robust Microbial Pangenome Analysis. *Appl. Environ. Microbiol.* **2013**, *79*, 7696–7701. [CrossRef] [PubMed]

45. Benedict, M.N.; Henriksen, J.R.; Metcalf, W.W.; Whitaker, R.J.; Price, N.D. ITEP: An Integrated Toolkit for Exploration of Microbial Pangenomes. *BMC Genom.* **2014**, *15*, 8. [CrossRef]

46. Zhao, Y.; Jia, X.; Yang, J.; Ling, Y.; Zhang, Z.; Yu, J.; Wu, J.; Xiao, J. PanGP: A Tool for Quickly Analyzing Bacterial Pangenome Profile. *Bioinformatics* **2014**, *30*, 1297–1299. [CrossRef] [PubMed]

47. Zhao, Y.; Wu, J.; Yang, J.; Sun, S.; Xiao, J.; Yu, J. PGAP: Pangenomes Analysis Pipeline. *Bioinformatics* **2012**, *28*, 416–418. [CrossRef] [PubMed]

48. Brittnacher, M.J.; Fong, C.; Hayden, H.S.; Jacobs, M.A.; Radey, M.; Rohmer, L. PGAT: A Multistrain Analysis Resource for Microbial Genomes. *Bioinformatics* **2011**, *27*, 2429–2430. [CrossRef]

49. Blom, J.; Albaum, S.P.; Doppmeier, D. EDGAR: A Software Framework for the Comparative Analysis of Prokaryotic Genomes. *BMC Bioinform.* **2009**, *10*, 1–14. [CrossRef] [PubMed]

50. Snipen, L.; Liland, K.H. Micropan: An R-Package for Microbial Pangenomics. *BMC Bioinform.* **2015**, *16*, 79. [CrossRef]

51. Marcus, S.; Lee, H.; Schatz, M.C. Split MEM: A Graphical Algorithm for Pangenome Analysis with Suffix Skips. *Bioinformatics* **2014**, *30*, 3476–3483. [CrossRef] [PubMed]

52. Ozer, E.A. AGE: A Tool for Clustering and Distribution Analysis of Bacterial Accessory Genomic Elements. *BMC Bioinform.* **2018**, *19*, 150. [CrossRef] [PubMed]

53. Thakur, S.; Guttman, D.S. A De-Novo Genome Analysis Pipeline (DeNoGAP) for Large-Scale Comparative Prokaryotic Genomics Studies. *BMC Bioinform.* **2016**, *17*, 260. [CrossRef]

54. Treangen, T.J.; Ondov, B.D.; Koren, S.; Phillippy, A.M. The Harvest Suite for Rapid Core-Genome Alignment and Visualization of Thousands of Intraspecific Microbial Genomes. *Genome Biol.* **2014**, *15*, 524. [CrossRef] [PubMed]

55. Sahl, J.W.; Caporaso, J.G.; Rasko, D.A.; Keim, P. The Large-Scale Blast Score Ratio (LS-BSR) Pipeline: A Method to Rapidly Compare Genetic Content between Bacterial Genomes. *PeerJ* **2014**, *2014*, e332. [CrossRef]

56. Kulsum, U.; Kapil, A.; Singh, H.; Kaur, P. NGSPanPipe: A Pipeline for Pangenome Identification in Microbial Strains from Experimental Reads. *Adv. Exp. Med. Biol.* **2018**, *1052*, 39–49. [PubMed]

57. Clarke, T.H.; Brinkac, L.M.; Inman, J.M.; Sutton, G.; Fouts, D.E. PanACEA: A Bioinformatics Tool for the Exploration and Visualization of Bacterial Pan-Chromosomes. *BMC Bioinform.* **2018**, *19*, 1–6. [CrossRef] [PubMed]

58. Ernst, C.; Rahmann, S. PanCake: A Data Structure for Pangenomes. German Conference on Bioinformatics. *Schloss Dagstuhl-Leibniz-Zent. Inform. Ger. Dagstuhl Publ.* **2013**, 35–45. [CrossRef]

