Revista de Informática Teórica e Aplicada - RITA - ISSN 2175-2745
Vol. 26, Num. 1 (2019) 36-46

RESEARCH ARTICLE

A gene based bacterial whole genome comparison toolkit

Um conjunto de ferramentas para a comparação de genomas completos de bactérias baseada em genes

Luciano Antonio Digiampietri[1]*, Vivian Mayumi Yamassaki Pereira[1], Geraldo José dos Santos Júnior[1], Giovani de Sousa Leite[1], Priscilla Koch Wagner[1], Leandro Márcio Moreira[2], Caio Santiago[3]

Abstract: Most of the computational biology analysis is made comparing genomic features. The nucleotide and amino acid sequence alignments are frequently used in gene function identification and genome comparison. Despite its widespread use, there are limitations in their analysis capabilities that need to be considered but are often overlooked or unknown by many researchers. This paper presents a gene based whole genome comparison toolkit which can be used not only as an alternative and more robust way to compare a set of whole genomes, but, also, to understand the tradeoff of the use of sequence local alignment in this kind of comparison. A study case was performed considering fifteen whole genomes of the *Xanthomonas* genus. The results were compared with the 16S rRNA-processing protein RimM phylogeny and some thresholds for the use of sequence alignments in this kind of analysis were discussed.
Keywords: Bioinformatics — Whole genome — Genome comparison — Phylogeny — Pangenome — Genome visualization

Resumo: Grande parte das análises realizadas na biologia computacional é feita comparando características genômicas. Os alinhamentos de nucleotídeos e de aminoácidos são frequentemente usados na identificação de funções gênicas e na comparação de genomas. Apesar de seu uso generalizado, há limitações em suas capacidades de análise que precisam ser consideradas, mas são frequentemente negligenciadas ou desconhecidas por muitos pesquisadores. Este artigo apresenta um conjunto de ferramentas de comparação de genomas completos baseado em genes que pode ser usado não somente como uma maneira alternativa e mais robusta de comparar um conjunto de genomas completos, mas também para entender as vantagens e desvantagens do uso do alinhamento local de sequências neste tipo de comparação. Um estudo de caso foi realizado considerando quinze genomas completos do gênero *Xanthomonas*. Os resultados foram comparados com a filogenia produzida utilizando a proteína *16S rRNA-processing protein RimM* e alguns limiares para o uso de alinhamentos de sequências neste tipo de análise foram discutidos.
Palavras-Chave: Bioinformática — Genome completo — Comparação de genomas — Filogenia — Pan-genoma — Visualização de genomas

[1] *Escola de Artes, Ciências e Humanidades, Universidade de São Paulo, Brasil*
[2] *Departamento de Ciências Biológicas, Universidade Federal de Ouro Preto, Brasil*
[3] *Bioinformática, Universidade de São Paulo, Brasil*
*__Corresponding author__: digiampietri@usp.br
DOI: https://doi.org/10.22456/2175-2745.84814 • **Received:** 15/07/2018 • **Accepted:** 10/12/2018

1. Introduction

Genetic studies date back to the early 20th-century [1], but it was only in the 1970s that a technique was introduced that made it possible to know the sequence of bases that make up DNA molecules of simple organisms [2]. The process was very costly from a financial and time-consuming point of view, making it impossible to massify it. With the advance of technologies, the mechanisms available for genetic sequencing have become more accessible, faster and cheaper, producing a proliferation of genomes or parts of sequenced genomes [1].

Given the massive amount of data available, comparative analysis becomes even more important for the discovery of new proteins functions and the understanding of genomes. Based on the comparison of sequences it is possible to infer the homology of sequences. Homologous sequences tend

to have their functions conserved [3, 4], and therefore the association of sequences with unknown functions with others which functions is known allows inferring the function of these sequences. It is a simpler and cheaper method than the verification made by laboratory experiments.

With the increasing speed with which new data arise, there are also new demands for analysis. Although the great volume of data seems to be an obstacle to the understanding the coding sequences and the genome as a whole, this brings opportunities for new comparative genomics studies. In the last years, a scenario with a model genome of some species [5] was replaced for another with populations of genomes of a certain group [6]. And having a population of related genomes as study material, it is possible to search for relations between specific differences of phenotype and differences in the genome [7].

Thus, the comparison of genomes is a very important task to identify the set of shared characteristics and the exclusive ones that can help in understanding the studied genomes [8, 7].

The principle of comparative genomics defines that genome features with similar characteristics probably have a conserved function [3, 4]. Thus, most of the computational biology analysis is made comparing different genomic features. The nucleotide and amino acid sequence alignments are frequently used in gene function identification and genome comparison [9].

Despite its widespread use, there are limitations in their analysis capabilities that need to be considered but are often overlooked or unknown by many researchers. For example, many works perform phylogenetic analysis using only one specific gene. Thus, if this gene is incorrectly annotated it may influence the resulting phylogeny. Moreover, a unique gene may not be able to discriminate genomes from close species, such as species of the same genus or pathovars from the same species, and will not be able to represent some important evolutionary phenomena, such as lateral gene transfer [10].

This paper presents a gene-based whole genome comparison toolkit which can be used not only as an alternative and more robust way to compare a set of prokaryote whole genomes, but, also, to understand the tradeoff of the use of sequence local alignment in this kind of comparison. Among the steps and features this toolkit there is a start with sequence clustering based on graph modeling, and others analysis anchored in the graph, as well as matrix, phylogenies and plot visualization.

A case study was performed considering fifteen *Xanthomonas* whole genomes. The results were compared with the 16S rRNA-processing protein RimM phylogeny and some thresholds for the use of sequence alignments in this kind of analysis were discussed. Moreover, a brief discussion about the use of only the gene sequences or the combination with the whole genome DNA sequence is presented.

The rest of this paper is organized as follows. Section 2 presents the developed toolkit. Section 3 contains the application of the toolkit in a study case. Section 4 summarizes the related work. Finally, Section 5 contains the conclusions and future work.

2. The developed toolkit

The toolkit is composed of three main types of tools: homologous gene identification, genome comparison, and gene network visualization and analysis. Figure 1 summarizes the toolkit process flowchart.

2.1 Homologous gene identification

In this work, the detection of possible homologies is made considering nucleotide and amino acid sequence alignments. The raw data used consists of whole genome nucleotide sequence (.fna files) and gene files (.faa files). From these raws data, two alignments are performed: genes versus genes (for example, using *BLASTP* program) and genes versus whole genome DNA (for example, using *TBLASTN* program). This second alignment is not required, but it is suggested because it allows the identification of non-annotated genes which can disturb the genome comparison using only annotated genes information. The alignment data format used by the toolkit is *BLAST m8 format*, thus, these alignments can be produced by different tools, such as BLAST [11] and bowtie [12].

There are two strategies implemented to identify the homologous genes: one genome versus the others and all genomes versus all genomes. In both, the query in the alignments corresponds to the genes from the genomes and the subject can be the genes from the genomes or the whole DNA sequence from them. The first strategy is used when the user wants to identify genes that are similar to the ones from a specific genome. To belong to a gene family, a gene must have an alignment (that satisfies the threshold values) with a gene from this specific genome. The second strategy is used in the comparison of several genomes (all versus all), and a gene will belong to a gene family if it aligns with any gene in this family.

