UNIVERSITA' DEGLI STUDI DI VERONA

DEPARTMENT OF Biotechnology

GRADUATE SCHOOL OF Natural Sciences and Engineering

DOCTORAL

PROGRAM IN

Biotechnology

XXXV cycle

TITLE OF THE DOCTORAL THESIS

Element-based graph pangenome

S.S.D. BIO/18

Coordinator: Prof. Matteo Ballottari

Tutor: Prof. Massimo Delledonne

Co-Tutor: Stephane Rombauts

Doctoral Student: Giulia Lopatriello

Element-based graph pangenome- Giulia Lopatriello

Tesi di dottorato

Verona,16/06/2023

SOMMARIO

Le analisi genomiche si basano sull'uso di un singolo genoma di riferimento, il quale però non rappresenta tutta la diversità genomica all'interno di una specie. Il pangenoma è stato ideato con lo scopo di descrivere l'intero contenuto genomico di una specie. I modelli presenti in letteratura si dividono in pangenomi lineari e pangenomi grafici basati sui nucleotidi. Il modello lineare contiene il genoma di riferimento insieme alle sequenze non rappresentate in esso (NRR). Questo modello permette di descrivere se un gene è presente in una determinata cultivar tramite l'analisi di presenza e assenza dei geni (analisi PAV). Il pangenoma grafico basato sui nucleotidi consente invece di visualizzare le somiglianze e le differenze locali delle regioni genomiche. I due modelli di pangenoma sono complementari tra loro: il pangenoma lineare crea un genoma consenso rappresentando presenza e assenza dei geni tra individui in una struttura tabellare, mentre il pangenoma grafico basato sui nucleotidi riporta graficamente tutte le possibili variazioni nucleotidiche, ma non può essere utilizzato per l'analisi di presenza e assenza dei geni.

Lo scopo di questa tesi è quello di creare un nuovo modello di pangenoma, chiamato pangenoma grafico basato su elementi, in cui è possibile rappresentare graficamente i geni ed effettuare l'analisi di presenza e assenza di essi. Il pangenoma grafico basato su elementi è stato costruito a partire dai geni identificati dall'l'annotazione automatica nei genomi di 5 diverse accessioni di *P. vulgaris* . I geni identificati sono stati convertiti in nodi di un grafo e sono stati collegati solo se adiacenti nel genoma. Successivamente, i geni ortologhi presenti nelle diverse cultivar sono stati identificati e collassati per rappresentare un singolo nodo del grafo.

I risultati hanno mostrato che, a causa della fusione di copie geniche originate da eventi di duplicazione (geni paraloghi), meno geni sono stati riportati nel pangenoma grafico basato sugli elementi rispetto al modello lineare. Inoltre, la visualizzazione delle regioni nel pangenoma grafico ad elementi è risultata più chiara rispetto a quella ottenuta nel modello grafico basato sui nucleotidi, poiché è stata rappresentata solo la presenza o l'assenza di geni tra le diverse cultivar. Rispetto agli altri modelli, il pangenoma grafico basato su elementi ha consentito di focalizzarsi sia a livello genico

sia a livello nucleotidico in una visualizzazione "zoom-in", mostrando anche le somiglianze e le dissomiglianze nucleotidiche locali.

In questo modo, il pangenoma grafico basato sugli elementi ha offerto una migliore interpretazione delle regioni genomiche, combinando i vantaggi dell'analisi della presenza/assenza dei geni con la visualizzazione grafica.

ABSTRACT

Genomic analyses are based on using a single reference genome that does not represent the whole intraspecies diversity. Instead, a pangenome contains the whole genome content of a species. State-of-art models for pangenome representations are divided into linear and nucleotide-based graph pangenomes. The linear model is composed by the reference genome with a set of non-representative reference (NRR) sequences, and it provides information about the presence of a gene in a certain cultivar through presence and absence analysis (PAV analysis). Nucleotide-based graph pangenome allows displaying of local similarities and dissimilarities of genomic regions. In this perspective, the two pangenome models are complementary to each other: the linear pangenome model creates a consensus genome reporting in a table representation the inter-individual gene presence and absence, whereas the nucleotide-based graph pangenome model displays graphically all possible nucleotide variations but cannot be used for gene presence and absence analysis.

The aim of this thesis was to create a new pangenome model, called element-based graph pangenome in which it is possible to graphically represent the genes and perform the analysis of the presence and absence of the reported genes. Briefly, genes were annotated in the genomes of 5 different cultivars of *Phaseolus vulgaris*, through automatic gene prediction. Genes representing nodes in the graph were linked only if they were adjacent in the genome. Then, orthologous genes between different cultivars were identified and merged to represent a single node in the graph. Developed element-based graph pangenome was compared to linear and nucleotide-based graph pangenome applied to the same bean accessions.

Results showed that due to the merging of gene copies derived by duplication event (paralogs), fewer genes were reported in element-based graph pangenome compared to linear model. Moreover, the visualization of regions was much clearer than that of nucleotide-based graph model, since only the presence or absence of genes across different cultivars was displayed. Different from other pangenome models, element-based graph pangenomes provided the advantage of moving information from the gene to the nucleotide in a" zoom-in" visualization, displaying local nucleotide similarities and dissimilarities.

In conclusion, the element-based graph pangenome offered a simpler interpretation of genomic regions, combining the advantages of analyzing the presence/absence of genes with the graphic visualization.

Sommario

INTRODUCTION

Discovery of pangenome

At the opening of the genomic era, it was thought that a single reference genome coming from an individual was sufficient to describe a species. However, in 2005, Tettelin et al.[1] questioned how many genomes were necessary to fully describe *Streptococcus agalactiae*, a bacterial species.

The sequencing of several strains of *Streptococcus agalactiae* [1], [2]revealed that ≈80% of any single genome was shared by all individuals (core genome) and the remaining part, the dispensable genome, was partially shared or private to individuals (Figure 1).

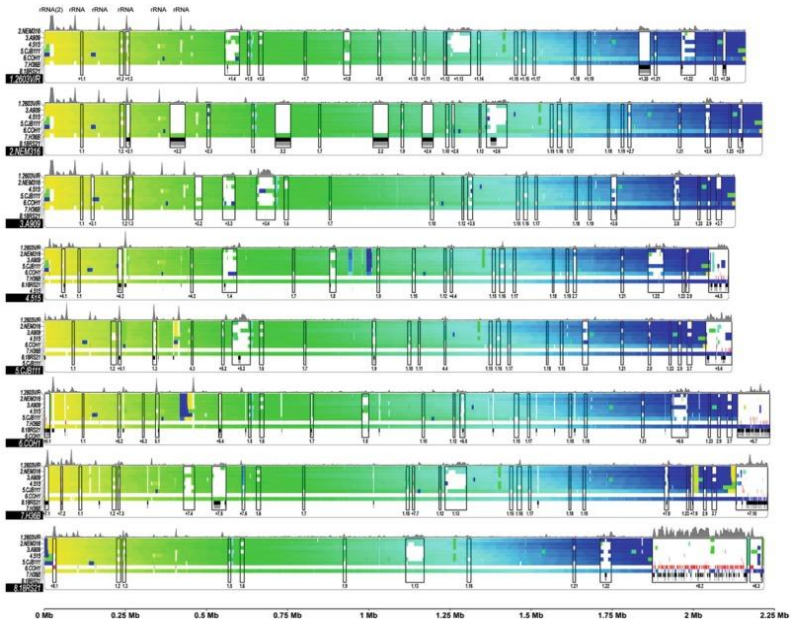

***Figure 1**. In silico comparative genomic analysis of Streptococcus agalactiae GBS genomes (Adapted from Tettelin et al., 2005)*

These findings underlined the low representativeness of the reference genome. Hence, to characterize the whole intraspecies, the term "Pangenome" has been defined by Tettelin et al. [1], [2]. This new genomic format describes a set of sequences which divides into:

- core sequences containing genes shared by all accessions,
- dispensable sequences with genes present in part of accessions,

- unique sequences with genes present in only one of the studied accessions (Figure 2A).

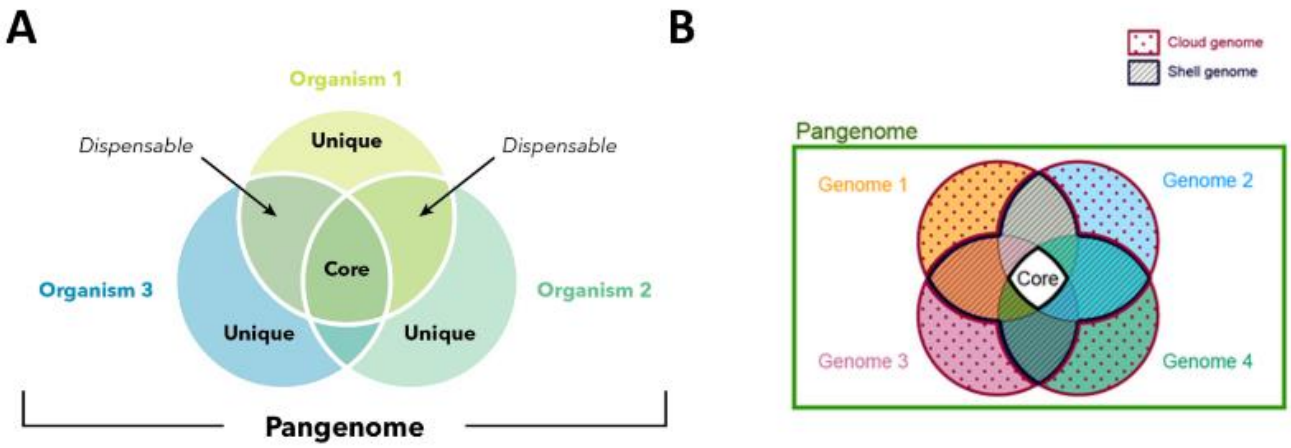

Figure 2. *Pangenome definition*

Alternatively, the pangenome could be subdivided into Softcore, Shell, and Cloud genome (Figure 2B). The softcore genome[3] comprises genomic regions or genes that are present in most individuals (>95%). Regions that are shared by at least 10% and less than 95% of individuals belong to the Shell genome. The remaining part, shared by a minor part of individuals (<10%) accounts for the Dispensable or Cloud genome.

The history of pangenome

In the same year of Tettelin, Morgante et al.[4] analysed regions derived from several allelic genome segments, in the maize inbreeds "Mo17" and "B73" (Figure 3A). The comparison of the two inbred lines revealed that 50% of the genome was shared, accounting for a size of 1.67 Gb (total genome size for each of the lines of 2.50 Gb). A dispensable genome of comparable size was found for each of the two lines: these genomic regions were made up mostly of transposable elements of different types.

Afterwards, Da Silva et al. [5]assembled the genome of the *Vitis vinifera* "Uruguayan Tannat clone UY11" and they performed an iterative mapping to reference against the "PN40024" reference genome[6], [7]. De novo transcriptome assembly found an amount of 1873 genes missing in "PN40024" reference genomes (Figure 3B).

Subsequently, the first plant pangenome was assembled in 2014 [7] where seven cultivars of *Glycine soya* were sequenced and de novo assembled. Li et al.[7] found that 80% of the genomes were shared by all individuals and the dispensable genome

contains more than 51% of gene families. Thus, the dispensable genome was containing genes involved in different biological processes and with new functions compared to genes present in all individuals.

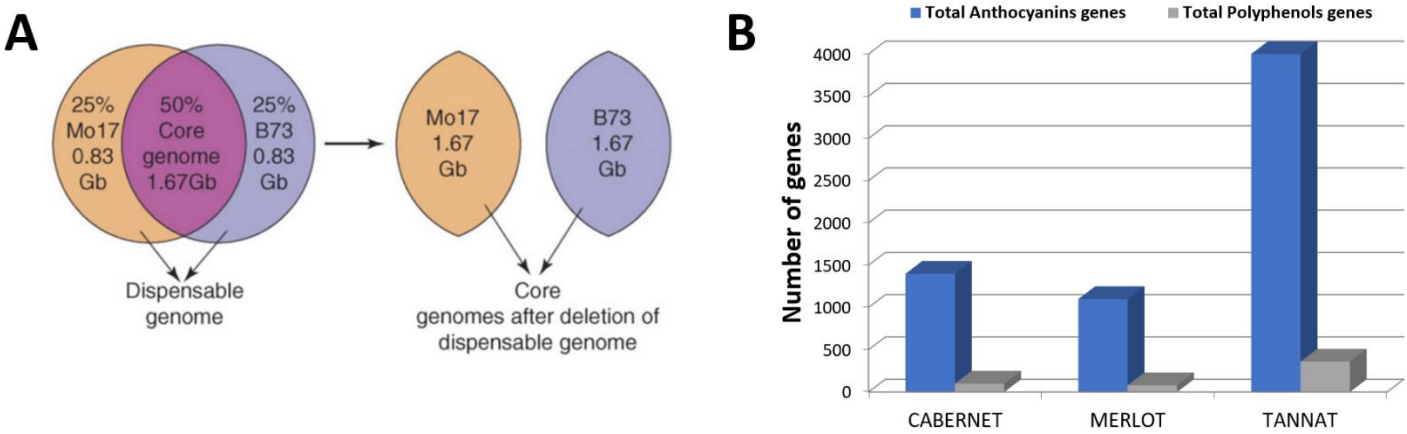

Figure 3. *(A) Pangenome of maize of lines Mo17 and B73 lines (Adapted from Morgante et al., 2005) (B) Pangenome of Vitis vinifera cultivars Cabernet, Merlot and Tannat*

Additionally, in 2017, a plant pan-genomic study involving 54 lines of grass *Brachypodium distachyon* [8] revealed that the rate of gene content identified with several individuals was twice the number of genes present in one individual. Dispensable genes identified by these studies were characterized to be involved in biotic stress response and development. Likewise, in 2019, the pan-genomic study applied to *Solanum lycopersicum L.* and its close wild relatives (*Solanum pimpinellifolium*, *Solanum cheesmaniae* and *Solanum galapagense*) reported 4,873 additional genes involved in disease-resistance biological processes [9] .

Although pangenomic studies have focused on bacterial and plant species, diversity among individuals was observed in animals (and humans) as well. Li et al. [10]identified in Asian and African individuals ~5 Mb of population-specific DNA which was not represented in the human reference genome. Consequently, these results suggested that the human pangenome would include an additional size of 19 to 40 Mb of novel information. Moreover, a study by Sherman et al. [11], published in 2018, through sequencing of 910 human individuals of African descent, confirmed that some regions (10% of total sequences) were missing from the human reference

genome, many of which contain protein-coding genes. Subsequently, The Human Pangenome Reference Consortium (HPRC) was established and in 2022 the HPRC published the first draft of the human pangenome [12]. This pangenome, which was generated starting from 47 phased, diploid assemblies from a cohort of individuals, adds 119 million base pairs of euchromatic polymorphic sequence and 1,529 gene duplications relative to the existing reference, GRCh38.

Types of pangenomes

All developed pangenomes can be divided into open or closed according to the growth of the number of genes per number of sequenced and analyzed genomes (Figure 4A).

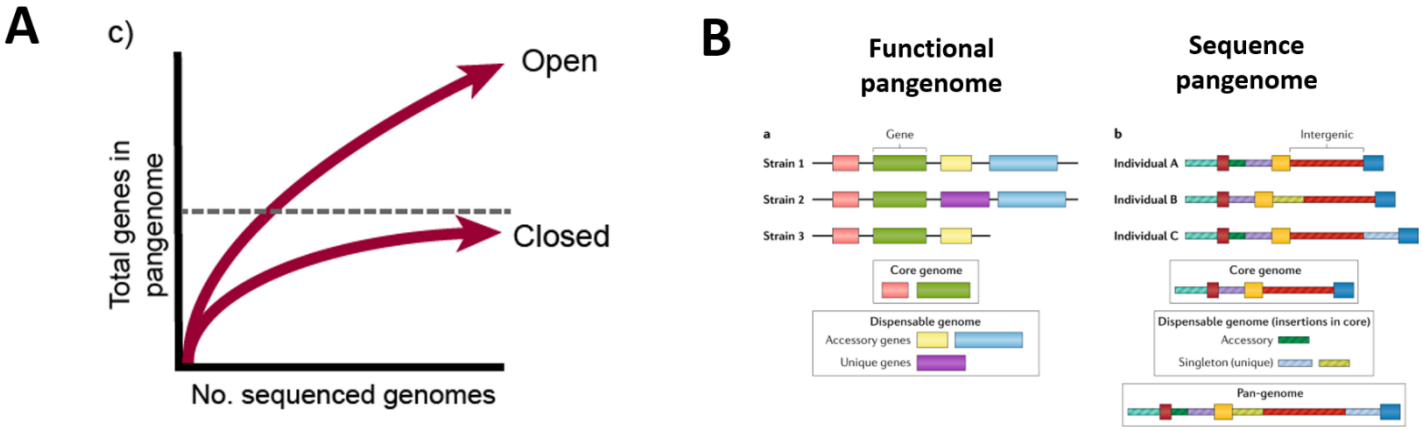

***Figure 4**. Classification of open and closed pangenomes (B) Functional and sequence pangenomes*

In open pangenomes[13], [14], the total number of sequences/genes increases as more genomes are added. Sympatric species which live in a community tend to have an open pangenome [15], [16].

In closed pangenomes[7]–[9], [17]–[21] ,the number of total sequences/genes will not increase with the increase of sequenced genomes. Closed pangenome are the ones that involved allopatric species since they live alone in their ecological niche[15], [16] .

Furthermore, genomes can be divided into functional and sequence pangenomes (Figure 4B).

A functional pangenome is described as a set of all genes for taxon representatives[13]. Such set contains redundant genes with the same function and

applying an additional analysis gene can be clustered into gene families[22]. However, information on the localization of genes in genomes is not reported. Usually, this form of pangenome has been largely applied to bacterial and other prokaryotic species.

In eukaryotic species, it is more interesting to study also intergenic and intronic regions. Hence, the pangenome has been extended to also non-coding genome. Hence functional pangenome became a sequence pangenome, which consists of a complete set of genomic sequences representing a species. Consequently, genomic sequences of individuals of the same species are compared with each other, generating a unique (non-redundant) set of DNA fragments and describing the structure of the pangenome [23], [24]

Additionally, state-of-art formats of pangenome divide into linear and nucleotide-based graph pangenomes.

Linear pangenome

Linear pangenome consists of a set of non-reference representative sequences collected with respect to the reference genome.

The construction of the linear pangenome divides into two phases. "Map-to-pan" is the first step of the generation of a linear pangenome, and it consists in assembling these extra sequences called non-reference representative (NRR) sequences. Several methods are available in the literature to assemble NRR sequences and they are divided into[25] (Figure 5):

- **metagenomic-like assembly of unaligned reads:** sequencing reads of each individual are aligned against the reference genome and unmapped reads of all individuals are pooled together to perform the assembly of NRR sequences[26]
- **independent assembly of unaligned reads:** unmapped reads of each individual are used to assemble a set of contigs specific for the individual. Contigs are then merged with the other individuals and are clustered at the DNA level to remove redundant sequences.
- **iterative assembly of unaligned reads:** iteratively, each individual is added to the pangenome, its sequencing reads are aligned against the pangenome,

and subsequently assembly of its deriving unmapped reads is performed and finally assembled contigs are added to the pangenome[27].

- **independent whole-genome assembly**: in this approach, each individual genome assembly is performed, and contigs are then mapped to reference genome to extract unaligned contigs. Subsequently, assembled contigs are then clustered together to remove redundancy and to construct the final set of NRR sequences[18], [28]–[30].

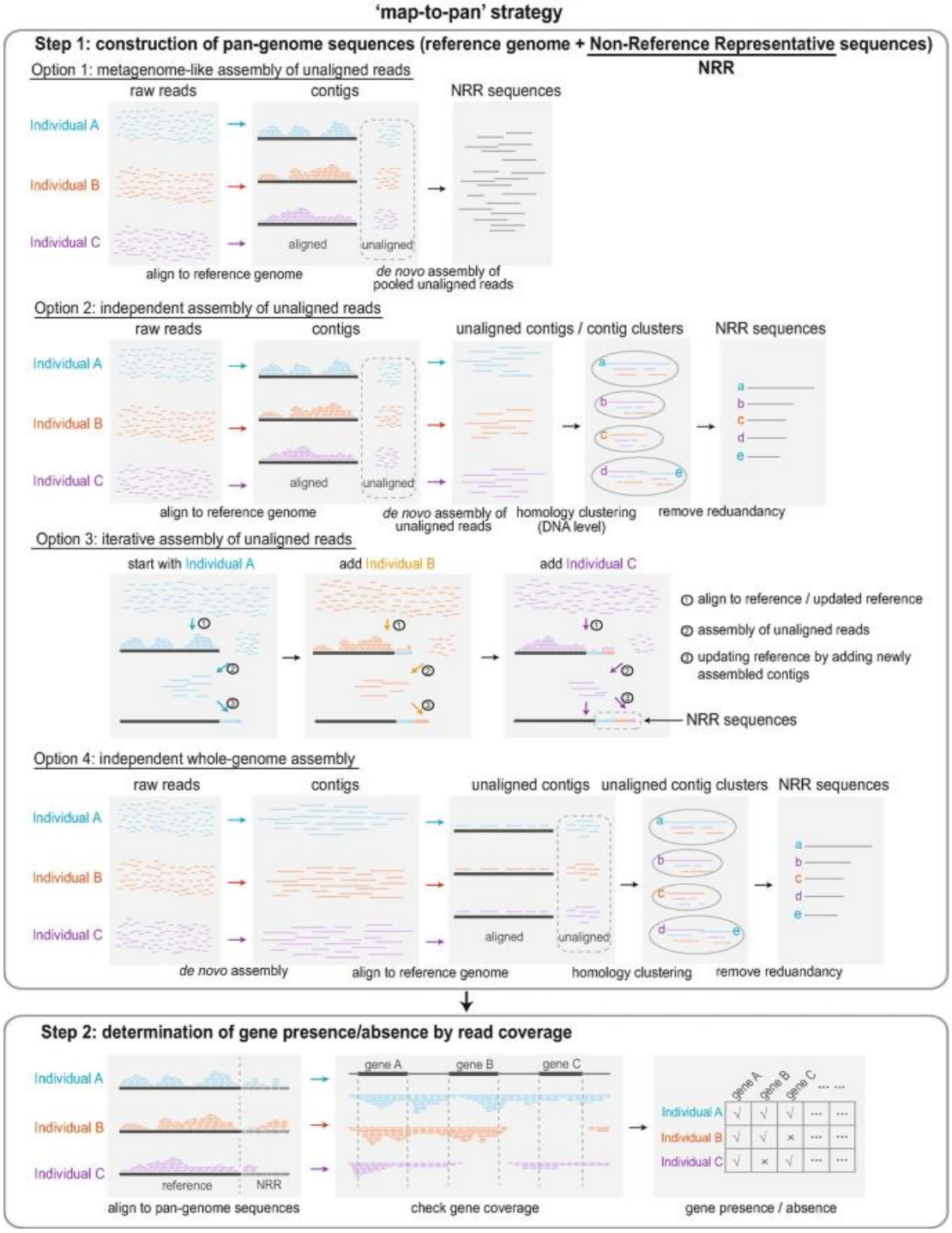

Figure 5. *Computational approaches to construct linear pangenome. In the upper panel, 4 approaches of map-to-pan setup are displayed in which NRR sequences of pangenome are generated. In the lower panel, determination of gene presence or absence by read coverage is displayed (Adapted from Hu et al., 2020)*

For better accuracy, all approaches of the map-to-pan phase require mapping the data against a reference genome with a low level of fragmentation.

