The Pharma Innovation Journal 2022; SP-11(10): 2189-2201

www.ThePharmaJournal.com

The Pharma Innovation

ISSN (E): 2277-7695
ISSN (P): 2349-8242
NAAS Rating: 5.23
TPI 2022; SP-11(10): 2189-2201

www.thepharmajournal.com
Received: 28-08-2022
Accepted: 30-09-2022

Dr. Divya Rajawat
Ph.D. Scholar, Division of Animal Genetics, Indian Veterinary Research Institute, Izatnagar, Bareilly Uttar Pradesh, India

Harshit Kumar
Division of Animal Genetics, Indian Veterinary Research Institute, Izatnagar, Bareilly Uttar Pradesh, India

Anuradha Panwar
Division of Animal Genetics, Indian Veterinary Research Institute, Izatnagar, Bareilly Uttar Pradesh, India

Sonali Sonejita Nayak
Division of Animal Genetics, Indian Veterinary Research Institute, Izatnagar, Bareilly Uttar Pradesh, India

Kanika Ghildiyal
Division of Animal Genetics, Indian Veterinary Research Institute, Izatnagar, Bareilly Uttar Pradesh, India

Anurodh Sharma
Division of Animal Genetics, Indian Veterinary Research Institute, Izatnagar, Bareilly Uttar Pradesh, India

Atul Singh Rajput
Livestock Production and Management Section, Indian Veterinary Research Institute, Izatnagar, Bareilly Uttar Pradesh, India

Gangula Lokavya Reddy
Division of Animal Reproduction, Indian Council of Agricultural Research (ICAR) -Indian Veterinary Research Institute (IVRI), Izatnagar, Uttar Pradesh Uttar Pradesh India

Manjit Panigrahi
Division of Animal Genetics, Indian Veterinary Research Institute, Izatnagar, Bareilly, Uttar Pradesh, India

Corresponding Author:
Dr. Divya Rajawat
Ph.D. Scholar, Division of Animal Genetics, Indian Veterinary Research Institute, Izatnagar, Bareilly Uttar Pradesh, India

Pan-genomics: A review of analysis, evolution, applications and future prospects

Dr. Divya Rajawat, Harshit Kumar, Anuradha Panwar, Sonali Sonejita Nayak, Kanika Ghildiyal, Anurodh Sharma, Atul Singh Rajput, Gangula Lokavya Reddy and Manjit Panigrahi

Abstract
A pan-genome is a collection of genetic sequences distributed throughout all species or groups. The term was first used by Tettelin and his team while he was working on *Streptococcus agalactiae* strains. The pan-genome can be split into the core, shell, and cloud pan-genome, and classified into two groups; open and close, based on Heap's law. While pan-genomic research began with bacteria, advances in genome sequencing and assembly techniques have enabled the production of pan-genomes for eukaryotic organisms such as fungi, plants, and mammals. The core genome includes genes crucial for the organism's survival and represented as housekeeping genes, while transmission, pathogenicity, and immunity are usually linked with Dispensable genes. As our knowledge of genomic diversity expanded, it became clear that a single reference sequence was insufficient to represent the range of genomic variation observed within species, leading to the development and expansion of the pan-genome concept. Extending pan-genomic studies to higher taxonomic groups will eventually provide the resources needed to investigate the differences between organisms allowing for a comprehensive description of genes, their evolutionary history, and function. Pan-genome s will undoubtedly become widespread in the next decade, rendering the single-reference approach obsolete to genomic research.

Keywords: Pan-genomics, *Streptococcus agalactiae*, eukaryotic

Introduction
The first complete sequence of the genome was performed in bacterial species (*Haemophilus influenzae*) in 1995 (Fleischmann *et al*., 1995) [28]. In 1996, the whole sequence of the first eukaryotic species was published in the organism *Saccharomyces cerevisiae*, followed by the Human genome in 2001 (Goffeau et a., 1996; Lander *et al*., 2001) [32, 52]. With the expanding number of genomes, it's time to reconsider the concept of a "reference" genome, and a new branch of genomics has come to light that is pan-genomics. The terminology 'pan-genome' comes from the word παν, which means 'whole,' whereas 'genome' refers to an organism's entire genes. A pan-genome is a collection of genetic sequences distributed throughout the entire species or group. The term was first used by Tettelin *et al*. in 2005 [105] while he was working on *Streptococcus agalactiae* strains. Although pan-genomics was first used to describe the genomic architecture of bacterial species, the notion of the pan-genome was quickly adopted by plant and animal scientists, resulting in more than 20 eukaryotic pan-genome studies to date (Richard, 2020) [87].
Bacterial genomes range from 0.6 to 8.0 megabases (Mb) and encode 600-6000 proteins on average. The presence of coding genes dominates the bacterial genome; that's why the bacterial genome is the choice for the pan-genomics analysis. However, the term evolved when pan-genome research expanded to include plants and animals. Two definitions of the pan-genome have evolved to handle the variations between bacterial and eukaryotic genomes. The sequence-centric approach entails that the pan-genome is the entire set of non-redundant sequences found in all individuals. In contrast, in the gene-centric approach, the pan-genome is the union of all the orthologous gene clusters (Mira *et al*., 2010) [66]. The pan-genome can be divided into core pan-genome, shell pan-genome, and cloud pan-genome (Medini *et al*., 2005) [63] (Figure 1). Core pan-genome comprises the genes or sequences found in all the individuals, shell pan-genome includes two or more strains, and cloud pan-genome consists of genes found only in one strain. The Cloud pan-genome is also represented as the dispensable genome by some authors that are the genes or sequences absent from one or more individuals (Vernikos *et al*., 2015) [110]. Cluster analysis, single-base polymorphism, copy number variations,

phylogenetic trees, and multiple sequence alignments are all classical depictions of the conservation and diversity of the genome. We can classify pan-genome into two groups, and this classification is based Heap's law: $N = kn^{-\alpha}$; where N denotes the number of gene families, n is the number of genomes, α is the exponential factor, and k is the proportionate constant (Costa *et al*., 2020) [20]. The interpretation is based on the α value. When the value of α is more than one, the pan-genome will be considered closed, and when the value of α is less and equal to one, the pan-genome will be called an open pan-genome. In the closed pan-genome, the size of the pan-genome can be forecasted theoretically, and when the new genes are added, only a few gene families will be added to the lineage. In contrast to that, when the pan-genome is open, it is impossible to predict the theoretical size of the pan-genome. *Escherichia coli* is an example of a species with an open pan-genome (Hyatt *et al*., 2010) [37].

The pan-genome can be defined in the group of genes rather than nucleotide sequences for prokaryotes because the maximum portion of DNA consists of the coding sequence. Genes do not only make up the majority of the portion in these species (usually 90% or more), but the content varies significantly in some bacterial species, and unique genes make up anywhere from 20% to 40% of the pan-genome. While pan-genomic research began with bacteria, advances in sequence analyses and assembly techniques have enabled the construction of pan-genome s for eukaryotic organisms such as fungi (McCarthy, 2019) [62], plants (Eizenga *et al*., 2020) [27], and mammals (Figure 2). Eukaryotes do not interchange DNA as freely as bacteria, resulting in a more stable gene composition. As a result, a eukaryotic pan-genome is typically defined as the collection of DNA sequences, not simply genes. For a species like humans, where coding sequences make up only 2% of the genome, a pan-genome made up entirely of coding sequences would reveal little about within-species differences (Francis and Wörheide, 2017) [29]. Pathogenicity, gene-mediated resistance, and other phenotypes relevant to human health are often influenced by genes-oriented change; consequently, examining the dispensable versus core genomes might help to explain these traits (Piovesan *et al*., 2019) [81].

Here, in this review, we present an overview of the concepts and development of pan-genomics in different species, including currently available bioinformatics tools. Then, we discussed various applications related to this field based on their cumulative citation by peer-reviewed scientific publications. Finally, we address the challenges and future directions related to pan-genomics.

Pan-genome analysis

The existence of genes from certain strains and the identification of a core and accessory pan-genome are the focus of routine pan-genomic investigations (Eizenga *et al*., 2020) [27]. The pan-genome assembly for various organisms has been made possible by the advanced techniques of genome sequencing. Currently, several methodologies for eukaryote pan-genome assembly are available, comprising *de novo* genome assembly, k-mer-based (de Bruijn graph) approach, and mapping and assembly (iterative assembly) (Figure 3).

Assembling an organism's genome from novel smaller sequenced genome fragments is known as *de novo* genome assembly. Identification of Gene orthology and alignments of whole-genome methods can be utilized to compare assembled genomes (McCarthy and Fitzpatrick, 2019) [62]. This approach allows retrieval of all individuals' entire genome, and it can resolve repetitive sequences and the presence or absence of variants (PNVs/CNVs). There are various tools to perform the *de novo* genome assembly approach, namely Velvet (Zerbino *et al*., 2009) [113], ALLPATHS (Butler *et al*., 2008) [11], SOA Pdenovo (Li *et al*., 2010) [55], and Ma Su RCA (Zimin *et al*., 2013) [116]. Often, *De novo* assemblies demand the construction of large and costly datasets, and technical defects, variations in assembly, and annotation can lead to false presence/absence of variants. The sequence segments are thoroughly mapped with the reference genome assembly in the iterative assembly. The remaining unmapped segments are extracted, assembled, further mapped, and added to the developing pan-genome. The updated reference sequence and iterative assembly between the mapping segments produce a unique pan-genome (Schatz *et al*., 2014) [61]. In this approach, there is no necessity for functional gene clusters (HGC) to find out the presence/absence of variants at each locus (Golicz *et al*., 2016) [33].