The next step in the process of homologous identification is the sequence clustering, that consists of define groups of sequences according to a specific strategy. In our strategy, to be considered homologous, two genes must have an alignment of their sequences which satisfy a threshold composed of seven alignment parameters: minimum identity percentage, minimum alignment percentage, minimum alignment length, maximum number of mismatched positions, maximum number of gap positions, maximum e-value, and minimum bit-score.

The homologous relation is modelled as an undirected and unweighted graph from the alignment results. Each node in the graph represents a gene and each edge represents an alignment which satisfied the threshold. Each connected component in the graph represents a homologous family. In this study, the values were defined as, in an automatic way, respectively, 96, 96, 60, 20, 5, 10^{-10}, and 100.

All these values can be chosen by the users or defined by a tool that automatically chooses the best match in a search space, in a process known as sequence clustering. In this work, we present a sequence clustering algorithm supported by the

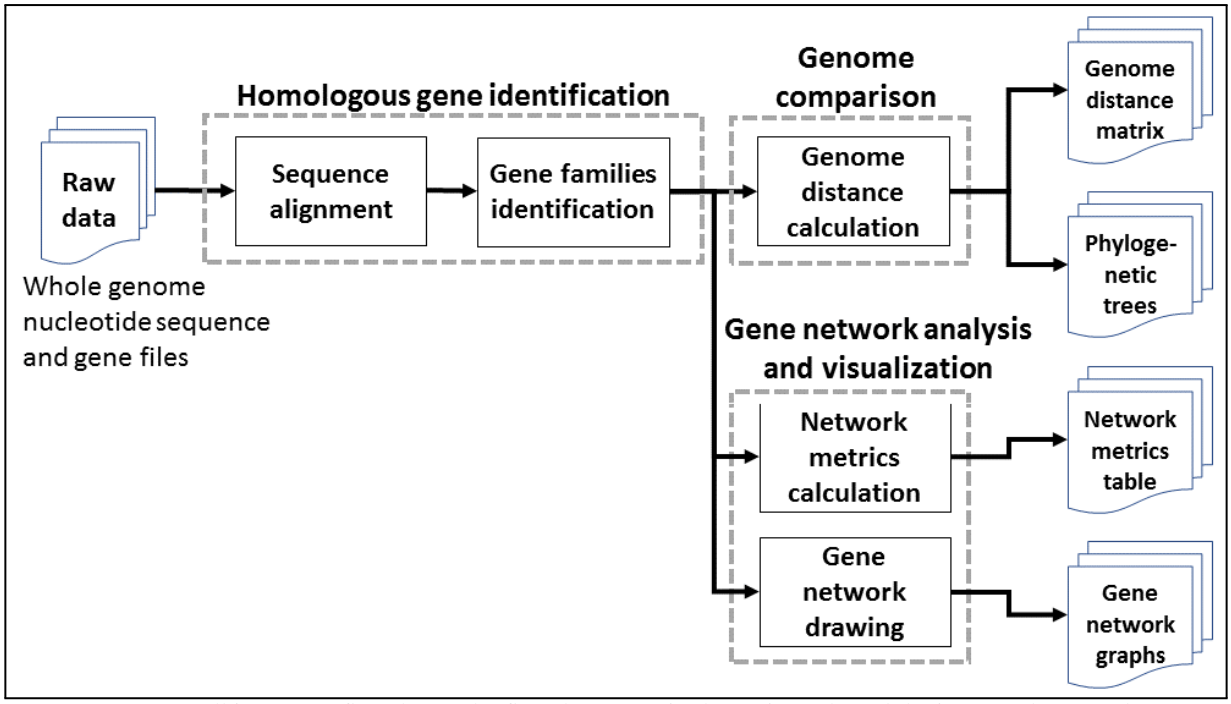

Figure 1. Toolkit process flowchart. The flowchart contain the main tools and the input and output data.

idea that the relation of homologous genes is transitive [13], and for this we maximize the clustering coefficient. This metric measures the transitivity of the relationships (edges) in the graph. For each subset of three connected nodes, the clustering coefficient measures the probability that these three nodes are a clique of size three. For example, given the nodes a, b, and c, if there is an edge between a and b, and another between b and c, the clustering coefficient will measure the probability of the existence of an edge between a and c. In the current context, if the sequence alignment threshold indicates that a and b are homologous, and b and c too, it is desired the existence of an edge between a and c since they are considered homologous because belong to the same connected component. Thus, the threshold used in the sequence alignment should provide high values of clustering coefficient in the corresponding gene homologous network. The threshold used in the study case are the ones that maximized the clustering coefficient in a set of bacteria that will be presented in the next section.

An important feature of this method is that as the edges are held or discarded, thus the original structure of the alignments is preserved. It diverges from others methods which ignore [14] or transform [15] the edges. It indicates the topology of families, and it will be useful in future researches to analyze the evolution the families or cluster again to make smaller groups.

2.2 Genome comparison

The genome comparison used the families produced by the homologous gene identification tools to compare the genomes and produce phylogenetic trees or cladograms.

There are two distance metrics implemented: Euclidean and Manhattan. These distances can be applied to compare the genomes using the homologous families considering the presence or absence of a gene in each family. We developed three variations of these distances: considering the total number o genes from each genome in each gene family, considering the binary information of presence or absence, or considering the normalized values of a number of genes in the respective family. The calculation of the distances produces a square matrix in which each line and column correspond to a genome, and each cell value contains the distance between the two genomes (the one from the line compared with the one from the column).

Moreover, a clustering algorithm is used to produce cladograms from this information. In this project the R (*https://www.r-project.org/*) *phytools* package clustering algorithm was used.

Besides the use of a cladogram or a phylogenic tree, the tools also map the genomes in a two-dimensional space considering the two principal components resulting from the Principal Component Analysis (PCA). This strategy maps a multidimensional data (in this case, genomes described with hundreds or thousands of genes in homologous families) in a new multidimensional data where each dimension (starting from the first one) maximizes the data variance. Thus, the two dimensions used to draw the genomes in a two-dimensional space are the ones that most represents the variance in the gene information when comparing these genomes.

2.3 Gene network analysis and visualization

About the clustering coefficient previously addressed, it is possible to get this metric for the whole genome, a specific family

or a node, but beyond the clustering coefficient, some others metrics can be calculated: number of connected components, distribution of the number of genes per connected component, distribution of the number of genomes with genes in each connected component, and gene degree distribution in the gene network. For each connected component, the most frequent annotated function is selected to represent the function of this homologous group.

A graphical representation of the gene homologous network is produced. The tool can, also, draw a figure with a specific component selected by the user. This tool uses a force-directed algorithm in order to approximate connected nodes and separate the ones that are not connected.

In addition, three .csv files are produced. Each one of these files corresponds to a table, where each column represents a genome and each line represents a family of homologous genes. The value of each table cell indicates if the genome contains or not a gene in the respective family, and/or the amount of these genes. In short, one file has information about all gene families; other about genes present in only one genome; and, the last one, about the genes present in all genomes.

3. Results/Discussion

In order to evaluate the developed toolkit and provide a discussion about some parameters and strategies, a case study was performed considering 15 whole genomes of *Xanthomonas* genus available at National Center for Biotechnology Information (NCBI) website (*http://www.ncbi.nlm.nih.gov/*).

Table 1 contains a brief description of these genomes, including the abbreviation and the colour used in the nodes from the graphs presented in this section. The *Xanthomonas* genus was selected because it contains bacteria which are economically important and, in the last two decades, they were intensively studied and different phylogenies were constructed, based on phylogenomics or specific genes [16, 17, 18].