59. Yuvaraj, I.; Sridhar, J.; Michael, D.; Sekar, K. PanGeT: Pangenomics Tool. *Gene* **2017**, *600*, 77–84. [CrossRef]

60. Chaudhari, N.M.; Gautam, A.; Gupta, V.K.; Kaur, G.; Dutta, C.; Paul, S. PanGFR-HM: A Dynamic Web Resource for Pan-Genomic and Functional Profiling of Human Microbiome with Comparative Features. *Front. Microbiol.* **2018**, *9*, 2322. [CrossRef]

61. Abudahab, K.; Prada, J.M.; Yang, Z.; Bentley, S.D.; Croucher, N.J.; Corander, J.; Aanensen, D.M. PANINI: Pangenome Neighbour Identification for Bacterial Populations. *Microb. Genom.* **2018**, *5*, 4. [CrossRef] [PubMed]

62. Santos, A.R.; Barbosa, E.; Fiaux, K.; Zurita-Turk, M.; Chaitankar, V.; Kamapantula, B.; Abdelzaher, A.; Ghosh, P.; Tiwari, S.; Barve, N.; et al. PANNOTATOR: An Automated Tool for Annotation of Pangenomes. *Genet. Mol. Res.* **2013**, *12*, 2982–2989. [CrossRef] [PubMed]

63. Fouts, D.E.; Brinkac, L.; Beck, E.; Inman, J.; Sutton, G. PanOCT: Automated Clustering of Orthologs Using Conserved Gene Neighborhood for Pan-Genomic Analysis of Bacterial Strains and Closely Related Species. *Nucleic Acids Res.* **2012**, *40*, e172. [CrossRef] [PubMed]

64. Hennig, A.; Bernhardt, J.; Nieselt, K. Pan-Tetris: An Interactive Visualisation for Pangenomes. *BMC Bioinform.* **2015**, *16*, 1–11. [CrossRef]

65. Sheikhizadeh, S.; Schranz, M.E.; Akdel, M.; de Ridder, D.; Smit, S. PanTools: Representation, Storage and Exploration of Pan-Genomic Data. *Bioinformatics* **2016**, *32*, 487–493. [CrossRef] [PubMed]

66. Pedersen, T.L.; Nookaew, I.; Wayne Ussery, D.; Månsson, M. PanViz: Interactive visualization of the structure of functionally annotated pangenomes. *Bioinformatics* **2017**, *33*, 1081–1082. [CrossRef]

67. Pantoja, Y.; Pinheiro, K.; Veras, A.; Araújo, F.; de Sousa, L.; Guimarães, L.C.; Silva, A.; Ramos, R.T. PanWeb: A Web Interface for Pan-Genomic Analysis. *PLoS ONE* **2017**, *12*, e0178154. [CrossRef] [PubMed]

68. Ding, W.; Baumdicker, F.; Neher, R.A. PanX: Pangenome Analysis and Exploration. *Nucleic Acids Res.* **2018**, *46*, e5. [CrossRef]

69. Liu, Y.Y.; Chiou, C.S.; Chen, C.C. PGAdb-Builder: A Web Service Tool for Creating Pangenome Allele Database for Molecular Fine Typing. *Sci. Rep.* **2016**, *6*, 36213. [CrossRef] [PubMed]

70. Thorpe, H.A.; Bayliss, S.C.; Sheppard, S.K.; Feil, E.J. Piggy: A Rapid, Large-Scale Pangenome Analysis Tool for Intergenic Regions in Bacteria. *Gigascience* **2018**, *7*, 1–11. [CrossRef] [PubMed]

71. Lees, J.A.; Galardini, M.; Bentley, S.D.; Weiser, J.N.; Corander, J. Pyseer: A Comprehensive Tool for Microbial Pangenome-Wide Association Studies. *Bioinformatics* **2018**, *34*, 4310–4312. [CrossRef]