Consequent to the map-to-pan phase, genes are annotated across the whole pangenome, and they are classified as core, variable and unique. In this analysis, called "Presence and absence analysis" (PAV analysis), usually short-read sequencing reads from several accessions are mapped against the pangenome and coverage of reads mapping in genic regions will define if a gene is present or not in a certain cultivar. To define the presence of the genes, different thresholds of coverage could be applied:

- having 95% of CDS covered by at least one read[18]
- having at least 85% of gene space covered [18]
- having 60% of CDS covered by at least one read[30]
- having at least 50% of gene space covered[21], [31], [32]

In the years, many pipelines have been created to automatically create linear pangenome on large-scale datasets. For instance, EUPAN[33] automatically performs the whole genome assembly approach in the "map to pan" phase, gene annotation and PAV analysis. Instead, the PSVCP[34] pipeline performs iterative assembly in the "map to pan" phase, PAV calling and SV identification. Panseq[35] which has been developed for bacterial pan-genomic studies, performed PAV analysis starting from a set of assembled genomes. PGAT[36] pipeline includes, besides the previously mentioned functionalities, also ortholog assignments, gene content analysis, SNP calling and enrichment analysis.

Although the widespread use in literature, the linear pangenome has the disadvantage of losing the information contained in reads not mapping to the reference and containing structural variants. In addition, information and coordinates of non-reference haplotypes are difficult to incorporate and consider in existing pipelines[37]. In conclusion, linear pangenome provides a consensus genome representation where gene presence and absence in different cultivars or individuals has not graphical visualization but it is represented in a table format.

Nucleotide-based graph pangenome

To represent inter-individual's nucleotide variation, a new graph format for the pangenome has been proposed in 2020 [38]. This data structure provides the advantage to simplify the representation of redundant sequences since conserved regions are compressed to represent a single haplotype in the pangenome (Figure 6A). Additionally, the main advantage is to directly infer similarities and dissimilarities of collected sequences in multiple alignments of the genomes.

Hence, while linear pangenome creates a sort of consensus model of the genome, the pangenome graph displays the whole inter-individual variation inside a population at nucleotide level.

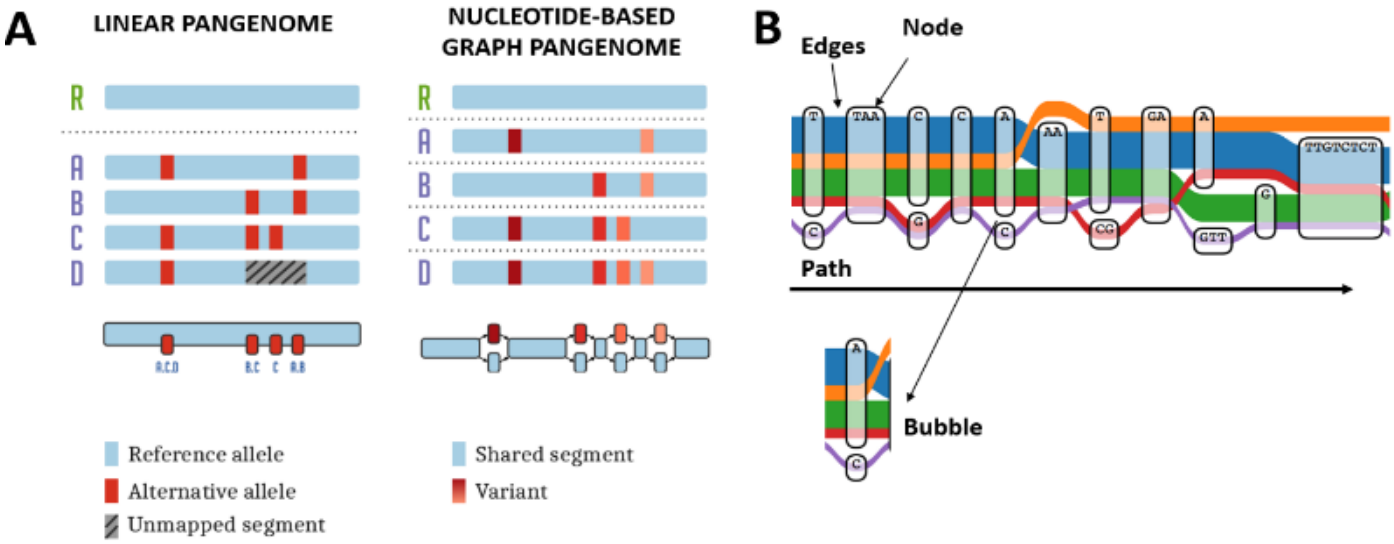

***Figure 6**. (A) Linear and nucleotide-based graph pangenomes (B) Nodes, edges and paths in nucleotide-based graph pangenome.*

The pangenome graph consists of two main components: nodes which represent genomic regions of individuals and edges that display the spatial connection of genomic regions (Figure 6B). Since the graph should be directed, a path in the graph describes the genome of one of the individuals involved in the pangenome. Hence, nodes in the path could describe reads, contigs, haplotypes or an entire genome. Moreover, bubbles (regions of the graph where multiple paths connect a common head and tail node) represent genomic regions which show divergence among the studied individuals.

As for linear format, a pangenome graph can be constructed starting from a set of genomes assemblies[39], [40] or from a set of resequencing data[41], [42].

Graphical Fragment Assembly (GFA) format was built to make a uniform format for representation purposes. However, visualization of pangenome graphs is adapted for each developed graph pangenome assembly tool. Nucleotide-based graph pangenome could be represented also in GAM (Graph Alignment/Map) format which includes alignment information deriving from SAM (Sequence Alignment/Map)/BAM (Binary Alignment/Map format. In addition, GAF (Graph Alignment Format) add to the present GFA format the text-based PAF (Pairwise Alignment Format). Recently graph mapping from GAF was also renamed as rGFA (Reference Graph Alignment Format.

Visualization tools as Bandage[43] globally display the graphs in terms of aspect and structures instead of base-level visualization (Figure 7). Instead, Sequence Tube Map [45]displays nucleotide variation at base-level and short-read mapping but at the local level.

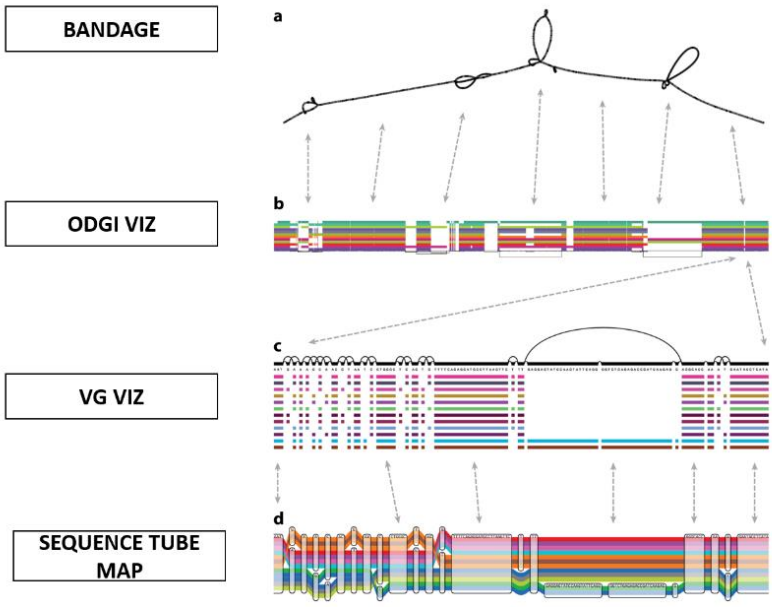

Figure 7. *Different graph visualization of region using Bandage(a), odgi viz(b), vg viz(c) and sequence-tube map(d)*

Nucleotide-based graph pangenome allows to integrate into the graph all detected structural variants as alternative paths in a genome graph. Reported structural variants improve the identification of other novel sequences in the pangenome graph

[46]. Recent studies demonstrated that graph pangenomes are more sensitive in identification of SVs with short reads with respect to the reference genome[41], [47]–[49].Since pangenome graphs were created to overcome the limitation of the linear reference genome, alignment tools have been created or adapted to work in graph pangenome format: some of these tools are GenomeMapper [50], Seven Bridges' Graph Genome Aligner[51], HISAT2 [52] (Hierarchical Indexing for Spliced Alignment of Transcripts 2) and V-MAP (Variant Map)[53].

To conclude, pangenome graph [54]can bring innovation to genomic studies biased by linear reference genome: decrease of the impact of reference bias, enhancement of mapping accuracy of sequencing data, increase of the sensibility of rare variant calling and improvement of de novo assembly of genomes of new individuals[37].

Particularly, higher precision and recall in variant identification (SNP, Indel, SV) was observed when using graph pangenome compared to linear pangenome in *Solanum lycopersicum* [55] species (Figure 8).

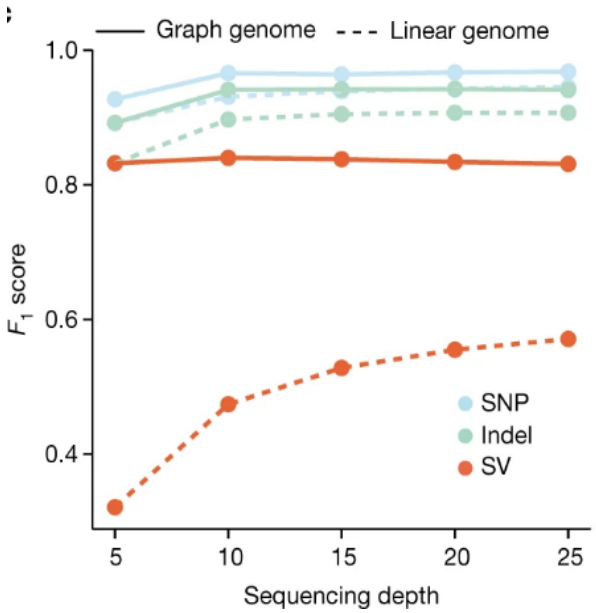

Figure 8. F1 scores (harmonic means of precision and recall) using simulated sequencing data from the genetic variants of 31 accessions with HiFi reads with different depths and genetic variants from the graph pangenome and the linear genome (Adapted from Zhou et al., 2020)

Gene annotation can be inserted into the nucleotide-based graph pangenome, but it cannot be directly applied to it. The gene annotation is applied to the sequences used

to generate the nucleotide-based graph pangenome. Then, the gene is transferred to a node if its coordinates intersect with the coordinates of the region represented by the node (Figure 9A).

Then, a node which describes nucleotide variation can partially or fully contain one or more genes and cannot describe exon boundaries (Figure 9B).

For visualization of the genic region, odgi [56] implemented a procedure to create a local graph starting from the coordinates of genes (Figure 9C).

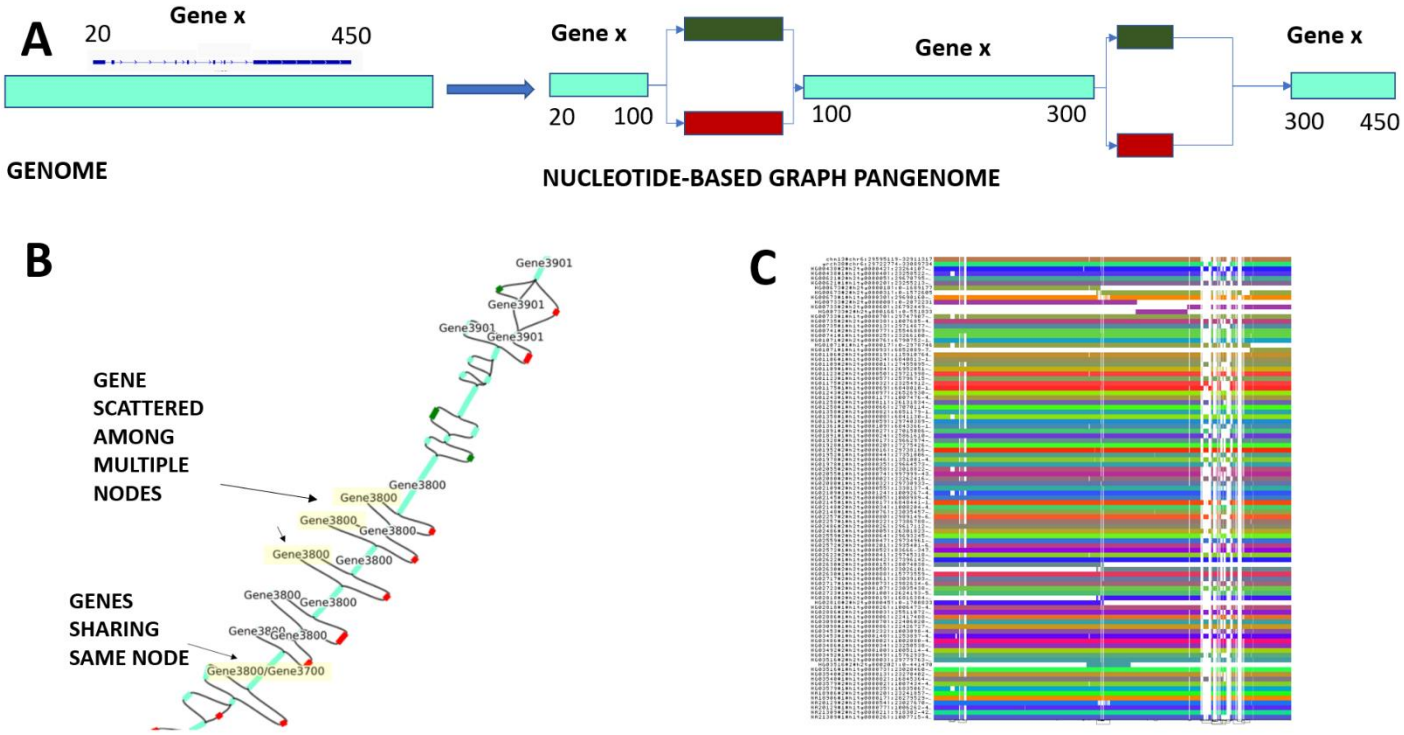

Figure 9 *(A) Gene annotation transfer to nodes of nucleotide-based graph pangenome (B) Bandage representation of a region including gene annotation (C)Subgraph created with odgi of MHC locus in human genome of 97 individuals. All contigs of the same haplotype are represented with the same colour . Most of the haplotypes has one contig covering the whole locus, meanwhile, in few of them, the locus is split in several contigs.*

However, as nucleotide-base graph pangenome represents the inter-individuals variation at nucleotide-level and genes are interspersed among multiple nodes, this model is not directly applicable for gene presence and absence analysis.

Automatic gene annotation

As pangenome analyses focus on gene content, identifying genes is an essential step in creating a pangenome model.

Gene annotation is the process of identifying protein-coding genes across the whole genome. Gene annotation accuracy has extremely high importance since errors introduced during gene annotation affect the downstream process.

Gene annotation can be performed in manual or automatic mode. Manual annotation exploits different types of experimental evidence to identify genes. However, this method requires big efforts in terms of time and costs. Instead, automatic gene annotation which relies on *ab initio* methods, reduces time and costs to obtain the final genome annotation. Considering that automatic gene prediction is not manually curated and a loss of accuracy is expected [57]the method achieving the best results in accuracy should be selected.

Two methods are mostly used to perform automatic gene annotation: the pure ab-initio method and the hint-based approach (Figure 10).

The pure *ab initio* method consists of exploiting a model describing the features of

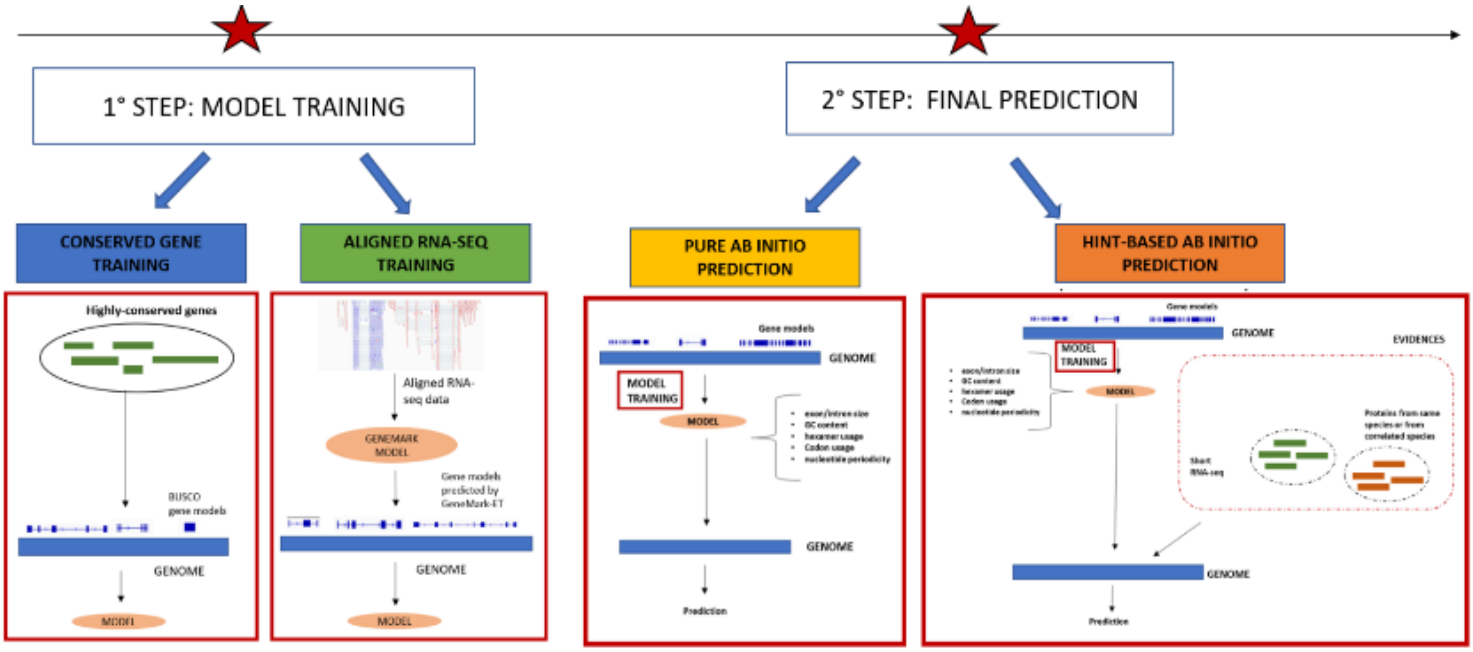

Figure 10. *Computational approaches for model training and final prediction in gene prediction*

the genes of the species of interest, to find the genes across the genome without the

use of experimental data. Conversely, in the hint-based approach, the ab-initio methods are integrated with experimental data to assist the predictor during the gene prediction.

Both approaches are based on Hidden Markov Model[57]which describes a transition between sub-models (Figure 11). These hidden models are associated with gene features (exon, intron, start and stop codon, 3' and 5' splice site, upstream and downstream intergenic region) and they are described by a probability of giving a particular observable nucleotide sequence.

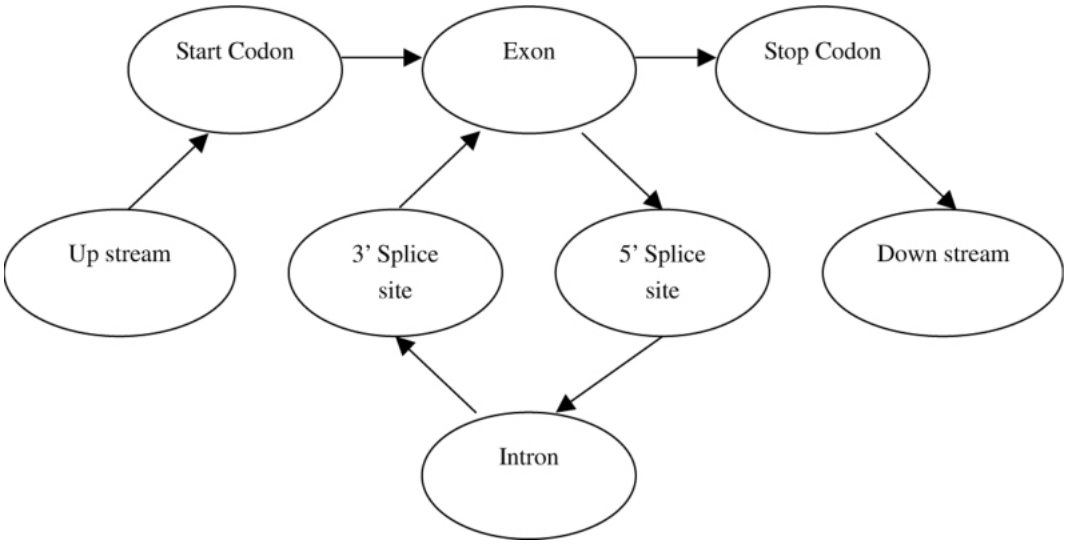

Figure 11. *Representation of Hidden Markov Model of gene prediction reporting the different gene features. (Adapted from Wang et al., 2004)*

However, as exon and intron length are important for gene detection and do not have a geometric distribution, a generalized Hidden Markov Model (GHMM) was developed to account also for the length distribution of different Markov submodels.

Hence, the gene model is a statistical model which relies on signal sensors and content sensors describing the intrinsic features of genes. The first type of sensor describes short sequence motifs belonging to the genic region, like for instance splice sites, branch points, polypyrimidine tracts, start codons and stop codons. On the other hand, content sensors describe nucleotide composition which is specific to the species of interest. Particularly, this feature allows for discriminating the open reading frame

(ORF) boundaries from the surrounding intergenic regions. In such a manner, being based on GHMM, *ab initio* gene predictors, like Genescan[58] GlimmerHMM[59], GeneID [60], FGENESH[61] Snap[62] Augustus[63]–[65] and GeneMark-ES [66] are able to identify unknown genes or genes that similarity-based approaches are not able to detect.

Additionally, in both approaches, model training is the most important step to achieve high-quality gene predictions. In this step, the model is trained on a set of genes contained in the organism of interest or correlated species.

The model training requires the use of bona fide gene structures: at least 200 genes should be used to train the model[67]. Bona fide gene structures can be constructed in a manual or automatic way. Two automatic procedures (BUSCO[68] and BRAKER [47], [48]) were implemented in the literature. The first method consists in automatically exploiting gene models that are highly conserved across species (BUSCO genes[68]–[71]). The second one consists in using the RNA-seq data to construct and predict gene models that will be used as a training set for the gene predictor.