In the k-mer-based approach, each sequence is disintegrated into fragments of different k lengths. The similarity or uniqueness of genes in a distinct clade of a species is represented as nodes, and the branches that connect the nodes are denoted as edges. A graph is constructed with the help of interconnected nodes and edges. A full pan-genome is visualized as a coloured de Bruijn graph, enabling the detection of regions that are shared across genomes or unique to a single individual (Iqbal *et al*., 2016) [38].

Different tools and software have been developed in the field of genomics related to various applications (Saravanan *et al*., 2019) [92]. Several genomic tools have been developed to analyze and visualize the pan-genomes across different species in the past few years. We have tried to summarise some of the important bioinformatics tools in Table 1.

Factors affecting pan-genome analysis

Pan genome analysis may influence by several factors such as annotation quality, orthologous gene detection, assembly quality, and selection of appropriate samples. Annotation of the complex eukaryotic genome is a difficult task. There are mainly two genome annotation techniques; evidence-based gene prediction and ar initio gene prediction. In the evidence-based gene prediction approach, genes are annotated based on their transcription unit (protein product). In contrast, in the *ab initio* gene prediction approach, genes are predicted based on the nearby signals or signs like consensus, or f, stop codon, etc. MAKER is a commonly used annotation pipeline that combines both approaches on a single platform (Holt and Yandell, 2011) [36]. Orthologous gene detection aims to find out the functional genes across the genome of different species. Ortho MCL, INPARANOID, and reciprocal best blast hit are some software to identify the functionally equivalent genes (Berglund *et al*., 2007; Chen *et al*., 2007) [7, 13]. Quality of assembly is also an essential factor in accurate analysis of pan-genome analysis. The total size of genome assembly is assessed by the approach (CEGMA) (Parra *et al*.,

2007) [76]. The optimum sample selection is the utmost important factor influencing the pan-genome analysis. The choice of unrelated individuals gives more authentic results, while the individuals sharing recent ancestry provide the false-positive results of the pan-genome size. Hence, to construct a successful pan-genome investigation design, it is critical to choose appropriate individuals who reflect the majority of variety within a population[22]. On a substantial stage, preferable attributes of a pan-genome include 'completeness' or possessing all essential features and sufficient sequence space to serve as a reference for the analysis of additional individuals; 'stability' or having uniquely recognizable features that can be investigated by various scientists and at distinct intervals; and 'comprehensibility' or facilitating understanding of the complexities of genome frameworks across several individuals or species (Thorvaldsdóttir *et al*., 2008).

Pan-genomes across trees of life

While working on the pathogenicity of neonatal infection in human infants, Tetellin *et al*. (2005) [63] observed that the genetic diversity that determines the virulence of a particular bacterial strain is not reflected in the sequence of a single genome and may limit the genome-wide screening of vaccine candidates. To resolve this problem, they compared the genomes of eight *Streptococcus agalactiae* strains (*Streptococcus agalactiae* Type-B). Tetellin *et al*. found that nearly 80% of the genome belongs to the strain-specific genome (Dispensable/cloud), and the remaining portion was shared by all the individuals (core genome). Multiple bacterial species' pan-genome analyses have been reported, and Bacteria have the most extended history of pan-genome research (Table 2). In Bacteria, the size of the core and cloud genomes is also linked to the manner of living. Bacteria associated with other bacteria of different phyla (Sympatric speciation) often consist of a small proportion of the core genome.In contrast, bacteria that live in an isolated condition (Allopatric speciation) possess a minor portion of the dispensable genome (Rouli *et al*., 2015) [89].

In Fungi, McCarthy (2019) [62] conducted the study of pan-genome by utilizing Pan sOCT software (perl based) and presented pan-genomes of four model fungal species, i.e., *Saccharomyces cerevisiae*, *Candida albicans*, *Cryptococcus neoformans*, *Aspergillus fumigatus* (Table 2). Different studies suggested that the fungal core genome is the compilation of genes with ancient ancestry related to the various metabolic and survival functions. In contrast, accessory genomes are leveraged with recent genes involved in molecule transport (Peter *et al*., 2018) [80]. Pan-genome analyses of fungal can trace links between isolates of varying virulence and discover new genes involved in infection and host response (Plissonneau *et al*., 2018) [82]. In plants, the pan-genome concept was initially proposed about transposable elements in maize (Morgante *et al*., 2007) [67]. Transposable elements (TEs) are often associated with dispensable genes. Still, nowadays scientific community is focusing on protein-coding genes, and to date, various plant species' pan-genome analyses have been reported (Table 2). In plants, direct comparison is difficult because plant species vary in ploidy level and type of breeding, and species with outcrossing breeding plans show a significant accessory genome content (Tao *et al*., 2019) [104]. The content of the pan-genome is also influenced by artificial selection and reproductive strategies (complementarity and heterosis) (Liu *et al*., 2020) [58].

The first study of pan-genomics in animals was conducted on Humans in that newly assembled *de novo* sequences of African and Asian populations were compared with already existing human reference genome assembly (Li *et al*., 2010) [55]. Identification of 162 National Center for Biotechnology Information (NCBI) human Ref Sew has been done, and 5 Mb of novel sequences was identified in both assemblies. Two more Human pan-genome studies were published with 275 individuals of Han Chinese populations (Duan *et al*., 2019) [24] and 910 of the African population (Sherman *et al*., 2019) [119]. These studies have concluded that three factors may influence the pan-genomic analysis: difference in the genetic makeup of individuals in the population, methodology used (*de novo* assemblies/ Iterative assemblies/ k-mer-based assemblies), and the number of sequences individuals in the population.

Besides humans, geneticists are also focusing on the pan-genome studies on different livestock. In this review, we have tried to compile all the studies on livestock that have been conducted to date.

Towards livestock pan-genome

Pig (*Sus scrofa*) was the first livestock species to conduct pan-genome analysis. Tian *et al*. (2019) [107] constructed the sequence map from a combined approach of Hi-C data and *de novo* assembly by compiling the 12 porcine breeds (Breeds originating in China and Europe) (Supplementary Table 1). As a result, they found that about 3% of constitutive sequences of the whole genome (~72 Mb) were missed in the reference genome assembly (Sscrofa11.1). The *TIG3* gene has been found in further analysis, which is associated with fatty acid metabolism. By combining the HTML, Java, and various scripts, one web-based interface pan-genome server (PIGPAN), was developed to utilize all the resources related to pig breeding, genetic diversity, and other biomedical research (animal.nwsuaf.edu.cn/code/index.php/panPig). In goats, a pan-genomic study was done by Li *et al*. in 2019 by retrieving the ten assemblies of different breeds from the publically available dataset NCBI caprine genome assembly (Supplementary Table 1). These assemblies were compared with pre-existing goat reference assembly ARS1 and constructed a pan-genome. As a result, it was found that 24,414 novel SNPs were recovered per individual, and the mapping rate was improved by 1.15%. A total of 38.3 Mb sequences were identified that were absent in ARS1. Further, for data visualization, one web-based database has been built (Goat pan-genome web interface) (http://animal.nwsuaf.edu.cn/panGoat) to retrieve various kinds of information like the diversity parameters, gene annotation, and expression level information of the pan-sequences as well as the whole pan-genome.

The advent of Next Generation Sequencing, SNP genotyping platforms and simultaneous reduction in the cost of sequencing had opened the door to genomic research in farm animals (Saravanan *et al*., 2022a) [95]. Since the reference assembly is constructed only by utilizing the single breed (Hereford), reference genomes are defective, deficient in informative SNPs, and inadequate to reveal the true genomic relevance of the population (Supplementary Table 1). The

Animal Genomics team at ETH Zurich constructed the first bovine pan-genome graph (Crysnanto, 2021) [21]. This cattle pan-genome combined the six reference bovine genomes (Angus, Brahman, Brown Swiss, Hereford, Highland, and Yak) and found the other 70,329,827 bases and 83,250 polymorphic sites. Further, the researchers explored the messenger RNAs (transcripts of the function genes) to find out the novel functional genes that are biologically significant. The number of genes was associated with immune function (Leukocyte immunoglobulin-like receptor A5). This pan-genome graph may lead to substantial refinement of existing reference genome assembly. Recently, Li *et al.* (2021) [54] attempted to assemble the ovine pan-genome graph from 13 genetically diverse sheep breeds utilizing the long-read sequencing (Pac Bio and Hi Fi). They demonstrated the 13,419 multi allelic variations, 7028 divergent alleles, and 142593 in dells. However, this paper is still in the preprint stage and has not yet been peer-reviewed.

Applications

The information acquired from reference genome assembly can be applied in several different genomic applications like Admixture analysis (Pal *et al*., 2022) [70, 71], breed-specific SNP panels (Kumar *et al*., 2019; Kumar *et al*., 2021a; 8.Kumar *et al*., 2021b) [45, 93, 97], copy number variations (Kumar *et al*., 2021c) [49], rare SNPs with intermediate frequencies (Kumar *et al*., 2021d) [50] and selection signature analysis (Saravanan *et al*., 2020a; Saravanan *et al*., 2021b, Rajawat *et al*., 2022a, Rajawat *et al*., 2022b) [118, 97, 72, 85]. Using different reference assemblies, different SNP chips have been developed in recent years (Panigrahi *et al*., 2022) [74]. These SNP BeadChips have versatile applications in the field of genomics *viz*. breed composition of different crossbred cattle (Ahmad *et al*., 2020; Chhotaray *et al*., 2020) [2, 15], analysis of different diversity parameters, and haplotype block structures (Chhotaray *et al*., 2021b; Saravanan *et al*., 2020b; Saravanan *et al*., 2021a) [17, 96, 93]. In Supplementary Table 2, we have mentioned the latest reference genome assemblies of important livestock species and their genome size and latest release.

