Send Orders for Reprints to reprints@benthamscience.ae

Current Genomics, 2018, *19*, 491-498

RESEARCH ARTICLE

An Effective Pipeline Based on Relative Coverage for the Genome Assembly of Phytoplasmas and Other Fastidious Prokaryotes

Cesare Polano and Giuseppe Firrao*

Department of Agricultural, Food, Environmental and Animal Sciences, University of Udine, Udine, Italy

ARTICLE HISTORY

Received: September 26, 2017
Revised: February 01, 2018
Accepted: March 05, 2018

DOI: 10.2174/1389202919666180314114628

Abstract: ***Background*:** For the plant pathogenic phytoplasmas, as well as for several fastidious prokaryotes, axenic cultivation is extremely difficult or not possible yet; therefore, even with second generation sequencing methods, obtaining the sequence of their genomes is challenging due to host sequence contamination.

***Objective*:** With the *Phytoassembly* pipeline here presented, we aim to provide a method to obtain high quality genome drafts for the phytoplasmas and other uncultivable plant pathogens, by exploiting the coverage differential in the ILLUMINA sequences from the pathogen and the host, and using the sequencing of a healthy, isogenic plant as a filter.

***Validation*:** The pipeline has been benchmarked using simulated and real ILLUMINA runs from phytoplasmas whose genome is known, and it was then used to obtain high quality drafts for three new phytoplasma genomes.

***Conclusion*:** For phytoplasma infected samples containing >2-4% of pathogen DNA and an isogenic reference healthy sample, the resulting assemblies can be next to complete. The *Phytoassembly* source code is available on GitHub at https://github.com/cpolano/phytoassembly.

Keywords: ILLUMINA, *Candidatus* phytoplasma, NGS, Second generation sequencing, Endophytes, Genome draft.

1. INTRODUCTION

Phytoplasmas are bacterial plant pathogens that cause disease in over 100 plant families [1]; they belong to the class *Mollicutes,* bacteria characterized by the absence of a cell wall, and are typically about 200-300 nm in size, with a genome of 0.5-1.2·10^6 nts [2]. They live in the host phloem cells and propagate by vectors such as insects (mainly *Cicadellidae, Fulgoroidea* and *Psyllidae;* [3]) and parasitic plants [4].

Genomics of fastidious prokaryotes are made challenging by the fact that they are difficult to cultivate *in vitro* [5]. For the phytoplasmas, protocols typically involve time-consuming isolation and purification of DNA from plant or insect-infected tissue using CsCl equilibrium buoyant density gradient in the presence of bisbenzimide [6], or physical isolation by Pulsed-Field Gel Electrophoresis (PFGE) of entire chromosomes [7]. Currently, only for four phytoplasmas the genomes have been sequenced to completion: '*Ca.* Phytoplasma asteris' Onion Yellows phytoplasma strain M [7], '*Ca.* P. asteris' Aster Yellows phytoplasma strain Witches' Broom ph. [8], '*Ca.* P. mali' strain AT [9] and '*Ca.* P. australiense' strains Paa and SLY [10, 11].

*Address correspondence to this author at the Department of Agricultural, Food, Environmental and Animal Sciences, University of Udine, Udine, Italy; Tel: + 39 0432 558531; Fax: + 39 0432 558501;
E-mail: firrao@uniud.it

Genomic surveys have also been published for multiple phytoplasmas [12-15]. With the introduction of New Generation Sequencing (NGS) methods, an emerging alternative, made possible by informatics tools, is to randomly sequence a large library of DNA extracted from diseased plants and then select the sequences of the pathogen. However, the pathogen sequence selection is not trivial and therefore, many genome drafts obtained with this approach so far are incomplete [16-22].

Improvements in drafting genomes of phytoplasmas might be obtained by applying newly proposed approaches. A technique based on single-cell sequencing was proposed by Chitsaz and coworkers [23], however, single cell sequencing is more complex than standard sequencing, and there are technical challenges [24] that can affect the quality of the data. Another strategy could be the use of software tools for the reconstruction of organelle genomes; however, many of these tools [25, 26] require a reference or a seed sequence for the microorganism itself or make assumptions related to the structure and number of the genomes that cannot be made for the phytoplasmas.

The pipeline here presented, named *Phytoassembly*, is an evolution of the procedure described in [17] and exploits on one hand the differential in coverage of the sequences originating from the pathogen and the host, which allows to discard a significant part of the (under-represented) sequences from the plant, and on the other hand, the mapping of the

1389-2029/18 $58.00+.00

remaining reads on a healthy plant reference, which filters out the rest of the plant sequences.