Toward the *Phaseolus vulgaris* element-based graph pangenome

Phaseolus vulgaris, known as the common bean, has been defined as the most important grain legume for human nutrition. Like other legumes, beans are able to fix atmospheric nitrogen. The origin of the common bean has been largely debated. Common bean varieties are organized into gene pools named Mesoamericans and Andean. Their event of divergence aged more than 100,000 years ago [72] together with their domestication and adaptation events have brought to originate different landraces (Figure 12). As speciation, domestication and adaptation led to morphological and functional changes, the presence or absence of genes' variation is expected among different accessions of *P. vulgaris* and hence, this species represents an optimal example to apply the pangenome.

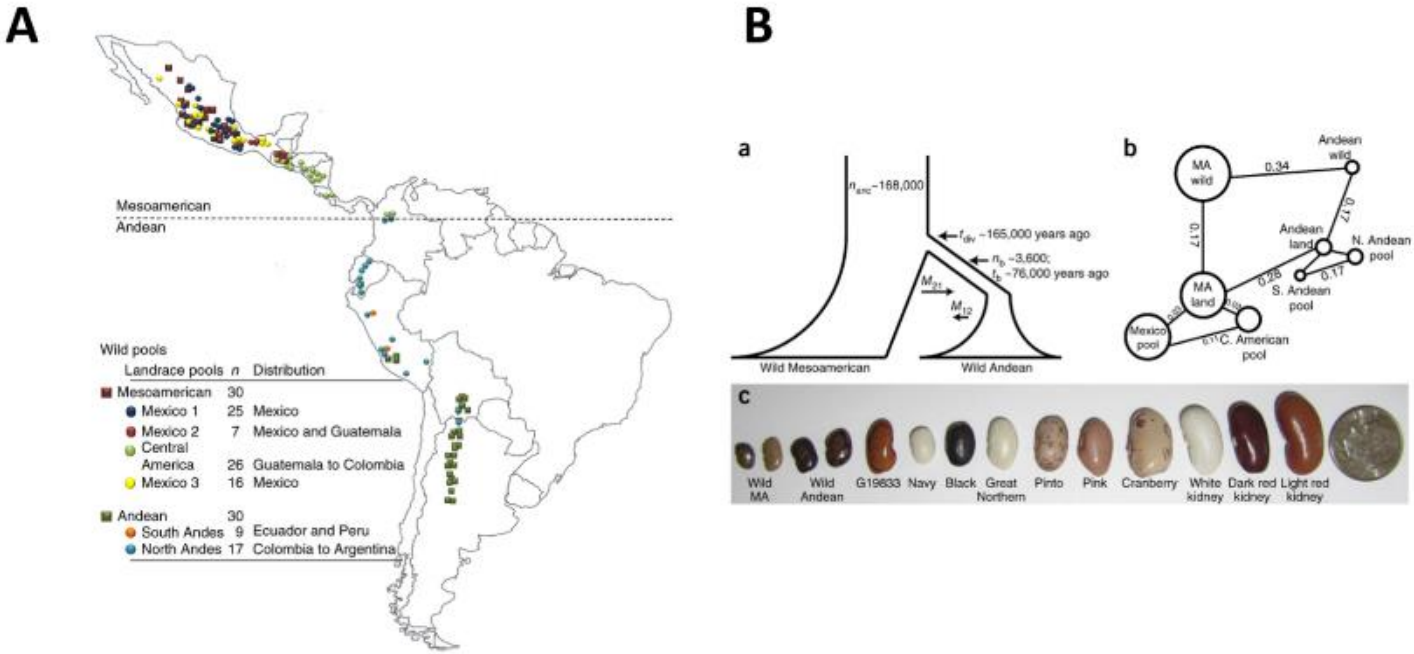

Figure 12. *(A) Geographic distribution of sampled genotypes (B) Divergence of the wild Mesoamerican and Andean. (Adapted from Schmutz et al., 2014)*

AIM OF THE THESIS

This thesis aims to address the limits of the linear and nucleotide-based graph pangenome, creating a new type of pangenome, named "element-based graph". In this pangenome format, we merge the advantages of linear pangenome (analysis of presence or absence of genes) with the advantages of nucleotide-based graph pangenome (graph visualization). The element-based graph pangenome was applied on 5 accessions of *P. vulgaris*. Prior to the development of element-based graph pangenome, a benchmark of approaches for each step (gene annotation and orthologous genes identification), was performed. The developed element-based graph pangenome was compared with linear and nucleotide-based pangenome to underline their strengths and weaknesses.

MATERIALS & METHODS

The reference genome of *Phaseolus vulgaris* (Pv442 and G19883 cultivar) [61] was downloaded from the Phytozome platform[73]Additionally, 4 other genomes corresponding to different cultivars were considered in the study: "MIDAS", "G12873", "BAT93" and "JaloEPP558" (Table 1). "BAT93" and "JaloEPP558" genomes have been provided by Institut National De La Recherche Agronomique whereas "MIDAS" and "G12873" have been sequenced and assembled by our laboratory.

	G19883 (Andean & Landrace)	**MIDAS (Andean & Landrace)**	**G12873 (Mesoamerican & Wild)**	**BAT93 (Mesoamerican & Landrace)**	**JALOEPP558 (Andean & Landrace)**
Total Assembly Size (bp)	**537,218,636**	509,180,482	584,993,346	637,803,808	606,487,223
Number of scaffolds	**478**	-	-	-	-
Number of contigs	**1,044**	1,913	6,293	1,441	1,061
Contigs average length (bp)	**509,156**	266,168	92,959	442,611	571,618
Contigs N50 (bp)	**1,885,876**	3,412,857	2,176,347	11,017,447	13,928,440
Contigs N90 (bp)	**377,857**	211,500	57,415	1,083,210	2,001,031
Longest contigs (bp)	**12,554,793**	24,636,533	20,321,960	36,599,242	45,469,680
Number of genes	**27,433**	-	-	-	-

Table 1. *Assembly statistics of 5 P. vulgaris acessions*

Automatic gene prediction in linear, element-based graph and nucleotide-based graph pangenomes

Training step

Two existing approaches were employed to perform automatic model training using bona-fide gene structures.

In approaches 1-2, conserved gene training was performed. Then training was performed using the software BUSCO [68] v4.1.4 and the specific database "fabales_odb10 genes" to train Augustus v3.3.3[63]–[65] model.

In approaches 3-4, training based on aligned RNA-seq data was performed. The training was based on BRAKER2 v2.1.4[69]–[71] software, which identifies gene models based on RNA-seq data of the selected species. To perform BRAKER2 [69]–[71] prediction, short RNA-seq data of 21 cultivars were aligned (unpublished data and from Bellucci et al[74]) against the genome using HISAT2[52] with the custom parameter of intron length (minimum intron length set of 23 kbp). Subsequently, BAM files were converted into hints and provided to BRAKER2 [69]–[71] software.

Prediction step

Prior to *ab initio* prediction, repetitive regions of the genomes have been soft-masked to avoid gene over-predictions. Repetitive regions were identified using RepeatMasker v 4.1.1 [75] software with a custom repeat library specific to each genome. Specific repeat libraries were identified using RepeatModeler2[76]with LTR module.

Gene prediction was conducted using 2 approaches: "pure" and hint-based prediction.

A pure *ab initio* approach was applied in approaches 1-3 where the trained model was used to perform a pure *ab initio* prediction using Augustus[63]–[65] on the soft-masked genomes.

In approaches 2-4, a hint-based approach using aligned RNA-seq data and protein was applied. Considering that aligned RNA-seq data could introduce noise, evidence introns supported by split RNA-seq reads with at least 20x coverage were provided to the Augustus predictor. As protein data, proteins annotated in the same species or

from correlated species (in *Phaseolus vulgaris*[77], *Medicago trunculata* [78]and *Glycine max*[79]) were aligned against the genome using GenomeThreader [80]

Quality metrics

Results for each procedure were evaluated in terms of:

- Filtering analysis
- BUSCO completeness or presence of highly conserved genes
- Gene sensitivity and specificity
- Fragmentation analysis

Filtering analysis

Filtering analysis was performed on the predictions of the four approaches to filter out transposon-related gene or artefacts gene prediction, falsely detected by the software. Predicted genes were scanned with InterProscan[81]v-5.46-81.0 for the presence of protein domains. Using a custom script, genes with transposons-related domains or without known protein domains were filtered out.

BUSCO completeness or presence of highly conserved genes

The presence of highly conserved genes was checked by performing BUSCO [68]completeness analysis using *fabales* database. The sequence of the predicted proteins of each tested approach was provided to BUSCO software v 4.1.4.

Gene sensitivity and specificity

Results were evaluated using gene sensitivity and specificity[82] metrics. The percentages of gene sensitivity and specificity were computed with the command evaluate_gtf by eval software[83] To have a ground truth, 2000 genes (not tagged as "highly-conserved" genes) were extracted from the current annotation of *P. vulgaris* reference annotation [77]This set of genes was divided randomly into 10 sets of testing genes to perform the evaluation ten times. To compute these metrics, obtained predictions were restricted to the gene locus of the 200 control genes, allowing 100 bp flanking regions for each gene locus (as suggested by the protocol in ref [67]).

Fragmentation analysis

Fragmentation analysis was performed using the protein length ratio (PLR) as a metric. PLR is computed as the ratio between a predicted protein and its respective protein present in *P. vulgaris* official annotation[77]. Predicted proteins were aligned

against the proteomes of *P. vulgaris* official annotation using blastp[84], [85]. Only best-hit for each predicted protein were considered according to bitscore (a metric provided by BLAST which measures sequence similarity independent of query sequence length and database size). Protein length was computed for all predicted proteins of the obtained annotation and occurrence of PLR values were plotted in a barplot.

Functional annotation

The filtered proteins were blasted against *Phaseolus vulgaris*[77], *Medicago trunculata* [78]and *Glycine max*[79] proteins with BLASTp[84], [85]v 2.12.0 and filtered by the best hits. The clustering of the predicted genes was performed with the proteins of all the species considered in the annotation using OrthoFinder [86], [87] v 2.5.4 and the functional annotation results were obtained through a custom script.

Development of linear pangenome

Map-to-pan phase

To build a linear pangenome, a non-iterative approach was applied using "MIDAS", "G12873", "BAT93" and "JaloEPP558" genome assemblies. Hence, the reference genome was independently mapped on each of the four genomes using minimap2 [87]v.2.17. Then, Assemblytics[88] v.1.2.1 was used to identify sequences that were considered deletions in the four alignments. Contigs, which were not aligned to the reference genome ("uncovered contigs"), were identified with samtools[89] depth v1.1.1. Both the deletions and the uncovered contigs were filtered for a minimum length of 1 Kb to keep only the sequences with significant length.

A clustering was performed to maintain only one orthologous sequence among the different accessions and to maintain all the paralogous. A sequence identity of 90% has been used as the threshold for the clustering phase and all the filtered sequences were clustered with CD-HIT-EST[90] v4.8.1.

Subsequently, the final accessory sequences obtained were then blasted blastn [84], [85] v2.9 against NCBI non-redundant nucleotide databases to remove the organellar contigs and the possible contaminants.

Automatic gene annotation

Gene annotation on non-representative reference (NRR) sequences has been applied using Approach 2. *P. vulgaris* of official annotation[77] was considered as annotation for reference genome.

PAV analysis

The presence/absence of the pangenome genes was defined with the realignment of "G19833", "MIDAS", "G12873", "BAT93" and "JaloEPP558" assemblies against the pangenome using minimap2[87] v.2.17 with option. A gene was called "present" in a cultivar if the coverage of the cultivar's assemblies computed with samtools [89]coverage v1.1.1 command was above the 5% of gene space region.

Development of nucleotide-based graph pangenome

Construction of nucleotide-based graph pangenome

A nucleotide-based graph pangenome was built from "G19833", "MIDAS", "G12873", "BAT93" and "JaloEPP558" genome assembliesusing the following software: minigraph[40], Pantools[91] and pggb[39] .

Nucleotide-based graph pangenome has been assembled using minigraph [40] with option -xggs. Pangenome graphs were generated using pggb [39] with option -p 90 -n 5 -t 20 -v -V 'G19833#'. Finally, Pantools [91] was also used to draft a nucleotide-based pangenome with kmer size of 255 nucleotides.

Bandage[43] odgi[56] and neo4j v3.5 [92] visualization tools were used to visualize the nucleotide-based pangenome generated with minigraph, pggb and Pantools, respectively.

Annotation of nucleotide-based graph pangenome

Genes were annotated in "MIDAS", "G12873", "BAT93" and "JaloEPP558" assemblies using Approach 2. *P. vulgaris* official annotation was used for "G19833" cultivar. The genes were reconnected to nodes of nucleotide-based graph pangenome, cross-checking to node coordinates and gene ones using bedtools[93] v2.26.0 intersect command and custom script.

Development of element-based graph pangenome

Automatic gene annotation

Gene annotation using Approach 2 was used to generate the nodes of the element-based graph pangenome using a custom script.

Annotation of spatial edges

Spatial edges were assembled by exploiting gene annotation in GTF[94] format using a custom script.

Annotation of orthologous edges and PAV analysis

Identification of orthologous genes was performed using both synteny and gene family analysis.

Genes contained in syntenic blocks were detected with i-ADHoRe[95]software. Internal to the analysis, homologous relationships between genes were detected through a protein alignment all-to-all as performed using blastall [96]using parameters recommended by MCScanX[97] Protein alignment of all cultivars and the gene annotation of all cultivars were provided for the analysis. Pairs of genes in synteny were extracted from the "multiplicon_pairs.txt" output file.

In the gene family approach, genes belonging to the same gene family were identified with Orthofinder [86]analysis using default parameters and the protein sequences of all cultivars. Orthologs genes of all cultivars were retrieved from the output "Orthologues" folder.

Subsequently, using a custom script, orthologs across different cultivars were collapsed to form single nodes. The same script was used to perform presence and absence analysis and to classify each gene as core, variable or unique.

Graphia [98] software was used to visualize local regions of element-based pangenome.

Gene Ontology (GO) enrichment analysis of unique genes have been performed using ClusterProfiler v3.18.1.

RESULTS

Automatic gene annotation was performed to linear, nucleotide-based graph and element-based graph pangenomes. Before construction of pangenome formats, an in-depth analysis of accuracy of approaches for automatic gene annotation was performed. After optimization of gene annotation, linear, nucleotide-based graph and element-based graph pangenomes were assembled and the obtained element-based graph format was compared with other two ones.

Benchmark of automatic gene annotation

Four automatic gene annotation approaches were benchmarked to assess the most accurate one. Approaches were applied to the *P. vulgaris* reference genome ("G19833" cultivar), and the results were compared with official annotation. The most accurate method was then applied to linear, nucleotide-based graph and element-based graph pangenomes.

The tested approaches were (Table 2):

- pure *ab initio* approach using a model trained with *fabales* highly conserved genes (Approach 1)
- hint-based approach using a model trained with *fabales* highly conserved genes (Approach 2)
- pure *ab initio* approach using a model trained with aligned RNA-seq data of 21 *P. vulgaris* cultivars (Approach 3)
- hint-based approach using a model trained with aligned RNA-seq data of 21 *P. vulgaris* cultivars (Approach 4)

<table>
<tr><th></th><th>APPROACH 1</th><th>APPROACH 2</th><th>APPROACH 3</th><th>APPROACH 4</th></tr>
<tr><td>1° STEP: MODEL TRAINING</td><td colspan="2">HIGHLY-CONSERVED GENE APPROACH (BUSCO)</td><td colspan="2">ALIGNED RNA-SEQ DATA (BRAKER)</td></tr>
<tr><td>2° STEP: FINAL PREDICTION</td><td>AB INITIO (AUGUSTUS)</td><td>HINT-BASED AB INITIO APPROACH (AUGUSTUS)</td><td>AB INITIO (AUGUSTUS)</td><td>HINT-BASED AB INITIO APPROACH (AUGUSTUS)</td></tr>
</table>

Table 2. *Computational approaches tested for automatic gene prediction.*

All tested approaches reported a higher number of predicted genes compared to the official annotation (Table 3).

	Official annotation	Approach 1	Approach 2	Approach 3	Approach 4
Number of initial predictions	27,433	30,053	32,387	34,786	36,713

Table 3. *Number of predictions for P. vulgaris official annotation and the four approaches*

Then, the quality of the four approaches was assessed according to:

- filtering analysis
- BUSCO completeness
- Gene sensitivity and specificity
- fragmentation analysis

Filtering analysis

Filtering analysis based on the presence of protein-coding domains was performed, to exclude potential artefacts ("unknown genes") or transposon-related genes.

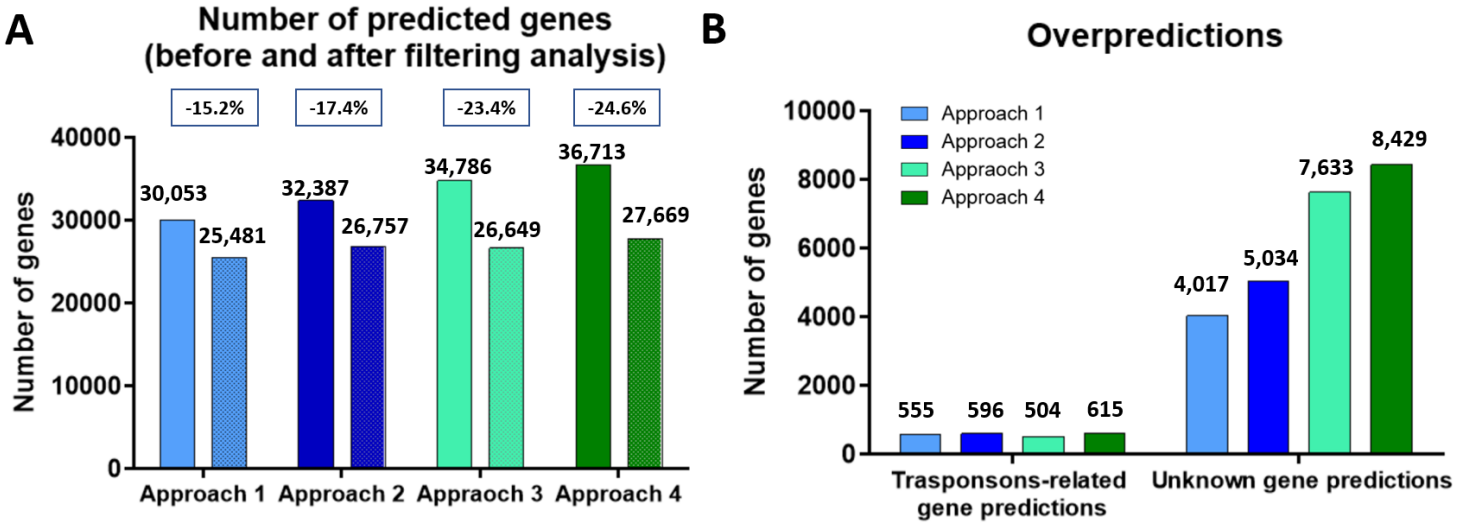

Figure 13. *(A) Number of predicted genes before and after filtering analysis (B) Number of predictions belonging to transposon-related and unknown genes.*

After filtering, the number of predicted genes of Approach 1 and 2 (BUSCO-training-based approaches) were reduced by 15.2% and 17.4% (Figure 13A, Table S1) whereas in Approaches 3-4, it was decreased by 23.4% and 24.6% (Figure 13A). Hence, a higher filtering prediction rate was observed in Approaches 3-4 (RNA-seq-training-based approaches), confirming the major presence of artefacts or transposon-related predictions.

Most of the excluded predictions in all approaches accounted for genes with no protein-coding domain in the ORF region ("unknown genes") while transposon-related genes account for a small amount (Figure 13B).

Finally, after the filtering analysis, Approaches 1, 2, 3 and 4 reported 25,481, 26,757, 26,649 and 27,669 final genes, respectively (Table S1). Therefore, the final annotation obtained in Approaches 2,3 and 4 showed a comparable number of genes related to *P. vulgaris* official annotation (27,433).

Hence, filtering analysis should be performed after automatic gene predictions to outcome the problems of over predictions.

BUSCO completeness or presence of complete genes highly conserved genes

The BUSCO completeness, known as presence of complete genes highly conserved (Figure 14A, TableS2) , is a commonly used quality metric for gene annotation [99]–[102].

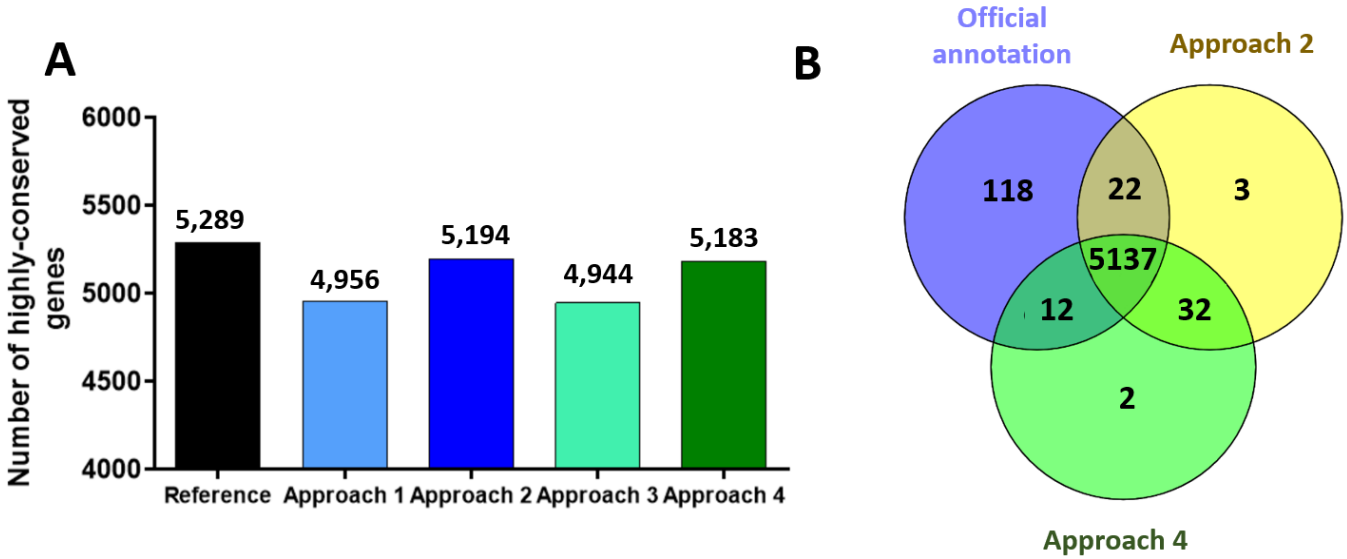

Figure 14. *(A) Presence of complete highly conserved genes of tested approaches (B) Venn diagram of complete highly conserved genes found by P. vulgaris official annotation and by Approaches 2-and 4*

Approach 1 and 3 predicted 4,956 and 4,944 highly-conserved genes (BUSCO complete genes), respectively, while hint-based approaches (Approaches 2 and 4) annotated 5,194 and 5,183, respectively. Thus, the integration of extrinsic evidence (RNA-seq and protein data) in hint-based approaches increased the sensitivity by approximately 238 genes. Additionally, Approach 2 marginally outperformed Approach 4 by 11 highly-conserved genes.