The Xanthomonas species are part of the Xanthomonadaceae family, which "consists of species of non-pathogenic and pathogenic γ-proteobacteria that infect different hosts, including humans and plant" [19]. This family is responsible for several diseases which results in heavy economic losses to agro-related industry [20]. In particular, *Xanthomonas campestris* is a species of particular interest because, besides causing different plant diseases, it is also used in the commercial production of xanthan gum, which is a water-soluble exo-polysaccharide used in the food industry [21].

The strategy used in this case study was the gene versus gene sequence alignments, comparing all genomes versus all genomes. The parameters used as thresholds for the sequence alignment were chosen automatically, and as previously discussed, that is those which maximize the clustering coefficient. Table 2 presents some of the tested values for two parameters (minimum identity percentage and minimum alignment percentage) and the resulting clustering coefficient. The line with highest clustering coefficient is highlighted in this table.

Figure 2 contains the genome distance matrix, based on the presence or absence of genes for each group of homologous genes using Euclidean distance. The genomes are grouped according to their distances.

The sequence alignment from the 65,119 genes in the sample resulted in 206,127 alignments (after applying the filters). 50,214 genes aligned with at least one other gene. 10,090 homologous groups were created with two or more genes. The most frequent groups have four genes (38.51% of the total); two genes (17.89%), three genes (16.82%), and five genes (9.87%). The giant group/component (connected component which contains the biggest number of genes) has 270 genes and corresponds to a group of ISXo8 transposases.

When considering the amount of genomes that have genes in each of the homologous groups, it was observed that only 33 groups have genes from all genomes (0.33%). The most frequent groups are the ones with genes from four genomes (38.97%), two (17.83%) and three (17.22%). There are 459 groups (4.55%) which have genes from 14 of the 15 genomes analyzed. This is justified by the fact that the genome of *Xanthomonas albilineans* presents a drastic reduction in the size of its genome compared to the other genomes of bacteria of this genus according to [22].

Figure 3 presents the homologous genes' network where each gene is colored according to the genome they belong (see Table 1). Isolated genes are not showed in this figure. As presented, there are 10,090 connected components (homologous families) and 50,214 nodes (genes). The zooming highlights a small region of Figure 3. It is possible to observe a dense component composed of dozens of genes (A); a component with a low clustering coefficient (B); a component composed of exact one gene from each genome (C); and a component with only four genes (D).

Figure 4 presents the giant component from the genes' network using as threshold 80% for minimum identity and minimum alignment on the left, and another one using 96% for these two parameters (default value) on the right. The component in the left has 354 genes, annotated as belonging to different types of transposases (such as IS1404, ISXoo3, and IS1403). It is possible to observe in the figure that this component connects at least two cohesive groups. In the component in the right, there are 270 genes (all of them are ISXo8 transposases). These data are supported by Salzberd and co-authors who have described the importance of these transposable elements, especially IsXo8, in the rapid evolution in the genome of *Xanthomonas oryzae* [23].

The resulting genome comparison was summarized in two graphical representations: a cladogram using a hierarchical clustering and the results from the PCA (using the two main components). These results were compared with the phylogeny produced by the multiple alignment of the 16S rRNA-processing protein RimM. As presented in the introduction, the phylogenetic analysis considering only one gene has some limitations and can lead to some mistakes.

The first limitation of the one gene phylogenetic analy-

Table 1. Genomes used in the case study

Genome	Abbrev.	Chromossome size	# of Plasmids	Plasmids total size	# of genes	Color legend
Xanthomonas albilineans GPE PC73	Xalbilineans	3768695	3	83604	3208	
Xanthomonas axonopodis pv. citri str. 306	Xacitri	5175554	2	98620	4427	
Xanthomonas axonopodis pv. citrumelo F1	Xacitrumelo	4967469	0	0	4181	
Xanthomonas axonopodis Xac29-1	Xac29-1	5153455	3	143070	4403	
Xanthomonas campestris pv. campestris str. 8004	Xcc8004	5148708	0	0	4271	
Xanthomonas campestris pv. campestris str. ATCC 33913	XccATCC33913	5076188	0	0	4179	
Xanthomonas campestris pv. campestris str. B100	XccB100	5079002	0	0	4466	
Xanthomonas campestris pv. raphani 756C	Xraphani	4941214	0	0	4516	
Xanthomonas campestris pv. vesicatoria str. 85-10	Xvesicatoria	5178466	4	241686	4726	
Xanthomonas citri subsp. citri Aw12879	Xccitri	5321499	2	77186	4760	
Xanthomonas fuscans subsp. fuscans	Xff	4981995	3	106688	4083	
Xanthomonas oryzae pv. oryzae KACC 10331	XooKACC10331	4941439	0	0	4065	
Xanthomonas oryzae pv. oryzae MAFF 311018	XooMAFF311018	4940217	0	0	4372	
Xanthomonas oryzae pv. oryzae PXO99A	XooPXO99A	5240075	0	0	4988	
Xanthomonas oryzae pv. oryzicola BLS256	Xooryzicola	4831739	0	0	4474	

	XccATCC33913	Xcc8004	XccB100	Xraphani	Xalbilineans	Xooryzicola	XooPXO99A	XooMAFF311018	XooKACC10331	Xacitrumelo	Xvesicatoria	Xff	Xccitri	Xacitri	Xac29-1
XccATCC33913	0,00	11,83	21,75	25,63	59,94	71,40	74,97	75,41	74,47	75,51	75,76	72,68	78,61	78,91	78,35
Xcc8004	11,83	0,00	20,17	26,85	60,46	71,83	75,49	75,93	74,99	75,91	75,92	73,08	78,56	79,42	78,87
XccB100	21,75	20,17	0,00	22,36	58,79	70,38	74,11	74,57	73,60	74,55	74,64	71,62	77,35	78,06	77,51
Xraphani	25,63	26,85	22,36	0,00	56,02	68,11	71,95	72,39	71,42	72,34	72,75	69,42	75,77	75,98	75,41
Xalbilineans	59,94	60,46	58,79	56,02	0,00	50,78	55,65	56,39	54,87	57,30	57,89	53,17	61,44	61,65	60,84
Xooryzicola	71,40	71,83	70,38	68,11	50,78	0,00	28,28	28,86	29,46	58,14	58,79	54,88	62,50	62,88	62,39
XooPXO99A	74,97	75,49	74,11	71,95	55,65	28,28	0,00	17,46	20,93	62,99	63,56	59,83	66,86	67,10	66,61
XooMAFF311018	75,41	75,93	74,57	72,39	56,39	28,86	17,46	0,00	18,63	63,54	64,04	60,42	67,40	67,58	67,16
XooKACC10331	74,47	74,99	73,60	71,42	54,87	29,46	20,93	18,63	0,00	62,56	63,10	59,41	66,51	66,74	66,26
Xacitrumelo	75,51	75,91	74,55	72,34	57,30	58,14	62,99	63,54	62,56	0,00	14,07	37,76	47,22	47,85	47,61
Xvesicatoria	75,76	75,92	74,64	72,75	57,89	58,79	63,56	64,04	63,10	14,07	0,00	38,21	47,27	48,27	48,01
Xff	72,68	73,08	71,62	69,42	53,17	54,88	59,83	60,42	59,41	37,76	38,21	0,00	34,73	35,38	35,14
Xccitri	78,61	78,56	77,35	75,77	61,44	62,50	66,86	67,40	66,51	47,22	47,27	34,73	0,00	19,18	21,52
Xacitri	78,91	79,42	78,06	75,98	61,65	62,88	67,10	67,58	66,74	47,85	48,27	35,38	19,18	0,00	16,52
Xac29-1	78,35	78,87	77,51	75,41	60,84	62,39	66,61	67,16	66,26	47,61	48,01	35,14	21,52	16,52	0,00

Figure 2. Genome distance matrix

sis is about the confidence in the gene annotation. It is not uncommon to find uncertain gene annotations. For example, considering the multiple amino acid alignment, performed by MUSCLE (*http://www.ebi.ac.uk/Tools/msa/muscle/*), from the *16S rRNA-processing protein RimM* from the genomes analysed in this case study, it is possible to see ten amino acids presents in only four of the genomes (left side of Figure 5). After these ten amino acids there is a start codon, thus, it could represent a possible annotation error.