2. MATERIALS AND METHODS

2.1. Design and Implementation of the Pipeline

The main strategy of the procedure presented here, consists in separating the plant sequences first by setting a cutoff point based on the differential coverage of the host and the pathogen contigs resulting from a pre-assembly. In samples collected from phytoplasma-infected plants, despite the prevalence of host DNA, the number of phytoplasma genome copies exceeds the number of host genome copies: phytoplasma genomes sizes range around 10^6 bp, while plant genomes are about 3 orders of magnitude larger [27]; therefore, when counting the reads in an ILLUMINA data-set obtained from a diseased plant sample containing *e.g.* 1% phytoplasma DNA, the coverage of phytoplasma DNA is expected to be 10 times greater than the coverage of the plant DNA. It is then possible to define a cutoff point of the contigs generated in a pre-assembly step (pre-contigs) so that the pre-contigs that belong to the pathogen and have a high reads coverage (per base) are retained, while the pre-contigs from the host, that have a low reads coverage are discarded. As in the case shown in Fig. (**1**), when the sample is obtained from a well-infected plant and it is, therefore, enriched in pathogen DNA, the definition of a cut-off point in the pre-contigs coverage graph that distinguishes the pre-contigs belonging to the pathogen (peaking on the right part of the graph) from those belonging to the host (on the left) is trivial. However, in many cases, when the pathogen DNA is scarce, there is an overlap between the phytoplasma and the host peaks, hence determining a convenient cutoff requires an estimation to ensure that all phytoplasma reads are retained during the selection. In the preliminary implementations of *Phytoassembly*, the definition of the optimal cutoff was achieved by carrying out quantitative PCRs to estimate the abundance of the pathogen DNA. It was found, however, that the qPCR analysis can be avoided as the information on pathogen abundance can be obtained from the ILLUMINA dataset. In the released version, *Phytoassembly* is structured to determine a convenient cutoff value without intervention from the user.

Thus, the first steps of the pipeline consist in a preassembly, the estimation of pre-contigs coverage and calculation of the cutoff value. Then the ILLUMINA reads belonging to contigs above the cutoff are selected and aligned against the healthy plant genome reference, so that those pertaining to the plant can be discarded and the non-plant reads can be assembled in preliminary phytoplasma assembly. Further polishing is carried out to filter out ambiguous contigs, originating from low-quality reads from the plant. This is based on the percentage of identity of BLAST matches against the healthy plant reference, the threshold being any match greater than 95%.

The standard procedure requires a reference genome from an uninfected plant in FASTA format and the sequence reads from an ILLUMINA MiSeq in FASTQ format. If necessary, the pipeline can also assemble reference genome reads in FASTQ format, and it is possible to also input the already assembled sequence reads in FASTA format. For best results, the healthy plant should be isogenic to, and grown in the same environment as the diseased specimen, so as to match the plant genome and include the same contaminants. The aforementioned BLAST verification becomes a necessity if the reference does not meet these qualities. On the other hand, it is possible to input a collection of reference genomes (simply by joining the relative FASTA files), *e.g.* to filter out known pathogens.

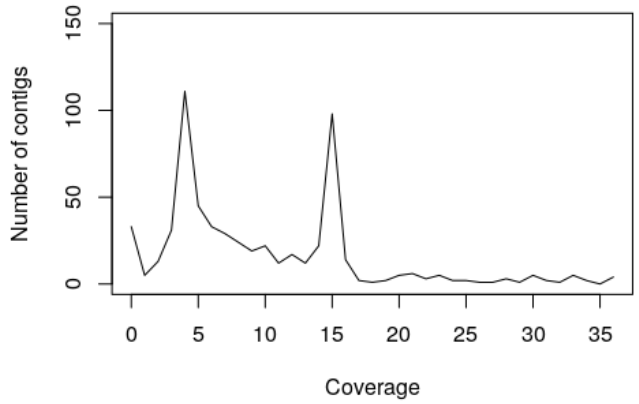

Fig. (1). Coverage graph of the artificial aster yellows phytoplasma strain witches' broom sample pre-assembly. The graph, from a dataset with 15% of phytoplasma reads, illustrates the position of the plant (left) and the phytoplasma (right) peaks. The cutoff value estimated by *Phytoassembly* falls between the two peaks. On the x-axis is the per-contig coverage, calculated as the number of reads aligned on the contig divided by the length of the contig, expressed as percent values. On the y-axis is the number of contigs with similar coverage. The plant peak has 111 contigs at coverage 4, the phytoplasma peak has 98 contigs at coverage 15.