Since Approaches 2 and 4 (hint-based approaches) had better performance, they were compared with *P. vulgaris* official annotation (Figure 14B). The two approaches did not reach the same completeness achieved by *P. vulgaris* official annotation which contained 5,289 highly conserved genes. This difference was imputable to 118 highly-conserved genes present in *P. vulgaris* official annotation but not identified by the two tested approaches. However, this difference accounted for 0.21% of the 5,289 highly-conserved genes present in the official annotation.

Presence of complete genes non-highly conserved

Performances of the four prediction methods were assessed using gene sensitivity and specificity. Given the numerousness of the testing genes, the analysis was performed on ten different testing sets containing 200 genes. Thus, the efficiency of different approaches in gene sensitivity and specificity was provided.

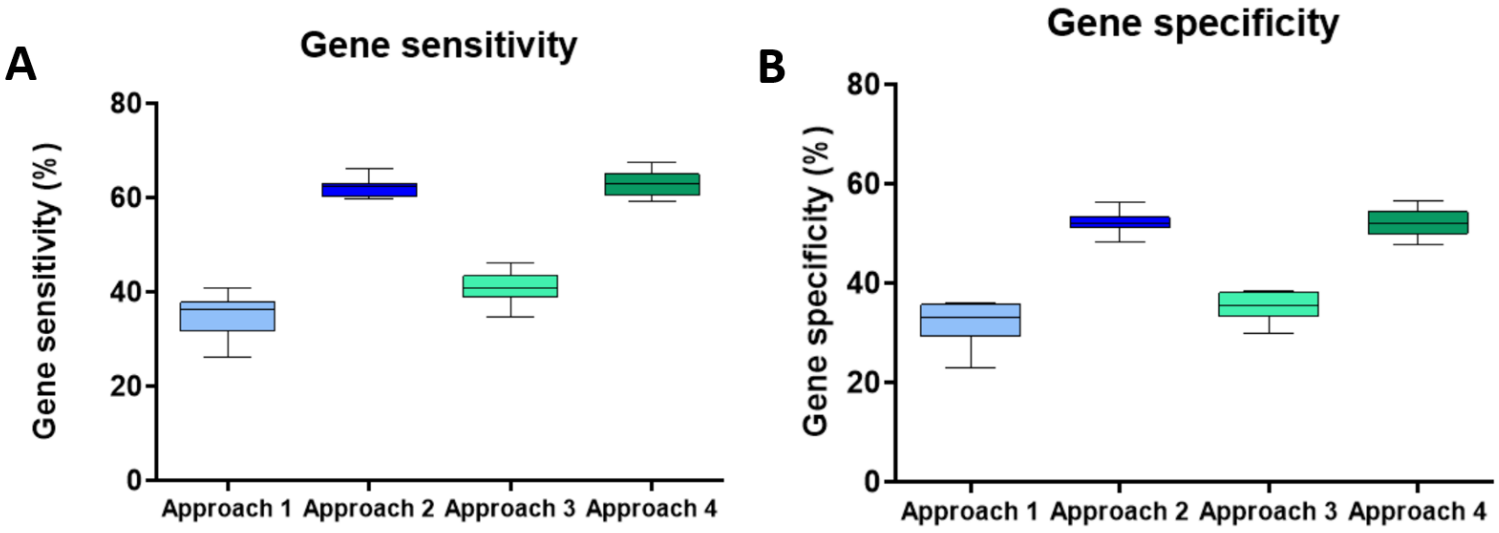

Figure 15. *Percentage of gene sensitivity (A) and gene specificity (B) of tested approaches 1-4 applied on 10 random sets of testing*

In terms of gene sensitivity (Figure 15A, Table S3), approaches 1 and 3 performed a median of 36.2% and 40.7%, respectively. Hence, for each testing set, less than 100 genes were accurately predicted by the software. Hint-based approaches (approaches 2 and 4) achieved a higher gene sensitivity (a median of 62.3% and 63.1%), related to the inclusion of RNA-seq and protein data.

Similar to gene sensitivity, gene specificity in hint-based approaches was superior to pure *ab-initio* approaches (Figure 15B, Table S3). While a gene specificity of median value 33%-35.4% was achieved in pure ab-initio approaches (Approaches 1 and 3), a median value of 52% was reached in hint-based approaches (Approaches 2 and 4). Thus, providing extrinsic evidence reduced the annotation of over-predictions by the predictor.

Additionally, the latter two approaches had similar performances in terms of gene sensitivity and specificity (Figure 15A-B). However, the distributions of gene sensitivity and specificity of Approach 4 were broader (Figure 15A-B) compared to

the ones relative to Approach 2, meaning that the performance of the latter mentioned one was less variable.

Fragmentation analysis

To evaluate a further quality parameter, the level of fragmentation in predicted genes was computed for all the four approaches. To perform this analysis, obtained predictions were compared to the *P. vulgaris* official annotation using Protein length ratio (PLR) metric (Figure 16).

Approximately 15,000 predicted proteins in both Approach 1 and 3 (pure *ab initio* methods) had a PLR of 1 (Figure 16). This means that 50% and 43% of total predicted proteins (30,053 and 34,786 in Approach 1 and 3 respectively) had lengths comparable with the respective proteins in *P. vulgaris* official annotation. Hint-based approaches (Approaches 2 and 4) predicted more than 20,000 proteins with a PLR of 1, showing a better performance than other approaches. Thus, the fragmentation of approximately 5,000 proteins was reduced using hints during the prediction.

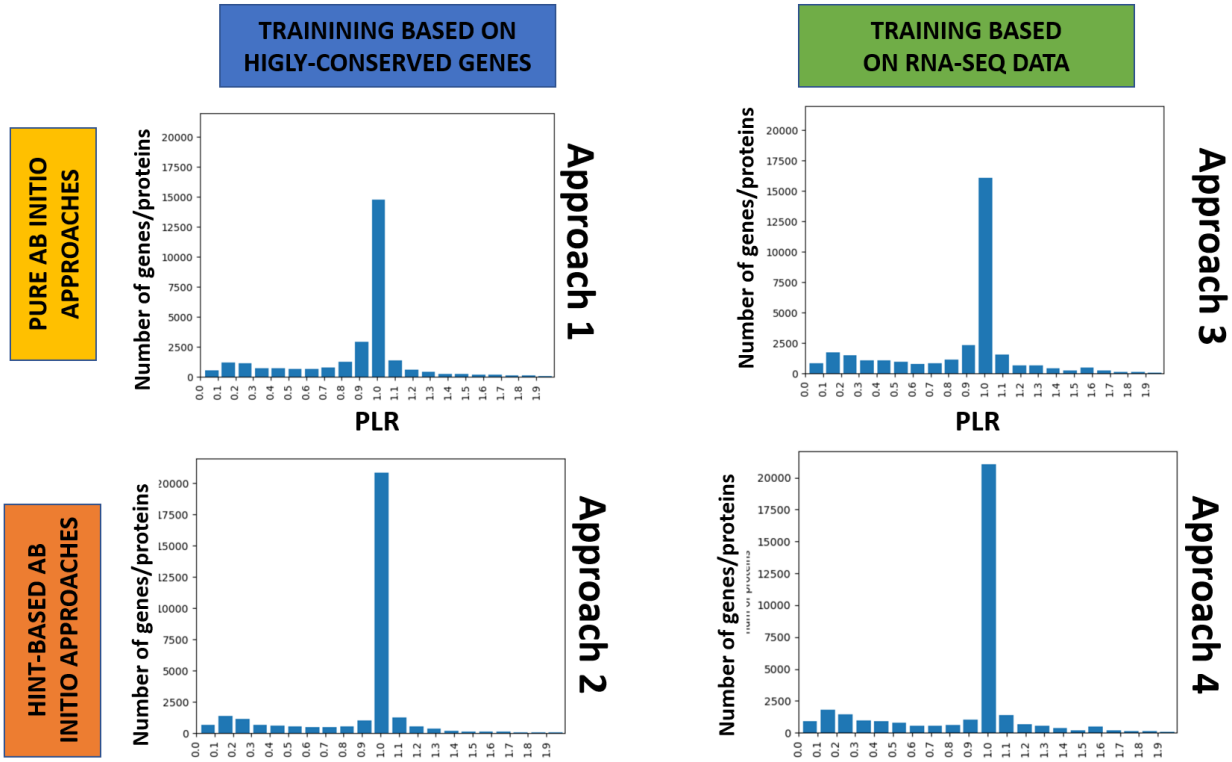

Figure 16. *Protein length ratio of gene predictions in approach 1(A), 2(B), 3(C), 4(D).*

Distributions of PLR values (Figure 17) showed that all approaches had most of the predictions with PLR values between 0.5 and 1.0. Approach 2 showed a compact distribution with higher values compared to other approaches (Figure 17). Approach 3 which is also a hint-based approach, had a wider distribution outlining an increased fragmentation (Figure 17).

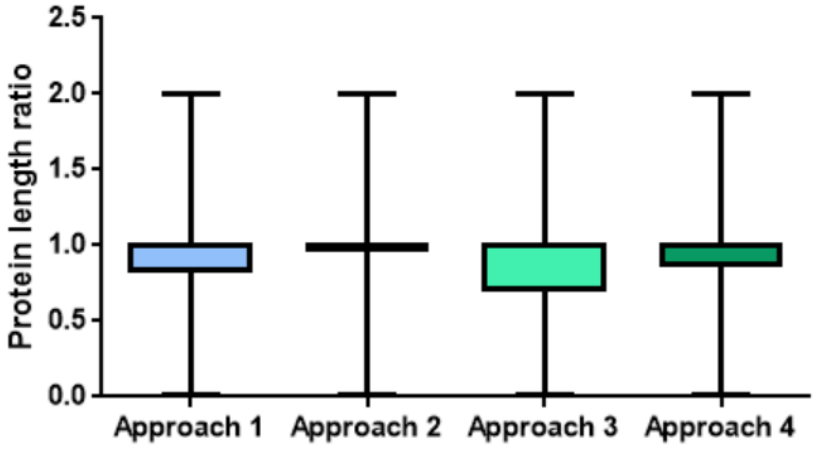

Figure 17. *Protein length ratio distribution for the four tested approaches*

Finally, the benchmarking analysis outlined that accurate results were achievable with hint-based approaches (Approaches 2 and 4) according to BUSCO completeness, gene sensitivity and specificity. Fragmentation analysis showed that performing a model training with highly conserved genes (Approach 2) reduced the level of fragmentation. Hence, Approach 2 coupled with filtering analysis was used for the development of the three pangenomes.

Development of pangenomes and automatic annotation on linear, nucleotide-based graph and element-based graph pangenomes

Automatic gene annotation was applied to the three pangenome formats (Figure 18).

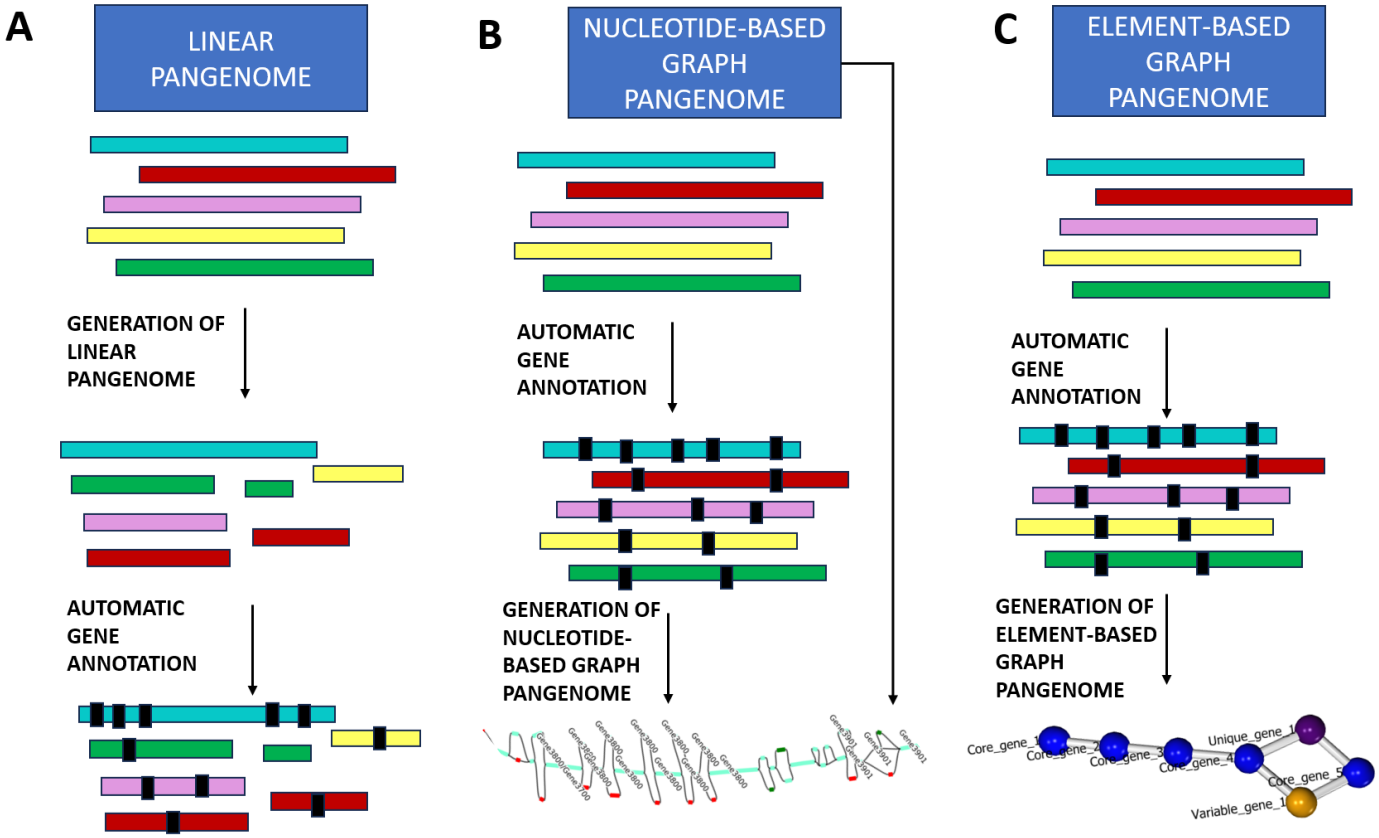

Figure 18. *Automatic gene annotation in linear (A), nucleotide-based graph (B), element-based graph (C)*

The annotation was directly applied to the generated NRR sequences in linear pangenome (Figure 18A).

Differently, in nucleotide-based graph pangenome, genes were annotated in the different genome assemblies which were then used to construct the graph (Figure 18B). After the generation of nucleotide-based graph pangenome, genes were connected to one, more or none of the nodes present in the graph, crosschecking the coordinates of nodes and genes.

In element-based graph pangenome, the assemblies were annotated with genes which then were used to generate the nodes of the graph (Figure 18C).

Development of linear pangenome

Map-to-pan phase

Linear pangenome was constructed through the independent whole-genome assembly approach using *P. vulgaris* cv. "G19833", "MIDAS", "G12873", "BAT93" and "JaloEPP558".

7,969 non-reference representative sequences accounting for a total length of 105,289,733 bp were extracted (Table 4) from the genomes of "non-reference" cultivars.

Subsequently, NRR sequences were added to the *P. vulgaris* reference genome. Thus, assembled *P. vulgaris* pangenome accounted for a total length of 642,508,369 bp formed by 9,491 sequences.

Reference total length	537,218,636
Number of reference sequences	1,522
NRR total length	105,289,733
NRR number of sequences	7,969
Pangenome total length	642,508,369
Number of pangenome sequences	9,491

Table 4. *Pangenome statistics for linear pangenome*

Automatic gene annotation

Automatic gene annotation (Approach 2+ Filtering analysis) which was performed on NRR sequences, predicted 2,376 genes. Thus, 30,549 genes were annotated in *P. vulgaris* pangenome.

PAV analysis

PAV analysis found 24,335 core, 3,483 variable and 1,991 unique genes out of the total of 29,809 genes.

Development of nucleotide-based graph pangenome

Construction of nucleotide-based graph pangenome

A nucleotide-based graph pangenome was constructed using the assemblies of P. vulgaris cv. "G19833", "MIDAS", "G12873", "BAT93" and "JaloEPP558" cultivars (Table 5). 3 different software were used to generate the nucleotide-based graph pangenome. The obtained pangenome with high level of contiguity was then chosen.

SOFTWARE	NUMBER OF NODES	NUMBER OF EDGES	TOTAL SEGMENT LENGTH
Minigraph	292,860	401,906	684,012,681
Pan-tools	5,144,904	8,829,433	2,862,405,752
Pggb	21,057,343	41,266,817	840,238,892

Table 5. *Number of nodes, edges and total segment length of assembled nucleotide-based pangenome generated using Minigraph, Pan-tools and Pggb softwares using P. vulgaris cv. "G19833", "MIDAS", "G12873", "BAT93" and "JaloEPP558"*

Minigraph pangenome showed the lowest number of nodes (292,860) compared to Pan-tools and pggb pangenome which were composed of 5,144,904 and 21,057,343 nodes respectively. In addition, graphs generated by Minigraph resulted to be the pangenome with the smallest size (684,012,681 bp) compared to Pan-tools (2,862,405,752 bp) and pggb (840,238,892 bp) pangenomes.

However, the size of the Minigraph pangenome assembly was comparable to the "BAT93" genome assembly (637,803,808 bp) which was the cultivar reported with a major genome length.

In conclusion, the Minigraph pangenome was chosen for the comparison as it was the most contiguous graph pangenomes in terms of nodes and total size compared to the other two pangenomes.

Automatic gene annotation of nucleotide-based graph pangenome

Automatic gene annotation (Approach 2 + Filtering analysis) was performed on "MIDAS", "G12873", "BAT93" and "JaloEPP558" genome assemblies where genes have not been identified yet. Then, automatic gene annotation was not applied for the "G19833" cultivar where the official gene annotation was present.

The number of annotated genes in "MIDAS", (27,101) was in line with the number of genes contained in *P. vulgaris* cv. "G19833" (27,433) (Table 6). Instead, in "G12873", 1000 more genes were annotated (28,469) compared to the official reference annotation. Furthermore, in "JaloEPP558", and "BAT93", the number of annotated genes was even higher (29,523 and 32,998 respectively).

Cultivar	Number of final genes	Number of genes in unique nodes	Number of genes in multiple nodes	Number of genes not assigned to nodes
G19833	27,433	19,643	7,790	0
MIDAS	27,101	1,197	332	25,572
G12873	28,469	3,507	1,711	23,251
Bat93	29,523	2,458	598	26,467
JaloEPP558	32,998	546	67	32,385

Table 6. *Automatic gene prediction on "MIDAS", "G12873", "BAT93" and "JaloEPP558" plus the official annotation of the reference genome ("G19833"). Number of genes in unique nodes, multiple nodes and not assigned to nodes in all the 5 cultivars.".*

Subsequently, automatic gene annotation was transferred into nucleotide-based graph pangenome: each gene was assigned to one or multiple nodes if their coordinates intersected with its gene space.

Observed transfer of gene annotation reported all genes annotated in the "G19833" cultivar. 25,572, 23,251, 26,467 and 32,385 genes respectively for" MIDAS", "G12873", "BAT93" and "JaloEPP558" cultivars were not assigned to nodes of the nucleotide-based graph. Manual inspection revealed that these genes were not assigned to any nodes since the software in conserved regions reported only the genes of the genome with lowest level of fragmentation (in this case "G19833"

genome).

Instead, 7,790, 332, 1,711, 598 and 67 of “G19833”,” MIDAS”, “G12873”, “BAT93” and “JaloEPP558” genes were assigned to multiple nodes in the pangenome graph.

PAV analysis

PAV analysis was not possible to perform since all nodes were assigned to only one cultivar among “G19833”, “MIDAS”, “G12873”, “BAT93” and “JaloEPP558” cultivar. Hence, shared nodes were not possible to assess, making it impossible to identify core, variable and unique genes.

Development of element-based graph pangenome

The method applied for the construction of element-based graph pangenome differs from the approaches used to assemble linear and nucleotide-based graph pangenome. The development of element-based graph pangenome is divided into four steps which are depicted in Figure 19A.

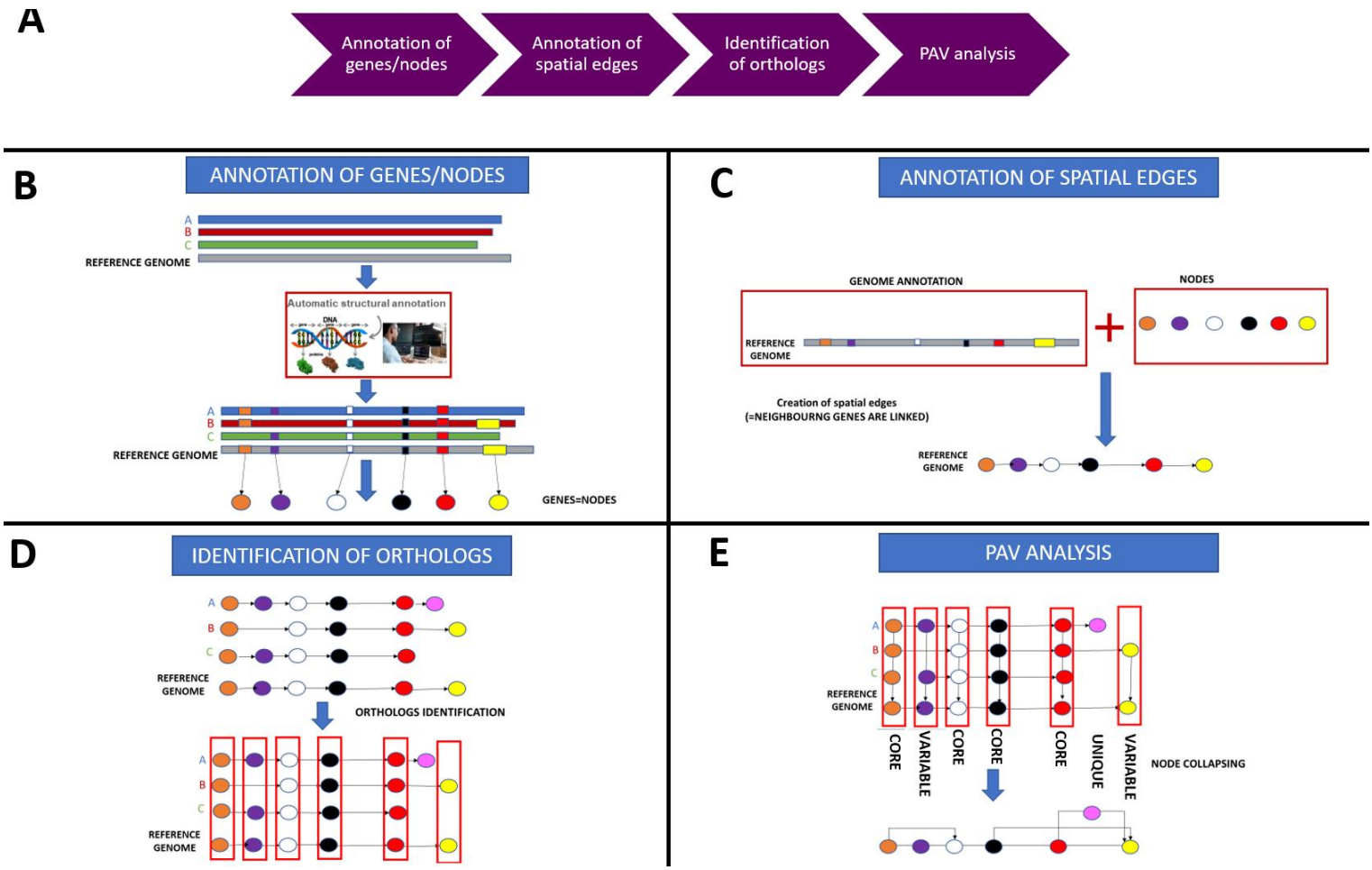

Figure 19. *(A) Process of development of element-based graph pangenome is divided into four steps: (B) Annotation of genes using automatic prediction; (C) Annotation of spatial edges; (D) Identification of ortholog; (E) Annotation of core, variable and unique genes through ortholog edges.*

The initial step in creating a pangenome is to define nodes (Figure 19B). Specifically, nodes are any object anchored to the genome, such as genes. To generate these nodes, automatic gene annotation is performed on a set of genome assemblies of different cultivars of the same species.