Different studies suggested that a single genome might not manifest the complete genomic complement and not be able to represent the full spectrum of divergent sequences of a species. Genomic studies in different individuals generally face the challenge of analyzing expanding genome sizes. In the case of other organisms, the number of the sequenced genome will exceed hundreds to thousands in the following years. Exploration of multiple genome studies rather than a single reference nullifies the sample bias and guarantees that the genetic diversity of particular species is fully reflected. In the subsequent segment, we will look at the different applications regarding pan-genomics and discuss them one by one.

Microbial genomics

Microbes are broadly applied and studied in different disciplines, including biology, biotechnology, and medicine. Complete knowledge of the evolutionary and functional genomics of microbes reveals the likelihood of developing therapeutic and preventive applications (Rogers *et al*., 2014; Saravanan *et al*., 2022b) [88, 74]. Millions of sequenced strains are available in sufficient detail for several clinically important bacterial species, allowing reference genome assemblies to be created (Liti *et al*., 2009) [57]. A new term has been 'Mobilome' evolved by Anani *et al*. (2020) [4]. The microbial mobilome is the collection of all the mobile genetic elements (MGEs) like bacteriophages, plasmids, and transposons (Anani *et al*., 2020) [4]. The genetic information conveyed by the mobile genetic elements can promote the emergence of new pathogens and drug resistance markers (*Bacillus anthracis*, *Escherichia coli*). Conventional vaccine production is time-consuming and antigenic drift may limit the accuracy and efficacy. The concept of reverse vaccinology was proposed by Rappuoli *et al*. (2000) [86] to predict the antigenic epitopes among selected vaccine candidates. This approach is based on the pan-genomic analysis of pathogenic microbes. The first vaccine production was carried out in *Neisseria meningitides* (Serotype B). Further efforts to generate vaccines via reverse vaccinology are ongoing on *Acinetobacter baumannii*, *Streptococcus pneumonia*, and *Escherichia coli* (Seib *et al*., 2012; Bidmos *et al*., 2018) [99, 9].

Traditionally, two main strategies have been followed for drug development: target-based screening and whole-cell screening (Hertzberg, 1993) [35]. A target must have a specific biological function related to survivability. Analyzing the pan-genome is a useful way of determining which genes code for important functions. As discussed previously, the core genes are collections of different drug targets. Several studies have demonstrated the pan-genome's opportunity to discover theoretical therapeutic targets in pathogens such as *Clostridium botulinum* (Bhardwaj and Somvanshi, 2017) [8], *Helicobacter pylori* (Ali *et al*., 2015) [3], *Leptospira* (Zeng *et al*., 2017) [112], and *Corynebacterium diphtheriae* (Jamal *et al*., 2017) [39]. The findings of all of these investigations could contribute to the emergence of therapeutic medicines, although more advanced pan-genomic research is required.

Genome-wide association studies (GWAS)

Next, pan-genome analysis may have the major impact on genome-wide association studies (GWAS). It is a complementary approach to the selection signatures and genomic selections (Kaisa *et al*., 2020) [43]. Genome-wide SNP panels are now widely available, which has facilitated genome-wide association studies (GWAS) for the discovery of novel markers linked to a range of animal attributes (Chhotaray *et al*., 2021a; Mehrotra *et al*., 2021a; Mehrotra *et al*., 2021b) [46, 64, 65]. The dispensable genomic regions connected with the critically important traits may be missing from the linear reference assembly and unwittingly excluded from correlation studies. So, for accurate analysis of association studies, it is important to incorporate the pan-genome sequence in resequencing analyses (Gege *et al*., 2019) [30]. It increases the possibility of exploring genotype-phenotype associations analyzed by the comprehensive variations between different breeds. Further, pan-sgenome analysis leads to an understanding of hereditary divergence and different evolutionary strategies, and in some situations, these findings may direct to even redefinition of species (Bayer *et al*., 2021) [5]. Genome-wide association studies (GWAS) are the new way to find genetic variables linked to important features like drug resistance or secondary metabolism. Individual variants such as structural variants

(SVs), single-nucleotide polymorphisms (SNPs), insertion/deletion, lack or existence of complete genes, annotated Gene oncology terms, and mobile genomic elements like integrons or prophages can all be studied in this way. At each level, pan-genomic techniques could be used (Manuweera *et al*., 2019) [60]. (Kehr *et al*., 2017) [44]. Understanding the functioning of novel genes and associated mutations in the breeding population makes it possible to avoid livestock loss due to embryonic lethality (Derks *et al*., 2019) [23]. Specific criteria, such as an accurate data processing pipeline to deliver variant calling, extracting genes from next-generation sequencing data, annotation of hypothesized genes and proteins, and determining the coordinate system of sequenced loci, must be encountered for the Genome-wide association study (GWAS) to be successful (Dutilh *et al*., 2013) [25]. At each level, pan-genomic techniques could be used to undertake bias-free analysis.

Metagenomics

The term "metagenome" refers to the genetic content of all microorganisms in a given environment. It has been widely used to research microbial diversity in various habitats, including air, soil, water, plants, animals, and humans (Parashar *et al*., 2021) [117]. Pan-genomics and metagenomics have made significant advances in studying genetic evolution and function in a specific taxon or community. Some microbiome studies have adopted this complementary strategy, and such integrated pan-genomics with metagenomics techniques have enabled researchers to study the diversity and dynamics of populations in microbial communities (Ma *et al*., 2020) [119] and find the solutions for various non-curable diseases (Panigrahi *et al*., 2022a) [95]. Metagenome-wide association studies, for example, attempt to link the microbial composition of the human to both pan-genome and metagenome (Qin *et al*., 2012) [83]. IMG/M is a genomic and metagenome-integrated comparative data analysis system including COG clusters, Pfam, InterPro domains, and KEGG pathways (Chen *et al*., 2017) [37]. PanPhlAn is a tool that uses metagenomes to identify the genetic content of particular strains. This program can determine the taxonomic profile, strain-level profile, functional profile, and phylogenetic profile from metagenomic samples and characterize pan-genome and metagenome at the strain level (Jun *et al*., 2015) [42]. Thus, Metagenomics has also been demonstrated to be capable of exposing entire species' genomes and tracing them across their surroundings.

Phylogenomics

Phylogenomics uses complete genome sequences to reconstruct the evolutionary history of a collection of species. It can take advantage of various signals such as sequence or gene content (Dutilh *et al*., 2007; Patra *et al*., 2021) [26, 78]. Depending on the relatedness of the included organisms, pan-genomics will allow quick extraction of genomic features with an evolutionary signal, such as gene content tables, sequence alignments of shared marker genes, genome-wide SNPs, or internal transcribed spacer sequences (Ciccarelli *et al*., 2006) [18]. Phylogenomic trees representing organism or cellular lineages provide valuable input data for various biological applications, such as mapping the evolutionary dynamics of mutation patterns. The pan-genome has a distinct advantage over traditional phylogenomics. It allows only the best matched and well-constructed residues from a multiple sequence alignment (MSA) (de Been *et al*., 2014) [14]. In the model of evolutionary events, the pan-genomic representation of several genomes provides for a precise encoding of the numerous genomic changes. This opens the door to significant new evolutionary findings in disciplines like the origin of complex life and animal evolution (Williams *et al*., 2013; Moroz *et al*., 2014). [111, 68]

Challenges and future directions

The availability of entire, well-annotated genome sequences is a major problem for pan-genomic research. However, by using machine learning, the accuracy of analysis is improved in different genomic analyses (Kumar *et al*., 2022) [91] but it is very difficult or even impossible, to furnish the repetitive regions assembly due to the short length of sequencing reads that leads to highly fragmented large repetitive genomes and limits the construction of near-complete genome sequences (Paten *et al*., 2017) [77]. The presentation and storage of the massive outputs of pan-genomic studies is also a constraint. Structural aberrations like indels and translocations may be unnoticed. Next, an improved annotation of the pan-genome products with critical functional and phenotypic information is another major hurdle.

Future pan-genomes will derive from biochemical alterations to the sequences. Extending the pan-genomic investigation to higher taxonomic groups will ultimately impart the different resources needed to analyze individuals' differences, enabling a comprehensive description of genes, evolutionary history, and functions (Tiwary *et al*., 2020) [108]. Simultaneously, appropriate data structures and computer algorithms are being constructed for better pan-genome data analysis across genera and species. SNPs, non-coding RNA, and indels are some of the additional elements that will require unique approaches in the future (Jha *et al*., 2022) [41]. Until now, there has been a significant advancement in orthology prediction methods. Additional approach for analysis of system biology is RNA-sea and transcriptomics. Transcriptomics is the study of the transcriptome i.e., the complete set of RNA transcripts that are produced by the genome, under specific circumstances or in a specific cell—using high-throughput methods, such as microarray analysis (Panigrahi *et al*., 2020) [15]. Several statistically robust techniques are required to distinguish orthologs and paralogs with the least false positives. A major challenge in this domain is better annotating the pan-genome with important functional and phenotypic information. Future pan-genomes will benefit from biochemical alterations to the sequences, such as hyper-methylated areas. A combination of gene expression and genomic data will also be necessary for linking the core genes with the shell and cloud genome and their associated expression levels. Further, the development of databases for pan-genome s will provide easier access to different datasets (Gerdol *et al*., 2020) [31].