The pipeline is written in the Bash and Perl languages and requires a working installation of *BioPerl* (http://bioperl.org/), *NCBI BLAST* + (https://blast.ncbi.nlm.nih.gov/Blast.cgi) and the *A5 pipeline* [28]. *Phytoassembly* has been tested on Linux Ubuntu 16.04 LTS and Mac OS X 10.11.6.

In detail, the pipeline includes the following steps.

Stage 0: Data preparation. Phytoassembly calls the A5 pipeline to assemble the healthy plant sequence reads (producing the file *Healthy.contigs.fasta*) unless an already assembled sequence is provided. Next, the diseased plant reads are assembled (producing the file *Diseased.contigs.fasta*). A step in the A5 pipeline produces error corrected reads (*Diseased.ec.fastq*), which are used in all the subsequent steps. The assembled reference sequence file is then indexed and aligned with the error corrected reads using the *BWA* tool [29]. The resulting file is converted to the *bam* format (*Diseased.mapped.bam*) and, using *samtools* (http://www.htslib.org/doc/samtools.html), a summary of statics is produced (*Diseased.sorted.csv*), consisting of the reference sequence name, sequence length, number of mapped and unmapped reads.

Stage 1: Cutoff. The pipeline estimates the cutoff value to be used by running once with cutoff 0, then using a fraction of the ratio between the sum of the lengths of the non-mapping reads at cutoff 0 (*Stage2.0.nonmatch.fastq,* see below) and the sum of the lengths of the error corrected reads (*Diseased.ec.fastq*) of the diseased plant, multiplied by 100.

Alternatively, if the user wants to supply a range of specifies fixed cutoff values, then the pipeline repeats the following steps from the lowest to the highest values provided (represented here as *$cutoffval*). From the summary of statistical data (*Diseased.sorted.csv*), per-contig coverages are calculated (as the ratio between the sum of the lengths of the mapped reads and the length of the contig, multiplied by 100), and saved in a text file (*Diseased.sorted.cov.csv*). The contigs with a coverage higher than *$cutoffval* are exported to a FASTA file (*Diseased.cutoff.$cutoffval.fasta*, where *$cutoffval* is *e.g*. "10"). The error-corrected reads from the diseased plant (*Assembly.ec.fastq*) are then aligned to the contigs in that, last file using BWA. From the alignment file (*Stage1.$cutoffval.match.sam*), the reads above the cutoff are extracted and exported in a FASTQ file (*Stage1.$cutoffval.match.fastq*).

Stage 2: Re-alignment and filtering. The reads from the cutoff (*Stage1.$cutoffval.match.fastq*) are now aligned with *BWA* against the healthy plant reference (*Healthy.contigs.fasta*) and a FASTQ file with the reads that do not align is exported (*Stage2.$cutoffval.nonmatch.fastq*). These non-aligned reads are assembled with the A5 pipeline (*Stage3.$cutoffval.contigs.fasta*).

Stage 3: BLAST. A BLAST nucleotide database is created from the reference healthy plant file (*Healthy.contigs.fasta,* which could also be a combination of different references) and used to query the contigs outputted by the previous stage (*Stage3.$cutoffval.contigs.fasta*) using *tblastx* (translated nucleotide query *vs*. translated nucleotide database BLAST). The results are saved in a text file (*Stage3.$cutoffval .contigs.csv*), which is then filtered according to the identity percentage (IP): entries with an IP greater than 95% are attributed to the plant (*Stage3.$cutoffval.contigs.plant.csv*), while those with a lower IP are attributed to the phytoplasma (*Stage3.$cutoffval.contigs.phyto.csv*). Using this last file the contigs pertaining to the phytoplasma are extracted from the query and saved in a FASTA file (*Stage3.$cutoffval.phyto.fasta*).

Stage 4: Clean-up. Lastly, the main outputs are compressed in the *gzip* format, moved to a folder (*Results_$timestamp*), statistical data such as contigs size and number are calculated, while the intermediate files are moved to a sub-folder (*Other_files*), which also contains the assembly of the reference and/or the diseased plant reads, unless skipped in Stage 0. If the user did not input a cutoff value, the *Results* folder will contain files for cutoff 0, the calculated maximum value and half of the maximum.

A flowchart of the *Phytoassembly* pipeline is provided as Supplementary Fig. (**S1**).