Once gene structures have been generated, interactions among nodes need to be defined (Figure 19C). Adjacent genes are connected through spatial edges, exploiting the gene annotation which allows defining the gene order.

Afterwards, orthologs are identified using synteny or/and gene family analysis and connected through edges (Figure 19D). At this stage, a raw "un-collapsed graph" has been created; in this setting, all genes are displayable, and all relationships (spatial and ortholog) are directly visualizable.

Core, variable and unique genes could be identified through relationships among orthologs of different cultivars (Figure 19E). These orthologs (core and variable genes) were collapsed and visualized as single nodes. Instead, unique genes remained unchanged. The obtained graph reported only annotation of core, variable and unique genes, resulting in a much simple interpretation (Figure 20).

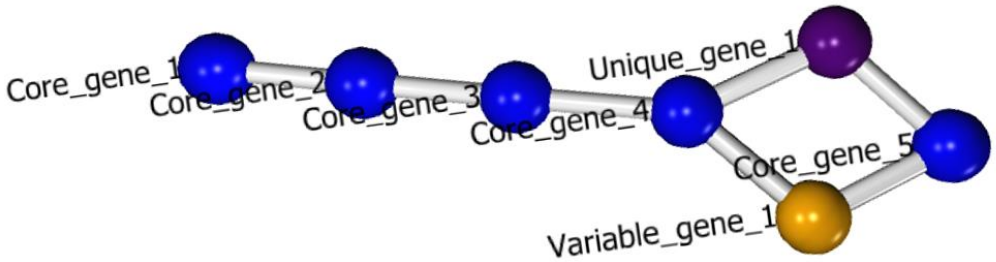

Figure 20. *Visualization of annotation of core, variable and unique genes in element-based graph pangenome*

Automatic gene annotation

Genes which were annotated on genome assemblies of" MIDAS", "G12873", "BAT93" and "JaloEPP558" cultivars were used to generate the nodes of element-based graph pangenome. The official P. *vulgaris* gene annotation was used for "G19833" cultivar.

Element-based graph pangenome accounted for 145,524 nodes corresponding to 145,524 annotated genes. This set of genes contained orthologs across cultivars or, in the same cultivar, copies of the same gene which have originated from gene duplication or unique genes.

Annotation of spatial edges

Spatial edges among 145,524 genes were created in element-based graph pangenome (Table 7).

CULTIVAR	**NUMBER OF ANNOTATED GENES**	**NUMBER OF GENES/NODES LINKED BY EDGES**
G19833	27,433	27,420 (99.953%)
MIDAS	27,101	26,813 (98.937%)
G12873	28,469	28,391 (99.726%)
Bat93	29,523	29,522 (99.997%)
JaloEPP558	32,998	32,997 (99.997%)

Table 7. *The number of annotated genes and the number of genes connected by edges for each cultivar in element-based graph pangenome.*

In all studied cultivars, not the totality of the annotated genes was connected through spatial relationships (Table 7). Manual inspection revealed that these were solitary "genes", without adjacent genes placed on shorter contigs. They were inserted in the graph as nodes without connection to other ones. (Figure 21).

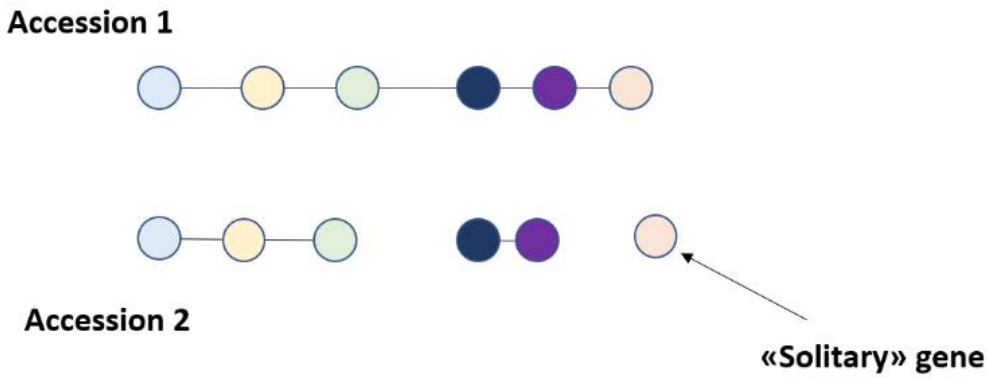

Figure 21. *Element-based graph pangenome with reconstructed spatial edges and solitary*

Annotation of orthologs relationships and PAV analysis

In Element-based graph pangenome, relationships between orthologs were identified and annotation of core, variable and unique genes was performed. For this analysis, three approaches were tested (Figure 22):

- Approach 1 based only on the synteny analysis.
- Approach 2 based only on gene family analysis.

- Approach 3 or combined analysis which integrates the results of synteny and gene family analysis.

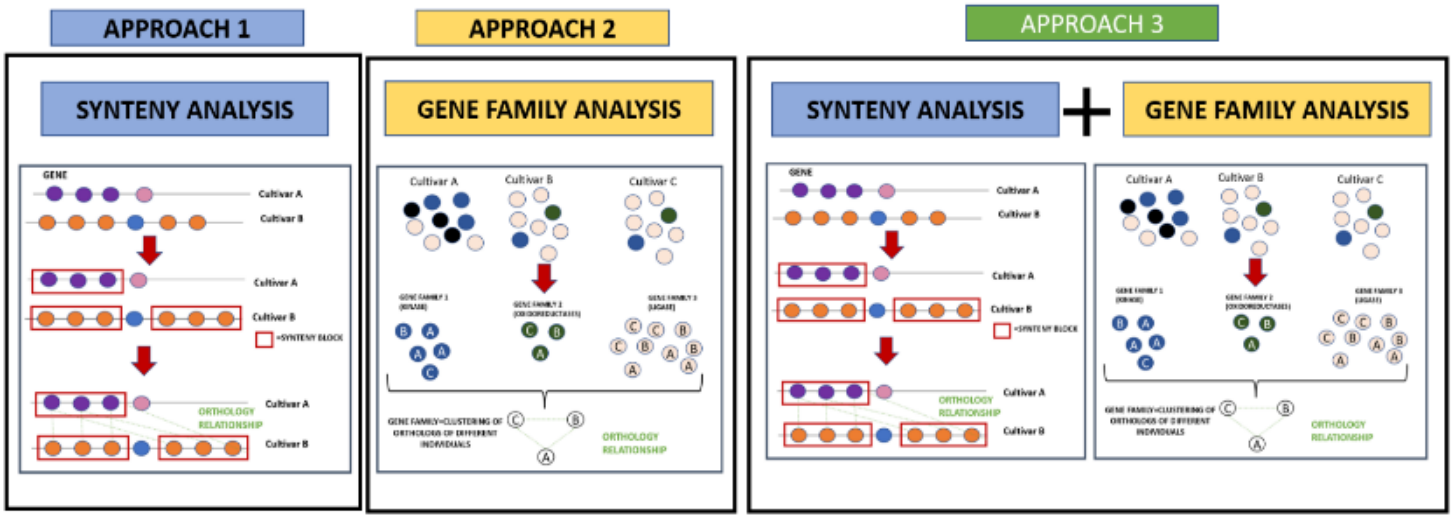

Figure 22. *Approaches tested to identify orthologous genes and annotate core, variable and unique genes.*

Number of total, core, variable and unique genes

The number of core, variable, unique and total genes were computed for the three tested approaches.

Approach 1 and 2 (Figure 23A, TableS4) identified 19,757 and 20,712 core genes, respectively. Approach 3 found a comparable number of core genes (20,784), respect to Approach 2.

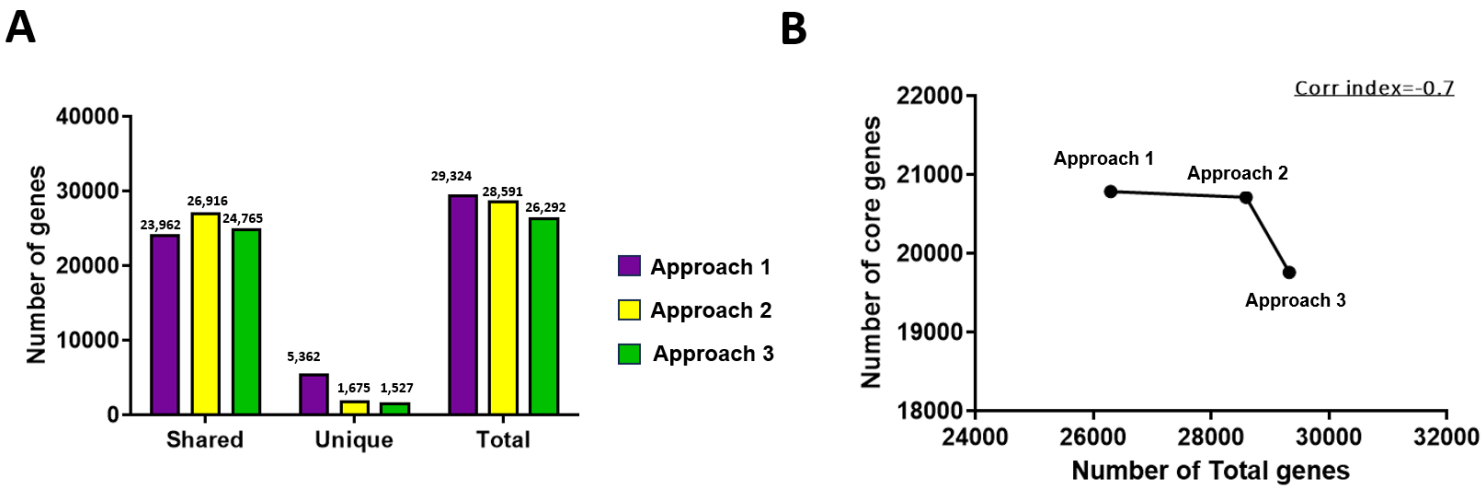

***Figure 23**. (A) Number of shared (core and variable), unique and total genes identified by synteny, gene family and combined approaches. (B) Graph reporting the number of genes against the number of total genes identified in synteny, gene family and combined approaches together with Pearson correlation coefficient*

Approaches 1 and 2 detected 4,205 and 6,204 variable genes (Figure 23A, Table S4). Approach 3 identified a lower number of variable genes (3,981) compared to the other two approaches.

Approach 1 identified the highest number of unique genes (5,362) among all approaches (Figure 23A, Table S4). Approach 2 identified only 1,675 unique genes. While Approach 3 still reported a lower number of unique genes (1,527).

Altogether, these results showed that Approach 3 seemed to have conservativeness in identifying shared genes (core and variable) compared to other approaches. In addition, the major contribution in identifying the ortholog relationship in Approach 3 is made by the integration of the results of gene-family analysis (Approach 2).

Approaches 1, 2 and 3 identified 29,324, 28,591 and 26,292 total genes, respectively (Figure 23A, TableS4). The more the approach is conservative, the more genes are identified as core and the fewer genes are reconstructed (Figure 23B). Additionally, a further reduction of total genes was imputable of collapsing of co-orthologs in element-based graph pangenomes.

Concordance of core, variable and unique genes

The three approaches were compared in terms of concordance of core, variable and unique genes (Figure 24). Orthologs across different cultivars that were previously merged, were un-collapsed and considered in the comparisons (Figure 24).

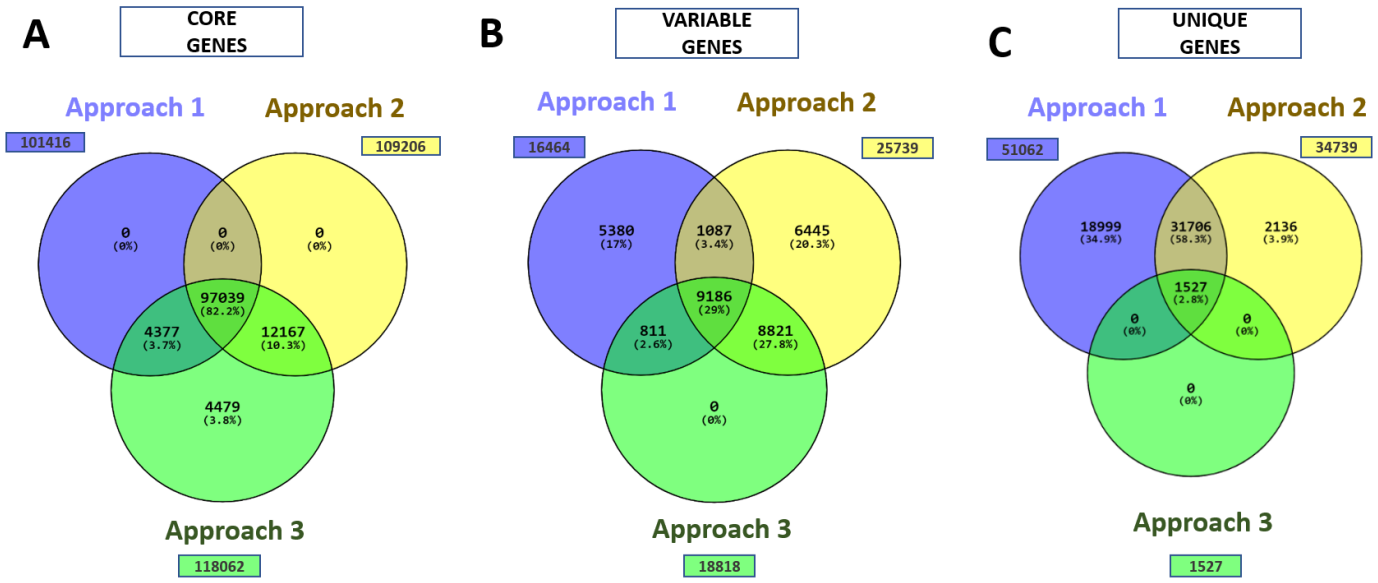

Figure 24. *Venn diagram of core (A), variable (B) and unique (C) genes identified in the three approaches: orthologous genes were considered separately for core and variable genes.*

All approach commonly identified 97,039 core genes (Figure 24A). Approach 3 detected 4,479 core genes which were not found by other approaches.

All three approaches commonly identified 9,186 variable and 1,527 unique genes (Figure 24B-C). Approach 3 identified no extra variable or unique genes.

All the previous findings confirmed the complementarity of Approach 1 and Approach 2, meaning that these approaches identified a part of homology relationships among orthologs. Thus, Approach 3 which showed the highest sensitivity in core and variable identification and the highest precision in the detection of variable and unique genes, was used for PAV analysis of element-based pangenome. In PAV analysis, 20,784 of 26,292 total genes were identified as core. 3,981 and 1,527 were annotated as variable and unique genes (Table 8).

1522 of 1527 unique genes (99.6%) were belonging to *P. vulgaris* cv "G19833". The remaining 5 unique genes belonged to "MIDAS" (2 genes), "BAT93" (2 genes) and "G12873" (1 gene) cultivars respectively.

Enrichment analysis showed that unique genes are significantly enriched in myosin complex (GO:0016459), cytoskeletal motor activity (GO:0003774) and actin cytoskeleton (GO:0015629) biological processes (Figure S1A).

Total genes	26,292
Core genes	20,784
Variable genes	3,981
Unique genes	1,527

Table 8. *Pangenome statistics for element-based graph pangenome*

Comparison of element-based graph pangenome with linear pangenome

Element-base graph pangenome was compared to linear pangenome in terms of core, variable and unique genes. Then, concordance in terms of core, variable and unique genes between the two pangenome was assessed to evaluate their differences.

Number of core, variable and unique genes in linear and element-based graph pangenomes

29,809 and 26,292 genes were annotated in linear and element-based graph pangenome, respectively (Figure 25A). Thus, linear pangenome reported a higher number of total genes since co-orthologs were not collapsed together as in element-based graph pangenome.

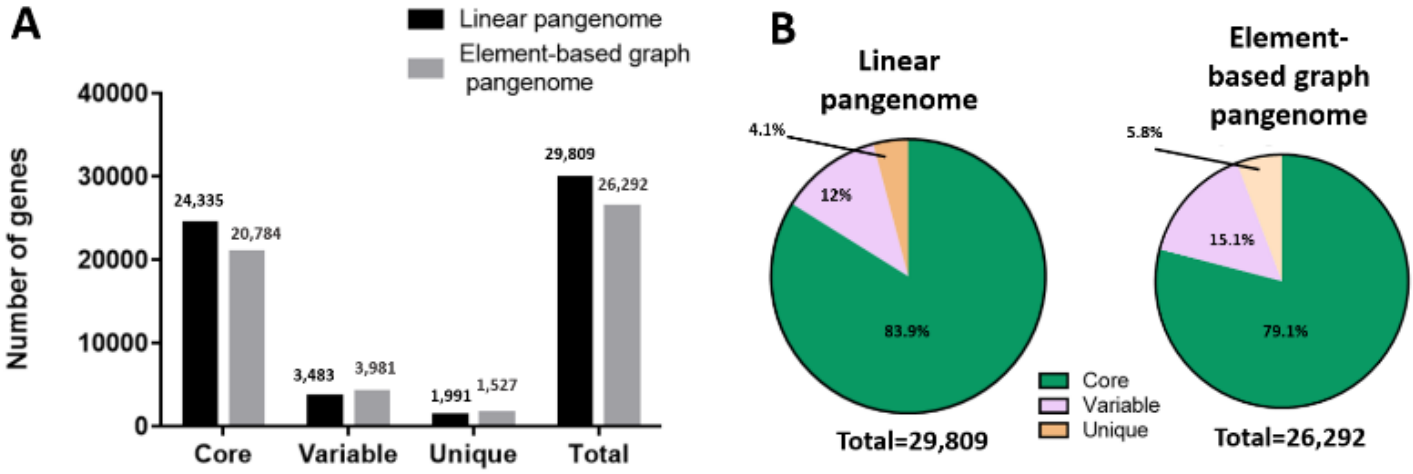

Figure 25*. (A) Number of core, variable and unique genes (B) Pie chart of core, variable and unique genes over total genes annotated in linear and element-based graph pangenomes*

Linear pangenome reported a higher number of core genes (24,335) compared to element-based graph pangenomes (20,784). Moreover, a comparable number of variable genes were identified by linear pangenome (3,483) and element-based graph pangenomes (3,981), respectively. A similar number of unique genes was also identified in the two pangenomes (3,483 and 3,981 in linear and element-based graph pangenomes, respectively).

In both pangenomes, the fraction of core, variable and unique genes was comparable accounting for approximately ≈80%, ≈15% and ≈5%, respectively (Figure 25B).

Concordance of core, variable and unique genes

Subsequently, concordance was computed to see if the same genes are predicted as core, variable and unique genes both in linear and element-graph-based pangenomes.

This analysis was restricted to only genes from the "G19833" cultivar as the gene models are equivalent in the two pangenomes (Figure 26).

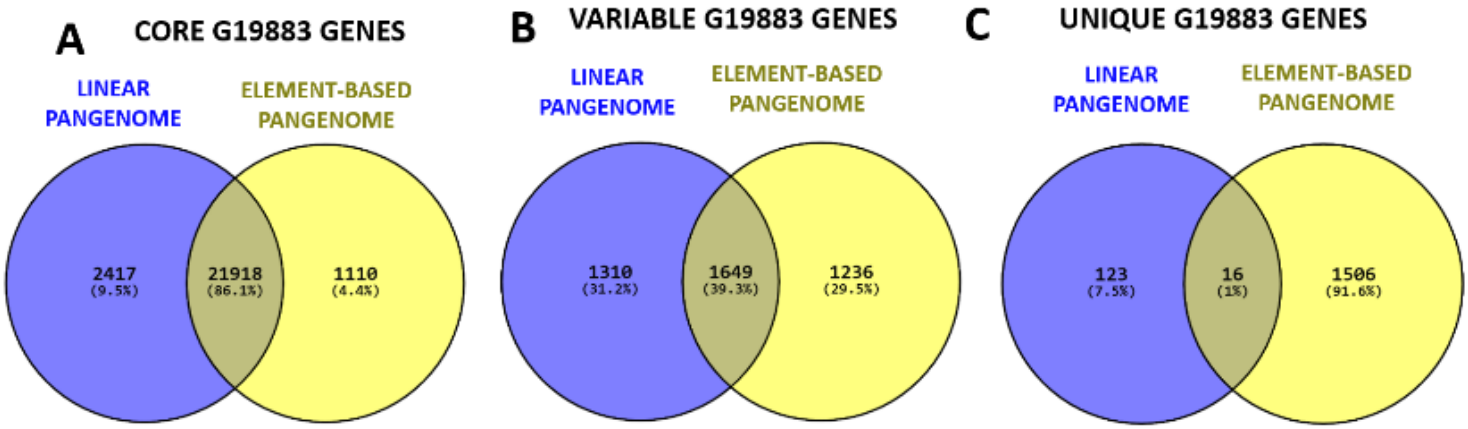

Figure 26. *Venn diagram of core(A), variable (B) and unique (C) genes present in G19873 cultivar present in linear and element-based graph pangenomes.*

21,918 core genes present in the "G19833" cultivar were commonly identified in the two pangenomes (Figure 26A), accounting for a high concordance of 86.1%. Additionally, in terms of variable genes, the concordance between the two pangenome is lower (39.3%): only 1,649 genes were found commonly in linear and element-based graph pangenome (Figure 26B). Finally, only 16 of the "G19833" genes were reported as unique in the two pangenomes (Figure 26C).

