MITOCHONDRIAL DNA PART A, 2018
https://doi.org/10.1080/24701394.2017.1419214

Taylor & Francis
Taylor & Francis Group

RESEARCH ARTICLE

Phylogenomics of tomato chloroplasts using assembly and alignment-free method

Raúl Martin Amado Cattáneo[a,b], Luis Diambra[a,b] and Andrés Norman McCarthy[a,c]

[a]Facultad de Ciencias Exactas-UNLP, CREG, La Plata, Argentina; [b]CONICET, Buenos Aires, Argentina; [c]CICPBA, La Plata, Argentina

ABSTRACT

Phylogenetics and population genetics are central disciplines in evolutionary biology. Both are based on the comparison of single DNA sequences, or a concatenation of a number of these. However, with the advent of next-generation DNA sequencing technologies, the approaches that consider large genomic data sets are of growing importance for the elucidation of evolutionary relationships among species. Among these approaches, the assembly and alignment-free methods which allow an efficient distance computation and phylogeny reconstruction are of great importance. However, it is not yet clear under what quality conditions and abundance of genomic data such methods are able to infer phylogenies accurately. In the present study we assess the method originally proposed by Fan et al. for whole genome data, in the elucidation of Tomatoes' chloroplast phylogenetics using short read sequences. We find that this assembly and alignment-free method is capable of reproducing previous results under conditions of high coverage, given that low frequency *k*-mers (i.e. error prone data) are effectively filtered out. Finally, we present a complete chloroplast phylogeny for the best data quality candidates of the recently published 360 tomato genomes.

ARTICLE HISTORY
Received 5 October 2017
Accepted 15 December 2017

KEYWORDS
Assembly and alignment-free method; phylogenomics; *k*-mers; tomato; chloroplast

1. Introduction

The evolutionary relationship among species, or populations, can be studied by inference methods based on a comparative analysis of genetic data under some models of DNA evolution. The set of such techniques is known as molecular phylogenetics analysis (or simply phylogenetics), and their product, the phylogenetic tree, is a diagrammatic model of the evolutionary history of a group of organisms. Nowadays, phylogenetics has become a principal tool in the understanding of both evolution and biodiversity.

In general, alcohol dehydrogenase and phytochrome genes (Small et al. 2004), chloroplast sequences corresponding to coding regions (e.g. *matK*, *rbcL*, *rpoB*, and *rpoC1*), non-coding spacers (e.g. *atpF-atpH*, *trnH-psbA,* and *psbK-psbI*) (Hollingsworth et al. 2009), and internal transcribed spacers (ITS) of the nuclear ribosomal DNA (Li et al. 2011), have been used for phylogenetic reconstruction in different taxonomic levels. However, single-sequence based approaches can fail when these fragment sequences have low variations in closely related species, or due to the absence of homologous nucleotide sequences in far related species. Additionally, the concatenation of many individual genes can be used to improve the resolution of the phylogenetic analysis (Chu et al. 2004; Qi et al. 2017). In the current genomic era, next-generation DNA sequencing technologies provide a large amount of genomic data which is readily available in gene-banks. Such data enable the use of phylogenomic approaches to establish evolutionary relationships.

The main distinction between phylogenetics and phylogenomics is scale. Phylogenomics lays at the union between evolutionary biology and genomic-scale studies (Chan and Ragan 2013). There have been numerous methods developed for performing phylogenetic analysis and, as the field calls for more ways to handle genome-scale data, these methods have improved and evolved to meet the challenge. Typical algorithms employed in phylogenetics scale poorly with the number of sequences; consequently high-quality phylogenomic analysis of large data sets can be computationally infeasible. In addition, next-generation sequences can be both incomplete and error prone. Analysis may also result complex due to the presence of genome rearrangement (fusion or deletion) or horizontal gene transfer. Thus, next-generation data require next-generation phylogenomics, including the presently assessed alignment-free approaches (Chan and Ragan 2013).

Chloroplast DNA is widely accepted as a primary source for phylogenetic analysis in plants. As a consequence, a pleiad of phylogenetic analyses have been performed based on comparison of sequences of (multiple) protein-coding genes in chloroplast genomes (Martin et al. 1998; Turmel et al. 1999, 2002; Adachi et al. 2000; Lemieux et al. 2000; De Las Rivas et al. 2002; Martin et al. 2002; Turmel et al. 2002). Alternatively, other methodologies for phylogenetic analysis of complete genomes have also been proposed (Sankoff et al. 1992; Fitz-Gibbon and House 1999; Tekaia et al. 1999; Lin and Gerstein 2000).

CONTACT Andrés Norman McCarthy amccarthy@exactas.unlp.edu.ar CICPBA, La Plata, Argentina

Nevertheless, all the previous approaches are based on the alignment of homologous sequences. This fact establishes that much information (such as gene rearrangement and insertions/deletions) in these data sets is lost after sequence alignment (without considering the intrinsic problems of alignment algorithms, Stuart et al. 2002; Li et al. 2011).

Every classical phylogenomic method consists of three main steps: clustering of homologous sequences, multiple sequence alignment (MSA) of each cluster, and inference of a phylogenetic tree based on the alignment. MSA is a crucial step in these approaches, and implicitly assumes that by inserting gaps and sliding blocks, we can generate a position-wise hypothesis of homology across the entire length of the sequences. This assumption may be unrealistic because genes and genomes are subject to other modification processes (i.e. recombination, inversion, rearrangement and lateral genetic transfer) (Chan and Ragan 2013).

The occurrence of such events would thus violate models of nucleotide substitution across sequence positions and lineages (specified in MSA-based approaches), thus biasing the subsequent phylogenetic inference (Wong et al. 2008; Stiller 2011).

Additionally, genomic data available through next-generation sequencing is growing geometrically. This, and the fact that MSA does not scale well with genome-scale data, establishes that classical phylogenetic methods will soon be impractical for large-scale comparative genome analysis. The assembly and alignment-free methods (although not yet fully developed or assessed) are potentially able to steer free from such limitations.

Alignment-free methodologies in phylogeny are techniques that can produce trees without the need to perform multiple sequence alignment (Haubold 2014). Such techniques are based on any number of statistical, computational, and biological principles. Recently, Fan et al. (2015) have developed an assembly and alignment-free (AAF) method for phylogeny reconstruction. This method first calculates pairwise genetic distances between two samples of short sequence reads. This distance between samples or species, is based on the estimate of the rate parameter from a Poisson process for a mutation occurring at a single nucleotide under the assumption (evolutionary model) that the mutation rate is the same for all nucleotides across the genomes. This also includes not only mutations caused by nucleotide substitutions, but also insertions and deletions (indels) (Fan et al. 2015). The phylogenetic relationships among the samples are then reconstructed from the pairwise distance matrix. However, it is not yet clear yet what degree of deepness and sequencing data quality is needed for a reliable phylogeny reconstruction. Direct analysis of unassembled genomic data has the potential to greatly increase the power of short read DNA sequencing technologies and allow comparative genomics of organisms without a completed reference available.

This paper has a two-fold aim. First, the validation of the AAF method using a well-known case study (i.e. Wu 2015), in order to establish the limits and conditions in which the method produces reliable results. Second, the application of this method to establish the phylogenomic relations for as many tomato chloroplasts as possible, whose sequences are currently available in genomic data banks.

In this study, we applied this AAF method to short sequence reads from a set of more than 40 wild and cultivated tomato species, taking advantage of the 360 genomes sequenced by Lin et al. (2014). The wild tomatoes present an excellent case study given the availability of genomic data sequences, and extensive analyses of morphology taxonomy (Peralta and Spooner 2005; Peralta et al. 2008) with different phylogenetic relationship methods such as plastid markers, low-copy nuclear markers, nuclear ribosomal ITS, and amplified fragment length polymorphisms (AFLP) (Peralta et al. 2008; Grandillo et al. 2011).

Four informal groups are accepted within the section Lycopersicon: (i) Lycopersicon group, the red and orange-fruited species clade which includes *Solanum lycopersicum, Solanum cheesmaniae, Solanum galapanse*, and *Solanum pimpinellifolium*. The taxa below the species level, most notably the small-fruited tomato known as *Solanum lycopersicum* var. *cerasiforme*, has been used to refer to putatively wild forms of *S. lycopersicum* that have been regarded as progenitors of the cultivated tomato. It is impossible to distinguish wild forms from cultivated forms or revertants from cultivation or possibly hybrids of wild and weedy taxa (Peralta et al. 2008). (ii) Arcanum group, the green fruit clade, with *Solanum arcanum*, *Solanum chmielewskii*, and *Solanum neorickii*. (iii) Eriopersicon group with *Solanum huaylasense*, *Solanum chilense*, *Solanum corneliomulleri*, *Solanum peruvianum* and *Solanum habrochaites*. (iv) Neolycopersicon group containing only *Solanum pennellii*, which was considered to be sister to the rest of the section based on its lack of the sterile anther that occurs as a morphological synapomorphy in *S. habrochaites* and the rest of the core tomatoes (Peralta et al. 2008). More recent studies using conserved orthologous sequence markers (COSII) (Rodriguez et al. 2009), genome-wide single nucleotide polymorphisms (SNPs) (Aflitos et al. 2014; Lin et al. 2014), and genomic repeat elements (Dodsworth et al. 2016) have largely supported previous hypotheses with respect to major clades within the tomatoes, although individual species relationships are less clear cut for some taxa. Thus, given the general acceptance of this informal classification, in the present study, we will often use it as reference in order to better clarify the results here presented.