2.2. Source of Data

Genome assemblies of "*Ca*. Phytoplasma asteris", strain Aster Yellows Witches'-Broom (AYWB; Bai *et al.,* 2006; accession number CP000061), Milkweed Yellows phytoplasma (MW1; [17]; accession number AKIL00000000), Italian Clover Phyllody phytoplasma (MA1; [17]; accession number AKIM00000000), Vaccinium Witches' Broom phytoplasma (VAC; [17]; accession number AKIN00000000) and Poinsettia branch-inducing phytoplasma strain JR1 (JR1; [17]; accession number AKIK00000000) were downloaded from the NCBI database. The ILLUMINA reads data-sets of published genomes are available for download from SRA, under accession number SAMN01041250 (MA1) and SAMN01041251 (MW1). The ILLUMINA dataset from Chicory Phyllody associated phytoplasma (ChiP; Martini *et al.,* in preparation) and the genome draft generated by *Phytoassembly* are available at links accessible from BioProject number PRJNA422968. The ILLUMINA data-set from "*Ca*. Phytoplasma aurantifolia" strain Witches' Broom of Lime 2034 (WBDL; Siqueira Alves *et al.,* submitted) and Cassava Frogskin Disease associated phytoplasma (CFSD; Neves *et al.,* manuscript in preparation) were provided by the authors of the cited papers.

2.3. Simulations and Further Data Analysis

Comparisons of the assemblies were carried out using BUSCO [30], *MUMmer* [31], and OMA [32]. To benchmark the pipeline, a sequencing experiment was simulated from an existing complete phytoplasma genome. Artificial sequence reads were generated from a complete sequencing of AYWB, using an ad-hoc Perl script that introduces reading errors and combines the phytoplasma and the plant reads. Reads obtained from a healthy periwinkle in a previous work ([17]; SRA accession number SRS356159) were combined with the artificially generated reads, so that phytoplasma reads resulted in adding 5%, 10% and 15% proportions to the plant reads.

3. RESULTS

3.1. Validation

As presented in the introduction, the procedure described here exploits the different coverages of pathogen and host contigs resulting for a preliminary assembly of the ILLUMINA reads. Fig. (**1**) shows a coverage graphs of the contigs resulting from a preassembly of an 'artificial' dataset generated from the genome of AYWB, and mixed in a proportion of 15% to real ILLUMINA reads from a healthy periwinkle. Although the two peaks corresponding to the host and pathogen contigs are clearly distinguishable in the graph, maximizing the recovery of the pathogen data in order to obtain the complete genome reconstruction requires the estimation and use of an inclusive, cautious cutoff value. We found that a convenient cutoff value can be estimated as 0.3 times the ratio between the sum of the lengths of the non-mapping reads at cutoff 0 and the sum of the lengths error corrected reads, multiplied by 100. To test the robustness of the pipeline with this estimate, we performed a number of tests using artificial and real datasets.

First, the pipeline was run for cutoff values between 0 and 15 with various simulated datasets and the size of the resulting final assemblies evaluated (Fig. **2**). With the estimated cutoff, the pipeline recovered 88.1% (with 5% of phytoplasma reads and cutoff 2), 94.2% (with 10% of phytoplasma reads and cutoff 4) and 93.9% (with 15% of phytoplasma reads and cutoff 5) of the original AYWB sequence. The number of reconstructed genes (including partials) was 711, 666 and 666, respectively, compared to 534 in the actual AYWB genome. The higher value of the gene number in the assemblies generated by the pipeline was due to the fragmentation of genes located at contigs ends.