Table 1. Summary of available pan-genome analysis tools

Tool	Description	Reference
Pan SEQ	For analysis of core and accessory genomic regions Make use of NCBI resources It has three main modules Novel Region Finder (NRF), Core and Accessory Genome Finder (CAGF) and Loci Selector (LS)	(Laing *et al*., 2010) [51]
PGAP PGAP X	Performs five analytical functions: cluster analysis of functional genes, pan-genome profile analysis, genetic variation analysis of functional genes, species evolution analysis and function enrichment of gene clusters	(Zhao *et al*., 2012) [99] (Zhao *et al*. 2012) [115]
Get Homologue	Performs customizable genome analysis and is targeted for non-bio informaticians Enables clustering of orthologous genes using multiple algorithms and filtering parameters	(Contreras-Moreira and Vinuesa, 2013) [19]
ITEP	Generates and curates protein families Curate protein families, compute similarities to externally defined domains Analyse gene gain and loss sand generate draft metabolic network	(Benedict *et al*., 2014) [6]
Split Mem	Generates compressed coloured de Bruijn graph of the pan-genome In the graph, nodes represent sequences which are common or unique within the population, and edges are the branch points between common or sample specific sequences	(Marcus *et al*., 2014) [61]
Pan GP	Performs scalable pan-genome analysis producing core genome, pan-genome and new genes curves Also implements two subsampling algorithms, which alleviate the computational burden of analysis of very high number of samples	(Zhao *et al*., 2014) [114]
LS-BSR (large-scale BLAST score ratio)	Calculates a score ratio (BSR value = query/reference bit score) BLAST (Altschul *et al*. 1997) or BLAT (Kent 2002). The output (bit score per CDS) can be visualized as a heatmap.	(Sahl *et al*., 2014) [90]
Harvest	It hosts three modules, namely Parsnip (core-genome analysis), Gingr (output visualization), and Harvest Tools (meta-analysis).	(Treangen *et al*., 2014) [109]
Micropan	Offers a set of tools designed for pan-genome analysis written in R Allows integration of pan-genome and additional analyses within a single programming language environment	(Snipen and Liland, 2015) [103]
Fri-Pan	Allows visualization of orthologous genes/gene clusters presence and absence for multiple strains Produces dendrogram and multidimensional scaling plots	(http://drpowell.github.io/ FriPan)
Pan-Tools	Supports the construction and visualization of pan-genome s Visual representation of the pan-genome is based on generalized De Bruijn graphs	(Sheikhizadeh *et al*., 2016) [100]
Pan *Viz*	Visualization tool with some analysis options. The input data needed is a pan-genome matrix as well as a gene ontology-based functional annotation of each gene group.	(Pedersen *et al*. 2017) [79]
SEQ-SEQ-pan	Workflow for the sequential alignment of sequences to build a pan-genome data structure and a whole-genome alignment.	(Jandrasits *et al*., 2018) [40]
PANINI	Implementing unsupervised machine learning with stochastic neighbour embedding based on the t-SNE (t-distributed stochastic neighbour embedding) algorithm;	(Abudahab *et al*., 2019) [1]
Pan GFR-HM	Web-based platform integrating functional and genomic analysis. Collection of ~1300 complete human-associated microbial genomes exploiting.	(Chaudhari *et al*., 2018) [12]
PAN2HGENE	Computational tool that allows identification of gene products missing from the original genome sequence	(Oliveira *et al*., 2021) [102]
Panakeia	Enables comparison of strains from different ecological niches. For diverse and highly clonal populations	https://github.com/BioSina/Panakeia

Table 2: Studies of pan-genomes in different organisms reported to date

Organism	Species	Genome size	Number	Core genome (%)	Detection method	Additional information
Bacteria	*Chlamydia trachomatis*	1.04 Mb	85	80	Homologous gene clustering	Allopatric
	Rickettsia prowazekii	1.1 Mb	10	8	Homologous gene clustering	Allopatric
	Mycobacterium tuberculosis	4.4 Mb	168	78	Homologous gene clustering	Allopatric
	Bacillus anthracis	5.2 Mb	50	51	Homologous gene clustering	Allopatric
	Streptococcus pneumoniae	2.2 Mb	52	31	Homologous gene clustering	Sympatric
	Hemophilus influenzae	1.8 Mb	55	33	Homologous gene clustering	Sympatric
	Escherichia coli	4.6 Mb	633	8	Homologous gene clustering	Sympatric
	Clostridium botulinum	3.9 Mb	46	5	Homologous gene clustering	Sympatric
Fungi	*Saccharomyces cerevisiae*	12 Mb	1011	63	ORF sequence similarity	Single-cell organism
	Candida albicans	15 Mb	34	91	Synteny	SCP commensalism
	Cryptococcus neoformans	9 Mb	25	8	Synteny	SCP commensalism

	Aspergillus fumigatus	29 Mb	2	83	Synteny	SCP commensalism
	Parastagonospora spp.	40 Mb	33	40	HCG	Plant pathogen
	Zymoseptoria tritici	40 Mb	5	58	HCG	Plant pathogen
Plants	*Brassica oleracea*	650 Mb	10	8	Read mapping	Outcrossing crop
	Glycine soya	1 Gb	7	49	HGC	Outcrossing crop
	Oryza sativa	430 Mb	3	92	Intersection of gene coordinates	Outcrossing crop
	Solanum lycopersicum	950 Mb	725	74	Read mapping	Outcrossing crop
	Triticum aestivum	17 Gb	19	64	Read mapping	Outcrossing crop
	Zea mays	2.4 Gb	503	39	Read mapping	Outcrossing crop
	Helianthus annus	3 Gb	493	83	Sequence similarity	Outcrossing crop
Human	*Homo sapiens*	3.2 Gb	2	86 additional gene	-	Asian and African genome
	Homo sapiens	3.2 Gb	910	-	-	African descent
	Homo sapiens	3.2 Gb	185	97	Read mapping	Han Chinese Individuals
Livestock	*Sus scrofa*	2.7 Gb	9	1737 additional gene	-	European and Chinese breeds
	Sus scrofa	2.7 Gb	12	-	*de novo* assembly and Hi-C	European and Chinese breeds
	Capra hircus	-	10 caprice assemblies	38.3 Mb additional sequences	*de novo* assembly and re sequencing	NCBI data
	Bos taurus	-	56	Additional 70,329,827 bases	Multi assembly graph	Brown Swiss along with five other breeds
	Ovis aries	-	684	142,422 indels	Pac Bio Hi Fi sequencing	13 diverse breeds