72. Jandrasits, C.; Dabrowski, P.W.; Fuchs, S.; Renard, B.Y. Seq-Seq-Pan: Building a Computational Pangenome Data Structure on Whole Genome Alignment. *BMC Genom.* **2018**, *19*, 47. [CrossRef] [PubMed]

73. Ozer, E.A.; Allen, J.P.; Hauser, A.R. Characterization of the Core and Accessory Genomes of Pseudomonas Aeruginosa Using Bioinformatic Tools Spine and AGEnt. *BMC Genom.* **2014**, *15*, 737. [CrossRef]

74. Chaudhari, N.M.; Gupta, V.K.; Dutta, C. BPGA- an Ultra-Fast Pangenome Analysis Pipeline. *Sci. Rep.* **2016**, *6*, 24373. [CrossRef] [PubMed]

75. Cheng, G.; Lu, Q.; Ma, L.; Zhang, G.; Xu, L.; Zhou, Z. BGDMdocker: A Docker a Workflow Base on Docker for Analysis and Visualization Pangenome and Biosynthetic Gene Clusters of Bacterial. *PeerJ* **2017**, *30*, e3948. [CrossRef]

76. Silva de Oliveira, M.; Thyeska Castro Alves, J.; Henrique Caracciolo Gomes de Sá, P.; Veras, A.A.D.O. PAN2HGENE–Tool for Comparative Analysis and Identifying New Gene Products. *PLoS ONE* **2021**, *16*, e0252414. [CrossRef] [PubMed]

77. Danilevicz, M.F.; Fernandez, C.G.T.; Marsh, J.I.; Bayer, P.E.; Edwards, D. Plant Pangenomics: Approaches, Applications and Advancements. *Curr. Opin. Plant Biol.* **2020**, *54*, 18–25. [CrossRef] [PubMed]

78. Beier, S.; Thomson, N.R. Panakeia—A Universal Tool for Bacterial Pangenome Analysis. *bioRxiv* **2021**. [CrossRef]

79. Duan, Z.; Qiao, Y.; Lu, J.; Lu, H.; Zhang, W.; Yan, F.; Sun, C.; Hu, Z.; Zhang, Z.; Li, G.; et al. HUPAN: A Pangenome Analysis Pipeline for Human Genomes. *Genome Biol.* **2019**, *20*, 1–11. [CrossRef] [PubMed]

80. Bosi, E.; Fani, R.; Fondi, M. Defining Orthologs and Pangenome Size Metrics. *Methods Mol. Biol.* **2015**, *1231*, 191–202. [PubMed]

81. Othoum, G.; Bougouffa, S.; Bokhari, A. Mining Biosynthetic Gene Clusters in Virgibacillus Genomes. *BMC Genom.* **2019**, *20*, 696. [CrossRef]

82. Othoum, G.; Prigent, S.; Derouiche, A. Comparative Genomics Study Reveals Red Sea Bacillus with Characteristics Associated with Potential Microbial Cell Factories (MCFs). *Sci. Rep.* **2019**, *9*, 19254. [CrossRef]

83. Kant, R.; Rintahaka, J.; Yu, X.; Sigvart-Mattila, P.; Paulin, L.; Mecklin, J.P.; Saarela, M.; Palva, A.; von Ossowski, I. A Comparative Pangenome Perspective of Niche-Adaptable Cell-Surface Protein Phenotypes in Lactobacillus Rhamnosus. *PLoS ONE* **2014**, *9*, e102762. [CrossRef] [PubMed]

84. McInerney, J.; McNally, A.; O'Connell, M. Why Prokaryotes Have Pangenomes. *Nat. Microbiol.* **2017**, *2*, 17040. [CrossRef] [PubMed]