Comparison of element-based graph pangenome with nucleotide-based graph pangenome

After comparing with linear pangenome, element-graph pangenome was compared to nucleotide-based graph pangenome. As reported in section of Results 2.2.3 PAV analysis was not possible to perform in nucleotide-based graph pangenome. Then, for comparison purpose, the visualization layout of element-based graph pangenomes was set side by side with the one of nucleotide-based graph pangenome. In addition, the growth of nodes and visualization complexity influenced by the number of input genomes was assessed in the two pangenomes.

Comparison of visualization

One locus of the region responsible for pod indehiscence trait [89] was visualized in both in nucleotide-based graph pangenomes (Figure 27A and in element-based graph pangenome (Figure 27B-C):the genetic locus with coordinates Chr05:38,307,142-38,324,025 have a length of 16.8 kbp and contains 3 genes.

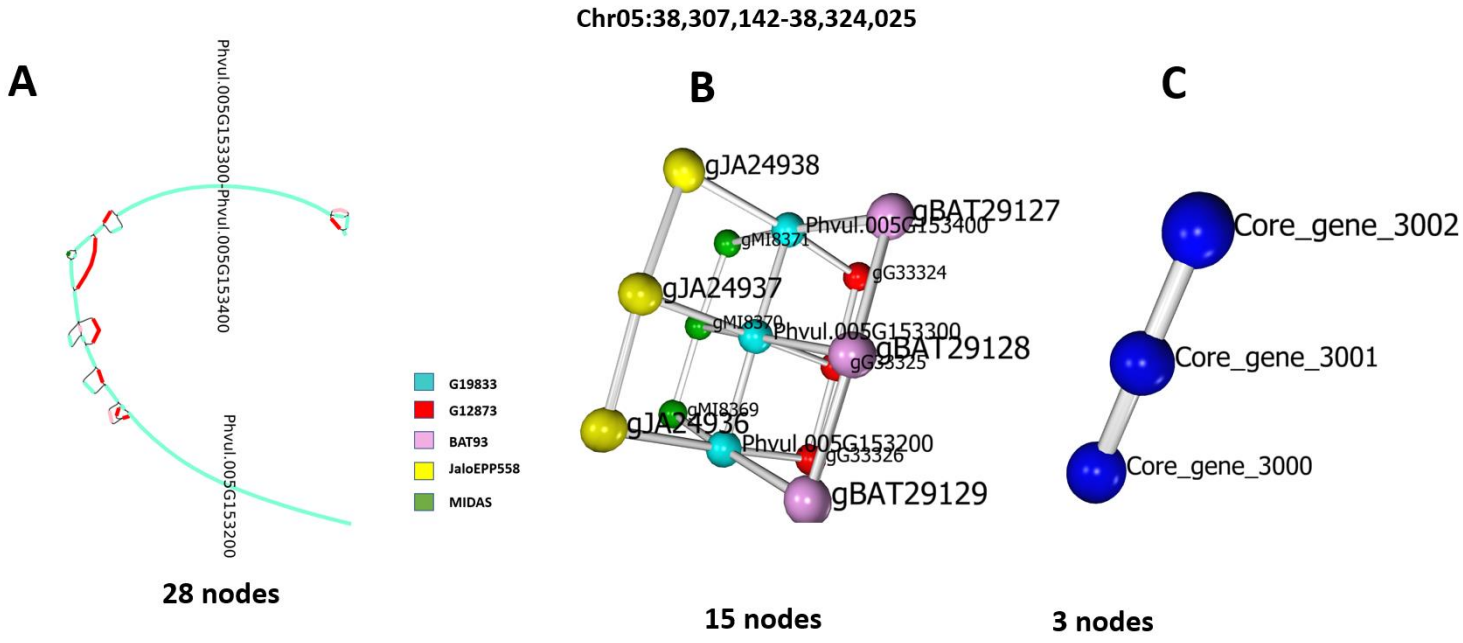

Figure 27. *Representation of region in Chr05:38,307,142-38,324,025 in nucleotide-based pangenome (A) in un-collapsed (B) and collapsed (C) element-based graph pangenome.*

Visualization in nucleotide-based pangenome displayed 28 nodes (Figure 27A). In this representation, the three genes belonging to the "G19833 "cultivar are

represented but none of their orthologs in the other cultivars are reported: two genes are reported together into one node whereas one gene is reported alone into another node. Moreover, intergenic regions contributed to the variation represented in the graph.

Visualization in element-based graph pangenome in un-collapsed fashion (before orthologs merging) displayed 15 nodes (Figure 27B): the genes of different cultivars are represented with their relationships. Hence, the presence and absence of genes in cultivars are easier to assess. Indeed, it is observable that all three genes were core.

With the collapse option set to ortholog across cultivars, the element-based graph pangenome was represented by only 3 nodes (Figure 27C). Visualization became clearer since it displayed only genes annotated as core, variable or unique. Overall, the collapsed representation in element-based graph pangenome allowed keeping global oversight compared to the nucleotide-based pangenome.

To assess differences in presence or absence of genes in both pangenomes, visualization was made in some genes known associated to phenotypic traits: Phvul.006G018800 [103], Phvul.007G171466 [104], Phvul.007G171333 [104], Phvul.002G300900 [105], Phvul.009G190100 [106] and Phvul.008G038400 [107](Figure 28).

Visualization of V, P and cbZIP genes (Phvul.006G018800, Phvul.007G171333 and Phvul.009G190100) in element-based graph pangenome confirmed the presence of these genes in all analyzed cultivars. Local visualization of nucleotide-based pangenome reported not all genes in the five cultivars, not correctly classifying V, P and cbZIP as core genes.

Additional copy of P gene (Phvul.007G171466) present in G19833 genome was correctlyclassified in both pangenomes as unique gene, meaning that MIDAS, G12873, Jalo and Bat cultivars have a single copy of this gene. SWEET4 gene (Phvul.002G300900) was classified variable and absent in Jalo cultivar, concordantly by the two pangenomes' representations.

Myb113 gene (Phvul.008G038400) was classified as variable genes in both element-based graph and nucleotide-based graph pangenomes. However, presence of Myb113 gene in cultivars was underestimated since this gene was classified absent in Jalo

cultivar in nucleotide-based graph pangenome. Finally, element-based showed that in G12873 genome, two gene copies are present (gG4907 and gG4909).

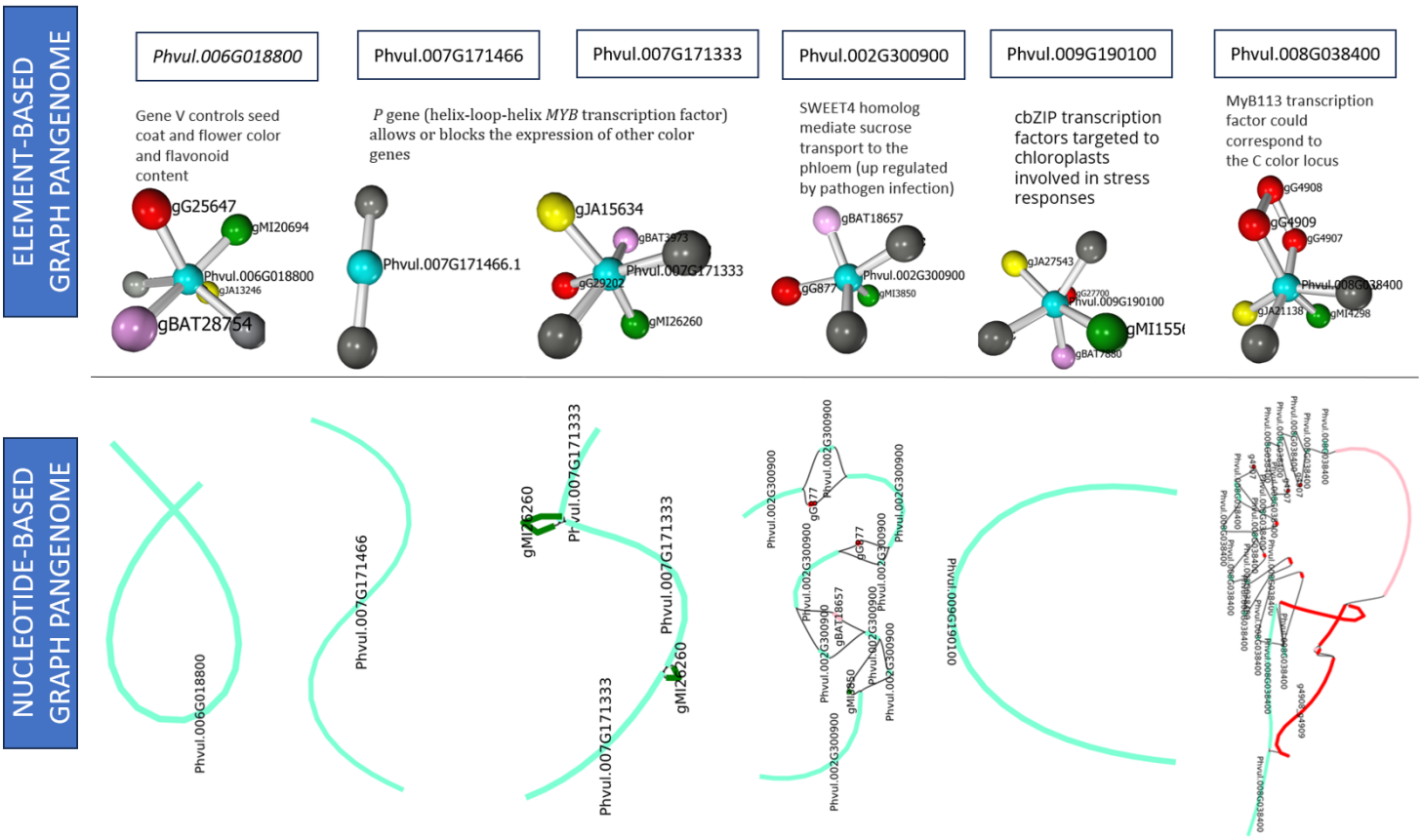

Figure 28. *Representation of V (Phvul.006G018800), P (Phvul.007G171466 and Phvul.007G171333), SWEET4 (Phvul.002G300900), cbZIP (Phvul.009G190100), MyB113 (Phvul.008G038400). Visualization in element-based graph pangenome (top part) and nucleotide-based graph pangenome (bottom part). Biological description and association to phenotypic traits is reported.*

Visualization complexity with the growth of input genomes

Visualization was inspected according to the growing of input genomes. Nucleotide and element-based graph pangenomes were assessed using iteratively 5 *P. vulgaris* genomes ("G19833", "MIDAS", "G12873", "BAT93" and "JaloEPP558"). The syntenic region, with coordinates Chr:06:30,666,467- 30,708,540 containing 7 genes, was visualized both in the two pangenomes (Figure 29).

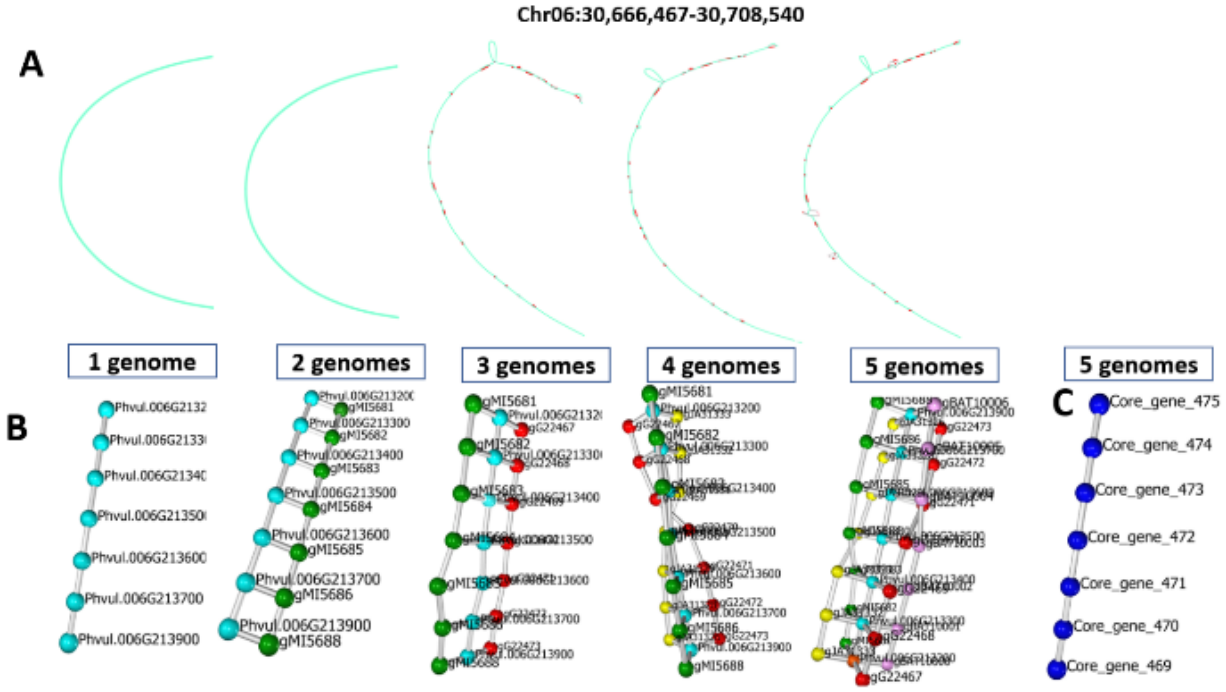

Figure 29. *Visualization of nucleotide-based pangenome (A) collapsed(B) and uncollapsed (C) element-based graph pangenome in region Chr06:30,666,467-30,708,540 using the 5 P. vulgaris genomes ("G19833", "MIDAS", "G12873", "BAT93" and "JaloEPP558")*

Visualization of the 42-kbp region in Chr06:30,666,467-30,708,540 led to an increase from 1 node to a total set of 70 nodes. On average, each addition of a new genome increased the complexity of the nucleotide-based pangenome with 17.5 nodes (Figure 29A, Table S7). In some cases, the addition of a genome was observed not to increase the nodes as new regions might have an identical sequence to the already present genomic regions.

Visualization of the same syntenic region in uncollapsed version of element-based graph pangenome led to an increase in the number of genes from 7 to 35, with all orthologs being displayed (Figure 29B). Even though all information is displayed, the visualization remains compact and easily interpretable thanks to the implemented "backbone" structure describing the spatial organization of genes. However, collapsing them led to a synthesized representation of the pangenome (Figure 29C) without loss of information relative to the genes involved.

Nodes' complexity with the growth of input genomes

The effect of the number of input genomes on the amount of nodes was assessed in nucleotide-based graph and element-based graph pangenomes. The "G19833", "MIDAS", "G12873", "BAT93" and "JaloEPP558" genomes were added iteratively in the pangenomes. The un-collapsed version of the element-based graph pangenome was included in the comparison since it represented different haplotypes of a region like nucleotide-based graph pangenome (Figure 30, Table S8).

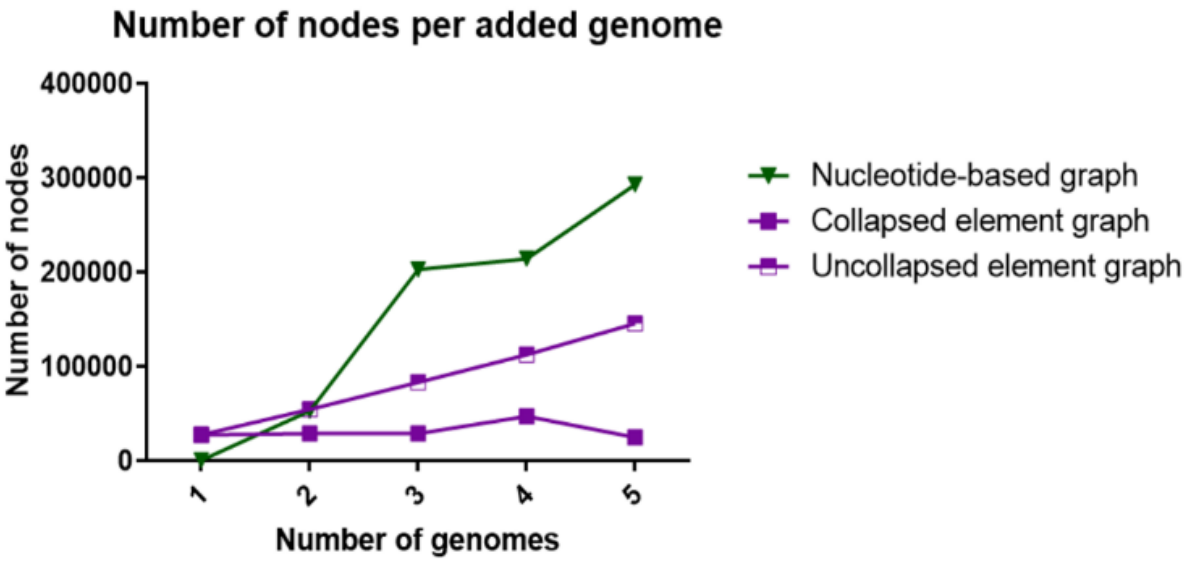

Figure 30. *The number of nodes in nucleotide-based, linear, collapsed-element and un-collapsed-element pangenomes using iteratively the 5 P. vulgaris genomes ("G19833", "MIDAS", "G12873", "BAT93" and "JaloEPP558").*

The growth of nodes in nucleotide-based graph pangenome was superior to the one of element-based graph pangenome (collapsed version), having on average an increase of 73,096 nodes for each added genome. The mean growth was 2.5-fold

superior to element-based graph pangenome in the un-collapsed version (29,523 nodes on average) where orthologous copies of the same genes or all possible orthologous genes are displayed.

DISCUSSION

State-of-art pangenome approaches [25], [27], [37], [108], [109]are gaining more and more importance and they will completely replace the reference genome in genomic studies. However, state-of-art pangenome have some limitations addressed in this thesis: linear pangenome does not provide a graphical representation of gene presence and absence among individuals whereas nucleotide-based graph pangenome does not include the whole gene content of studied cultivars as well as cannot be used for gene presence and absence analysis. Due to the limitations of present approaches, we proposed a new type of pangenome, called element-based graph pangenome.

In element-based graph pangenome, PAV analysis is represented in graph format and information about cultivars can be compressed and decompressed (Figure 31A). This type of pangenome becomes a consensus genome in which shared genes maintain spatial relationships and where orphan genes are inserted between their adjacent genes. Hence, in such visualization, variable and unique genes in element-based pangenome generate bubbles, which represent polymorphic loci in nucleotide-based graphs [108], [110]. Additionally, syntenic regions (Figure 31B) in element-based pangenome are easily detectable as two or more strings of orthologs interconnected to each other.

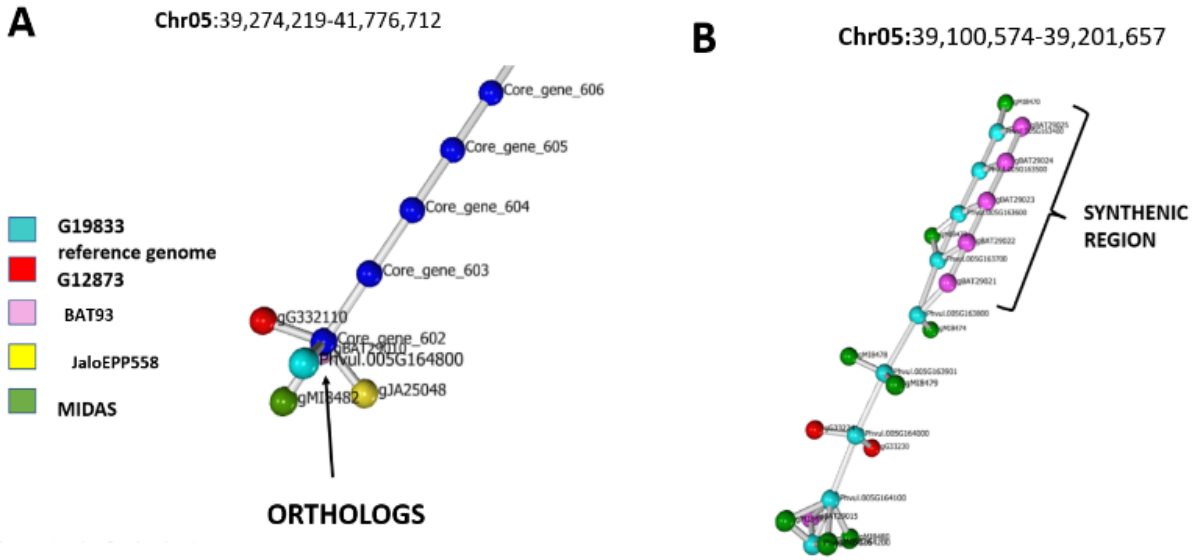

***Figure 31**. (A) Open visualization of one gene in collapsed version of element-based graph pangenome. (B)Visualization of syntenic region in un-collapsed version of element-based graph pangenome*

Nucleotide similarities and dissimilarities in coding regions are displayed in a zoomed-in representation through local nucleotide-based graph pangenome [108] (Figure 32). The element-based graph pangenome, on the other hand, is a zoomed-out representation of a graphical nucleotide-based pangenome.

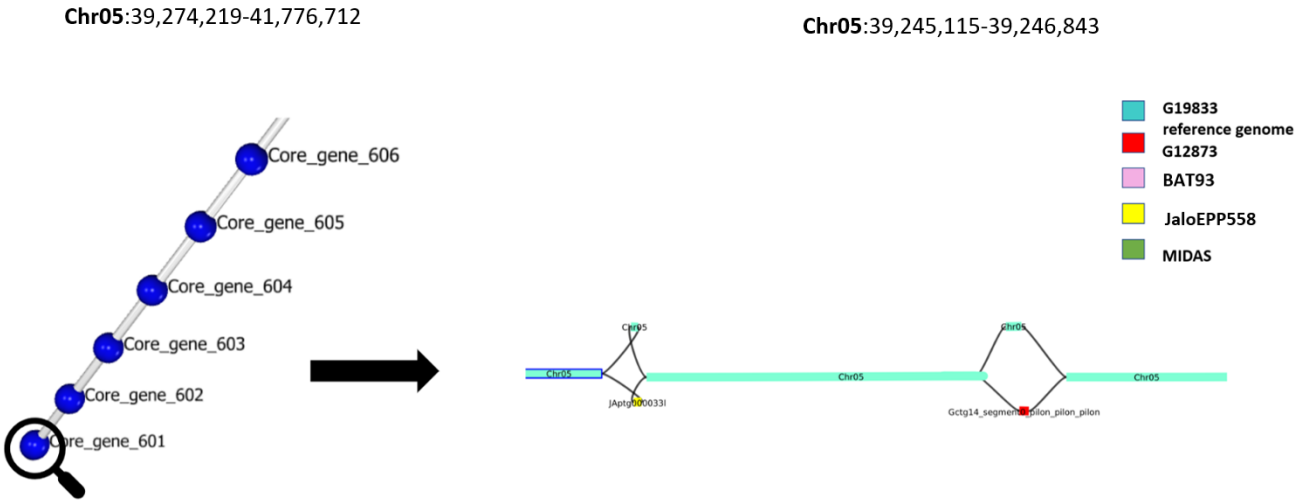

Figure 32. *Conversion of element-based graph pangenome into local nucleotide-based pangenome to visualize nucleotide variation of one core gene.*

Comparison analysis with other pangenome formats demonstrated additional features of this newly proposed approach. First, visualization in element-based graph pangenome allowed an easy interpretation of gene presence or absence in cultivars compared to nucleotide-based graph pangenome, where genes are not present in nodes or dispersed into multiple nodes. Hence, element-based graph pangenome provided a full representation of gene annotation which is not fully included in nucleotide-based graph pangenome.