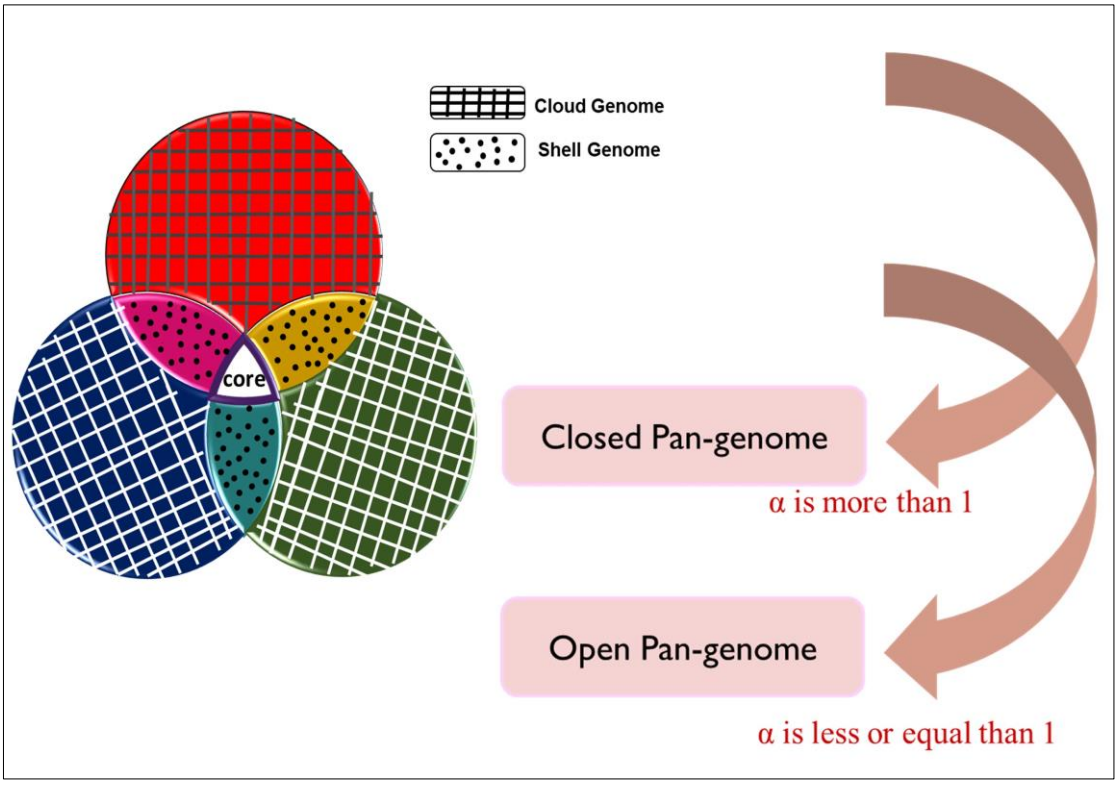

Fig 1: Different components and types of the pan-genome

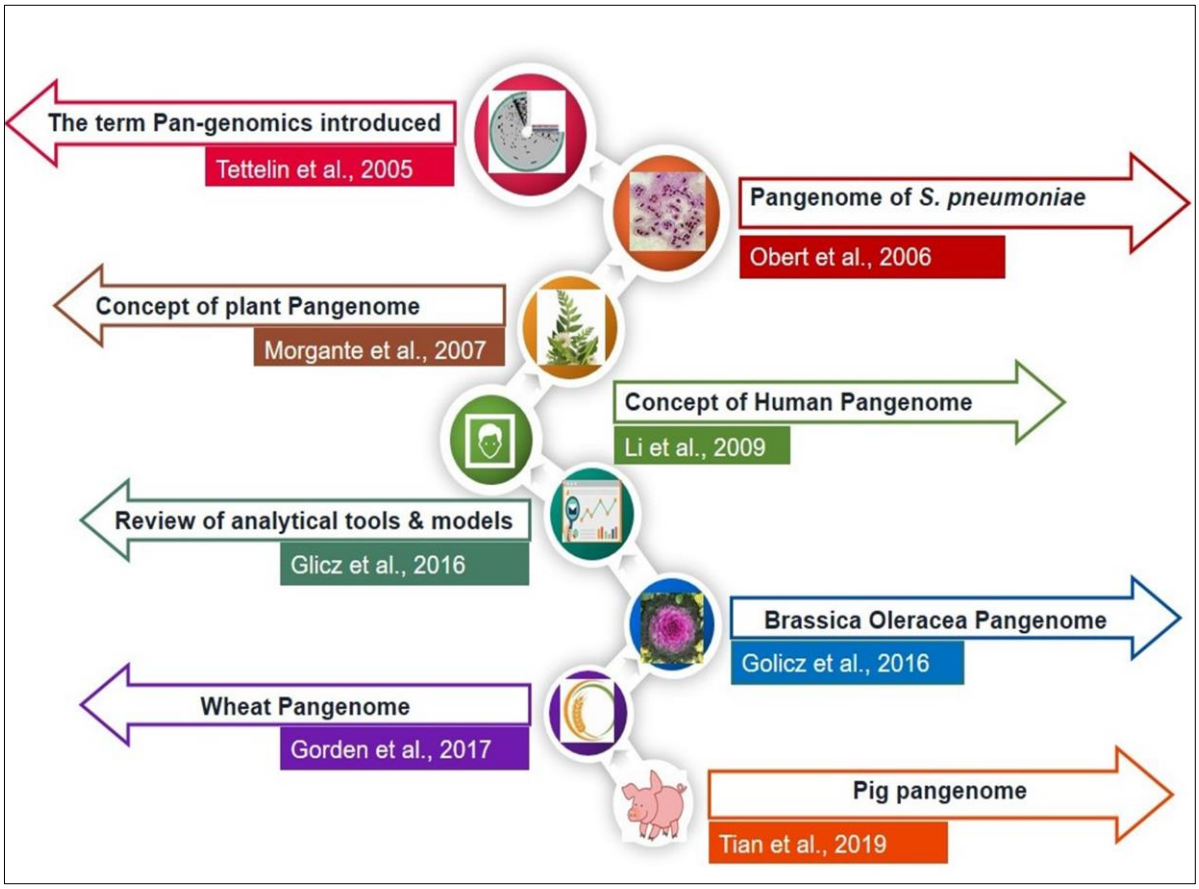

Fig 2: Timeline of development in pan-genomic research

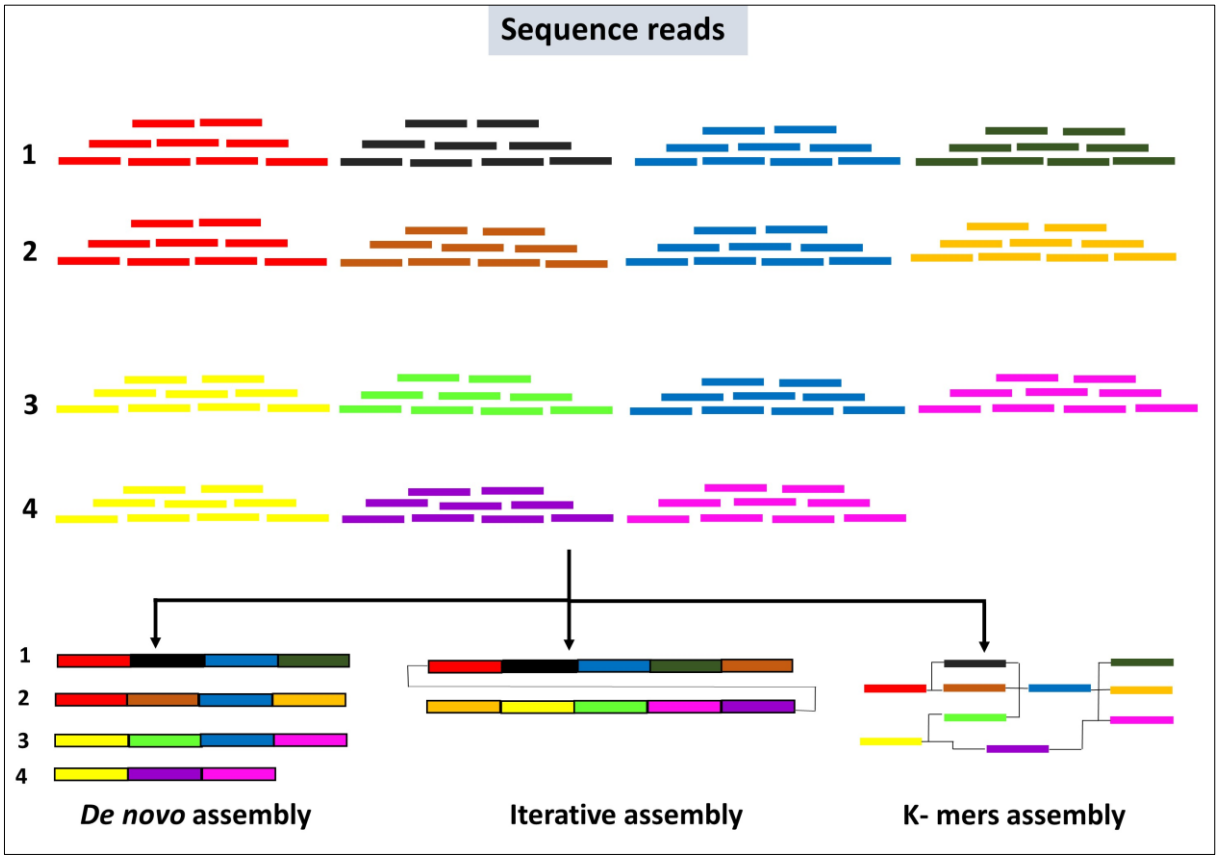

Fig 3: Different approaches to pan-genome assembly

Supplementary Table 1: The Reference assemblies used for pan-genome in different Livestock

S. No	The assemblies used for pig pan-genome construction
1	Chinese Pig breeds (6)
2	Tibetan
3	Wuzhishan
4	Jinhua
5	Meishan
6	Bamei
7	Rongchang
8	Europian Pig breeds (5)
9	Landrace
10	Large White Yorkshire
11	Pietran
12	Berkshire
13	Hampshire
The *de novo* assemblies used for goat pan-genome construction	
1	*Capra hircus* (ARS1)
2	*Capra hircus* (CHIR2.0)
3	*Capra Siberica* (CSI1.0)
4	*Ovis ammon* (Argali1.0)
5	*Ovis musimon* (Oori1)
6	*Ovis aries* (Oar4.0)
7	*Capra aegagrus* (CapAeg 1.0)
8	*Capra aegagrus* (Caeg1)
9	*Ammotragus lervia* (ALER1.0)
10	*Pseudois nayaur* (ASM318257v1)
The assemblies used for pig pan-genome construction	
1	Hereford
2	Angus
3	Highland
4	Original Braunvieh
5	Brahman
6	Yak

The assemblies used for pig pan-genome construction

Supplementary Table 2: The Reference genome assemblies of important livestock species

Animal	Species	First release	Genome size	Latest release	Sequencing centre	Reference
Cattle[57]	Bos taurus	2009	2.86 Gb	UMD3.1.1 (2014)	University of Maryland	Rosen *et al*., 2020 [124]
				Btau_5.0.1 (2015)	Cattle Genome Sequencing International Consortium, Baylor College of Medicine, Texas	
				ARS-UCD1.2 (2018)	University of Maryland	
Sheep[58,59]	Ovis aries	2010	2.71 Gb	OARv4.0 (2015)	International Sheep Genomics Consortium	International Sheep Genomics Consortium, 2010 [125]
				Oar_rambouillet_v1.0(2017) Oar_rambouillet_v2.0(2021)	Baylor College of Medicine Human Genome Sequencing Center	Davenport *et al*., 2022 [120]
Pig[60]	Sus scrofa	2011	2.7 Gb	Sscrofa11.1	The Swine Genome Sequencing Consortium (SGSC)	Archibald *et al*., 2010 [121]
Goat[61,62]	Capra hircus	2013	2.58 Gb	CHIR_2.0 (2015)	International Goat Genome Consortium, Beijing Genomics Institute	Dong *et al*., 2013 [122]
			2.63Gb	Saanen_v1 (2021)	Northwest A&F University	Li *et al*., 2021 [123]

Conclusion

In this new era, no doubt that pan genomic analysis will provide an excellent opportunity to find out the novel sequence, and there are vast applications. Pan-genome can help construct an improved reference genome in the different organisms for the thorough investigation of the core genetic variants. However, we are at the preliminary stage. Some significant concerns regarding the pan-genome analysis include low sequence reading accuracy, storage of massive data sets, reference-allele biases, and false-positive and negative structural variant prediction. Further advancements in next-generation sequencing, bioinformatics tools, and different biotechnological approaches will eventually resolve these concerns. Pan-genomic studies will become a mainstream approach in the next decade, leaving the single-reference approach outdated to genomic research.