Instead of dealing with data from all three organelles (chloroplast, mitochondrion, and nucleus), we concentrate on sequence data from chloroplast only. Chloroplast (cp) DNA sequences are a useful tool for plant identification and determination of the phylogeny relationship among species (Kress and Erickson 2008; Lahaye et al. 2008). This technique for the identification of close relatives has the potential of gene discovery for crop improvement (Daniell et al. 2010). Different chloroplast loci have been used for calculating close and distant phylogenetic relationships between plants but, up to date, the effectiveness of the proposed combinations to be used as chloroplast barcodes for the plantae kingdom has resulted far less effective than those used for mitochondria in the animal kingdom (Hollingsworth et al. 2011; Li et al. 2012).

2. Materials and methods

2.1. Genomic data set

In this paper, the AAF method was implemented over chloroplast sequences in two different ways. In the first step, we applied it to simulated pair-end (PE) Illumina data for comparison purposes with a previous phylogenetics analysis obtained by means of the neighbour-joining method over whole chloroplast genomes (Wu 2015). In the second step, AAF method was applied over real PE Illumina data from 45 wild and cultivated tomato species listed in Table 1.

To generate the simulated sequences we downloaded 10 complete tomato chloroplast genomes (GenBank accession nos. KP117020-KP117027, NC_007898, and NC_024584) and two potatoes chloroplast genomes as outgroup (GenBank accession nos. NC_007943 and NC_008096). We used the GemSIM package (McElroy et al. 2012) to generate PE reads of 100 bp, reaching coverages of 5× and 1000× for the downloaded chloroplast genomes. These reads have associated insert sizes of 500 bp, with 60 bp standard deviation, with a standard sequencing error model. The simulated sequences are indicated in Table 1 by with an (*). This selection of controlled data sets allows us to establish comparisons between the procedure presently proposed and previously published phylogenetic analysis of reference (Wu 2015).

For the second step, we used PE reads (Illumina Inc.) from Lin et al. (2014). These data sets are publicly available in the NCBI Short Read Archive (SRA) database. This series is the result of single run sequencing (Illumina HiSeq 2000) of 360 wild and cultivated tomato species (Lin et al. 2014). From this set, we have selected those sequences that present the highest overall coverage ratio and depth for each variety. Thus, a new subgroup of 45 tomatoes was selected, composed of 24 *S. lycopersicum*, 6 *S. lycopersicum* var. *cerasiforme*, 5 *S. pimpinellifolium*, completed with 3 *S. cheismaniae*, 3 *S. peruvianum*, 1 *S. chilense*, 1 *S. neorikii*, 1 *S. galapagense*, and 1 *S. habrochiates*. The accession numbers for the selected tomatoes are listed in Table 2. This table also includes the tomato species used by Wu (2015). As in Table 1, the (*) indicates that real PE reads are not available and simulated PE data was used, whilst (**) indicate the chloroplast genomes assembled by Wu (2015), and (***) indicates accessions with PI CGN code.

As we apply the AAF method over chloroplast sequences only, and not over the complete run, we need a preliminary processing of the sequence data sets. To select the reads of interest we map each sequence data set against the complete chloroplast genome of *S. Lycopersicum* LA3023 (accession no. NC_007898) using Bowtie2 software (Langmead and Salzberg 2012). All PE reads that align concordantly at least once to the reference above, with a maximum PE fragment alignment length of 500 bp, were used. The average coverage of the chloroplast sequences aligned is shown in Table 2.

Finally, we reduce these processed sequences to the minimum chloroplast coverage present in the selected samples (800×), in order to obtain comparable data.

We employed SPLITSTREE4 (Huson and Bryant 2006) to create a filtered supernetwork from 10,000 bootstrap trees produced by maximum parsimony analysis, with filtering set at 10% of all the input trees in all the analysis, except in the final supernetwork shown in Figure 8 that we used 10% of 5000 non-parametric bootstrap trees.

Table 1. Informal taxonomy groups within the section *Lycopersicon*.

Informal taxonomy group	Botanical variety	Number of taxa
Lycopersicon	*S. lycopersicum*	24
	S. pimpinellifolium	5
	S. lycopersicum var. *cerasiforme*	6
	S. cheesmaniae	3
	S. galapagense	1
Neolicopersicon	*S. pennellii**	1
Arcanum	*S. neorickii*	1
Eriopersicon	*S. habrochaites*	1
	S. chilense	1
	S. peruvianum	3
Outgroup	*S. tuberosum**	1
	S. bulbocastanum *	1

2.2. Assembly and alignment-free method

The AAF method used here is based on counting all possible *k*-mers, for each set of genomic data. A *k*-mer is a substring of nucleotides A, C, T, and G of length *k*. As the number of *k*-mers counted depends on the sequencing coverage and the distribution of the reads on the genome, this frequency table is converted to a table of presence/absence of *k*-mers among taxa. Then, the phylogenetic distance D between two species is estimated using the metric (Fan et al. 2015):

$$D = -1/k \cdot n_s/n_t \quad (1)$$

where n_s is the number of *k*-mers that are shared between taxa, and n_t is the total number of *k*-mers (Fan et al. 2015). The number of occurrence for each *k*-mer within the reads sequence data is n_r, a threshold θ for the number of repeats can be set to remove most random errors in the reads. When this filtering is set on, a *k*-mer is only recorded as present if it occurs more θ times in the same species.

Before computing distances, it is mandatory to choose an adequate frequency threshold and optimal length for the *k*-mers to be used in the statistics above. On one hand, if a *k*-mer covers, for example, multiple substitutions, it will count equally as one carrying only a single substitution. Consequently, shorter *k*-mers are more likely to have greater sensitivity to single evolutionary events. On the other hand, identical *k*-mers could be derived from physically, functionally, or evolutionary different regions of the genome and are therefore not homologous (*k*-mer homoplasy). Longer *k*-mers are less likely to suffer from *k*-mer homoplasy (Fan et al. 2015). Thus, the selection of *k*-mer length is a tradeoff between the problem of sensitivity (which requires a smaller *k*) and *k*-mer homoplasy (which requires a larger *k*).

2.2.1. Optimal k-mer length

We compute the frequency distribution for *k*-mer occurrences using the simulated Illumina read sequences of 12 cp genomes for low and high coverage (5× and 1000×). In Figure 1, we show the frequency distribution for *S. lycopersicum* LA3023 as an illustrative example. For low coverage, a short *k*-mer, such as 7 nucleotides, is incapable of differentiating the first peak corresponding to singletons, due mostly to sequencing errors, and the second peak of sound data (Figure

Table 2. Summary of the sampled collection of tomato.