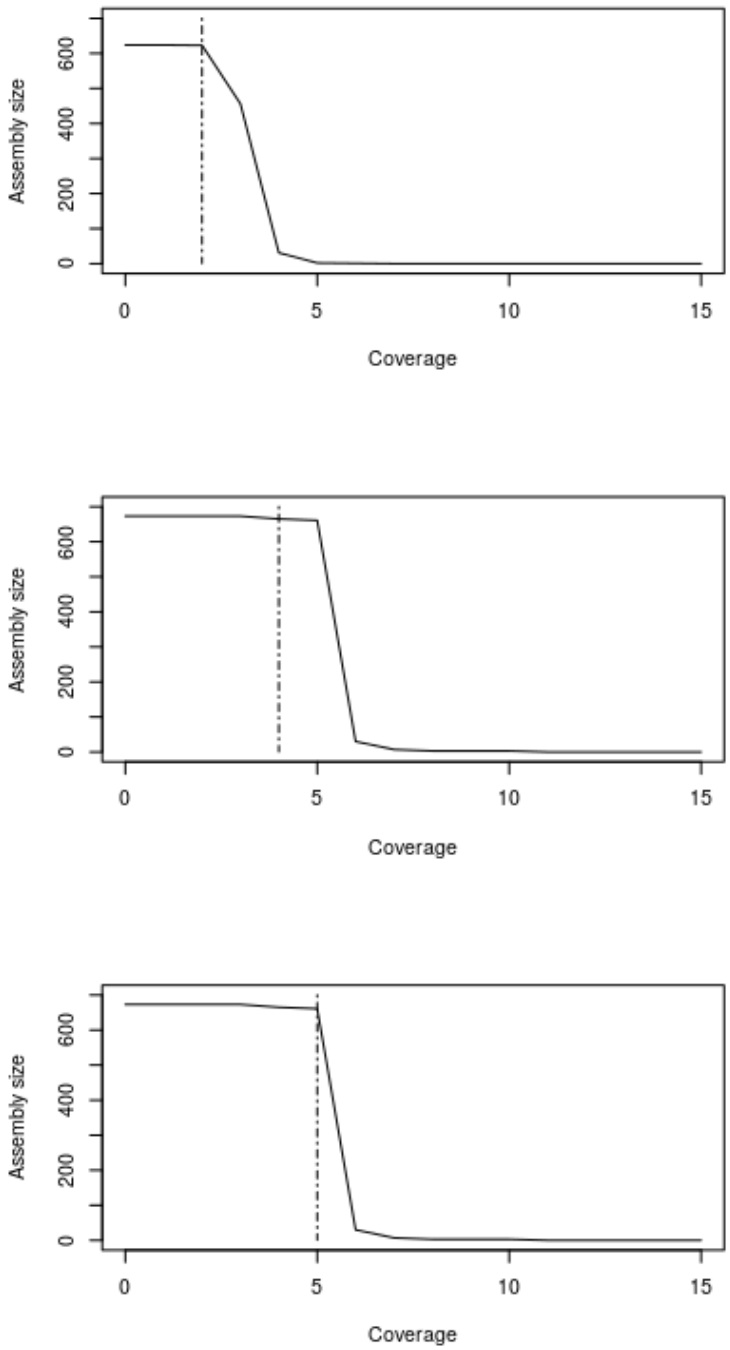

Fig. (2). Size (in knts) of the artificial aster yellows phytoplasma strain witches' broom (AYWB) sequences resulting from the use of different cutoff values. Datasets have phytoplasma/plant reads ratio of 5% (top), 10% (middle), or 15% (bottom). The vertical line shows the optimal cutoff determined by *Pythoassembly.* BLAST filtering did not remove any sequence from the output.

As a quality evaluation, we compared the genes found in the complete AYWB genome with those in the assembly generated by the pipeline from the dataset with 10% of phytoplasma reads and cutoff 4 using OMA. According to the results, 59 genes of AYWB did not have an identical counterpart in the *Phytoassembly* reconstructed genome. However, 20 of those genes showed >95% identity with a gene in the AYWB genome, the differences being due to misassembly of genes that are present in multiple, nonidentical, copies. The remaining 39 genes (7%) were all annotated as hypothetical proteins or phage associated proteins, and were characterized by low complexity in sequence. In conclusion, the pipeline provided suitable data for the complete reconstruction of the genetic features of the AYWB phytoplasma, failing only in areas of the genome with low complexity likely associated with phage integrations.

A second test used actual ILLUMINA reads of MW1 and MA1, and the results were compared with the previously obtained assemblies [17]. The reference genome used was a *Velvet* (https://www.ebi.ac.uk/~zerbino/velvet/) assembly from ILLUMINA reads of periwinkle. The reconstructed assembly of MW1 was 632,844 nts long without cutoff and

631,878 nts long with a 10 cutoff (222 contigs), while the 2012 assembly comprised 583,806 nts (158 contigs) (Table **1**); the minimum size of the contigs in the *Phytoassembly* reconstructions is 307 nts (N50 6,099 nts), while in 2012 one is 231 nts (N50 7,972 nts). The reconstructed assembly of MA1 was 710,075 nts long without cutoff and 708,886 nts long with a 10 cutoff (299 contigs), while the previously obtained assembly comprised 597,245 nts (197 contigs); the minimum size of the contigs in the *Phytoassembly* assembly is 188 nts and 184 nts (N50 10,390 nts and 10,407 nts), while in 2012, one is 230 nts (N50 12,309 nts). The MW1 assemblies differ on 128 contigs, 308-5477 nts in size; MA1 assemblies differ on 35 contigs, 299-1227 nts in size.