The three *Xanthomonas oryzae pv. oryzae* 16S rRNA-processing protein RimM are so similar that the phylogenetic tree produced by PhyML (*http://www.atgc-montpellier.fr/phyml/*) was not able to phylogenetically distinguish them. The same occurs with the three species of *Xanthomonas citri*. Moreover, the use of this gene was not able to identify with a high *bootstrap* value the positions of the *Xanthomonas albilineans* and *Xanthomonas campestris raphani*.

The strategy developed, considering the presence or absence of all the homologous groups (and without the need of multiple sequence alignments) was able to differentiate the three pathovars from *Xanthomonas oryzae pv. oryzae*, and the three species of *Xanthomonas citri*. Moreover, it was able to group the four pathovars from *Xanthomonas campestris* and to identify that *Xanthomonas albilineans* is probably more related to the *Xanthomonas oryzae* than the other species, and the *Xanthomonas campestris raphani* is more related to the *Xanthomonas campestris* species (see Figure 6). Indeed, the latter case makes all sense since raphani patovar is included in the campetris species. This information is consistent with the information presented in Figure 2.

The result of the PCA analysis of the presence and absence of genes is shown in Figure 7. The main component is mapped int the x-axis and it is able to represent 77.49% of the variance. The second is mapped in the y-axis and it is able to represent 59.53% of the remaining variance.

We also used the homologous genes identified by the

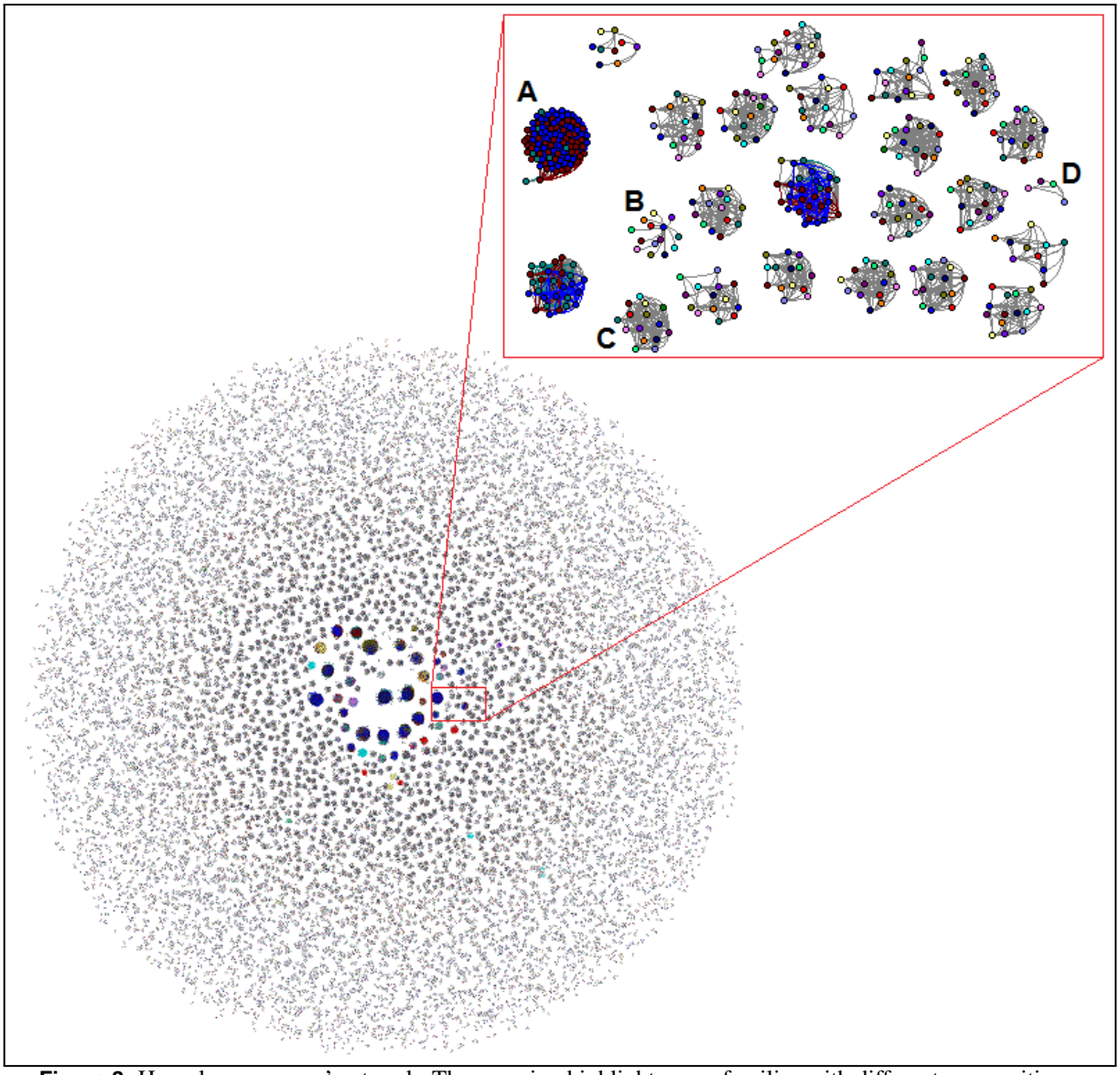

Figure 3. Homologous genes' network. The zooming highlights gene families with different compositions.

toolkit in another study which considered only the core genome. Table 3 shows the list of homologous gene families with only one gene presented in all genomes and we analysed if these families could be used to differentiate closely related genomes. We highlight that the small number of genes in the core genome may have occurred due to two factors. The first one is the default values considered in this experiment to filter the alignments, as they are very restrict in order to maximize the clustering coefficient. The second one is the fact that we considered the genome of the *Xanthomonas albilineans GPE PC73* which is known to be very different of the others Xanthomonas studied in this experiment [23].

In order to make the analysis, multiple alignments were performed with MUSCLE tool for each of these gene families. Based on the resulting alignments, phylogenetic reconstructions were produced using PhyML.

We observed that none of the 30 different phylogenies distinguished all the closely related genomes as well as were observed in the phylogeny with 16S rRNA gene. Some phylogenies were capable to group the *Xanthomonas oryzae* and the *Xanthomonas campestris* as can be observed in the phylogeny based on the gene family acetyl-CoA carboxylase biotin carboxylase subunit presented in the Figure **??**. However none of the phylogenies, including the phylogeny of the Figure **??**, distinguished all the different pathovars of the species studied.

The worst case was observed when the phylogeny was constructed based on the homologous gene families 50S ribosomal protein L35 and integration host factor subunit alpha, which presented the topology shown in Figure **??**. In this phylogeny, only *Xanthomonas albilineans GPE PC73* was considered different from the others.