Individual code	TGRC code	Botanical variety	SRA accession number	Coverage chloroplast	Coverage whole genome	Informal group
TS-420	LA2184	*S. pimpinellifolium*	SRR1572276	963.72	5.8	Lycopersicon
TS-267	LA2660	*S. pimpinellifolium*	SRR1572259-60-61	6084.02	18.1	Lycopersicon
TS-433	–	*S. pimpinellifolium*	SRR1572285-86	2442.7	5.33	Lycopersicon
TS-432	–	*S. pimpinellifolium*	SRR1572283-84	3272.3	5.4	Lycopersicon
TS-415**	LA1596	*S. pimpinellifolium*	SRR1572271	2615.59	7.7	Lycopersicon
TS-299	LA2131	*S. lycopersicum* var. *cerasiforme*	SRR1572435	836	5.5	Lycopersicon
TS-72	–	*S. lycopersicum* var. *cerasiforme*	SRR1572344	2655.29	5.8	Lycopersicon
TS-91	–	*S. lycopersicum* var. *cerasiforme*	SRR1572349-50	1822.36	6.9	Lycopersicon
TS-105	–	*S. lycopersicum* var. *cerasiforme*	SRR1572361	6178.2	4.8	Lycopersicon
TS-131	LA1162	*S. lycopersicum* var. *cerasiforme*	SRR1572373	1352.44	5.5	Lycopersicon
TS-129	LA2845	*S. lycopersicum* var. *cerasiforme*	SRR1572372	1577.4	5.9	Lycopersicon
TS-44	–	*S. lycopersicum*	SRR1572467	4024.4	6.95	Lycopersicon
TS-100	–	*S. lycopersicum*	SRR1572499	4864.82	9.16	Lycopersicon
TS-132	LA3903	*S. lycopersicum*	SRR1572527	1000.69	6.19	Lycopersicon
TS-135	LA0466	*S. lycopersicum*	SRR1572530	4266.05	6.65	Lycopersicon
TS-137	–	*S. lycopersicum*	SRR1572532	4050.88	5.44	Lycopersicon
TS-152	LA1021	*S. lycopersicum*	SRR1572545	1047.16	5.8	Lycopersicon
TS-168	–	*S. lycopersicum*	SRR1572559	4959.9	4.8	Lycopersicon
TS-172	–	*S. lycopersicum*	SRR1572564	3562.4	5.89	Lycopersicon
TS-178	–	*S. lycopersicum*	SRR1572570	3899.5	3.56	Lycopersicon
TS-184	LA2283	*S. lycopersicum*	SRR1572575	900.1	4.8	Lycopersicon
TS-190	–	*S. lycopersicum*	SRR1572582	3907.68	5.68	Lycopersicon
TS-237	LA3243	*S. lycopersicum*	SRR1572619	4451.5	4.03	Lycopersicon
TS-249	LA1462	*S. lycopersicum*	SRR1572626	1112.88	6.06	Lycopersicon
TS-251	–	*S. lycopersicum*	SRR1572627	3038.3	5.2	Lycopersicon
TS-253	LA4345	*S. lycopersicum*	SRR1572628	1512.2	4.5	Lycopersicon
TS-256	LA2260	*S. lycopersicum*	SRR1572630	3782.76	7.44	Lycopersicon
TS-282	–	*S. lycopersicum*	SRR1572654	1903.15	6.1	Lycopersicon
TS-321**	–	*S. lycopersicum*	SRR1572684	4523.09	8.5	Lycopersicon
TS-409	–	*S. lycopersicum*	SRR1572666	5262.07	7.89	Lycopersicon
TS-242	LA0134C	*S. lycopersicum*	SRR1572623	885.9	5.35	Lycopersicon
TS-191	–	*S. lycopersicum*	SRR1572583	2865.8	6.1	Lycopersicon
TS-192	–	*S. lycopersicum*	SRR1572584	3321.36	5.8	Lycopersicon
TS-193	–	*S. lycopersicum*	SRR1572585	2613.36	5.6	Lycopersicon
TS-203	–	*S. lycopersicum*	SRR1572594	2599.2	5.2	Lycopersicon
TS-408**	LA1969	*S. chilense*	SRR1572696	5308.99	3.23	Eriopersicon
TS-407**	–	*S. habrochaites*	SRR1572697	1133.56	2.57	Eriopersicon
TS-404**	–	*S. peruvianum*	SRR1572695	3556.01	3.17	Eriopersicon
TS-403	PI 128650***	*S. peruvianum*	SRR1572694	1342.74	2.83	Eriopersicon
TS-402	–	*S. peruvianum*	SRR1572692-93	1113.13	5.88	Arcanum
TS-146**	LA2133	*S. neorickii*	SRR1572685	2924.97	3.46	Arcanum
TS-208**	LA0528	*S. galapagense*	SRR1572686	1791.71	2.26	Lycopersicon
TS-199**	LA0746	*S. cheesmaniae*	SRR1572688	3981	3.29	Lycopersicon
TS-207	LA1037	*S. cheesmaniae*	SRR1572689	4171.3	3.1	Lycopersicon
TS-217	LA0429	*S. cheesmaniae*	SRR1572690-91	2045.5	2.44	Lycopersicon

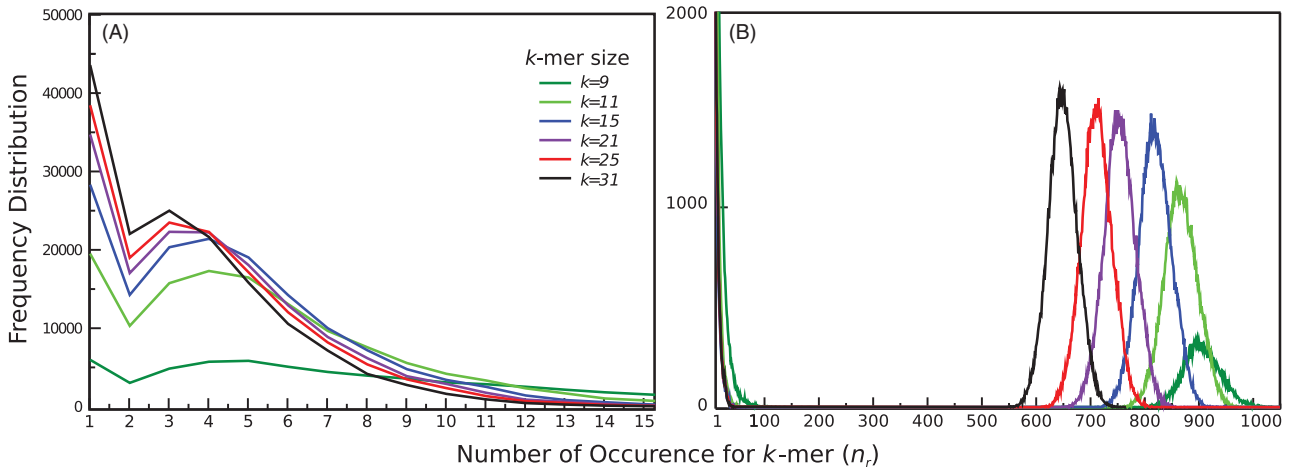

Figure 1. Frequency distribution of *S. lycopersicum* LA3023 for different *k*-mers and coverages, (A) 5× and (B) 1000×.

1(A)). This limitation is gradually overcome as the *k*-mer length increases. Namely, two distinct peaks appear as from *k*-mer of length 9. Although for low coverage, these two peaks are always overlapped, the height and position of the second peak becomes optimal for $k = 25$. For the high coverage case (Figure 1(B)), the first peak (error prone) is completely isolated from the second one for *k*-mers with length greater than 9. The area under the second peak grows with

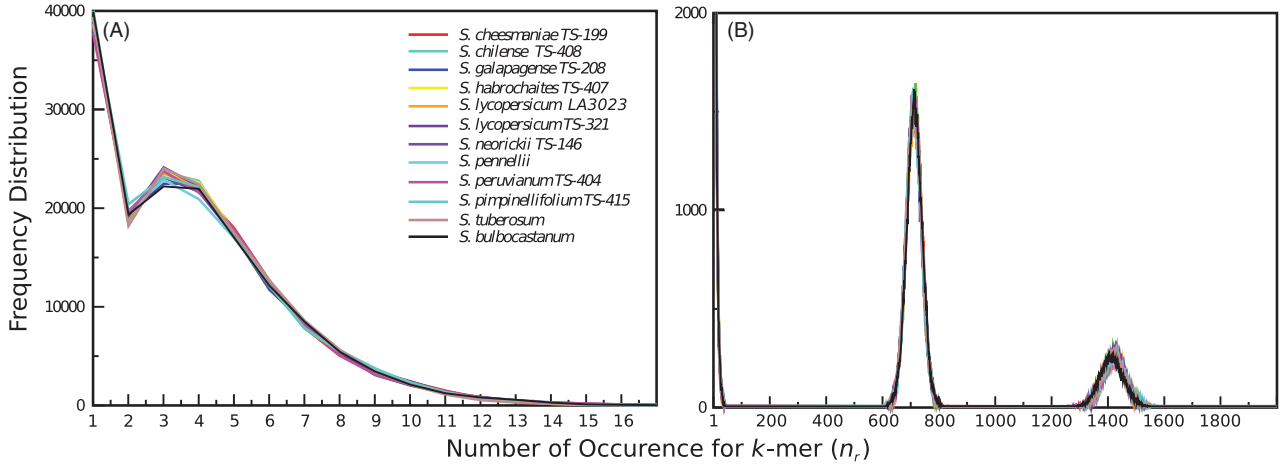

Figure 2. Frequency distribution for diferent chloroplast simulated sequences for *k*-mer 25 and coverages, (A) 5× and (B) 1000×.