To assess the completeness of the MA1 and MW1 genome reconstructions by *Phytoassembly*, the assemblies were checked for missing conserved genes, using BUSCO. Running the program with the set of 14 phytoplasma genome drafts used in [33], we generated an *ad hoc* list comprising a subset of 77 BUSCOs (conserved genes) that are common to all phytoplasma genomes. As shown in Table **2**, one gene was missing in the assembly of MW1 and two genes were missing in the assembly of MA1. It was, therefore, estimated that *Phytoassembly* can recover >95% of the coding information of the sampled genomes.

3.2. Novel Drafts

Using this pipeline, high-quality draft assemblies of the WBDL, CFSD, and ChiP were obtained. The size of the assemblies varied from about 550,000 to about 800,000 nts (Table **1**).

Each of the phytoplasma genomes reconstructed by *Phytoassembly* was analyzed along with the four complete phytoplasma genomes available [7-10] using standalone OMA, in order to identify shared orthologs. 274 'shared' orthologs are present in all of the four phytoplasma genomes.

The CFSD sample was processed using a healthy cassava sample, obtaining a phytoplasma genome assembly of 818,980 nts in 293 contigs, ranging from 311 to 35,791 nts in length (see Table **1** for a full comparison between the samples). This sample shares 457 orthologs with at least one of the four phytoplasmas, and 247 with all of them.

The WBDL sample was processed with an ensemble of *Citrus sinensis* and *Citrus clementina,* because an isogenic reference was not available. After annotation, the phytoplasma genome assembly was 794,372 nts long divided into 182 contigs, ranging from 602 to 56,244 nts. This sample shares 479 orthologs with at least one of the four phytoplasmas, and 220 with all of them. An additional about 1,000,000 nts long set of small contigs could not be attributed to the phytoplasma nor to the plant, as they were not represented in the available *Citrus* genomes, but are assumed to be specific lime repeated sequences.

The ChiP sample was processed using the healthy periwinkle specimen (see MW1 and MA1 above), obtaining an assembly of 1,931,149 nts. The output of the pipeline was consistently oversized for a phytoplasma, which rarely exceeds 10^6 nts. It was, therefore annotated using RAST [34], and the result showed that 1,338,982 nts (69.3%) actually belonged to a spiroplasma, while the true phytoplasma genome was 547,918 nts (28.4%), assembled in 138 contigs, ranging from 605 to 25,180 nts.

The check for draft completeness, carried out with BUSCO and the *ad hoc* conserved gene list revealed, as shown in Table **2**, that no conserved genes were missing in the CSFD assembly, one gene was missing in the assembly of WBDL, and two genes were missing in the assembly of ChiP.

4. DISCUSSION

The *Phytoassembly* pipeline successfully addresses the problem of obtaining the genomic sequences of phytoplasmas, by selectively excluding the reads of the host plant from an infected plant ILLUMINA sequence data-set. It does so by first filtering out reads with low coverage, which can be assumed to belong to the plant, because of the vast disparity in coverage between the plant and the pathogen genome; then by removing the reads that can be aligned on the healthy plant genome.

As an improvement of the procedure developed in [17], which required *ad hoc* tuning and various manual or external steps for the *de novo* assembly, *Phytoassembly* can carry out autonomously, the complete analysis and relies on an assembler (the A5 pipeline) which doesn't require additional input from the user. The assembler is tailored for ILLUMINA reads, and works with paired-ends.

The sequences that pass the re-alignment step are those that do not map on the healthy plant reference, therefore they can only belong to genes not attributable to the plant host. While the main aim of the *Phytoassembly* procedure is the isolation of phytoplasma genes, by virtue of the mechanism employed, it can also isolate other non-culturable pathogens, or mask specific pathogens by adding their genomes to the healthy plant reference.

The pipeline attempts to determine a cutoff value using the ratio between the sum of the lengths of the non-mapping reads at cutoff 0 and the sum of the lengths of the error corrected reads of the diseased plant, multiplied by 100. This ratio was chosen because the error corrected reads exclude any ambiguous or unreliable data from the estimation, and the non-mapping reads represent a fraction roughly proportional to the pathogen quota in the sequencing. Using the value as it is, however, leads to an excessive cutoff. Plotting the nucleotide count of the phytoplasma reconstructions at various cutoffs (Fig. **2**), a common feature is a significant drop after a value that appears correlated to the percentage of pathogen genome in the diseased plant specimen. Based on the results of the artificial reads test, a more conservative estimation is obtained by using 0.3 times the aforementioned ratio.