Disclosure statement

The authors declare that they have no conflict of interest.

Funding

The author(s) reported there is no funding associated with the work featured in this article.

References

1. Abudahab K, Prada JM, Yang Z, *et al*. PANINI: pan-

genome neighbour identification for bacterial populations. Microb. Genom; c2019, 5(4).
2. Ahmad, SF, Panigrahi M, Chhotaray S, Pal D, Parida S, Bhushan B. *et al*. Revelation of genomic breed composition in a crossbred cattle of India with the help of Bovine50K Bead Chip. Genomics. 2020 Mar 1;112(2):1531-5. https://doi.org/10.1016/j.ygeno.2019.08.025
3. Ali A, Naz A, Soares SC, and *et al*. Pan-genome analysis of human gastric pathogen H. pylori: comparative genomics and pathogenomics approaches to identify regions associated with pathogenicity and prediction of potential core therapeutic targets. Biomed Res. In t; c2015.
4. Anani H, Zgheib R, Hasni I, *et al*. Interest of bacterial pan-genome analyses in clinical microbiology. Microb. Pathog. 2020;1(149):104275.
5. Bayer PE, Scheben A, Golicz AA, *et al*. modelling of gene loss propensity in the pan-genome s of three Brassica species suggests different mechanisms between polyploids and diploids. Plant Biotechnol. J 2021;19(12):2488-500.
6. Benedict MN, Henriksen JR, Metcalf WW, *et al*. ITEP: an integrated toolkit for exploration of microbial pan-genomes. BMC Genomics. 2014;15(1):1-1.
7. Berglund AC, Sjölund E, Östlund G, *et al*. InParanoid 6: eukaryotic ortholog clusters with inparalogs. Nucleic Acids Res. 2007;36(suppl-1):D263-6.
8. Bhardwaj T, Somvanshi P. Pan-genome analysis of Clostridium botulinum reveals unique targets for drug development. Gene. 2017;623:48-62.
9. Bidmos FA, Siris S, Gladstone CA, *et al*. Bacterial vaccine antigen discovery in the reverse vaccinology 2.0 Era: Progress and challenges. Front. Immunol. 2018;9:2315.
10. Bovine HapMap Consortium, Gibbs RA, Taylor JF, *et al*. Genome-wide survey of SNP variation uncovers the genetic structure of cattle breeds. Science. 2009;324(5926):528-32.
11. Butler J, MacCallum I, Kleber M, *et al*. ALLPATHS: de novo assembly of whole-genome shotgun microreads. Genome Res. 2008;18(5):810-20.
12. Chaudhari NM, Gautam A, Gupta VK, *et al*. PanGFR-HM: a dynamic web resource for pan-genomic and functional profiling of human microbiome with comparative features. Front. Microbiol. 2018;9:2322.
13. Chen F, Mackey AJ, Vermunt JK, *et al*. Assessing performance of orthology detection strategies applied to eukaryotic genomes. PloS One. 2007;2(4):e383.
14. Chen IM, Chu K, Palaniappan K, *et al*. IMG/M v. 5.0: an integrated data management and comparative analysis system for microbial genomes and micro biomes. Nucleic Acids Res. 2019;47(D1):D666-77.
15. Chhotaray S, Panigrahi M, Pal D, Ahmad SF, Bhushan B, Gaur GK, *et al*. Ancestry informative markers derived from discriminant analysis of principal components provide important insights into the composition of crossbred cattle. Genomics. 2020;112:1726-1733. https://doi.org/10.1016/j.ygeno.2019.10.008
16. Chhotaray S, Panigrahi M, Bhushan B, Gaur GK, Dutt T, Mishra BP. *Et al*. Genome-wide association study reveals genes crucial for coat color production in Vrindavani cattle. Livest. Sci. 2021a;247:104476. https://doi.org/10.1016/j.livsci.2021.104476
17. Chhotaray S, Panigrahi M, Pal D, Ahmad SF, Bhanuprakash V, *et al*. Genome-wide estimation of inbreeding coefficient, effective population size and haplotype blocks in Vrindavani crossbred cattle strain of India. Biol. Rhythm Res. 2021b;52(5):666-79. https://doi.org/10.1080/09291016.2019.1600266
18. Ciccarelli FD, Doerks T, Von Mering C, *et al*. Toward automatic reconstruction of a highly resolved tree of life. Science. 2006;311(5765):1283-7.
19. Contreras-Moreira B, Vinuesa P. GET_HOMOLOGUES, a versatile software package for scalable and robust microbial pan-genome analysis. Appl. Environ. Microbiol. 2013;79(24):7696-701.
20. Costa SS, Guimarães LC, Silva A, *et al*. First steps in the analysis of prokaryotic pan-genomes. Bioinsssform. Biol. Insights. 2020;14:1177932220938064.
21. Crysnanto D. Establishing Bovine Pan-genome *Graphs* (Doctoral dissertation, ETH Zurich); c2021.
22. De Been M, Lanza VF, de Toro M, *et al*. Dissemination of cephalosporin resistance genes between Escherichia coli strains from farm animals and humans by specific plasmid lineages. PLoS Genet. 2014;10(12):e1004776.
23. Derks MF, Gjuvsland AB, Bosse M, *et al*. Loss of function mutations in essential genes cause embryonic lethality in pigs. *PLoS* Genetics. 2019;15(3):e1008055.
24. Duan Z, Qiao Y, Lu J, *et al*. HUPAN: a pan-genome analysis pipeline for human genomes. Genome Biol. 2019;20(1):1-1.
25. Dutilh BE, Backus L, Edwards RA, *et al*. Explaining microbial phenotypes on a genomic scale: GWAS for microbes. Brief. Funct. Genomics. 2013;12(4):366-80.
26. Dutilh BE, van Noort V, van der Heijden RT, *et al*. Assessment of phylogenomic and orthology approaches for phylogenetic inference. Bioinform. 2007;23(7):815-24.
27. Eizenga JM, Novak AM, Sibbesen JA, *et al*. Pan-genome graphs. Annu. Rev. Genom. Hum. Genet. 2020;21:139-62.
28. Fleischmann RD, Adams MD, White O, *et al*. Whole-genome random sequencing and assembly of Haemophilus influenzae Rd. Science. 1995;269(5223):496-512.
29. Francis WR, Wörheide G. Similar ratios of introns to intergenic sequence across animal genomes. *Genome* Biol. Evol. 2017;9(6):1582-98.
30. Gege C, Hambruch E, Hambruch N, *et al*. Nonsteroidal FXR ligands: current status and clinical applications. Bile Acids and Their Receptors. 2019:167-205.
31. Gerdol M, Moreira R, Cruz F, *et al*. Massive gene presence-absence variation shapes an open pan-genome in the Mediterranean mussel. Genome Biol. 2020;21(1):1-21.
32. Goffeau A, Barrell BG, Bussey H, *et al*. Life with 6000 genes. Sci. 1996;274(5287):546-67.
33. Golicz AA, Batley J, Edwards D. Towards plant pangenomics. Plant Biotechnol. J. 2016;14(4):1099-105.
34. Hellerstein M. Stable Isotopes in Drug Development and Personalized Medicine: Biomarkers that Reveal Causal Pathway Fluxes and the Dynamics of Biochemical Networks. Stable Isotope Standards for Clinical Mass Spectrometry.97.
35. Hertzberg RP. Whole cell assays in screening for biologically active substances. Curr. Opin. Biotech.