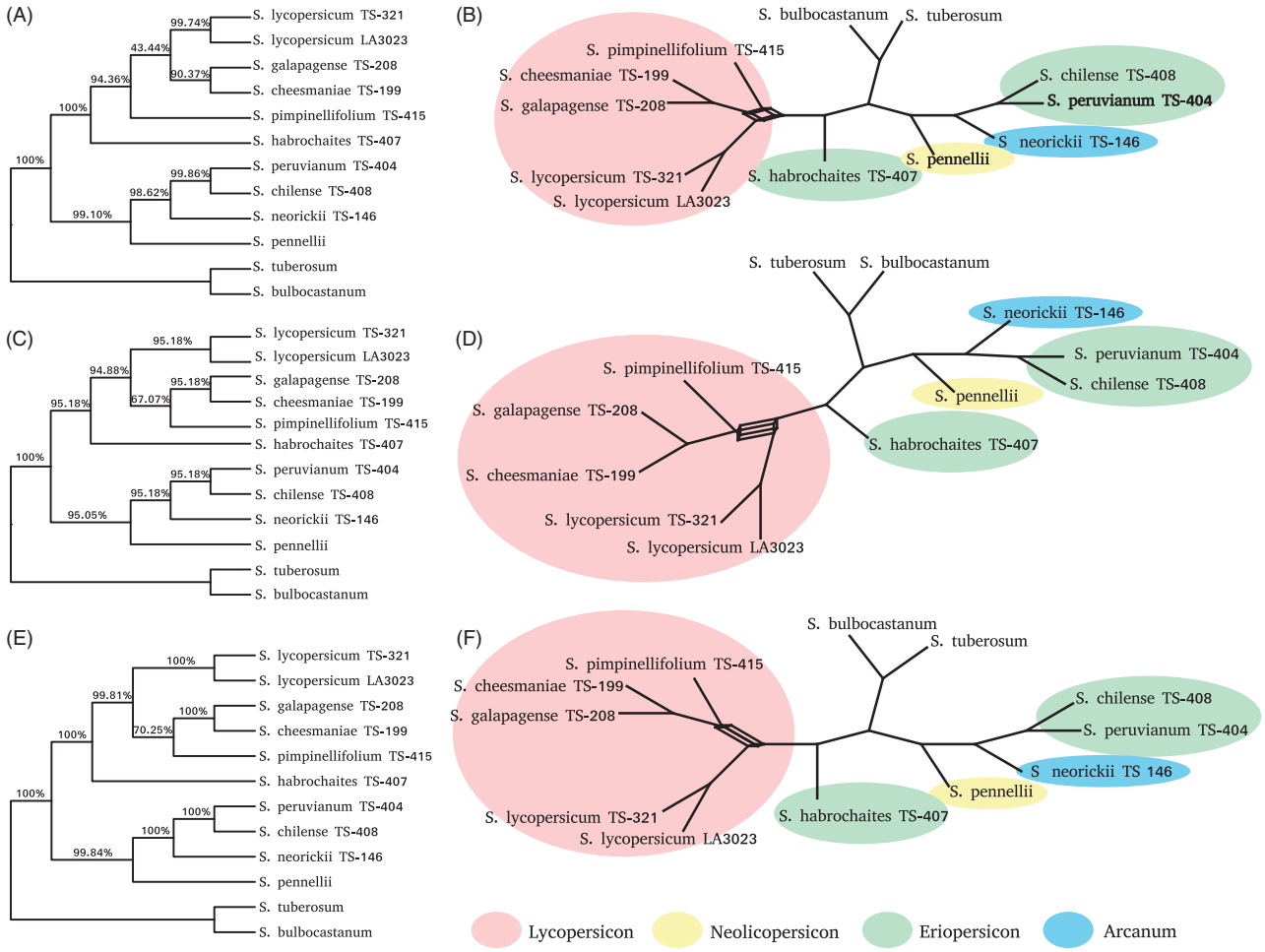

Figure 3. First case study phylogenomic trees and filtered supernetwork. (A) Phylogenomic tree of whole chloroplast genome sequences simulated with a coverage of 5×, *k*-mer length of 25 ($k = 25$) and filter singletons ($\theta = 2$). (B) Network showed as a filtered supernetwork of whole chloroplast genome sequences simulated with a coverage of 5×, $k = 25$ and $\theta = 2$. Splits present in 10% of all the bootstrap tree are displayed. (C) Phylogenomic tree of whole chloroplast genome sequences simulated with a coverage of 1000×, $k = 25$ and a filter of $\theta = 550$. (D) Network showed as a filtered supernetwork of whole chloroplast genome sequences simulated with a coverage of 1000×, $k = 25$ and $\theta = 550$. Splits presents in 10% of all the bootstrap tree are displayed. (E) Phylogenomic tree of whole chloroplast genome sequences simulated with a coverage of 1000×, $k = 25$ and $\theta = 100$. (F) Network showed as a filtered supernetwork of whole chloroplast genome sequences simulated with a coverage of 1000×, $k = 25$ and $\theta = 100$. Splits present in 10% of all the bootstrap trees are displayed. Numbers above the branches of the cladograms are the bootstrap values.

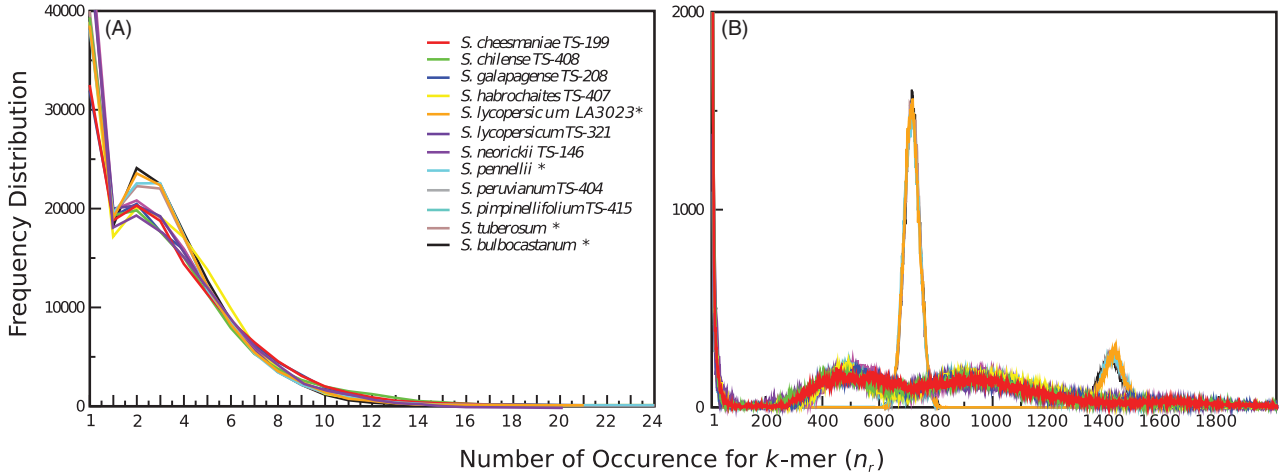

Figure 4. Frequency distribution of simulated genomes data (*) and real sequences data, *k*-mer 25 and coverages (A) 5× and (B) 1000×.

the *k*-mer length, reaching an optimal value for *k*-mer 25. Thus, 25 were selected as the optimal *k*-mer length for the subsequent analysis.

2.2.2. Low-frequency k-mers filter out

The possible errors introduced by lack of alignment are related with the inference of the actual evolutionary relationships among species. Additionally, the lack of assembly mainly generates sampling errors caused by low genome coverage and sequencing errors (Song et al. 2013; Yi and Jin 2013). Some studies have proposed to filter out all *k*-mers which frequencies are below a given threshold θ. For example by removing *k*-mers that present less than three copies ($\theta = 3$) can reduce the impact of the sequencing errors (Fan et al. 2015). However, as sequencing coverage decreases, a larger fraction of real *k*-mers will be singletons in the dataset, and, therefore, filtering will remove real *k*-mers. As a consequence, although filtering will be beneficial at high coverage, at low-coverage filtering will become detrimental.

Filtering out singletons can correct the sequencing error effect with low coverage (between 5× and 8×), according to genome size (Fan et al. 2015). In Figure 2(A), 5× coverage Illumina sequencing simulation of the 12 cp genomes and with a *k*-mer length of 25, two peaks may be observed. The first one corresponding to singletons (naturally expected sequencing errors) and that of $n_r = 3$, which reasonably mostly corresponds to correct genome sequence information. Therefore, in this case the threshold value was set to 2. Figure 2(B) with a 1000× coverage shows three distinct peaks. The first one, which becomes extinguished well under $n_r = 100$, represents the error prone sequencing data. A second peak, corresponding to data which exists as a single copy within chloroplast DNA, which shows a maximum around roughly $n_r = 700$, and a third peak which corresponds to the inverted repeat chloroplast DNA zone (IRa and IRb regions), which shows a maximum at frequencies around and above values of $n_r = 1400$. High coverage conditions enable for complete resolution between the error prone and sound data peaks. Therefore, these conditions are expected to offer more sensitive results.