A method to further optimize the cutoff value would be to run the pipeline at cutoff 0, increasing the value until the last estimation has a significant drop (in the order of more than 1000 nts) in the reconstructed genome size. Although this would increase the computation time significantly, while the method with cutoff estimation repeats the procedure only once, *Phytoassembly* provides an option for the non-automatic, ex post search of the optimal cutoff value. To this end, a bash script is provided (*phytoiterative.sh*) that runs

Table 1. Data relative to draft phytoplasma assemblies obtained with *Phytoassembly*.

-	Nucleotides	Contigs	Min. size	Max. size	N50 size	N50 contigs	G+C
AYWB reference	706,569	1	706,569	706,569	706,569	1	27%
AYWB 5% cutoff 0	624,492	242	398	21,808	3,987	47	27%
AYWB 5% cutoff 2	622,737	243	398	21,808	3,845	47	27%
AYWB 10% cutoff 0	673,019	111	407	137,058	30,483	7	27%
AYWB 10% cutoff 4	665,375	95	559	137,058	30,472	7	27%
AYWB 15% cutoff 0	664,899	95	512	90,316	28,048	8	27%
AYWB 15% cutoff 5	663,628	97	500	87,545	25,058	9	27%
Milkweed Yellows ph. (MW1) reference	583,806	158	231	22,485	7,972	26	27%
Phytoassembly MW1, cutoff 0	632,844	224	308	22,483	6,099	32	28%
Phytoassembly MW1, cutoff 10	631,878	222	307	22,483	6,099	32	28%
Italian Clover Phyllody ph. (MA1) reference	597,245	197	230	40,778	12,309	16	27%
Phytoassembly MA1, cutoff 0	710,075	296	188	39,685	10,390	20	27%
Phytoassembly MA1, cutoff 10	708,886	299	184	39,685	10,407	20	27%
Cassava Frogskin Disease (CFSD)	818,980	293	311	35,791	7,796	28	29%
Ca. Phytoplasma Aurantifolia (WBDL)	794,372	182	602	56,244	13,769	17	28%
Chicory Phyllody (ChiP) raw	1,931,149	370	605	83,360	11,391	35	26%
Chicory Phyllody (ChiP)	547,918	138	605	25,180	4,832	30	25%

Table 2. Conserved genes missing from new genome drafts built by *Phytoassembly*.

Assembly	Missing BUSCOs	Description
MA1	POG090A00A0	tRNA uridine 5-carboxymethylaminomethyl modification protein
MA1	POG090A001V	Ribosomal protein S15
MW1	POG090A019O	Signal recognition particle protein Srp54
CSFD	None	-
CHIP	POG090A00VB	Transcription termination/antitermination factor NusG
CHIP	POG090A012Q	Ribosomal protein L35
WBDL	POG090A00FL	Elongation factor G

iteratively *Phytoassembly* with different cutoff values and outputs the results that maximize the length and quality of the pathogen genome draft.

CONCLUSION

Phytoassembly is a focused tool that allows a user-friendly and performant processing of ILLUMINA sequence data from a pair of samples, a phytoplasma infected plant sample and its uninfected reference sample, outputting a high-quality genome draft of the pathogen. Given the increasing availability of access to ILLUMINA technology, *Phytoassembly* is expected to be a valuable help in the characterization of the genomes of the large, diverse and economically relevant group of plant pathogens that belong to the genus "*Ca*. Phytoplasma".

The *Phytoassembly* source code is available on GitHub at https://github.com/cpolano/phytoassembly.

ETHICS APPROVAL AND CONSENT TO PARTICIPATE

Not applicable.

HUMAN AND ANIMAL RIGHTS

No Animals/Humans were used for studies that are the basis of this research.

CONSENT FOR PUBLICATION

Not applicable.

CONFLICT OF INTEREST

The authors declare no conflict of interest, financial or otherwise.

ACKNOWLEDGEMENTS

Declared none.

SUPPLEMENTARY MATERIAL

Supplementary material is available on the publisher's website along with the published article.

REFERENCES

[1] Lee, I.-M.; Davis, R.E.; Gundersen-Rindal, D.E. Phytoplasma: *Phytopathogenic mollicutes*. *Annu. Rev. Microbiol.*, **2000**, *54*(1), 221-255.

[2] Zhao, Y.; Davis, R.E.; Lee, I.-M. Phylogenetic positions of 'Candidatus *Phytoplasma asteris'* and *Spiroplasma kunkelii* as inferred from multiple sets of concatenated core housekeeping proteins. *Int. J. Syst. Evol. Microbiol.*, **2005**, *55*(5), 2131-2141.