As a conclusion to this experiment, we observed that

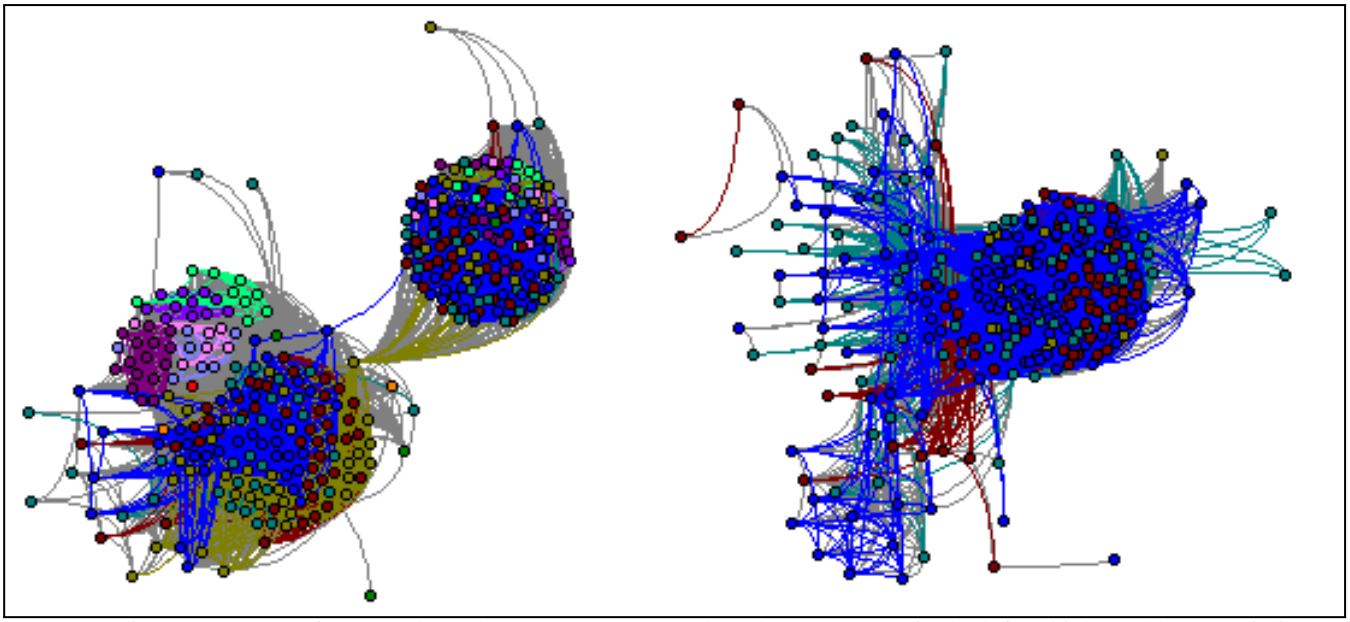

Figure 4. Giant components from two homologous genes' networks. The component in the left, which used less restrictive thresholds, has 354 genes from different types of transposases. On the other hand, the component in the right contains 270 genes of only one type of transposase.

Table 2. Clustering coefficient considering different thresholds

Minimum identity	Minimum alignment	Clustering coefficient
80%	80%	0.903
85%	85%	0.905
90%	90%	0.914
91%	91%	0.915
92%	92%	0.919
93%	93%	0.924
94%	94%	0.925
95%	95%	0.937
96%	**96%**	**0.940**
97%	97%	0.937
98%	98%	0.908
99%	99%	0.884

the use of homologous gene families individually is not a good approach to produced phylogenies of closely related genomes. No phylogeny based on these genes, individually, was able to, at the same time, distinguish the three main groups of *Xanthomonas* (the group of *Xanthomonas oryzae*, *Xanthomonas campestris* and the other group of the remaining *Xanthomonas*), and all the pathovars presented in the dataset.

It is worth mentioning the Multilocus Sequence Analysis approach (MLSA) is also able to solve some of the comparative genome challenges, especially when comparing phylogenetically close related genomes [24]. Instead of producing a phylogenetic tree using the aminoacid sequences of one gene, this approach produces phylogenetic tree from a concatenation of selected gene sequences. One of the main challenges of this approach is to select the correct genes for this analysis.

Young et al [25] applied the MLSA approach in the genomic comparison of the Xanthomonas genus. The phylogenetic tree was produced considering the following genes: dnaK, fyuA, gyrB, and rpoD.

We applied the same approach, considering the same genes, for the genomes used in our case study. The resulting phylogenetic tree is presented in Figure 10. We highlight this approach was able to distinguish some of the main groups of Xanthomonas, but the *Xanthomonas axonopodis* pv. citrumelo F1 was inserted apart of the other *Xanthomonas axonopodis* genomes and the *Xanthomonas campestris* pv. vesicatoria str. 85-10 genome. It occurred because one of the four genes used in the MLSA (the gyrB gene) is significative different in the *Xanthomonas axonopodis* pv. citrumelo F1 genome.

As discussed, the information presented in this section was based on the amino acid sequence alignment of the genes from 15 genomes. The use of annotated genes information may suffer from some problems that also occur in the phylogenetic analysis based on one specific gene, such as: genes that were not identified/annotated and genes that were incorrectly annotated. The strategy that considers the alignment of genes with the DNA from the whole genomes can minimize these problems. Only to exemplify, when comparing the *Xanthomonas campestris pv. campestris str. B100* alignment results versus all the DNA (using *TBLASTN*) and all the genes (using *BLASTP*), the first one was able to identify differences in 272 homologous groups, summing up 461 possible genes that are not present in the genes annotated files. Most of them are hypothetical (68%), but different types of genes were iden-

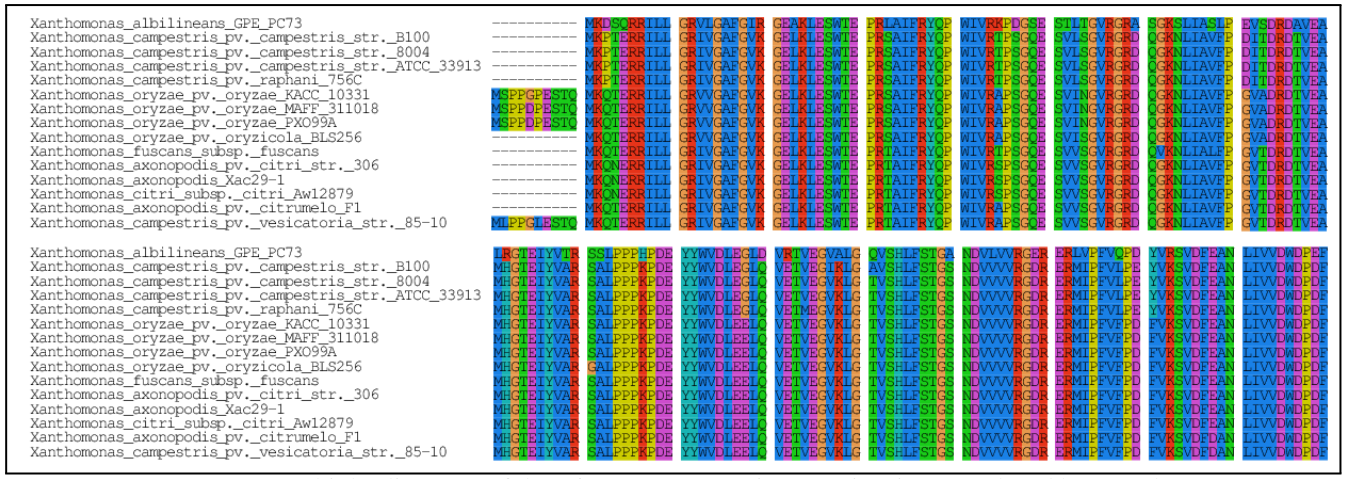

Figure 5. Multiple alignment of the 16S rRNA-processing protein RimM produced by Muscle