3. Results

As stated previously, prior to the phylogenomic analysis of the 45 tomato accessions, we conduct a study over a subset of 10 tomatoes, whose chloroplast genome sequences have already been assembled. This subset offers the opportunity to optimize parameters of the AAF method and contrast the resulting cladogram with the one previously published by Wu, using whole chloroplast genome comparison (Wu 2015). For straightforward analysis and interpretation of the results, we use an informal, although generally accepted, taxonomy classification by Peralta et al. (2008) which is summarized in Table 1.

3.1. Phylogenomics of real and simulated reads from 12 chloroplast sequencing data

We used AAF method to calculate the cladogram and supernetwork of the 10 tomatoes and two potatoes studied by Wu (2015), in two different ways. In the first case, we used exclusively simulated Illumina sequencing data, produced from the corresponding assembled 12 chloroplast sequences. In the second case study, we used the real sequencing data, when available. Thus, the second case was finally composed of eight real sequencing data (Lin et al. 2014) as well as four simulated data.

3.1.1. First case study

All-simulated chloroplast sequencing data: Figure 3 shows the most parsimonious tree from our analysis of simulated sequencing data from 12 cp with AAF method, using *k*-mer length of 25 (see Section 2.2.1). Figure 3(A) is the result for low coverage (5×) and the filtering out of singletons ($\theta = 1$) (Section 2.2.2). The results for high coverage (1000×) calculated with $\theta = 550$ are shown in Figure 3(C), and with $\theta = 100$ in Figure 3(E) (Section 2.2.2).

Two members of the Eriopersicon group are recovered within the same clade (*S. peruvianum* and *S. chilense*), with high branch support (>95%) in the three cladograms. The third member, *S. habrochaites*, is separated from this group as is observed by Wu (2015). The Arcanum group (*S. neorickii*) is recovered as sister of the main members of the Eriopersicon

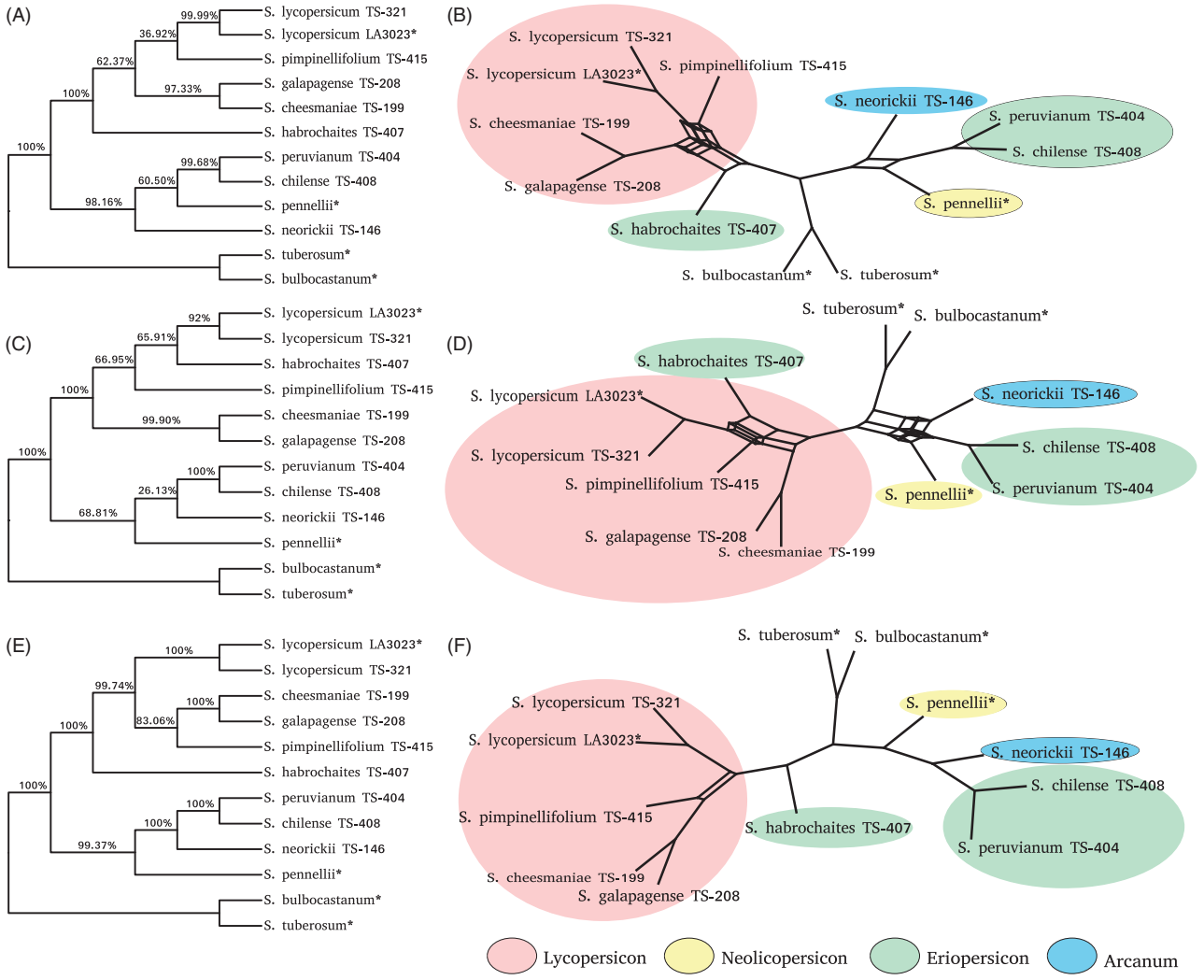

Figure 5. Second case study phylogenomic trees and filtered supernetwork. (A) Phylogenomic tree of whole chloroplast genome Illumina and simulated sequences, both with a coverage of 5×, k-mer length of 25 and filter singletons ($\theta = 2$). (B) Network showed as a filtered supernetwork of cp genome Illumina and simulated sequences with a coverage of 5×, $k = 25$ and $\theta = 2$. Splits presents in 10% of all the bootstrap tree are displayed. (C) Phylogenomic tree of chloroplast genome illumina and simulated sequences, with a coverage of 1000X, $k = 25$ and filter of $\theta = 550$. (D) Network showed as a filtered supernetwork of whole chloroplast genome Illumina and simulated sequences with a coverage of 1000×, $k = 25$ and filter of $\theta = 550$. Splits presents in 10% of all the bootstrap tree are displayed. (E) Phylogenomic tree of whole chloroplast genome Illumina and simulated sequences, with a coverage of 1000X, $k = 25$ and $\theta = 100$. (F) Network showed as a filtered supernetwork of whole chloroplast genome Illumina and simulated sequences with a coverage of 1000×, $k = 25$ and $\theta = 100$. Splits presents in 10% of all the bootstrap tree are displayed. Number above the branches of the cladograms are the bootstrap values.

group, with a bootstrap support of over 95%. The Neolicopersicon group, conformed only by *S. pennellii*, is sister to the Arcanum and Eriopersicon groups with 95% support.

AAF method recovers the red-orange fruited clade, the Lycopersicon group, with *S. lycopersicum*, *S. pimpinellifolium, S. galapagense*, *and S. cheesmaniae* in a strong support (>94%) for all the trees. In Figure 3(A), we observed that *S. pimpinellifolium* is sister to this group, whilst in Figure 3(C,E), *S. pimpinellifolium* appears as sister to the group formed by *S. galapagense* and *S. cheesmaniae*. This last result comes as the sole difference between the present results and those obtained by Wu (2015), although they are in complete accordance with the reference chloroplast phylogenomic results originally published by Palmer and Zamir (1982), which, differing with the results from Wu, establish the same phylogenomic relations for *S. pimpinellifolium* as those presented in Figure 3(C,E). Nevertheless, this discrepancy between Palmer and Wu results minor and may be accounted for when taking into consideration the results shown by the corresponding supernetworks for all three conditions. Namely, that the three supernetworks in Figure 3(B,D,F) show the same overall topology, with evidence of a common reticulation node in the Lycopersicon group clade. When comparing Figure 3(D,F), it comes apparent that both study cases recover the same supernetwork topology. Thus, for the case of simulated data, establishing a cutoff immediately after the first peak ($\theta = 100$) or immediately before the second peak ($\theta = 550$) results equivalent. This may be readily explained by the fact that, for the case of simulated data, the distance that separates the first and second peaks carries literally no *k*-mer data, either sound or error prone.

The results corresponding to both high and low coverages are in great correspondence with the tomato chloroplast phylogeny obtained by Wu (2015) and we do not observe differences regarding the supernetworks obtained using different θ values.