85. Vos, M.; Eyre-Walker, A. Are Pangenomes Adaptive or Not? *Nat. Microbiol.* **2017**, *2*, 1576. [CrossRef] [PubMed]

86. Vos, M.; Hesselman, M.C.; Te Beek, T.A.; van Passel, M.W.; Eyre-Walker, A. Rates of Lateral Gene Transfer in Prokaryotes: High but Why? *Trend Microbiol.* **2015**, *23*, 598–605. [CrossRef] [PubMed]

87. Livingstone, P.G.; Morphew, R.M.; Whitworth, D.E. Genome Sequencing and Pangenome Analysis of 23 *Corallococcus* Spp. Strains Reveal Unexpected Diversity, with Particular Plasticity of Predatory Gene Sets. *Front. Microbiol.* **2018**, *9*, 3187. [CrossRef] [PubMed]

88. Huang, X.; Kurata, N.; Wang, Z.X.; Wang, A.; Zhao, Q.; Zhao, Y.; Liu, K.; Lu, H.; Li, W.; Guo, Y.; et al. A Map of Rice Genome Variation Reveals the Origin of Cultivated Rice. *Nature* **2012**, *490*, 497–501. [CrossRef] [PubMed]

89. Zhang, B.; Zhu, W.; Diao, S. The Poplar Pangenome Provides Insights into the Evolutionary History of the Genus. *Commun. Biol.* **2019**, *2*, 215. [CrossRef] [PubMed]

90. Barchi, L.; Rabanus-Wallace, M.T.; Prohens, J.; Toppino, L.; Padmarasu, S.; Portis, E.; Rotino, G.L.; Stein, N.; Lanteri, S.; Giuliano, G. Improved Genome Assembly and Pan-genome Provide Key Insights on Eggplant Domestication and Breeding. *Plant J.* **2021**, *107*, 579–596. [CrossRef]

91. Monat, C.; Sabot, F. *Pangenomics in Crop Plants*; Population Genomics; Springer: Cham, Switzerland, 2020. [CrossRef]

92. Lei, L.; Goltsman, E.; Goodstein, D.; Wu, G.A.; Rokhsar, D.S.; Vogel, J.P. Plant Pangenomics Comes of Age. *Ann. Rev. Plant Biol.* **2021**, *72*, 411–435. [CrossRef] [PubMed]

93. Della Coletta, R.; Qiu, Y.; Ou, S.; Hufford, M.B.; Hirsch, C.N. How the Pangenome Is Changing Crop Genomics and Improvement. *Genome Biol.* **2021**, *22*, 3. [CrossRef] [PubMed]

94. Bayer, P.E.; Petereit, J.; Danilevicz, M.F.; Anderson, R.; Batley, J.; Edwards, D. The Application of Pangenomics and Machine Learning in Genomic Selection in Plants. *Plant Genome* **2021**, *14*, e20112. [CrossRef]

95. Hu, B.; Xie, G.; Lo, C.; Starkenburg, S.R.; Chain, P.S.G. Pathogen Comparative Genomics in the Next-Generation Sequencing Era: Genome Alignments, Pangenomics and Metagenomics. *Brief. Funct. Genom.* **2011**, *10*, 322–333. [CrossRef]

96. Casa-Esperón, E. Horizontal Transfer and the Evolution of Host-Pathogen Interactions. *Int. J. Evol. Biol.* **2012**, *2012*, 679045. [CrossRef] [PubMed]

97. Perna, N.T.; Plunkett, G.; Burland, V.; Mau, B.; Glasner, J.D.; Rose, D.J.; Mayhew, G.F.; Evans, P.S.; Gregor, J.; Kirkpatrick, H.A.; et al. Genome Sequence of Enterohaemorrhagic Escherichia Coli O157: H7. *Nature* **2001**, *409*, 529–533. [CrossRef] [PubMed]

98. Rasko, D.A.; Rosovitz, M.J.; Myers, G.S.; Mongodin, E.F.; Fricke, W.F.; Gajer, P.; Crabtree, J.; Sebaihia, M.; Thomson, N.R.; Chaudhuri, R.; et al. The Pangenome Structuree of *Escherichia coli*: Comparative Genomic Analysis of *E. coli* Commensal and Pathogenic Isolates. *J. Bacteriol.* **2008**, *190*, 6881–6893. [CrossRef]