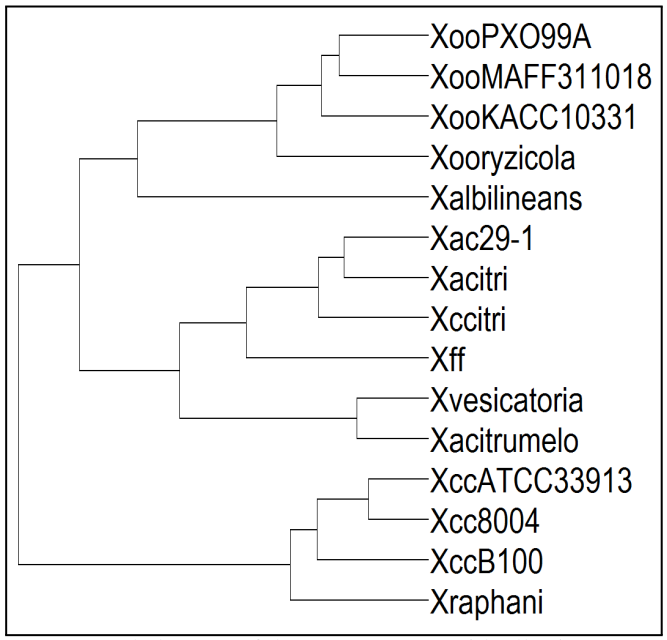

Figure 6. Cladogram of the core genome with all 15 genomes analyzed, considering the presence or absence of genes in the homologous groups.

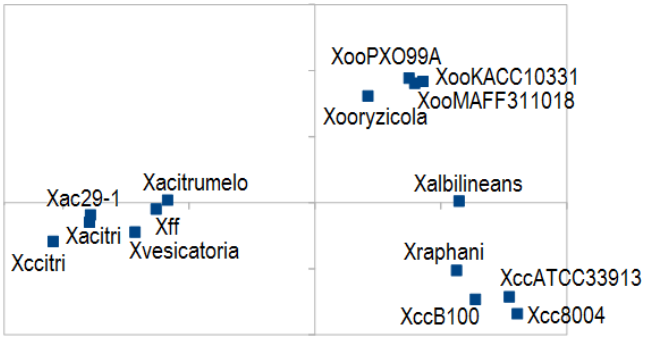

Figure 7. Mapping of the genomes' genes information in the two principal components. The mapping grouped the genomes in three groups and let the *X. albilineans* apart of these groups.

tified (including different types of proteins from the secretion system that may be involved in the pathogenic characteristics of these bacteria).

An additional case study with 69 genomes of the Xanthomonadacea family was performed. The toolkit was able to automatically identify eight families of orthologous proteins in 99.3% of the phytoassociated genomes, allowing the identification of proteins potentially associated with adaptation and virulence in plant tissue. This result confirms data from the literature [19]. The toolkit is also being used in other case study, comparing 55 *Streptococcus pyogens* genomes.

4. Related Work

There is an extensive literature about whole genome comparison and identification of homologous groups of genes [26, 27].

The identification of homologous, i.e., the sequence clustering are, typically, based on pairwise alignment and graph-based model [15, 28, 29, 30, 14]. Between them, the most widely accepted algorithm for is the MCL that uses Hidden Markov Model for clustering biological relations [15]. This is a semi-automatic algorithm because depends on granularity indicator (guided by inflation parameter). However, choosing the inflation parameter do not have so clear consequences than define thresholds to alignments. An another point, the proposed solution (using threshold) is finer granularity than MCL, resulting in smaller components than MCL.

One of the most related works with this toolkit is the Roary, which uses pairwise local alignments to make clusters of sequences, and posteriorly visualizations [31]. The Roary uses the MCL for clustering, on the other hand, we use an own clustering algorithm that is substantantily more restrictive than MCL and non-dependent of the inflation-like parameters. Some exported data are very similar to Roary's

Table 3. List of homologous gene families presented in all genomes

#	Homologous gene family
1	*30S ribosomal protein S1*
2	*30S ribosomal protein S7*
3	*30S ribosomal protein S9*
4	*30S ribosomal protein S11*
5	*30S ribosomal protein S12*
6	*50S ribosomal protein L13*
7	*50S ribosomal protein L14*
8	*50S ribosomal protein L18*
9	*50S ribosomal protein L22*
10	*50S ribosomal protein L35*
11	*acetyl-CoA carboxylase biotin carboxylase subunit*
12	*acyl carrier protein*
13	*adenylosuccinate synthetase*
14	*ATP-dependent Clp protease proteolytic subunit*
15	*ATP-dependent protease ATP-binding subunit ClpX*
16	*cell division protein*
17	*chemotaxis response regulator*
18	*glycine cleavage system transcriptional repressor*
19	*integration host factor subunit alpha*
20	*NADH-ubiquinone oxidoreductase 20 kda subunit*
21	*rod shape-determining protein MreB*
22	*transcription antitermination protein NusG*
23	*transcription regulator protein*
24	*transcription termination factor Rho*
25	*translation initiation factor IF-3*
26	*twitching motility protein*
27	*two-component system regulatory protein*
28	*two-component system regulatory protein (response regulator) required for AvrXa21*
29	*two-component system response regulatory protein (PilG)*
30	*type II citrate synthase*

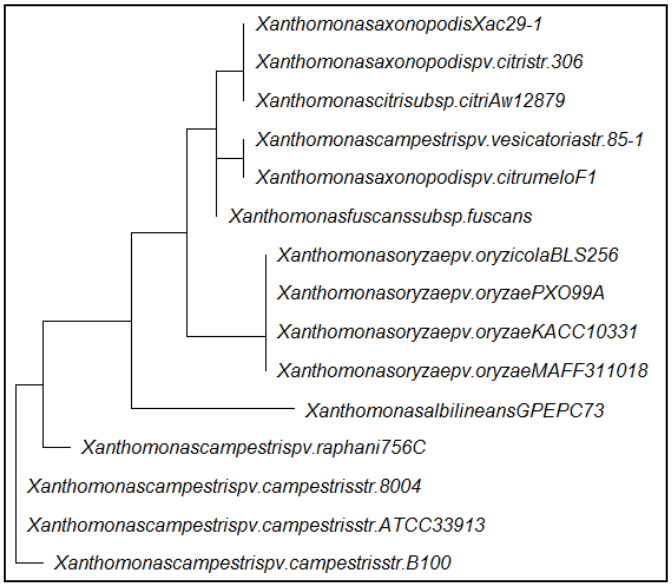

Figure 8. Phylogeny based on homologous gene family acetyl-CoA carboxylase biotin carboxylase subunit.

Figure 9. Phylogeny based on homologous gene family 50S ribosomal protein L35. The phylogeny of the family integration host factor subunit alpha presented the same topology: all the genomes are grouped together except the *Xanthomonas albilineans* one.