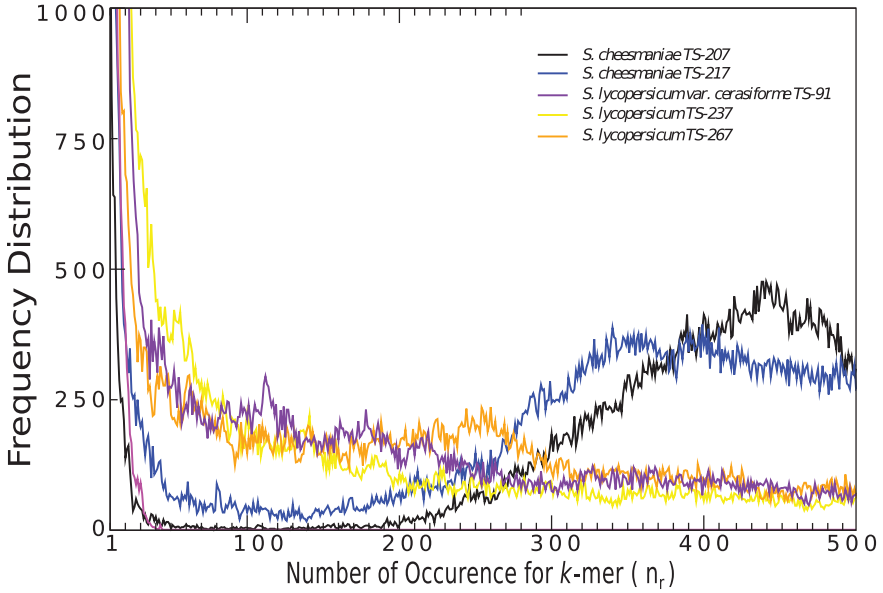

Figure 6. Frequency distribution of five real genomes, that shows the comparison of resolved (*S. cheesmaniae* TS-207) and non-resolved (rest) error prone from sound data. All datasets correspond to 800× coverage and *k*-mer 25.

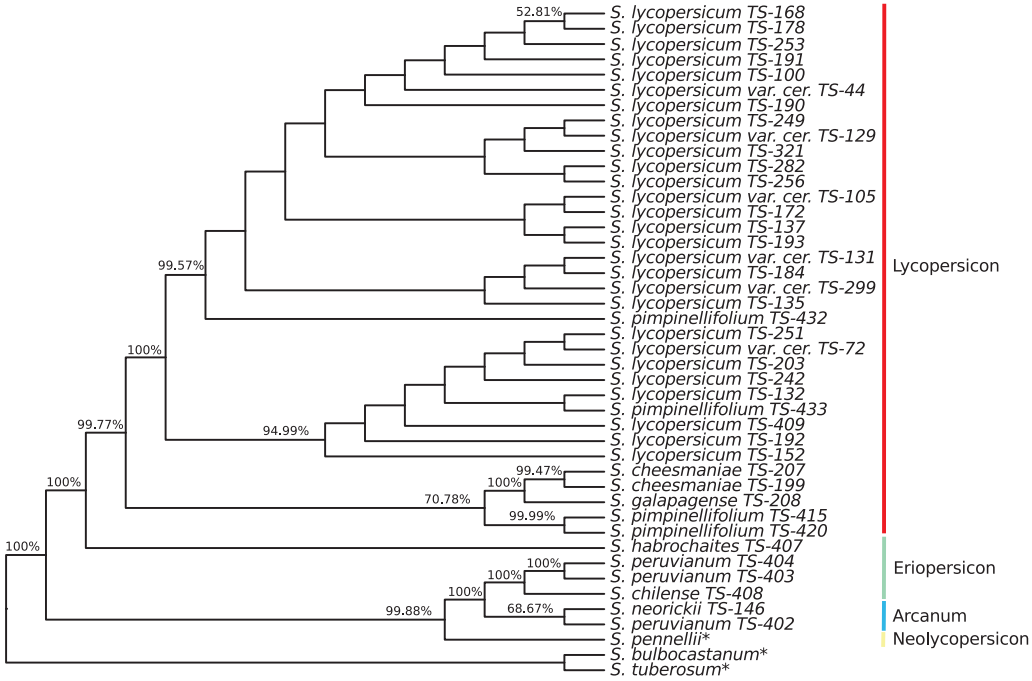

Figure 7. Phylogenomics relationships in 42 Solanum, section Lycopersicon, calculated with AAF method. Bootstrap values higher than 50% are shown (non-parametric bootstrap with 10,000 resampling of each total *k*-mer table).

3.1.2. Second case study

Eight real and four simulated chloroplast sequencing data: When dealing with real read sequences, certain constraints appear which must be taken into account in order to adequately choose the filtering parameter (θ). Figure 4 illustrates the frequency distribution of 4 simulated and 8 real sequences obtained for coverages of 5× (Figure 4(A)) and 1000× (Figure 4(B)), with *k*-mer length of 25. Although the simulated data in Figure 4(B) is the same as in the previous case, in the case of the real data one can observe that the second and third peaks (corresponding to the sound data) are broader an their maximum is shifted towards lower values. For comparison, we use the same three filtering conditions used in the case of all simulated data, i.e. $\theta = 2$ for 5× coverage, and $\theta = 550$ and $\theta = 100$ for the 1000× coverage case.

The trees and supernetworks calculated with real and simulated Illumina sequences shown in Figure 5 recover the Lycopersicon group (red-orange fruited clade), with *S. lycopersicum*, *S. pimpinellifolium, S. galapagense,* and *S. cheesmaniae* in Figure 5(A) with a 62% bootstrap and (E) with a strong support of 99.74%. In this group *S. pimpinellifolium* TS-415 in (A) is sister to *S. lycopersicum* LA3023 and *S. lycopersicum*

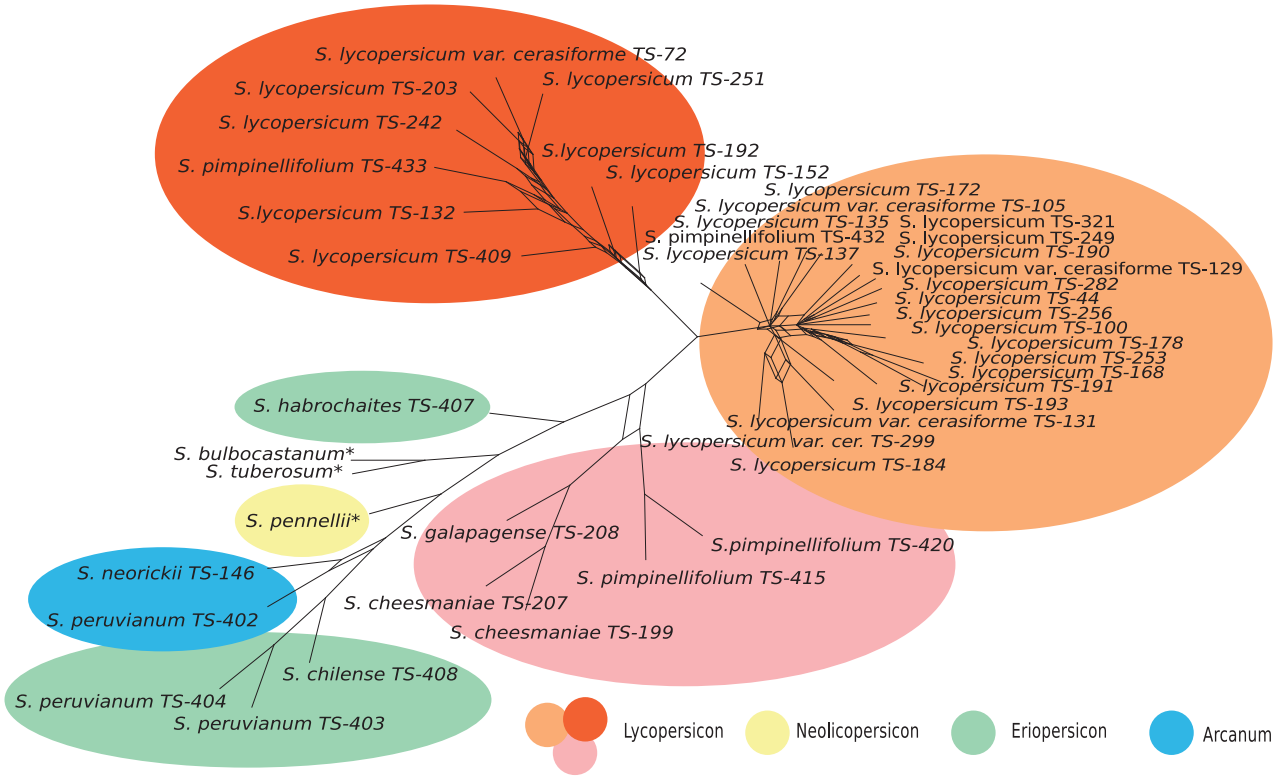

Figure 8. Relationship in chloroplast genomes shown as a filtered supernetwork. Splits present in 10% of 5000 nonparametric bootstrap trees are displayed. Four clusters of conict appear in the supernetwork. The first tree correspond one to each Lycopersicon subgroup clade, whilst the fourth appears between the Arcanum and the Eriopersicon.