1993;4(1):80-4.
36. Holt C, Yandell M. MAKER2: an annotation pipeline and genome-database management tool for second-generation genome projects. *BMC* Bioinform. 2011;12(1):1-4.
37. Hyatt D, Chen G-L, Locascio PF, et a. Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC* Bioinform. 2017;11(1):119. Available from: http://dx.doi.org/10.1186/1471-2105-11-119
38. Iqbal Z, Caccamo M, Turner I, *et al*. De novo assembly and genotyping of variants using colored de Bruijn graphs. *Nat. Genet.* 2016;44(2):226-32.
39. Jamal SB, Tiwari S, Silva A, *et al*. Pathogenesis of Corynebacterium diphtheriae and available vaccines: An Overview. Glob. j infect. Dis. clin. res. 2017;3(1):020-4.
40. Jandrasits C, Dabrowski PW, and Fuchs S, *et al*. Seq-seq-pan: building a computational pan-genome data structure on whole genome alignment. *BMC* Genome. 2018;19(1):1-2.
41. Jha UC, Nayyar H, PARIDA SK, *et al*. Progress of genomics-driven approaches for sustaining underutilized legume crops in the post-genomic era. Front. Genet. 2022 Apr 7:536.
42. Jun SR, Robeson MS, Hauser LJ, *et al*. PanFP: pan-genome -based functional profiles for microbial communities. *BMC* Res. Notes. 2015;8(1):1-7.
43. Kaisa K, Kumar H, Saravanan K, Rajawat D, Bhushan B, Kumar P. Concepts of Genomic Selection in Poultry and its Applications. Int. J. Livest. Res. 2020;10(10):32-42. https://doi.org/10.5455/ijlr.20200427022022
44. Kehr B, Helgadottir A, Melsted P, *et al*. Diversity in non-repetitive human sequences not found in the reference genome. Nat. Genet. 2017;49(4):588-93.
45. Kumar H, Panigrahi M, Chh0otaray S, Pal D, Bhanuprakash V, Saravanan KA. *et al*. Identification of breed-specific SNP panel in nine different cattle genomes. Biomed. Res. 2019;30(1). https://doi.org/10.35841/biomedicalresearch.30-18-1195
46. Kumar H, Panigrahi M, Chhotaray S, Parida S, Chauhan A, Bhushan B *et al*. Comparative analysis of five different methods to design a breed-specific SNP panel for cattle. Anim. Biotechnol. 2021a;32(1):130-6. https://doi.org/10.1080/10495398.2019.1646266
47. Kumar H, Panigrahi M, Panwar A, Rajawat D, Nayak S, S. Machine-Learning Prospects for Detecting Selection Signatures Using Population Genomics Data. J Comput. Bio; c2022. https://doi.org/10.1089/cmb.2021.0447
48. Kumar H, Panigrahi M, Rajawat D, Panwar A, Nayak S, Kaisa, K. Selection of breed-specific SNPs in three Indian sheep breeds using ovine 50 K array. Small Rumin. Res. c2021b;205:106545. https://doi.org/10.1016/j.smallrumres.2021.106545
49. Kumar H, Panigrahi M, Saravanan KA, Rajawat D, Parida S, *et al*. Genome-wide detection of copy number variations in Tharparkar cattle. Anim. Biotechnol; c2021c 1-8. https://doi.org/10.1080/10495398.2021.1942027
50. Kumar H, Panigrahi M, Saravanan KA, Paida S, Bhushan. Gaur GK, Dutt T. SNPs with intermediate minor allele frequencies facilitate accurate breed assignment of Indian Tharparkar cattle. Gene. 2021d;777:145473. https://doi.org/10.1016/j.gene.2021.145473
51. Laing C, Buchanan C, Taboada EN, *et al*. Pan-genome sequence analysis using Panseq: an online tool for the rapid analysis of core and accessory genomic regions. *BMC* Bioinform. 2010;11(1):1-4.
52. Lander ES, Linton LM, Birren B, *et al*. Initial sequencing and analysis of the human genome; c2001.p. 409-860
53. Li R, Fu W, Su R, *et al*. Towards the complete goat pan-genome by recovering missing genomic segments from the reference genome. Front. Genet. 2019;10:1169.
54. Li R, Gong M, Zhang X, *et al*. The first sheep graph pan-genome reveals the spectrum of structural variations and their effects on different tail phenotypes. BioRxiv; c2021.
55. Li R, Li Y, Zheng H, *et al*. Building the sequence map of the human pan-genome. Nat. Biotechnol. 2010;28(1):57-63.
56. Li Y, Hu Y, Bolund L, Wang J. State of the art de novoassembly of human genomes from massively parallel sequencing data. *Hum. Genom.* 2010;4(4):1-7.
57. Liti G, Carter DM, Moses AM, *et al*. Population genomics of domestic and wild yeasts. *Nature*. 2009;458(7236):337-41.
58. Liu Y, Du H, Li P, *et al*. Pan-genome of wild and cultivated soybeans. Cell. 2020;182(1):162-76.
59. Ma C, Lo PK, Xu J, *et al*. Molecular mechanisms underlying lignocellulose degradation and antibiotic resistance genes removal revealed via metagenomics analysis during different agricultural wastes composting. Bioresour. Technol. 2020;314:123731.
60. Manuweera B, Mudge J, Kahanda I, *et al*. Pan-genome -wide association studies with frequented regions. InProceedings of the 10th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics; c2019. p. 627-632.
61. Marcus S, Lee H, Schatz MC. SplitMEM: a graphical algorithm for pan-genome analysis with suffix skips. Bioinform. 2014;30(24):3476-83.
62. McCarthy CG, Fitzpatrick DA. Pan-genome analyses of model fungal species. Microb. Genom; c2019.
63. Medini D, Donati C, Tettelin H, *et al*. The microbial pan-genome. Curr. Opin. Genet. Dev. 2005;15(6):589-94.
64. Mehrotra A, Bhushan B, Karthikeyan A, Singh A, Panda S, Bhati M, Panigrahi M *et al*. Genome-wide SNP data unravel the ancestry and signatures of divergent selection in Ghurrah pigs of India. Livest Sci. 2021a;250:104587. https://doi.org/10.1016/j.livsci.2021.104587.
65. Mehrotra A, Bhushan B, Kumar A, Panigrahi MAK Singh A, Tiwari AK. A 1.6 Mb region on SSC2 is associated with antibody response to classical swine fever vaccination in a mixed pig population. Anim. Biotechnol. 2021b;1-6. Advance online publication. https://doi.org/10.1080/10495398.2021.1873145
66. Mira A, Martín-Cuadrado AB, D'Auria G, *et al*. The bacterial pan-genome: a new paradigm in microbiology. Int Microbiol. 2010;13(2):45-57.
67. Morgante M, De Paoli E, Radovic S. Transposable elements and the plant pan-genomes. Curr. Opin. Plant Biol. 2007;10(2):149-55.
68. Moroz LL, Kocot KM, Citarella MR, *et al*. The ctenophore genome and the evolutionary origins of neural systems. Nature. 2014;510(7503):109-14.
69. National Research Council. The new science of metagenomics: revealing the secrets of our microbial planet.
70. Pal D, Panigrahi M, Chhotaray S, *et al*. Unraveling genetic admixture in the Indian crossbred cattle by