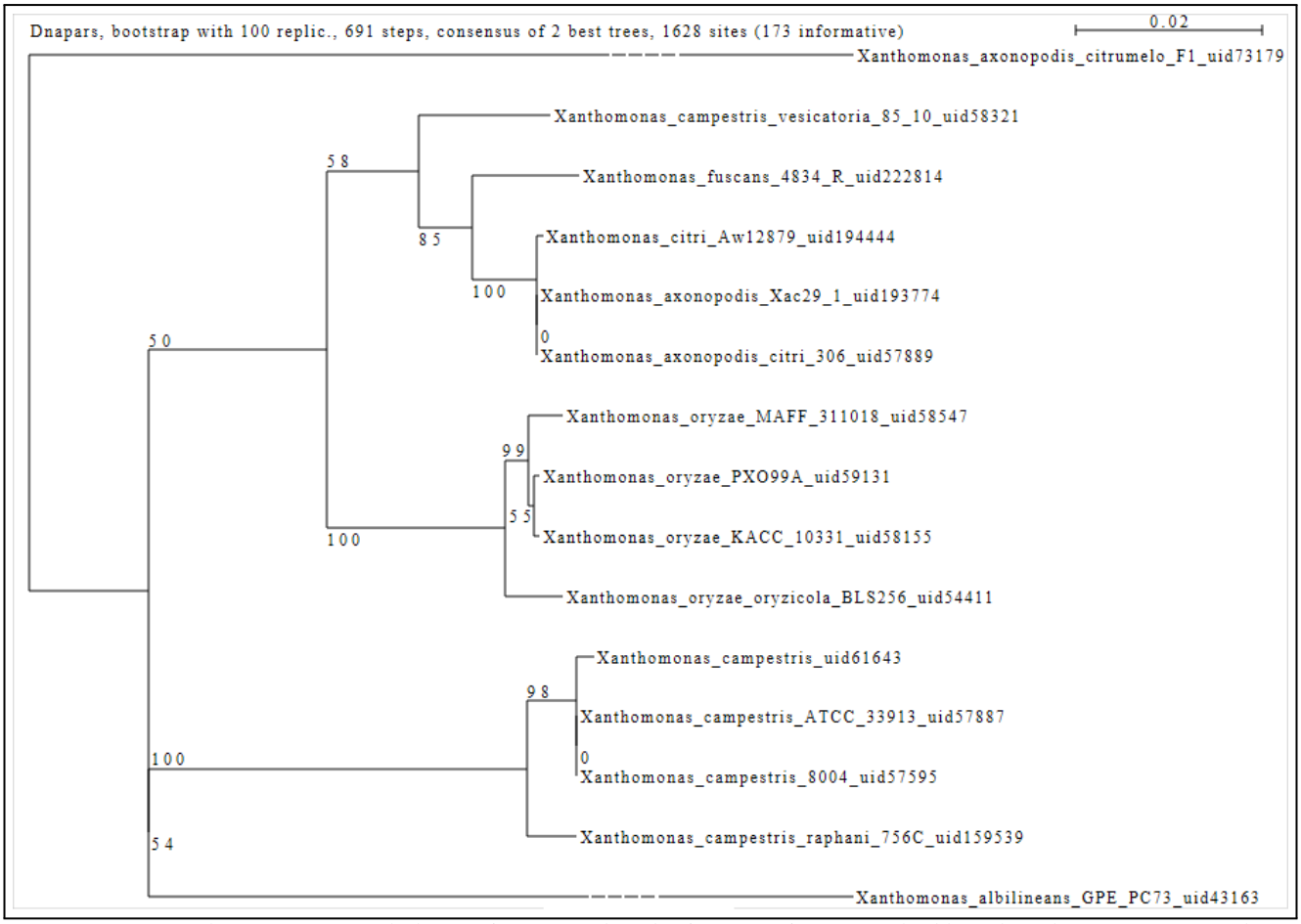

Figure 10. Xanthomonas genus phylogenetic tree produced using a parsimony method based on the MLSA approach.

output files making partially compatible with related tools, with the addiction of new tools like 2D plot using PCA and a graphical pan-genome network visualization.

There are wide range of comparative genomics toolkits, the majority of them follow the model of Roary, and consequently of our work too. This means clustering sequences in families (oftentimes using MCL) and based in these families present different results. Each toolkit differs from which type of results are presented, for example, ITEP is able to make a metabolic reconstruction [32], BPGA present a distribution of families based on KEGG pathway [33].

An alternative approach to discovery orthologous relations, as well as phylogeny, is through whole-genome alignment [34]. This approach considers that aligned regions can be a consequence of a speciation phenomenon, and the phylogeny can be obtained using a distance metric. This approach was already used for pairwise genome comparison for bacteria and eukaryotes [35, 36], where the application and the visualization shown good results. The pairwise alignment allows achieved satisfactory results in the identification of speciation events [37, 38]. However, the pairwise alignment does not provide the understanding of the whole population of a species. On the other hand, there are some open questions about multiple genome alignments. Two of these questions concern the alignment of the reverse sequence and how to analyze duplicated regions [39, 40]. These limitations are particularly problematic in bacteria populations that present more rearrangement and duplication events, for example, in the case study analyzed in the current paper, one component has 270 genes (more than 15 copies per genome).

5. Conclusions

This paper presented a toolkit for helping the activities of whole genome comparison based on gene information.

The developed tools aim to provide alternative ways to explore phylogenetic characteristics from prokaryote whole genomes, but also to help in the understanding of the advantages and disadvantages of the different approaches.

The data produced by the tools can be used as basic information for the core-genome and pan-genome analysis. Besides the use in whole genome comparison, the developed tools can also be applied in the assessment of the annotation of genes comparing the annotation of genes from a genome that is being annotated with the annotation of similar ones. The tools can provide insights about annotations errors and identify genes that should be annotated.

In order to download the presented toolkit, the reader is invited to contact the authors. An example of the results produced by the toolkit can be visualized in the following website: ⟨http://143.107.58.250/reportStrep2/⟩. It contains the comparison of 55 *Streptococcus pyogenes* genomes.

As future work we intend to improve the toolkit with different strategies of specific gene phylogenetic analysis, and tools to automatically compare these results with the ones produced by the tools presented in this paper. Moreover, we intend to develop tools to perform the core and pan-genome analysis of a group of related genomes. We also developed an alternative algorithm for homologous gene identification which will be compared with the algorithm presented in this paper [41].

AUTHOR CONTRIBUTIONS

All authors contributed equally to this work.

References

[1] FIETTO, J. L. R.; MACIEL, T. E. F. Sequenciando genomas. In: *Ciências genômicas: fundamentos e aplicações*. 1. ed. Porto alegre, Brazil: Sociedade Brasileira de Computação, 2015. v. 1, p. 27–64.

[2] SANGER, F.; NICKLEN, S.; COULSON, A. R. Dna sequencing with chain-terminating inhibitors. *Proc. Natl. Acad. Sci. U. S. A.*, v. 74, n. 12, p. 5463–5467, 12 1977.

[3] HARDISON, R. C. Comparative genomics. *PLOS Biol.*, v. 1, n. 2, p. e58, 11 2003.

[4] XIA, X. *Comparative Genomics*. 1. ed. Berlin, Heidelberg: Springer Berlin Heidelberg, 2013. v. 1. (SpringerBriefs in Genetics, v. 1).

[5] LANDER, E. S. et al. Initial sequencing and analysis of the human genome. *Nature*, v. 409, n. 6822, p. 860–921, 2 2001.