TS-321, and in (C) *S. pimpinellifolium* TS-415 is sister to the *S. galapagense* TS-208 and *S. cheesmaniae* TS-199 group. In the tree shown in Figure 5(C), the Lycopersicon group is nested together with *S. habrochaites* TS-407 from the Neolicopersicon group. The Eriopersicom group, with *S. peruvianum* TS-404 and *S. chilense* TS-408, is recovered with a high branch support (>99.68%) in the three cladograms. The *S. neorickii* TS-146 (Arcanum group) is recovered as sister of the Eriopersicon group in all the trees, although with different level of support. In (A) and (E), bootstrap is over 98.16%, whilst in (C), it presents a modest value of 26.13%.

As regard the supernetworks, (F) is the most resolved net; notwithstanding a non-resolved node in the Lycopersicon group clade, which also appears in all three cases of the previous case study. Opposingly, (B) presents two non-resolved nodes (points), whilst (D) presents a clear case of errors due to the loss of sound data, secondary to the use of an incorrect θ value.

In conclusion, as regard real chloroplast data, AAF method requires of high coverage and viable *k*-mer data in the range of $k = 25$, as well as the use of an adequate cutoff value for low-frequency (error prone) *k*-mers. In the present study this comes as no real limitation, due to the fact that every genomic study here utilized complies with such requirements.

3.2. Phylogenomic study of 41 real and three simulated chloroplast sequencing data

We used AAF method to calculate the cladogram and supernetwork of 42 tomatoes and two potatoes as an outgroup. For a preliminary analysis, we considered 45 real sequencing data candidates from the 360 tomato consortium recently published data as described in Section 2.

First, we examine the quality of all data by means of the *k*-mer frequency distribution. In this sense, we discard those data sets whose associated distribution show that the error prone and sound data peaks are overlapped. Figure 6 depicts frequency distribution of four data sets (*S. cheesmaniae* TS-217, *S. lycopersicum* TS-237, *S. lycopersicum* TS-267, and *S. lycopersicum* var. *cerasiforme* TS-91) with error prone data, compared with the sound sequence of *S. cheesmaniae* TS-207. Thus, following the above criteria, these were discarded for the analysis. Thus, we finally selected 41 real sequencing data candidates from the 360 tomato consortium recently published data (Lin et al. 2014), as well as one simulated tomato sequencing data. Likewise, two simulated potato sequencing data were selected as an outgroup.

The single most parsimonious chloroplast tree from this analysis (Figure 7) recovers the complete Lycopersicon group with *S. lycopersicum*, *S. pimpinellifolium*, *S. galapagense,* and *S. cheesmaniae*, with a high branch support of 99.77% bootstrap. In this group, we can distinguish three sub-groups. The first one, with a 99.57% bootstrap support, and sister to *S. pimpinellifolium* TS-432, is composed by two thirds of the 22 *S. lycopersicum* and all but one of the *S. lycopersicum* var. *cerasiforme* studied. The second subgroup contains *S. pimpinellifolium* TS-433, the remaining third of the *S. lycopersicum* studied as well as the last *S. lycopersicum* var. *cerasiforme*. The third subgroup, with a 70.78% bootstrap support, collects both *S. cheesmaniae* studied, *S. galapagense* and the two remaining *S. pimpinellifolium* studied, TS-415 and TS-420.

Both *S. cheesmaniae* TS-207 and *S. cheesmaniae* TS-199 appear together with a 99.47% bootstrap value, and are sisters to *S. galapagense* with a 100% bootstrap support. All three subgroups are related to at least one of the four *S. pimpinellifolium* varieties studied. This last result is in concordance with the variability of the phylogenetic relationships for *S. pimpinellifollium* available in the present literature (Peralta et al. 2008), and comes in reasonable accordance with the chloroplast phylogenomics results presented both by Wu (2015) and Palmer and Zamir (1982).

With 98.88% support, the Neolicopersicon group, composed only by *S. pennellii*, the Arcanum group (*S. neorickii* TS-146, *S. peruvianum* TS-402), and the Eriopersicon group (*S. chilense* TS-408, *S. peruvianum* TS-404, *and S. peruvianum* TS-403) are claded together. Additionally, *S. habrochaites* TS-407 appears as sister to the Lycopersicon group with a 100% bootstrap support. This is in complete correspondence with Wu (2015) as well as with our previous validation study cases. The Eriopersicon group, with *S. peruvianum* TS-404, *S. peruvianum* TS-403, and *S. chilense* TS-408, is recovered with high branch support (100%). The *S. neorickii* TS-146 and *S. peruvianum* TS-404, which conform the Arcanum group, are recovered as sister to the Eriopersicon group with 100% bootstrap support. The *S. peruvianum* may be divided between North and South varieties, according to their intercrossing capabilities. Peralta and Spooner (2005) included *S. peruvianum* North within the Arcanum group and *S. peruvianum* South within the Eriopersicon group. This comes as no surprise given the fact that TS-404 and TS-403 correspond to the *S. peruvianum* South variety, whilst TS-402 corresponds to the *S. peruvianum* North variety (Jablonska et al. 2007).

In the filtered supernetwork (Figure 8), the three principal Lycopersicon subgroups are clearly separated, with the *S. pimpinellifolium* distributed among the three, and all four of them closely connected to the nodes separating these three subgroups, which consequently indicates a close relationship between them. Additionally, every connection point is also an indetermination cluster in the supernetwork, which could reasonably account for the ubiquity of *S. pimpinellifolium* in current literature (Peralta et al. 2008). Likewise, the supernetwork shows a clear separation between all groups considered; i.e. the three Lycopersicon subgroups, Neolicopersicon, Eriopersicon, and Arcanum groups. Finally, *S. habrochiates* appears connected to the supernetwork, between the third Lycopersicon subgroup and the potato outgroup, as described by Wu (2015).

4. Discussion

In the present study, we have tested the capabilities of the AAF method to establish reliable phylogenomic relationships for tomato chloroplasts. The method produces accurate results when applied to ideal sequencing data (i.e. simulated data), for both low- and high-coverage conditions. Nevertheless, when analyzing real sequencing datasets certain issues arise that need to be taken into consideration. First, high-coverage conditions produce better results than low coverage ones. Second, a certain degree of data curation is needed before the AAF method is applied. Namely, the *k*-mer frequency distribution histograms must be previously checked in order to verify complete resolution between the first and second peaks (i.e. the error prone and sound data peaks). Finally, an optimal cutoff value for θ, common to all datasets, must be correctly established in order to discard error prone data without the loss of sound data. Here, it has been established that the AAF method is able to correctly establish tomato chloroplast phylogenomic relationships, which opens the possibilities for further phylogenomic studies using more comprehensive raw genomic sequencing data.

Under the conditions previously established, we studied the phylogenetic relationships for 42 tomato chloroplast, using two potatoes chloroplasts as outgroup. We hereby obtained a general phylogenetic tree structure compatible with the data established by previous studies. Namely, that every member of the four informal groups presently studied cluster together, maintaining the expected relationships between different groups. Nevertheless, certain interesting observations may further established, such as the fact that the Lycopersicon group appears in three distinct sub-clusters, two of which account for all the *S. lycopersicum* and *S. lycopersicum* var. *cerasiforme* studied, whilst the third sub-cluster is composed by the *S. galapagense* and both *S. cheesmaniae* studied, as would be expected according to data previously published. Additionally, the four *S. pimpinellifolium* studied appear scattered across these three Lycopersicon sub-clusters, which may partially explain the lack of consensus in previous literature as to their precise phylogenetic relationship within the Lycopersicon informal group.

Nevertheless, it is important to bear in mind that in order to preserve methodological consistency, the present study has been restricted only to the taxons analyzed by the 360 genome consortium and that show sound SRA data. It could be expected that certain phylogenetic relationships may change as more taxons are further added in future studies using this method.

Finally, it is interesting to observe that *S. peruvianum* TS-403 lies within the Eriopersicon informal group, whilst *S. peruvianum* TS-402 lies within the Arcanum informal group, which can be readily accounted for when considering that the first corresponds to *S. peruvianum* (north) variety, whilst the second corresponds to *S. peruvianum* (south) variety (Jablonska et al. 2007).

Summarizing, we believe that the present study has established that the ability to perform reliable phylogenomic studies without the need for assembly or alignment of raw sequencing data is not only a great advantage but a real necessity when dealing with the ever-growing sequencing data produced worldwide.

Disclosure statement

No potential conflict of interest was reported by the authors.

Funding

This work was funded by the Agencia Nacional de Promoción Científica y Tecnológica [PICT-2012-1970].