different approaches using Bovine 50K BeadChip. Trop. Anim. Health Prod. 2022;54(2):1-8.
71. Pal D, Panigrahi M, Chhotaray S, Kumar H, Nayak SS, Rajawat D *et al*. Unraveling genetic admixture in the Indian crossbred cattle by different approaches using Bovine 50K BeadChip. Trop. Anim. Health Prod. 2022;54(2):135. https://doi.org/10.1007/s11250-022-03133-7
72. Panigrahi M, Kumar H, Nayak SS, Rajawat D, Parida S, Bhushan B. Molecular characterization of CRBR2 fragment of TLR4 gene in association with mastitis in Vrindavani cattle. Microb. Pathog. 2022a;165:105483. https://doi.org/10.1016/j.micpath.2022.105483
73. Panigrahi M, Kumar H, Sah V, Dillipkumar Verma A, Bhushan B, Parida S. Transcriptome profiling of buffalo endometrium reveals molecular signature distinct to early pregnancy. Gene. 2020;743:144614. https://doi.org/10.1016/j.gene.2020.144614
74. Panigrahi M, Kumar H, Saravanan KA, Rajawat D, Sonejita Nayak S, Ghildiyal K. Trajectory of livestock genomics in South Asia: A comprehensive review. *Gene*, *843*, 146808. Advance online publication; c2022b. https://doi.org/10.1016/j.gene.2022.146808
75. Parasar P, Bhushan B, Panigrahi M, Kumar H, Kaisa K, Dutt T. Characterization of BoLA class II DQA and DQB by PCR-RFLP, cloning, and sequencing reveals sequence diversity in crossbred cattle. Anim. Biotechnol. Advance online publication; c2021.p.1-11. https://doi.org/10.1080/10495398.2021.2006205
76. Parra G, Bradnam K, Korf I. CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes. Bioinformatics. 2007;23(9):1061-7.
77. Paten B, Novak AM, Eizenga JM, *et al*. Genome graphs and the evolution of genome inference. Genome Res. 2017;27(5):665-76.
78. Patra B, Panigrahi M, Kumar H, Kaisa K, Dutt T, Bhushan B. Molecular and phylogenetic analysis of MHC class I exons 7-8 in a variety of cattle and buffalo breeds. Anim. Biotechnol. Advance online publication; c2021.p.1-7. https://doi.org/10.1080/10495398.2021.1999969
79. Pedersen TL, Nookaew I, Wayne Ussery D, *et al*. Pan *Viz*: interactive visualization of the structure of functionally annotated pan-genome s. Bioinform. 2017;33(7):1081-2.
80. Peter J, De Chiara M, Friedrich A, *et al*. Genome evolution across 1,011 Saccharomyces cerevisiae isolates. *Nature.* 2018;556(7701):339-44.
81. Piovesan A, Antonaros F, Vitale L, *et al*. Human protein-coding genes and gene feature statistics in 2019. *BMC Res. Notes.* 2019;12(1):1-5.
82. Plissonneau C, Hartmann FE, Croll D. Pan-genome analyses of the wheat pathogen Zymoseptoria tritici reveal the structural basis of a highly plastic eukaryotic genome. BMC Biol. 2018;16(1):1-6.
83. Qin J, Li Y, Cai Z, Li S, *et al*. A metagenome-wide association study of gut microbiota in type 2 diabetes. Nature. 2012;490(7418):55-60.
84. Rajawat D, Panigrahi M, Kumar H, *et al*. Identification of important genomic footprints using eight different selection signature statistics in domestic cattle breeds. *Gene.* 2022a;11:146165.
85. Rajawat D, Panigrahi M, Kumar H, *et al*. Revealing Genomic Footprints of Selection for Fiber and Production Traits in Three Indian Sheep Breeds. J Nat. Fibers. 2022b;2:1-2.
86. Rappuoli R. Reverse vaccinology. Curr. Opin. Microbiol. 2000;3(5):445-50.
87. Richard GF. Eukaryotic pan-genome s. The *Pan-genome.* 2020:253-91.
88. Rogers SJ, Vismara L, Wagner AL, *et al*. Autism treatment in the first year of life: a pilot study of infant start, a parent-implemented intervention for symptomatic infants. J Autism Dev. Disord. 2014;44(12):2981-95.
89. Rouli L, Merhej V, Fournier PE, *et al*. The bacterial pan-genome as a new tool for analysing pathogenic bacteria. New Microbes New Infect. 2015;7:72-85.
90. Sahl JW, Caporaso JG, Rasko DA, *et al*. The large-scale blast score ratio (LS-BSR) pipeline: a method to rapidly compare genetic content between bacterial genomes. Peer J. 2014;2:e332.
91. Saravanan KA, Panigrahi M, Kumar H, Rajawat D, Nayak SS, Bhushan B. Role of genomics in combating COVID-19pandemic. Gene. 2022b;823:146387. https://doi.org/10.1016/j.gene.2022.146387
92. Saravanan KA, Panigrahi M, Kumar H, Bhushan B. Advanced software programs for the analysis of genetic diversity in livestock genomics: a mini review. Biol. Rhythm Res; c2019. p. 1-11. https://doi.org/10.1080/09291016.2019.1642650
93. Saravanan KA, Panigrahi M, Kumar H, Bhushan B, Dutt T, Mishra BP. Genome-wide analysis of genetic diversity and selection signatures in three Indian sheep breeds. Livest. Sci. 2021a;243:104367. https://doi.org/10.1016/j.livsci.2020.104367
94. Saravanan KA, Panigrahi M, Kumar H, Bhushan B, Dut T, Mishra BP. Selection signatures in livestock genome: A review of concepts, approaches and applications. Livest. Sci. 2020a;241:104257. https://doi.org/10.1016/j.livsci.2020.104257.
95. Saravanan KA, Panigrahi M, Kumar H, Nayak SS, Rajawat D, Bhushan B, Dutt T. Progress and future perspectives of livestock genomics in India: a mini review. Anim. Biotechnol; c2022a. https://doi.org/10.1080/10495398.2022.2056046
96. Saravanan KA, Panigrahi M, Kumar H, Parida S, Bhushan B, Gaur GK *et al*. Genome-wide assessment of genetic diversity, linkage disequilibrium and haplotype block structure in Tharparkar cattle breed of India. Anim. Biotechnol; c2020b.p.1-15. https://doi.org/10.1080/10495398.2020.1796696
97. Saravanan KA, Panigrahi M, Kumar H, Parida S, Bhushan B, Gaur GK *et al*. Genomic scans for selection signatures revealed candidate genes for adaptation and production traits in a variety of cattle breeds. Genomics. 2021b;113:955-963. https://doi.org/10.1016/j.ygeno.2021.02.009
98. Schatz MC, Maron LG, Stein JC, *et al*. Whole genome de novo assemblies of three divergent strains of rice, Oryza sativa, document novel gene space of au's and indica. ss. 2014;15(11):1-6.
99. Seib KL, Zhao X, Rappuoli R. Developing vaccines in the era of genomics: a decade of reverse vaccinology. Clin. Microbiol. Infect. 2012;18:109-16.
100. Sheikhizadeh S, Schranz ME, Akdel M, *et al*. PanTools: representation, storage and exploration of pan-genomic data. Bioinform. 2016;32(17):i487-93.

101. Sherman RM, Forman J, Antonescu V, *et al*. Assembly of a pan-genome from deep sequencing of 910 humans of African descent. *Nat. Genet.* 2019;51(1):30-5.
102. Silva de Oliveira M, Thyeska Castro Alves J, Henrique Caracciolo Gomes de Sá P, *et al*. PAN2HGENE–tool for comparative analysis and identifying new gene products. *Plos one.* 2021;16(5):e0252414.
103. Snipen L, Liland KH. micropan: an R-package for microbial pan-genomics. *BMC Bioinfor.* 2015;16(1):1-8.
104. Tao Y, Zhao X, Mace E, *et al*. Exploring and exploiting pan-genomics for crop improvement. *Molec. Plant.* 2019;12(2):156-69.
105. Tettelin H, Masignani V, Cieslewicz MJ, *et al*. Genome analysis of multiple pathogenic isolates of Streptococcus agalactiae: implications for the microbial "pan-genome". Proc. Natl. Acad. Sci. U.S.A. 2005;102(39):13950-5.
106. Thorvaldsdóttir H, Robinson JT, Mesirov JP. Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. *Brief.* Bioinform. 2013;14(2):178-92.
107. Tian X, Li R, Fu W, *et al*. Building a sequence map of the pig pan-genome from multiple de novo assemblies and Hi-C data. Sci. China Life Sci. 2019;63(5):750-63.
108. Tiwary BK. Evolutionary pan-genomics and applications. InPan-genomics: Applications, Challenges, and Future Prospects. Academic Press; c2020.p. 65-80.
109. Treangen TJ, Ondov BD, Koren S, *et al*. The Harvest suite for rapid core-genome alignment and visualization of thousands of intraspecific microbial genomes. Genome Biol. 2014;15(11):1-5.
110. Vernikos G, Medini D, Riley DR, *et al*. Ten years of pan-genome analyses. Curr. Opin. Microbiol. 2015;23:148-54.
111. Williams TJ, Wilkins D, Long E, *et al*. The role of planktonic F lavobacteria in processing algal organic matter in coastal East Antarctica revealed using metagenomics and metaproteomics. Environ. Microbiol. 2013;15(5):1302-17.
112. Zeng L, Wang D, Hu N, *et al*. A novel pan-genome reverse vaccinology approach employing a negative-selection strategy for screening surface-exposed antigens against leptospirosis. *Front.* Microbial. c2017;8:396.
113. Zerbino DR, McEwen GK, Margulies EH, *et al*. Pebble and rock band: heuristic resolution of repeats and scaffolding in the velvet short-read de novo assembler. *PloS One.* 2009;4(12):e8407.
114. Zhao Y, Jia X, Yang J, *et al*. PanGP: a tool for quickly analyzing bacterial pan-genome profile. Bioinform. 2014;30(9):1297-9.
115. Zhao Y, Wu J, Yang J, *et al*. PGAP: pan-genomes analysis pipeline. Bioinform. 2012;28(3):416-8.
116. Zimin AV, Marçais G, Puiu D, *et al*. The MaSuRCA genome assembler. Bioinformatics. 2013;29(21):2669-77
117. Parashar N, Hait S. Plastics in the time of COVID-19 pandemic: Protector or polluter?. Science of the Total Environment. 2021 Mar 10;759:144274.
118. Yaashikaa PR, Kumar PS, Varjani S, Saravanan A. A critical review on the biochar production techniques, characterization, stability and applications for circular bioeconomy. Biotechnology Reports. 2020a Dec 1;28:e00570.
119. Bashir MF, Ma B, Komal B, Bashir MA, Tan D, Bashir M. Correlation between climate indicators and COVID-19 pandemic in New York, USA. Science of the Total Environment. 2020 Aug 1;728:138835.
120. Struyf T, Deeks JJ, Dinnes J, Takwoingi Y, Davenport C, Leeflang MM, Spijker R, Hooft L, Emperador D, Domen J, Tans A. Signs and symptoms to determine if a patient presenting in primary care or hospital outpatient settings has COVID-19. Cochrane Database of Systematic Reviews; c2022.p.5.
121. Archibald S, Scholes RJ, Roy DP, Roberts G, Boschetti L. Southern African fire regimes as revealed by remote sensing. International Journal of Wildland Fire. 2010 Nov 5;19(7):861-78.
122. Dong Y, Wu X, Martini A. Atomic roughness enhanced friction on hydrogenated graphene. Nanotechnology. 2013 Aug 21;24(37):375701.
123. Bo HX, Li W, Yang Y, Wang Y, Zhang Q, Cheung T, Wu X, Xiang YT. Posttraumatic stress symptoms and attitude toward crisis mental health services among clinically stable patients with COVID-19 in China. Psychological medicine. 2021 Apr;51(6):1052-3.
124. Piccoli L, Park YJ, Tortorici MA, Czudnochowski N, Walls AC, Beltramello M, Silacci-Fregni C, Pinto D, Rosen LE, Bowen JE, Acton OJ. Mapping neutralizing and immunodominant sites on the SARS-CoV-2 spike receptor-binding domain by structure-guided high-resolution serology. Cell. 2020 Nov 12;183(4):1024-42.
125. International Sheep Genomics Consortium, Archibald AL, Cockett NE, Dalrymple BP, Faraut T, Kijas JW, Maddox JF, McEwan JC, Hutton Oddy V, Raadsma HW, Wade C. The sheep genome reference sequence: a work in progress. Animal genetics. 2010 Oct;41(5):449-53.