[6] KEHDY, F. S. G. et al. Origin and dynamics of admixture in brazilians and its effect on the pattern of deleterious mutations. *Proc. Natl. Acad. Sci.*, v. 112, n. 28, p. 8696–8701, 7 2015.

[7] ILINA, E. N. et al. Comparative genomic analysis of mycobacterium tuberculosis drug resistant strains from russia. *PLoS ONE*, v. 8, n. 2, p. e56577, 2 2013.

[8] LU, Y. et al. Omics data reveal the unusual asexual-fruiting nature and secondary metabolic potentials of the medicinal fungus cordyceps cicadae. *BMC Genom.*, v. 18, n. 1, p. 668, 2017.

[9] TATUSOV, R. L.; KOONIN, E. V.; LIPMAN, D. J. A genomic perspective on protein families. *Science*, v. 278, n. 5338, p. 631–637, 1997.

[10] DALQUEN, D. A. et al. The impact of gene duplication, insertion, deletion, lateral gene transfer and sequencing error on orthology inference: A simulation study. *PLOS ONE*, v. 8, n. 2, p. 1–11, 2 2013.

[11] ALTSCHUL, S. et al. Basic local alignment search tool. *J. Mol. Biol.*, v. 215, n. 3, p. 403–410, 1990.

[12] LANGMEAD, B. et al. Ultrafast and memory-efficient alignment of short dna sequences to the human genome. *Genome Biol.*, v. 10, n. 3, p. R25, 2009.

[13] SASSON, O.; LINIAL, N.; LINIAL, M. The metric space of proteins– comparative study of clustering algorithms. *BIOINFORMATICS*, v. 18, n. 1, p. 14–21, 2002.

[14] BOLTEN, E. et al. Clustering protein sequences–structure prediction by transitive homology. *Bioinform. (Oxf. Engl.)*, v. 17, n. 10, p. 935–41, 10 2001.

[15] ENRIGHT, A. J.; DONGEN, S. V.; OUZOUNIS, C. A. An efficient algorithm for large-scale detection of protein families. *Nucleic acids res.*, v. 30, n. 7, p. 1575–84, 4 2002.

[16] RYAN, R. P. et al. Pathogenomics of xanthomonas: understanding bacterium-plant interactions. *Nat. rev., Microbiol.*, v. 9, n. 5, p. 344–55, 5 2011.

[17] JALAN, N. et al. Comparative genomic and transcriptome analyses of pathotypes of xanthomonas citri subsp. citri provide insights into mechanisms of bacterial virulence and host range. *BMC genom.*, v. 14, n. 14, p. 551, 2013.

[18] ZHANG, Y. et al. Positive selection is the main driving force for evolution of citrus canker-causing xanthomonas. *ISME J.*, v. 9, n. 10, p. 2128–2138, 2015.

[19] ASSIS, R. de A. B. et al. Identification and analysis of seven effector protein families with different adaptive and evolutionary histories in plant-associated members of the xanthomonadaceae. *Sci. Reportsvolume*, v. 7, n. 23, p. 16133, 2017.

[20] TENNANT, P. F. et al. Diseases and pests of citrus (citrus spp.). *Tree Sci Biotech*, v. 3, n. 1, p. 81–107, 2009.

[21] PALANIRAJ, A.; JAYARAMAN, V. Production, recovery and applications of xanthan gum by xanthomonas campestris. *J. Food Eng.*, v. 106, n. 1, p. 1–12, 2011.

[22] PIERETTI, I. et al. The complete genome sequence of xanthomonas albilineans provides new insights into the reductive genome evolution of the xylem-limited xanthomonadaceae. *BMC Genom.*, v. 10, n. 1, p. 616, 12 2009.

[23] SALZBERG, S. L. et al. Genome sequence and rapid evolution of the rice pathogen xanthomonas oryzae pv. oryzae pxo99a. *BMC genom.*, v. 9, n. 204, p. 204, 2008.

[24] GEVERS, D. et al. Re-evaluating prokaryotic species. *Nat. Rev. Microbiol.*, v. 3, n. 9, p. 733–739, 2005.

[25] YOUNG, J. et al. A multilocus sequence analysis of the genus Xanthomonas. *Syst. Appl. Microbiol.*, v. 31, n. 5, p. 366 – 377, 2008.

[26] ALTENHOFF, A. M.; DESSIMOZ, C. Phylogenetic and functional assessment of orthologs inference projects and methods. *PLoS comput. biol.*, v. 5, n. 1, p. e1000262, 1 2009.

[27] LAING, C. R. et al. Everything at once: comparative analysis of the genomes of bacterial pathogens. *Vet. microbiol.*, v. 153, n. 1-2, p. 13–26, 11 2011.

[28] ENRIGHT, A. J.; OUZOUNIS, C. A. Generage: a robust algorithm for sequence clustering and domain detection. *BIOINFORMATICS*, v. 16, n. 5, p. 451–457, 2000.

[29] PROCLUST: improved clustering of protein sequences with an extended graph-based approach. *BIOINFORMATICS*, v. 18, n. 2, p. 182–191, 2002.

[30] ABASCAL, F.; VALENCIA, A. Clustering of proximal sequence space for the identification of protein families. *BIOINFORMATICS*, v. 18, n. 7, p. 908–921, 2002.

[31] PAGE, A. J. et al. Roary: rapid large-scale prokaryote pan genome analysis. *Bioinformatics*, v. 31, n. 22, p. 3691–3693, 2015.

[32] BENEDICT, M. N. et al. Itep: An integrated toolkit for exploration of microbial pan-genomes. *BMC Genom.*, v. 15, n. 1, p. 8, 2014.

[33] CHAUDHARI, N. M.; GUPTA, V. K.; DUTTA, C. Bpga-an ultra-fast pan-genome analysis pipeline. *Sci. Rep.*, v. 6, n. April, p. 1–10, 2016.

[34] COURONNE, O. Strategies and tools for whole-genome alignments. *Genome Res.*, v. 13, n. 1, p. 73–80, 1 2003.

[35] FLEISCHMANN, R. D. et al. Whole-genome comparison of mycobacterium tuberculosis clinical and laboratory strains. *J. Bacteriol.*, v. 184, n. 19, p. 5479–5490, 10 2002.

[36] ROUCHKA, E. C.; GISH, W.; STATES, D. J. Comparison of whole genome assemblies of the human genome. *Nucleic acids res.*, v. 30, n. 22, p. 5004–14, 11 2002.

[37] SILVA, A. C. R. da et al. Comparison of the genomes of two xanthomonas pathogens with differing host specificities. *Nature*, v. 417, n. 6887, p. 459–463, 5 2002.

[38] QIAN, W. Comparative and functional genomic analyses of the pathogenicity of phytopathogen xanthomonas campestris pv. campestris. *Genome Res.*, v. 15, n. 6, p. 757–767, 5 2005.

[39] DARLING, A. E.; MAU, B.; PERNA, N. T. progressivemauve: Multiple genome alignment with gene gain, loss and rearrangement. *PLoS ONE*, v. 5, n. 6, p. e11147, 6 2010.

[40] DUBCHAK, I. et al. Multiple whole-genome alignments without a reference organism. *Genome Res.*, v. 19, n. 4, p. 682–689, 4 2009.

[41] SANTIAGO, C.; PEREIRA, V.; DIGIAMPIETRI, L. Homology detection using multilayer maximum clustering coefficient. *J. Comput. Biol.*, v. 25, n. 2, p. 1328–1338, 2018.