Moreover, a reduced number of nodes compared to nucleotide-based graph pangenome confirmed the contribution given by the diversity of non-coding genome [111], in particular the transposable elements[4]. Future studies which include a large set of individuals[112], will benefit from this advantage.

All these analyzed features (compact visualization, full representation of genes and the low number of nodes) made the element-based graph pangenome more efficient in terms of PAV analysis and visualization compared to nucleotide-based pangenomes.

Nevertheless, compared to linear pangenome, element-based graph pangenome reported a lower number of total genes due to collapsing of paralogs. These genes normally are separately reported in linear pangenome. However, consistent with the literature [113], [114], paralog genes should be collapsed together only if they share the same genomic context.

Linear[25] and nucleotide-based graph pangenomes[40] are influenced by the stringency of identity or coverage threshold chosen to collapse similar or identical regions. On the contrary, element-based graph pangenome has strong annotation dependency for which the accuracy of automatic gene prediction has a main influence on the discovery of core, variable and unique genes in element-based graph pangenome [115]. Indeed, applying a low-conservative approach will increase the rate of true negative genes and it will consequently reduce the sensitivity of finding core genes whereas using an approach with low precision will increase the amount of false positive genes (over-prediction) and the number of private genes in element-based graph pangenome. Hence, in this work, the most accurate automatic gene prediction was assessed through benchmarking, considering both conservativeness and precision in gene finding.

Benchmark results confirmed the non-optimal performances of automatic gene prediction found in the literature [82], [116]. Reported high values of completeness of highly conserved genes[117] did not outlined the best approach. Instead, gene sensitivity and specificity analysis[82]supported the non-optimal accuracy of automatic prediction in specific-organism genes where low sensitivity and precision usually occur [118]–[120]. However, results confirmed that providing extrinsic evidence during the prediction through hint-based approaches (approaches 2 and 4) increases the likelihood of annotating real genes [121]. Fragmentation analysis outlined that noise and inaccurate mapping of extrinsic data led to low precision and fragmentation in the RNA-seq data training approach (Approach 4). Hence, in contrast with the previous findings [122] reporting a good performance of model training based on RNA-seq data, the hint-based approach, which exploits the training of conserved genes (Approach 2), outperformed the others. Additionally, the filtering analysis allowed the exclusion of artefacts or unidentified repeats which were not properly masked before genome annotation [119]

Besides automatic gene annotation and like other types of pangenomes, element-based graph pangenome is influenced by orthologous identification[115], [123], [124] which could be performed through synteny [125], [126] and/or gene family analysis[86], [127]. Expectedly, synteny analysis (Approach 1) overestimated unique genes due to assembly fragmentation [128]. In addition, gene duplication or translocation events [129]–[132] may have decreased the sensitivity of this approach in finding orthologous genes, falsely annotated as unique. Instead, evidence showed that clustering of protein sequences in gene family-based analysis (Approach 2) allowed to overcome such limitations, observing a higher sensitivity in orthologs detection. Nevertheless, synteny analysis (Approach 1) was more sensitive in finding orthologs for some genes[133].Then singularly, the two approaches (Approach 1 and 2) partially identified the connection among core genes, overestimating variable genes. Hence, the integration of results from both analyses in a combined approach (Approach 3) resulted in having the highest conservativeness.

In conclusion, we propose a new type of pangenome called element-based graph pangenome providing the advantage to detect the presence or absence of elements annotated in the genome. In this case, we considered only the coding part since most of the pan-genomic studies did not examine other regions. Hereafter, in the future, element-based graph pangenome could include conserved noncoding elements (CNEs) which have been reported to be organized in clusters [134], [135] or transposable elements whose insertion into genic regions creates relevant phenotypic traits[136]–[138] . Thus, in this way, an element-based graph pangenome could indicate the presence and the genomic positions of coding and non-coding elements across genomes, allowing extrapolation of biological information among studied individuals.

REFERENCES

[1] H. Tettelin *et al.*, "Genome analysis of multiple pathogenic isolates of *Streptococcus agalactiae* : Implications for the microbial 'pan-genome,'" *Proceedings of the National Academy of Sciences*, vol. 102, no. 39, pp. 13950–13955, Sep. 2005, doi: 10.1073/pnas.0506758102.

[2] D. Medini, C. Donati, H. Tettelin, V. Masignani, and R. Rappuoli, "The microbial pan-genome," *Curr Opin Genet Dev*, vol. 15, no. 6, pp. 589–594, Dec. 2005, doi: 10.1016/j.gde.2005.09.006.

[3] R. A. Blaustein, A. G. McFarland, S. Ben Maamar, A. Lopez, S. Castro-Wallace, and E. M. Hartmann, "Pangenomic Approach To Understanding Microbial Adaptations within a Model Built Environment, the International Space Station, Relative to Human Hosts and Soil.," *mSystems*, vol. 4, no. 1, 2019, doi: 10.1128/mSystems.00281-18.

[4] M. Morgante, E. De Paoli, and S. Radovic, "Transposable elements and the plant pan-genomes.," *Curr Opin Plant Biol*, vol. 10, no. 2, pp. 149–55, Apr. 2007, doi: 10.1016/j.pbi.2007.02.001.

[5] C. Da Silva *et al.*, "The high polyphenol content of grapevine cultivar tannat berries is conferred primarily by genes that are not shared with the reference genome.," *Plant Cell*, vol. 25, no. 12, pp. 4777–88, Dec. 2013, doi: 10.1105/tpc.113.118810.

[6] "The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla," *Nature*, vol. 449, no. 7161, pp. 463–467, Sep. 2007, doi: 10.1038/nature06148.

[7] Y. Li *et al.*, "De novo assembly of soybean wild relatives for pan-genome analysis of diversity and agronomic traits," *Nat Biotechnol*, vol. 32, no. 10, pp. 1045–1052, Oct. 2014, doi: 10.1038/nbt.2979.

[8] S. P. Gordon *et al.*, "Extensive gene content variation in the Brachypodium distachyon pan-genome correlates with population structure," *Nat Commun*, vol. 8, no. 1, p. 2184, Dec. 2017, doi: 10.1038/s41467-017-02292-8.

[9] L. Gao *et al.*, "The tomato pan-genome uncovers new genes and a rare allele regulating fruit flavor," *Nat Genet*, vol. 51, no. 6, pp. 1044–1051, Jun. 2019, doi: 10.1038/s41588-019-0410-2.

[10] R. Li *et al.*, "Building the sequence map of the human pan-genome," *Nat Biotechnol*, vol. 28, no. 1, pp. 57–63, Jan. 2010, doi: 10.1038/nbt.1596.

[11] R. M. Sherman *et al.*, "Assembly of a pan-genome from deep sequencing of 910 humans of African descent," *Nat Genet*, vol. 51, no. 1, pp. 30–35, Jan. 2019, doi: 10.1038/s41588-018-0273-y.

[12] W.-W. Liao *et al.*, “A draft human pangenome reference,” *Nature*, vol. 617, no. 7960, pp. 312–324, May 2023, doi: 10.1038/s41586-023-05896-x.

[13] C. Plissonneau, F. E. Hartmann, and D. Croll, “Pangenome analyses of the wheat pathogen Zymoseptoria tritici reveal the structural basis of a highly plastic eukaryotic genome,” *BMC Biol*, vol. 16, no. 1, p. 5, Dec. 2018, doi: 10.1186/s12915-017-0457-4.

[14] L. Zhou, T. Zhang, S. Tang, X. Fu, and S. Yu, “Pan-genome analysis of Paenibacillus polymyxa strains reveals the mechanism of plant growth promotion and biocontrol,” *Antonie Van Leeuwenhoek*, vol. 113, no. 11, pp. 1539–1558, Nov. 2020, doi: 10.1007/s10482-020-01461-y.

[15] S. M. Diene *et al.*, “The Rhizome of the Multidrug-Resistant Enterobacter aerogenes Genome Reveals How New ‘Killer Bugs’ Are Created because of a Sympatric Lifestyle,” *Mol Biol Evol*, vol. 30, no. 2, pp. 369–383, Feb. 2013, doi: 10.1093/molbev/mss236.

[16] K. Georgiades and D. Raoult, “Defining Pathogenic Bacterial Species in the Genomic Era,” *Front Microbiol*, vol. 1, 2011, doi: 10.3389/fmicb.2010.00151.

[17] C. N. Hirsch *et al.*, “Insights into the Maize Pan-Genome and Pan-Transcriptome ,” *Plant Cell*, vol. 26, no. 1, pp. 121–135, Feb. 2014, doi: 10.1105/tpc.113.119982.

[18] W. Wang *et al.*, “Genomic variation in 3,010 diverse accessions of Asian cultivated rice,” *Nature*, vol. 557, no. 7703, pp. 43–49, May 2018, doi: 10.1038/s41586-018-0063-9.

[19] S. Hübner *et al.*, “Sunflower pan-genome analysis shows that hybridization altered gene content and disease resistance,” *Nat Plants*, vol. 5, no. 1, pp. 54–62, Dec. 2018, doi: 10.1038/s41477-018-0329-0.

[20] B. Hurgobin *et al.*, “Homoeologous exchange is a major cause of gene presence/absence variation in the amphidiploid *Brassica napus*,” *Plant Biotechnol J*, vol. 16, no. 7, pp. 1265–1274, Jul. 2018, doi: 10.1111/pbi.12867.

[21] A. A. Golicz *et al.*, “The pangenome of an agronomically important crop plant Brassica oleracea,” *Nat Commun*, vol. 7, no. 1, p. 13390, Nov. 2016, doi: 10.1038/ncomms13390.

[22] S. Sun *et al.*, “An extended set of yeast-based functional assays accurately identifies human disease mutations,” *Genome Res*, vol. 26, no. 5, pp. 670–680, May 2016, doi: 10.1101/gr.192526.115.

[23] L. D. Alcaraz, G. Moreno-Hagelsieb, L. E. Eguiarte, V. Souza, L. Herrera-Estrella, and G. Olmedo, “Understanding the evolutionary relationships and major traits of Bacillus through comparative genomics,” *BMC*

Genomics, vol. 11, no. 1, p. 332, Dec. 2010, doi: 10.1186/1471-2164-11-332.

[24] L. Snipen, T. Almøy, and D. W. Ussery, "Microbial comparative pan-genomics using binomial mixture models," *BMC Genomics*, vol. 10, no. 1, p. 385, 2009, doi: 10.1186/1471-2164-10-385.

[25] Z. Hu, C. Wei, and Z. Li, "Computational Strategies for Eukaryotic Pangenome Analyses," in *The Pangenome*, Cham: Springer International Publishing, 2020, pp. 293–307. doi: 10.1007/978-3-030-38281-0_13.

[26] W. Yao, G. Li, H. Zhao, G. Wang, X. Lian, and W. Xie, "Exploring the rice dispensable genome using a metagenome-like assembly strategy," *Genome Biol*, vol. 16, no. 1, p. 187, Dec. 2015, doi: 10.1186/s13059-015-0757-3.

[27] A. A. Golicz, J. Batley, and D. Edwards, "Towards plant pangenomics," *Plant Biotechnol J*, vol. 14, no. 4, pp. 1099–1105, Apr. 2016, doi: 10.1111/pbi.12499.

[28] Z. Hu *et al.*, "Novel sequences, structural variations and gene presence variations of Asian cultivated rice," *Sci Data*, vol. 5, no. 1, p. 180079, May 2018, doi: 10.1038/sdata.2018.79.

[29] C. Sun *et al.*, "RPAN: rice pan-genome browser for ∼3000 rice genomes," *Nucleic Acids Res*, vol. 45, no. 2, pp. 597–605, Jan. 2017, doi: 10.1093/nar/gkw958.

[30] L. Ou *et al.*, "Pan-genome of cultivated pepper (*Capsicum*) and its use in gene presence-absence variation analyses," *New Phytologist*, vol. 220, no. 2, pp. 360–363, Oct. 2018, doi: 10.1111/nph.15413.

[31] B. A. Read *et al.*, "Pan genome of the phytoplankton Emiliania underpins its global distribution," *Nature*, vol. 499, no. 7457, pp. 209–213, Jul. 2013, doi: 10.1038/nature12221.

[32] J. D. Montenegro *et al.*, "The pangenome of hexaploid bread wheat," *The Plant Journal*, vol. 90, no. 5, pp. 1007–1013, Jun. 2017, doi: 10.1111/tpj.13515.

[33] Z. Hu *et al.*, "EUPAN enables pan-genome studies of a large number of eukaryotic genomes," *Bioinformatics*, vol. 33, no. 15, pp. 2408–2409, Aug. 2017, doi: 10.1093/bioinformatics/btx170.

[34] J. Wang *et al.*, "A pangenome analysis pipeline provides insights into functional gene identification in rice," *Genome Biol*, vol. 24, no. 1, p. 19, Jan. 2023, doi: 10.1186/s13059-023-02861-9.

[35] C. Laing *et al.*, "Pan-genome sequence analysis using Panseq: an online tool for the rapid analysis of core and accessory genomic regions," *BMC*

Bioinformatics, vol. 11, no. 1, p. 461, Dec. 2010, doi: 10.1186/1471-2105-11-461.

[36] M. J. Brittnacher, C. Fong, H. S. Hayden, M. A. Jacobs, M. Radey, and L. Rohmer, "PGAT: a multistrain analysis resource for microbial genomes," *Bioinformatics*, vol. 27, no. 17, pp. 2429–2430, Sep. 2011, doi: 10.1093/bioinformatics/btr418.

[37] "Computational pan-genomics: status, promises and challenges," *Brief Bioinform*, p. bbw089, Oct. 2016, doi: 10.1093/bib/bbw089.

[38] J. M. Eizenga *et al.*, "Pangenome Graphs," *Annu Rev Genomics Hum Genet*, vol. 21, no. 1, pp. 139–162, Aug. 2020, doi: 10.1146/annurev-genom-120219-080406.

[39] "Building pangenome graphs".

[40] H. Li, X. Feng, and C. Chu, "The design and construction of reference pangenome graphs with minigraph.," *Genome Biol*, vol. 21, no. 1, p. 265, Oct. 2020, doi: 10.1186/s13059-020-02168-z.

[41] J. Sirén *et al.*, "Pangenomics enables genotyping of known structural variants in 5202 diverse genomes," *Science (1979)*, vol. 374, no. 6574, Dec. 2021, doi: 10.1126/science.abg8871.

[42] G. Hickey *et al.*, "Genotyping structural variants in pangenome graphs using the vg toolkit," *Genome Biol*, vol. 21, no. 1, p. 35, Dec. 2020, doi: 10.1186/s13059-020-1941-7.

[43] R. R. Wick, M. B. Schultz, J. Zobel, and K. E. Holt, "Bandage: interactive visualization of de novo genome assemblies.," *Bioinformatics*, vol. 31, no. 20, pp. 3350–2, Oct. 2015, doi: 10.1093/bioinformatics/btv383.

[44] G. Gonnella, N. Niehus, and S. Kurtz, "GfaViz: flexible and interactive visualization of GFA sequence graphs," *Bioinformatics*, vol. 35, no. 16, pp. 2853–2855, Aug. 2019, doi: 10.1093/bioinformatics/bty1046.

[45] W. Beyer *et al.*, "Sequence tube maps: making graph genomes intuitive to commuters," *Bioinformatics*, vol. 35, no. 24, pp. 5318–5320, Dec. 2019, doi: 10.1093/bioinformatics/btz597.

[46] P. E. Bayer, A. A. Golicz, A. Scheben, J. Batley, and D. Edwards, "Plant pan-genomes are the new reference," *Nat Plants*, vol. 6, no. 8, pp. 914–920, Jul. 2020, doi: 10.1038/s41477-020-0733-0.

[47] Y. Liu *et al.*, "Pan-Genome of Wild and Cultivated Soybeans," *Cell*, vol. 182, no. 1, pp. 162-176.e13, Jul. 2020, doi: 10.1016/j.cell.2020.05.023.

[48] P. Qin *et al.*, "Pan-genome analysis of 33 genetically diverse rice accessions reveals hidden genomic variations," *Cell*, vol. 184, no. 13, pp. 3542-3558.e16, Jun. 2021, doi: 10.1016/j.cell.2021.04.046.

[49] P. Ebert *et al.*, “Haplotype-resolved diverse human genomes and integrated analysis of structural variation,” *Science (1979)*, vol. 372, no. 6537, Apr. 2021, doi: 10.1126/science.abf7117.

[50] K. Schneeberger *et al.*, “Simultaneous alignment of short reads against multiple genomes,” *Genome Biol*, vol. 10, no. 9, p. R98, 2009, doi: 10.1186/gb-2009-10-9-r98.

[51] G. Rakocevic *et al.*, “Fast and accurate genomic analyses using genome graphs,” *Nat Genet*, vol. 51, no. 2, pp. 354–362, Feb. 2019, doi: 10.1038/s41588-018-0316-4.

[52] D. Kim, J. M. Paggi, C. Park, C. Bennett, and S. L. Salzberg, “Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype,” *Nat Biotechnol*, vol. 37, no. 8, pp. 907–915, Aug. 2019, doi: 10.1038/s41587-019-0201-4.

[53] K. Vaddadi, R. Srinivasan, and N. Sivadasan, “Read mapping on genome variation graphs,” in *Leibniz International Proceedings in Informatics, LIPIcs*, Schloss Dagstuhl- Leibniz-Zentrum fur Informatik GmbH, Dagstuhl Publishing, Sep. 2019. doi: 10.4230/LIPIcs.WABI.2019.7.

[54] S. Ballouz, A. Dobin, and J. A. Gillis, “Is it time to change the reference genome?,” *Genome Biol*, vol. 20, no. 1, p. 159, Dec. 2019, doi: 10.1186/s13059-019-1774-4.

[55] Y. Zhou *et al.*, “Graph pangenome captures missing heritability and empowers tomato breeding,” *Nature*, vol. 606, no. 7914, pp. 527–534, Jun. 2022, doi: 10.1038/s41586-022-04808-9.

[56] A. Guarracino, S. Heumos, S. Nahnsen, P. Prins, and E. Garrison, “ODGI: understanding pangenome graphs,” *Bioinformatics*, vol. 38, no. 13, pp. 3319–3326, Jun. 2022, doi: 10.1093/bioinformatics/btac308.

[57] R. Guigó, P. Agarwal, J. F. Abril, M. Burset, and J. W. Fickett, “An assessment of gene prediction accuracy in large DNA sequences.,” *Genome Res*, vol. 10, no. 10, pp. 1631–42, Oct. 2000, doi: 10.1101/gr.122800.

[58] A. M. Lynn *et al.*, “An automated annotation tool for genomic DNA sequences using GeneScan and BLAST,” *J Genet*, vol. 80, no. 1, pp. 9–16, Apr. 2001, doi: 10.1007/BF02811413.

[59] W. H. Majoros, M. Pertea, and S. L. Salzberg, “TigrScan and GlimmerHMM: two open source ab initio eukaryotic gene-finders.,” *Bioinformatics*, vol. 20, no. 16, pp. 2878–9, Nov. 2004, doi: 10.1093/bioinformatics/bth315.

[60] E. Blanco, G. Parra, and R. Guigó, “Using geneid to identify genes.,” *Curr Protoc Bioinformatics*, vol. Chapter 4, p. Unit 4.3, Jun. 2007, doi: 10.1002/0471250953.bi0403s18.

[61] A. A. Salamov and V. V. Solovyev, “Ab initio Gene Finding in *Drosophila* Genomic DNA,” *Genome Res*, vol. 10, no. 4, pp. 516–522, Apr. 2000, doi: 10.1101/gr.10.4.516.

[62] I. Korf, “Gene finding in novel genomes,” *BMC Bioinformatics*, vol. 5, no. 1, p. 59, 2004, doi: 10.1186/1471-2105-5-59.

[63] M. Stanke, O. Keller, I. Gunduz, A. Hayes, S. Waack, and B. Morgenstern, “AUGUSTUS: ab initio prediction of alternative transcripts,” *Nucleic Acids Res*, vol. 34, no. Web Server, pp. W435–W439, Jul. 2006, doi: 10.1093/nar/gkl200.

[64] K. J. Hoff and M. Stanke, “Predicting Genes in Single Genomes with AUGUSTUS,” *Curr Protoc Bioinformatics*, p. e57, Nov. 2018, doi: 10.1002/cpbi.57.

[65] M. Stanke and B. Morgenstern, “AUGUSTUS: a web server for gene prediction in eukaryotes that allows user-defined constraints,” *Nucleic Acids Res*, vol. 33, no. Web Server, pp. W465–W467, Jul. 2005, doi: 10.1093/nar/gki458.

[66] M. Borodovsky and A. Lomsadze, “Eukaryotic Gene Prediction Using GeneMark.hmm-E and GeneMark-ES,” *Curr Protoc Bioinformatics*, vol. 35, no. 1, Sep. 2011, doi: 10.1002/0471250953.bi0406s35.

[67] “Training Augutus.”

[68] M. Manni, M. R. Berkeley, M. Seppey, and E. M. Zdobnov, “BUSCO: Assessing Genomic Data Quality and Beyond,” *Curr Protoc*, vol. 1, no. 12, Dec. 2021, doi: 10.1002/cpz1.323.

[69] K. J. Hoff, A. Lomsadze, M. Borodovsky, and M. Stanke, “Whole-Genome Annotation with BRAKER,” 2019, pp. 65–95. doi: 10.1007/978-1-4939-9173-0_5.

[70] K. J. Hoff, A. Lomsadze, M. Stanke, and M. Borodovsky, “BRAKER2: Incorporating Protein Homology Information into Gene Prediction with GeneMark-EP and AUGUSTUS.”

[71] K. J. Hoff, S. Lange, A. Lomsadze, M. Borodovsky, and M. Stanke, “BRAKER1: Unsupervised RNA-Seq-Based Genome Annotation with GeneMark-ET and AUGUSTUS,” *Bioinformatics*, vol. 32, no. 5, pp. 767–769, Mar. 2016, doi: 10.1093/bioinformatics/btv661.

[72] S. Mamidi *et al.*, “Demographic factors shaped diversity in the two gene pools of wild common bean Phaseolus vulgaris L.,” *Heredity (Edinb)*, vol. 110, no. 3, pp. 267–276, Mar. 2013, doi: 10.1038/hdy.2012.82.