99. Badet, T.; Oggenfuss, U.; Abraham, L. A 19-Isolate Reference-Quality Global Pangenome for the Fungal Wheat Pathogen Zymoseptoria Tritici. *BMC Biol.* **2020**, *18*, 12. [CrossRef] [PubMed]

100. Plissonneau, C.; Hartmann, F.E.; Croll, D. Pangenome Analyses of the Wheat Pathogen Zymoseptoria Tritici Reveal the Structural Basis of a Highly Plastic Eukaryotic Genome. *BMC Biol.* **2018**, *16*, 5. [CrossRef] [PubMed]

101. Agarwal, G.; Gitaitis, R.D.; Dutta, B. Pangenome of Novel Pantoea Stewartii Subsp. Indologenes Reveals Genes Involved in Onion Pathogenicity and Evidence of Lateral Gene Transfer. *Microorganisms* **2021**, *9*, 1761. [CrossRef] [PubMed]

102. Gonzalez, V.; Aventin, N.; Centeno, E.; Puigdomenech, P. High Presence/Absence Gene Variability in Defense-Related Gene Clusters of Cucumis Melo. *BMC Genom.* **2013**, *14*, 1–8. [CrossRef] [PubMed]

103. Shen, J.; Araki, H.; Chen, L.; Chen, J.Q.; Tian, D. Unique Evolutionary Mechanism in R-Genes under the Presence/Absence Polymorphism in Arabidopsis Thaliana. *Genetics* **2006**, *172*, 1243–1250. [CrossRef] [PubMed]

104. Winzer, T.; Gazda, V.; He, Z.; Kaminski, F.; Kern, M.; Larson, T.R.; Li, Y.; Meade, F.; Teodor, R.; Vaistij, F.E.; et al. A Papaver Somniferum 10-Gene Cluster for Synthesis of the Anticancer Alkaloid Noscapine. *Science* **2012**, *336*, 1704–1708. [CrossRef]

105. Swanson-Wagner, R.A.; Eichten, S.R.; Kumari, S.; Tiffin, P.; Stein, J.C.; Ware, D.; Springer, N.M. Pervasive Gene Content Variation and Copy Number Variation in Maize and Its Undomesticated Progenitor. *Genome Res.* **2010**, *20*, 1689–1699. [CrossRef] [PubMed]

106. Rouli, L.; Merhej, V.; Fournier, P.E.; Raoult, D. The Bacterial Pangenome as a New Tool for Analysing Pathogenic Bacteria. *New Microbes New Infect.* **2015**, *7*, 72–85. [CrossRef]

107. de Souza, G.A.; Arntzen, M.Ø.; Wiker, H.G. MSMSpdbb: Providing Protein Databases of Closely Related Organisms to Improve Proteomic Characterization of Prokaryotic Microbes. *Bioinformatics* **2010**, *26*, 698–699. [CrossRef] [PubMed]

108. Caputo, A.; Fournier, P.E.; Raoult, D. Genome and Pangenome Analysis to Classify Emerging Bacteria. *Biol. Direct* **2019**, *14*, 5. [CrossRef]

109. Naz, K.; Naz, A.; Ashraf, S.T.; Rizwan, M.; Ahmad, J.; Baumbach, J.; Ali, A. PanRV: Pangenome-Reverse Vaccinology Approach for Identifications of Potential Vaccine Candidates in Microbial Pangenome. *BMC Bioinform.* **2019**, *20*, 1–10. [CrossRef]

110. Dalsass, M.; Brozzi, A.; Medini, D.; Rappuoli, R. Comparison of Open-Source Reverse Vaccinology Programs for Bacterial Vaccine Antigen Discovery. *Front. Immunol.* **2019**, *10*, 113. [CrossRef]