References

Adachi J, Waddell PJ, Martin W, Hasegawa M. 2000. Plastid genome phylogeny and a model of amino acid substitution for proteins encoded by chloroplast DNA. J Mol Evol. 50:348–358.

Aflitos S, Schijlen E, De Jong H, De Ridder D, Smit S, Finkers R, Wang J, Zhang G, Li N, Mao L, et al. 2014. Exploring genetic variation in the tomato (*Solanum* section *Lycopersicon*) clade by whole-genome sequencing. Plant J. 80:136–148.

Chan CX, Ragan MA. 2013. Next-generation phylogenomics. Biol Direct. 8:3.

Chu KH, Qi J, Yu ZG, Anh VO. 2004. Origin and phylogeny of chloroplasts revealed by a simple correlation analysis of complete genomes. Mol Biol Evol. 21:200–206.

Daniell H, Kumar S, Dufourmantel N. 2010. Breakthrough in chloroplast genetic engineering of agronomically important crops. TRENDS Biotechnol. 23:238–245.

De Las Rivas J, Lozano JJ, Ortiz AR. 2002. Comparative analysis of chloroplast genomes: functional annotation, genome-based phylogeny, and deduced evolutionary patterns. Genome Res. 12:567–583.

Dodsworth S, Chase MW, Arkinen TS, Knapp S, Leitch AR. 2016. Using genomic repeats for phylogenomics: a case study in wild tomatoes (*Solanum* section *Lycopersicon: Solanaceae*). Biol J Linn Soc. 117:96–105.

Fan H, Ives AR, Surget-Groba Y, Cannon CH. 2015. An assembly and alignment-free method of phylogeny reconstruction from next-generation sequencing data. BMC Genomics. 16:522.

Fitz-Gibbon ST, House CH. 1999. Whole genome-based phylogenetic analysis of free-living microorganisms. Nucleic Acids Res. 27:4218–4222.

Grandillo S, Chetelat R, Knapp S, Spooner D, Peralta I, Cammareri M, Perez O, Termolino P, Tripodi P, Chiusano ML, et al. 2011. Solanum sect. Lycopersicon. In: Wild crop relatives: genomic and breeding resources. Berlin (Germany): Springer; p. 129–215.

Haubold B. 2014. Alignment-free phylogenetics and population genetics. Brief Bioinformatics. 15:407–418.

Hollingsworth PM, Forrest LL, Spouge JL, Hajibabaei M, Ratnasingham S, van der Bank M, Chase MW, Cowan RS, Erickson DL, Fazekas AJ, et al. 2009. A DNA barcode for land plants. Proc Natl Acad Sci. 106:12794–12797.

Hollingsworth PM, Graham SW, Little DP. 2011. Choosing and using a plant DNA barcode. PLoS One. 6:e19254.

Huson DH, Bryant D. 2006. Application of phylogenetic networks in evolutionary studies. Mol Biol Evol. 23:254–267.

Jablonska B, Ammiraju JS, Bhattarai KK, Mantelin S, de Ilarduya O, Roberts PA, Kaloshian I. 2007. The Mi-9 gene from Solanum arcanum conferring heat-stable resistance to root-knot nematodes is a homolog of Mi-1. Plant Physiol. 143:1044–1054.

Kress WJ, Erickson DL. 2008. DNA barcodes: genes, genomics, and bioinformatics. Proc Natl Acad Sci USA. 105:2761–2762.

Lahaye R, van der Bank M, Bogarin D, Warner J, Pupulin F, Gigot G, Maurin O, Duthoit S, Barraclough TG, Savolainen V. 2008. DNA barcoding the oras of biodiversity hotspots. Proc Natl Acad Sci USA. 105:2923–2928.

Langmead B, Salzberg SL. 2012. Fast gapped-read alignment with Bowtie 2. Nat Methods. 9:357–359.

Lemieux C, Otis C, Turmel M. 2000. Ancestral chloroplast genome in Mesostigma viride reveals an early branch of green plant evolution. Nature. 403:649–652.

Li CP, Han GS, Chu KH. 2012. Analyzing multi-locus plant barcoding datasets with a composition vector method based on adjustable weighted distance. PLoS One. 7:e42154.

Li D-Z, Gao L-M, Li H-T, Wang H, Ge X-J, Liu J-Q, Liu JQ, Chen ZD, Zhou SL, Chen SL, et al. 2011. From the cover: comparative analysis of a large dataset indicates that internal transcribed spacer (ITS) should be incorporated into the core barcode for seed plants. Proc Natl Acad Sci. 108:19641–19646.

Lin J, Gerstein M. 2000. Whole-genome trees based on the occurrence of folds and orthologs: implications for comparing genomes on different levels. Genome Res. 10:808–818.

Lin T, Zhu G, Zhang J, Xu X, Yu Q, Zheng Z, Zhang Z, Lun Y, Li S, Wang X, et al. 2014. Genomic analyses provide insights into the history of tomato breeding. Nat Genet. 46:1220–1226.

Martin W, Rujan T, Richly E, Hansen A, Cornelsen S, Lins T, Leister D, Stoebe B, Hasegawa M, Penny D. 2002. Evolutionary analysis of Arabidopsis, cyanobacterial, and chloroplast genomes reveals plastid phylogeny and thousands of cyanobacterial genes in the nucleus. Proc Natl Acad Sci. 99:12246–12251.

Martin W, Stoebe B, Goremykin V, Hansmann S, Hasegawa M, Kowallik KV. 1998. Gene transfer to the nucleus and the evolution of chloroplasts. Nature. 393:162–165.

McElroy KE, Luciani F, Thomas T. 2012. GemSIM: general, error-model based simulator of next-generation sequencing data. BMC Genomics. 13:74.

Palmer JD, Zamir D. 1982. Chloroplast DNA evolution and phylogenetic relationships *in Lycopersicon*. Proc Natl Acad Sci U S A. 79:5006–5010. America.

Peralta IE, Knapp S, Spooner DM. 2008. Taxonomy of wild tomatoes and their relatives. (*Solanum* sect. *Lycopersoides*, sect. *Juglandifolia*, sect. *Lycopersicon*; Solanaceae). Syst Bot Monographs. 84:1–186.

Peralta IE, Spooner DM. 2005. Morphological characterization and relationships of wild tomatoes (*Solanum L.* sect. *Lycopersicon*). Monographs Syst Bot. 104:227–257.

Qi W, Yu Z-G, Yang J. 2017. DLTree: a web server for phylogeny reconstruction using dynamical language method. Bioinformatics. 33:2214–2215.

Rodriguez F, Wu F, Ané C, Tanksley S, Spooner DM. 2009. Do potatoes and tomatoes have a single evolutionary history, and what proportion of the genome supports this history? BMC Evol Biol. 9:191.

Sankoff D, Leduc G, Antoine N, Paquin B, Lang BF, Cedergren R. 1992. Gene order comparisons for phylogenetic inference: evolution of the mitochondrial genome. Proc Natl Acad Sci. 89:6575–6579.

Small RL, Cronn RC, Wendel JF. 2004. L. A. S. JOHNSON REVIEW No. 2. Use of nuclear genes for phylogeny reconstruction in plants. Aust Syst Bot. 17:145–170.

Song K, Ren J, Zhai Z, Liu X, Deng M, Sun F. 2013. Alignment-free sequence comparison based on next-generation sequencing reads. J Comput Biol. 20:64–79.

Stiller JW. 2011. Experimental design and statistical rigor in phylogenomics of horizontal and endosymbiotic gene transfer. BMC Evol Biol. 11:259.

Stuart GW, Moffett K, Baker S. 2002. Integrated gene and species phylogenies from unaligned whole genome protein sequences. Bioinformatics. 18:100–108.

Tekaia F, Lazcano A, Dujon B. 1999. The genomic tree as revealed from whole proteome comparisons. Genome Res. 9:550–557.

Turmel M, Otis C, Lemieux C. 1999. The complete chloroplast DNA sequence of the green alga Nephroselmis olivacea: insights into the architecture of ancestral chloroplast genomes. Proc Natl Acad Sci. 96:10248–10253.

Turmel M, Otis C, Lemieux C. 2002. The chloroplast and mitochondrial genome sequences of the charophyte Chaetosphaeridium globosum: insights into the timing of the events that restructured organelle DNAs within the green algal lineage that led to land plants. Proc Natl Acad Sci. 99:11275–11280.

Wong KM, Suchard MA, Huelsenbeck JP. 2008. Alignment uncertainty and genomic analysis. Science. 319:473–476.

Wu Z. 2015. The completed eight chloroplast genomes of tomato from *Solanum* genus. Mitochondrial DNA. 27:1–3.

Yi H, Jin L. 2013. Co-phylog: an assembly-free phylogenomic approach for closely related organisms. Nucleic Acids Res. 41:e75.