[73] D. M. Goodstein *et al.*, “Phytozome: a comparative platform for green plant genomics,” *Nucleic Acids Res*, vol. 40, no. D1, pp. D1178–D1186, Jan. 2012, doi: 10.1093/nar/gkr944.

[74] E. Bellucci *et al.*, "Decreased Nucleotide and Expression Diversity and Modified Coexpression Patterns Characterize Domestication in the Common Bean," *Plant Cell*, vol. 26, no. 5, pp. 1901–1912, May 2014, doi: 10.1105/tpc.114.124040.

[75] M. Tarailo-Graovac and N. Chen, "Using RepeatMasker to Identify Repetitive Elements in Genomic Sequences," *Curr Protoc Bioinformatics*, vol. 25, no. 1, Mar. 2009, doi: 10.1002/0471250953.bi0410s25.

[76] J. M. Flynn *et al.*, "RepeatModeler2 for automated genomic discovery of transposable element families," *Proceedings of the National Academy of Sciences*, vol. 117, no. 17, pp. 9451–9457, Apr. 2020, doi: 10.1073/pnas.1921046117.

[77] J. Schmutz *et al.*, "A reference genome for common bean and genome-wide analysis of dual domestications," *Nat Genet*, vol. 46, no. 7, pp. 707–713, Jul. 2014, doi: 10.1038/ng.3008.

[78] H. Tang *et al.*, "An improved genome release (version Mt4.0) for the model legume Medicago truncatula," *BMC Genomics*, vol. 15, no. 1, p. 312, Dec. 2014, doi: 10.1186/1471-2164-15-312.

[79] J. Schmutz *et al.*, "Genome sequence of the palaeopolyploid soybean," *Nature*, vol. 463, no. 7278, pp. 178–183, Jan. 2010, doi: 10.1038/nature08670.

[80] G. Gremme, V. Brendel, M. E. Sparks, and S. Kurtz, "Engineering a software tool for gene structure prediction in higher organisms," *Inf Softw Technol*, vol. 47, no. 15, pp. 965–978, Dec. 2005, doi: 10.1016/j.infsof.2005.09.005.

[81] P. Jones *et al.*, "InterProScan 5: genome-scale protein function classification," *Bioinformatics*, vol. 30, no. 9, pp. 1236–1240, May 2014, doi: 10.1093/bioinformatics/btu031.

[82] M. Burset and R. Guigó, "Evaluation of Gene Structure Prediction Programs," *Genomics*, vol. 34, no. 3, pp. 353–367, Jun. 1996, doi: 10.1006/geno.1996.0298.

[83] E. Keibler and M. R. Brent, "Eval: A software package for analysis of genome annotations," *BMC Bioinformatics*, vol. 4, no. 1, p. 50, 2003, doi: 10.1186/1471-2105-4-50.

[84] S. F. Altschul, W. Gish, W. Miller, E. W. Myers, and D. J. Lipman, "Basic local alignment search tool," *J Mol Biol*, vol. 215, no. 3, pp. 403–410, Oct. 1990, doi: 10.1016/S0022-2836(05)80360-2.

[85] D. W. Mount, "Using the Basic Local Alignment Search Tool (BLAST)," *Cold Spring Harb Protoc*, vol. 2007, no. 7, p. pdb.top17, Jul. 2007, doi: 10.1101/pdb.top17.

[86] D. M. Emms and S. Kelly, “OrthoFinder: phylogenetic orthology inference for comparative genomics,” *Genome Biol*, vol. 20, no. 1, p. 238, Dec. 2019, doi: 10.1186/s13059-019-1832-y.

[87] H. Li, “Minimap2: pairwise alignment for nucleotide sequences,” *Bioinformatics*, vol. 34, no. 18, pp. 3094–3100, Sep. 2018, doi: 10.1093/bioinformatics/bty191.

[88] M. Nattestad and M. C. Schatz, “Assemblytics: a web analytics tool for the detection of variants from an assembly,” *Bioinformatics*, vol. 32, no. 19, pp. 3021–3023, Oct. 2016, doi: 10.1093/bioinformatics/btw369.

[89] H. Li *et al.*, “The Sequence Alignment/Map format and SAMtools.,” *Bioinformatics*, vol. 25, no. 16, pp. 2078–9, Aug. 2009, doi: 10.1093/bioinformatics/btp352.

[90] W. Li and A. Godzik, “Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences,” *Bioinformatics*, vol. 22, no. 13, pp. 1658–1659, Jul. 2006, doi: 10.1093/bioinformatics/btl158.

[91] E. M. Jonkheer *et al.*, “PanTools v3: functional annotation, classification and phylogenomics.,” *Bioinformatics*, vol. 38, no. 18, pp. 4403–4405, Sep. 2022, doi: 10.1093/bioinformatics/btac506.

[92] “https://neo4j.com/.”

[93] A. R. Quinlan and I. M. Hall, “BEDTools: a flexible suite of utilities for comparing genomic features,” *Bioinformatics*, vol. 26, no. 6, pp. 841–842, Mar. 2010, doi: 10.1093/bioinformatics/btq033.

[94] “https://www.ensembl.org/info/website/upload/gff.html”.

[95] S. Proost *et al.*, “i-ADHoRe 3.0—fast and sensitive detection of genomic homology in extremely large data sets,” *Nucleic Acids Res*, vol. 40, no. 2, pp. e11–e11, Jan. 2012, doi: 10.1093/nar/gkr955.

[96] S. A. Shiryev, J. S. Papadopoulos, A. A. Schäffer, and R. Agarwala, “Improved BLAST searches using longer words for protein seeding.,” *Bioinformatics*, vol. 23, no. 21, pp. 2949–51, Nov. 2007, doi: 10.1093/bioinformatics/btm479.

[97] Y. Wang *et al.*, “MCScanX: a toolkit for detection and evolutionary analysis of gene synteny and collinearity.,” *Nucleic Acids Res*, vol. 40, no. 7, p. e49, Apr. 2012, doi: 10.1093/nar/gkr1293.

[98] T. C. Freeman *et al.*, “Graphia: A platform for the graph-based visualisation and analysis of high dimensional data,” *PLoS Comput Biol*, vol. 18, no. 7, p. e1010310, Jul. 2022, doi: 10.1371/journal.pcbi.1010310.

[99] “The UniProt Consortium 2019”.

[100] E. W. Sayers *et al.*, “Database resources of the National Center for Biotechnology Information,” *Nucleic Acids Res*, vol. 49, no. D1, pp. D10–D17, Jan. 2021, doi: 10.1093/nar/gkaa892.

[101] P. Ewels, M. Magnusson, S. Lundin, and M. Käller, “MultiQC: summarize analysis results for multiple tools and samples in a single report,” *Bioinformatics*, vol. 32, no. 19, pp. 3047–3048, Oct. 2016, doi: 10.1093/bioinformatics/btw354.

[102] R. Challis, E. Richards, J. Rajan, G. Cochrane, and M. Blaxter, “BlobToolKit – Interactive Quality Assessment of Genome Assemblies,” *G3 Genes|Genomes|Genetics*, vol. 10, no. 4, pp. 1361–1374, Apr. 2020, doi: 10.1534/g3.119.400908.

[103] P. E. McClean *et al.*, “The Common Bean V Gene Encodes Flavonoid 3’5’ Hydroxylase: A Major Mutational Target for Flavonoid Diversity in Angiosperms,” *Front Plant Sci*, vol. 13, Mar. 2022, doi: 10.3389/fpls.2022.869582.

[104] P. E. McClean *et al.*, “White seed color in common bean (*Phaseolus vulgaris*) results from convergent evolution in the *P* (*pigment*) gene,” *New Phytologist*, vol. 219, no. 3, pp. 1112–1123, Aug. 2018, doi: 10.1111/nph.15259.

[105] J. A. O’Rourke *et al.*, “An RNA-Seq based gene expression atlas of the common bean,” *BMC Genomics*, vol. 15, no. 1, p. 866, 2014, doi: 10.1186/1471-2164-15-866.

[106] Y. Reinprecht, Y. Qi, F. Shahmir, T. H. Smith, and K. P. Pauls, “Yield and antiyield genes in common bean (<scp> *Phaseolus vulgaris* </scp> L.),” *Legume Science*, vol. 3, no. 3, Sep. 2021, doi: 10.1002/leg3.91.

[107] C. García-Fernández, A. Campa, and J. J. Ferreira, “Dissecting the genetic control of seed coat color in a RIL population of common bean (Phaseolus vulgaris L.),” *Theoretical and Applied Genetics*, vol. 134, no. 11, pp. 3687–3698, Nov. 2021, doi: 10.1007/s00122-021-03922-y.

[108] J. M. Eizenga *et al.*, “Pangenome Graphs.,” *Annu Rev Genomics Hum Genet*, vol. 21, pp. 139–162, Aug. 2020, doi: 10.1146/annurev-genom-120219-080406.

[109] Z. Wu *et al.*, “Graph pangenome reveals functional, evolutionary, and phenotypic significance of human nonreference sequences”, doi: 10.1101/2022.09.05.506692.

[110] B. Paten, J. M. Eizenga, Y. M. Rosen, A. M. Novak, E. Garrison, and G. Hickey, “Superbubbles, Ultrabubbles, and Cacti,” *Journal of Computational Biology*, vol. 25, no. 7, pp. 649–663, Jul. 2018, doi: 10.1089/cmb.2017.0251.

[111] Y.-F. Jiang *et al.*, “Pangenome obtained by long-read sequencing of 11 genomes reveal hidden functional structural variants in pigs,” *iScience*, vol. 26, no. 3, p. 106119, Mar. 2023, doi: 10.1016/j.isci.2023.106119.

[112] F. Zhang *et al.*, “Long-read sequencing of 111 rice genomes reveals significantly larger pan-genomes,” *Genome Res*, Apr. 2022, doi: 10.1101/gr.276015.121.

[113] D. E. Fouts, L. Brinkac, E. Beck, J. Inman, and G. Sutton, “PanOCT: automated clustering of orthologs using conserved gene neighborhood for pan-genomic analysis of bacterial strains and closely related species,” *Nucleic Acids Res*, vol. 40, no. 22, pp. e172–e172, Dec. 2012, doi: 10.1093/nar/gks757.

[114] M. A. Huynen and P. Bork, “Measuring genome evolution,” *Proceedings of the National Academy of Sciences*, vol. 95, no. 11, pp. 5849–5856, May 1998, doi: 10.1073/pnas.95.11.5849.

[115] J. T. Lovell *et al.*, “The genomic landscape of molecular responses to natural drought stress in Panicum hallii,” *Nat Commun*, vol. 9, no. 1, p. 5213, Dec. 2018, doi: 10.1038/s41467-018-07669-x.

[116] S. Rogic, A. K. Mackworth, and F. B. F. Ouellette, “Evaluation of Gene-Finding Programs on Mammalian Sequences,” *Genome Res*, vol. 11, no. 5, pp. 817–832, May 2001, doi: 10.1101/gr.147901.

[117] E. Veeckman, T. Ruttink, and K. Vandepoele, “Are We There Yet? Reliably Estimating the Completeness of Plant Genome Sequences,” *Plant Cell*, vol. 28, no. 8, pp. 1759–1768, Aug. 2016, doi: 10.1105/tpc.16.00349.

[118] S. J. Goodswen, P. J. Kennedy, and J. T. Ellis, “Evaluating High-Throughput Ab Initio Gene Finders to Discover Proteins Encoded in Eukaryotic Pathogen Genomes Missed by Laboratory Techniques,” *PLoS One*, vol. 7, no. 11, p. e50609, Nov. 2012, doi: 10.1371/journal.pone.0050609.

[119] M. Yandell and D. Ence, “A beginner’s guide to eukaryotic genome annotation,” *Nat Rev Genet*, vol. 13, no. 5, pp. 329–342, May 2012, doi: 10.1038/nrg3174.

[120] V. Ter-Hovhannisyan, A. Lomsadze, Y. O. Chernoff, and M. Borodovsky, “Gene prediction in novel fungal genomes using an ab initio algorithm with unsupervised training,” *Genome Res*, vol. 18, no. 12, pp. 1979–1990, Dec. 2008, doi: 10.1101/gr.081612.108.

[121] D. Monyak *et al.*, “Welcome to the big leaves: best practices for improving genome annotation in non-model plant 1 genomes 2 3 Vidya S Vuruputoor”, doi: 10.1101/2022.10.03.510643.

[122] T. Kwon, E. R. Hanschen, and B. T. Hovde, “Addressing the pervasive scarcity of structural annotation in eukaryotic algae,” *Sci Rep*, vol. 13, no. 1, p. 1687, Jan. 2023, doi: 10.1038/s41598-023-27881-0.

[123] Y. Gong, Y. Li, X. Liu, Y. Ma, and L. Jiang, “A review of the pangenome: how it affects our understanding of genomic variation, selection and breeding in domestic animals?,” *J Anim Sci Biotechnol*, vol. 14, no. 1, p. 73, May 2023, doi: 10.1186/s40104-023-00860-1.

[124] G. Vernikos, D. Medini, D. R. Riley, and H. Tettelin, “Ten years of pan-genome analyses,” *Curr Opin Microbiol*, vol. 23, pp. 148–154, Feb. 2015, doi: 10.1016/j.mib.2014.11.016.

[125] N. Walden and M. E. Schranz, “Synteny Identifies Reliable Orthologs for Phylogenomics and Comparative Genomics of the Brassicaceae,” *Genome Biol Evol*, vol. 15, no. 3, Mar. 2023, doi: 10.1093/gbe/evad034.

[126] N. Cochetel, A. Minio, M. Massonnet, A. M. Vondras, R. Figueroa-Balderas, and D. Cantu, “Diploid chromosome-scale assembly of the *Muscadinia rotundifolia* genome supports chromosome fusion and disease resistance gene expansion during *Vitis* and *Muscadinia* divergence,” *G3 Genes|Genomes|Genetics*, vol. 11, no. 4, Apr. 2021, doi: 10.1093/g3journal/jkab033.

[127] L. Li, C. J. Stoeckert, and D. S. Roos, “OrthoMCL: Identification of Ortholog Groups for Eukaryotic Genomes,” *Genome Res*, vol. 13, no. 9, pp. 2178–2189, Sep. 2003, doi: 10.1101/gr.1224503.

[128] D. Liu, M. Hunt, and I. J. Tsai, “Inferring synteny between genome assemblies: a systematic evaluation,” *BMC Bioinformatics*, vol. 19, no. 1, p. 26, Dec. 2018, doi: 10.1186/s12859-018-2026-4.

[129] P. Avdeyev, S. Jiang, S. Aganezov, F. Hu, and M. A. Alekseyev, “Reconstruction of Ancestral Genomes in Presence of Gene Gain and Loss,” *Journal of Computational Biology*, vol. 23, no. 3, pp. 150–164, Mar. 2016, doi: 10.1089/cmb.2015.0160.

[130] P. Feijão, F. V. Martinez, and A. Thévenin, “On the distribution of cycles and paths in multichromosomal breakpoint graphs and the expected value of rearrangement distance,” *BMC Bioinformatics*, vol. 16, no. S19, p. S1, Dec. 2015, doi: 10.1186/1471-2105-16-S19-S1.

[131] M. A. Alekseyev and P. A. Pevzner, “Breakpoint graphs and ancestral genome reconstructions,” *Genome Res*, vol. 19, no. 5, pp. 943–957, May 2009, doi: 10.1101/gr.082784.108.

[132] J. Ma *et al.*, “Reconstructing contiguous regions of an ancestral genome,” *Genome Res*, vol. 16, no. 12, pp. 1557–1565, Dec. 2006, doi: 10.1101/gr.5383506.

[133] N. Walden and M. E. Schranz, “Synteny Identifies Reliable Orthologs for Phylogenomics and Comparative Genomics of the Brassicaceae,” *Genome Biol Evol*, vol. 15, no. 3, Mar. 2023, doi: 10.1093/gbe/evad034.

[134] A. Sandelin *et al.*, “Arrays of ultraconserved non-coding regions span the loci of key developmental genes in vertebrate genomes,” *BMC Genomics*, vol. 5, no. 1, p. 99, Dec. 2004, doi: 10.1186/1471-2164-5-99.

[135] S. Y. Kim and J. K. Pritchard, “Adaptive Evolution of Conserved Noncoding Elements in Mammals,” *PLoS Genet*, vol. 3, no. 9, p. e147, Sep. 2007, doi: 10.1371/journal.pgen.0030147.

[136] W. Z. A. W. E. H. R. B. M. P. A. M. J. S. D. M. T. M. S. Y. K. J. D. B. I. Y. Bo Xia, “The genetic basis of tail-loss evolution in humans and apes,” *bioRxiv*, 2021.

[137] P. This, T. Lacombe, M. Cadle-Davidson, and C. L. Owens, “Wine grape (Vitis vinifera L.) color associates with allelic variation in the domestication gene VvmybA1,” *Theoretical and Applied Genetics*, vol. 114, no. 4, pp. 723–730, Feb. 2007, doi: 10.1007/s00122-006-0472-2.

[138] E. Butelli *et al.*, “Retrotransposons Control Fruit-Specific, Cold-Dependent Accumulation of Anthocyanins in Blood Oranges,” *Plant Cell*, vol. 24, no. 3, pp. 1242–1255, Mar. 2012, doi: 10.1105/tpc.111.095232.

SUPPLEMENTARY DATA

	Reference	Approach1	Approach2	Approach3	Approach4
Number of inital predictions	27433	30053	32387	34786	36713
Number of final predictions		25481	26757	26649	27669
Reduction rate of predictions		15.2	17.4	23.4	24.6

Table S1. *Number of initial and final predictions for each tested approach. The reduction rate reported as percentage of filtered prediction against the total ones.*

	Approach1	Approach2	Approach3	Approach4
Complete	92.4%	96.7%	92.2%	96.6%
Fragmented	2.3%	0.9%	2.6%	1.0%
Missing	5.3%	2.4%	5.2%	2.4%

Table S2. *Complete,fragmented and missing fabales conserved genes for tested approach*

GENE SENSITIVITY (%)										
	Test1	Test2	Test3	Test4	Test5	Test6	Test7	Test8	Test9	Test10
Approach 1	32.32	37.69	34.5	30.15	38.38	37.69	40.7	35	37.37	26.13
Approach 2	61.11	62.31	66	60.3	60.1	62.81	62.31	63	63.13	59.8
Approach 3	39.39	44.72	39	39.2	42.42	38.69	46.23	42	42.93	34.67
Approach 4	60.61	64.32	67.5	60.8	59.09	62.81	63.32	66	64.65	60.3
GENE SPECIFICITY (%)										
	Test1	Test2	Test3	Test4	Test5	Test6	Test7	Test8	Test9	Test10
Approach 1	30.33	34.4	31.22	26.09	35.85	35.55	36	31.67	35.24	23.01
Approach 2	52.38	51.88	56.17	48.39	51.52	51.44	52.54	52.72	54.82	50.64
Approach 3	34.98	38.53	33.77	31.84	37.84	33.77	38.02	35.9	38.12	29.74
Approach 4	51.72	52.24	56.49	47.83	49.79	51.44	52.28	54.1	55.17	50

Table S3. *Gene specificity and sensitivity using ten testing sets of random 200 P.vulgaris official annotation*

	Synteny	Gene Family	Synteny+Gene Family
Core genes	19757	20712	20784
Variable genes	4205	6204	3981
Unique genes	5362	1675	1527
Total	29324	28591	26292

***Table S4**. Core, Variable, Unique and Total genes identified by Synteny, Gene family and Combined approaches.*

	Linear pangenome	Element-based graph pangenome
Core	20396	20784
Variable	10080	3981
Unique	100	1527
Total	30549	26292

***Table S5.** Core, Variable, Unique and Total genes identified in linear and element-based graph pangenomes.*

Number of genomes	Nodes in Nucleotide-based pangenome
1	1
2	1
3	62
4	62
5	70

***Table S7.** Complexity of nucleotide-based pangenome in a local region. Number of nodes is reported per number of added genome.*

	Uncollapsed element graph	Collapsed element graph	Nucleotide-based graph
1	27433	27433	478
2	54534	28723	52919
3	83003	28934	202942
4	112526	40842	214120
5	145524	26292	292860
AVERAGE GROWTH	29523	-285	73096

Table S8. *Complexity of nucleotide-based pangenome and element-based graph pangenome in terms of nodes per added genomes*

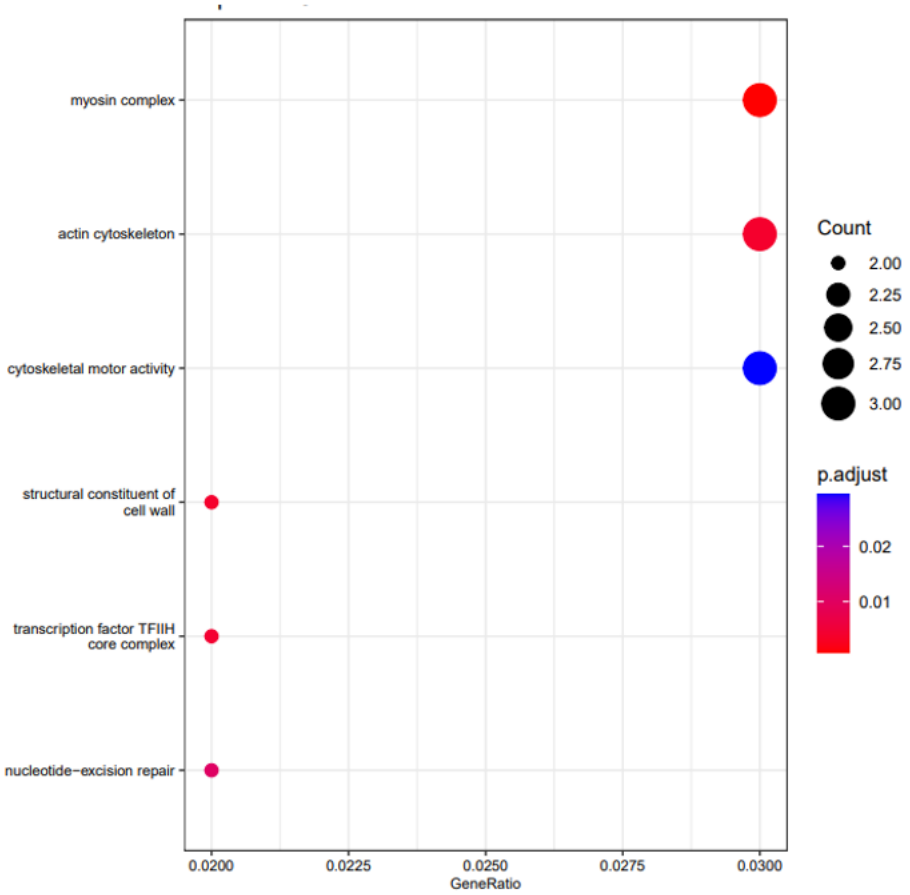

Figure S1. *Functional enrichment of 1,527 unique genes*