[3] Weintraub, P.G.; Beanland, L. Insect vectors of phytoplasmas. *Annu. Rev. Entomol.*, **2006**, *51*(1), 91-111.

[4] Marcone, C.; Ragozzino, A.; Seemuller, E. Dodder transmission of alder yellows phytoplasma to the experimental host *Catharanthus roseus* (periwinkle). *For. Pathol.*, **1997**, *27*(6), 347-350.

[5] Tran-Nguyen, L.T.T.; Gibb, K.S. Optimizing phytoplasma DNA purification for genome analysis. *J. Biomol. Tech.*, **2007**, *18*(2), 104-112.

[6] Saeed, E.; Seemüller, E.; Schneider, B.; Saillard, C.; Blanchard, B.; Bertheau, Y.; Cousin, M.T. Molecular cloning, detection of chromosomal DNA of the Mycoplasma Like Organism (MLO) associated with Faba Bean (*Vicia faba* L.) Phyllody by southern blot hybridization and the Polymerase Chain Reaction (PCR). *J. Phytopathol.*, **1994**, *142*(2), 97-106.

[7] Oshima, K.; Kakizawa, S.; Nishigawa, H.; Jung, H.-Y.; Wei, W.; Suzuki, S.; Arashida, R.; Nakata, D.; Miyata, S.; Ugaki, M.; Namba, S. Reductive evolution suggested from the complete genome sequence of a plant-pathogenic phytoplasma. *Nat. Genet.*, **2004**, *36*(1), 27-29.

[8] Bai, X.; Zhang, J.; Ewing, A.; Miller, S.A.; Jancso Radek, A.; Shevchenko, D.V; Tsukerman, K.; Walunas, T.; Lapidus, A.; Campbell, J.W.; Hogenhout, S.A. Living with genome instability: The adaptation of phytoplasmas to diverse environments of their insect and plant hosts. *J. Bacteriol.,* **2006**, *188*(10), 3682-3696.

[9] Kube, M.; Schneider, B.; Kuhl, H.; Dandekar, T.; Heitmann, K.; Migdoll, A.M.; Reinhardt, R.; Seemüller, E. The linear chromosome of the plant-pathogenic mycoplasma 'Candidatus *Phytoplasma mali'*. *BMC Genomics*, **2008**, *9*(1), 306. Available from: https://bmcgenomics.biomedcentral.com/articles/10.1186/1471-2164-9-306

[10] Tran-Nguyen, L.T.T.; Kube, M.; Schneider, B.; Reinhardt, R.; Gibb, K.S. Comparative genome analysis of 'Candidatus *Phytoplasma australiense'* (subgroup tuf-Australia I; rp-A) and 'Ca. *Phytoplasma asteris'* strains OY-M and AY-WB. *J. Bacteriol.*, **2008**, *190*(11), 3979-3991.

[11] Andersen, M.T.; Liefting, L.W.; Havukkala, I.; Beever, R.E. Comparison of the complete genome sequence of two closely related isolates of 'Candidatus *Phytoplasma australiense*' reveals genome plasticity. *BMC Genomics*, **2013**, *14*(1), 529. Available from: https://bmcgenomics.biomedcentral.com/articles/10.1186/1471-2164-14-529

[12] Liefting, L.W.; Kirkpatrick, B.C. Cosmid cloning and sample sequencing of the genome of the uncultivable mollicute, Western X-disease phytoplasma, using DNA purified by pulsed-field gel electrophoresis. *FEMS Microbiol. Lett.*, **2003**, *221*(2), 203-211.

[13] Garcia-Chapa, M.; Batlle, A.; Rekab, D.; Rosquete, M.R.; Firrao, G. PCR-mediated whole genome amplification of phytoplasmas. *J. Microbiol. Meth*., **2004**, *56*(2), 231-242.

[14] Cimerman, A.; Arnaud, G.; Foissac, X. Stolbur phytoplasma genome survey achieved using a suppression subtractive hybridization approach with high specificity. *Appl. Environ. Microbiol.,* **2006**, *72*(5), 3274-3283.

[15] Kawar, P.G.; Pagariya, M.C.; Dixit, G.B.; Prasad, D.T. Identification and isolation of SCGS phytoplasma-specific fragments by riboprofiling and development of specific diagnostic tool. *J. Plant Biochem. Biotechnol.,* **2010**, *19*(2), 185-194.

[16] Casati, P.; Quaglino, F.; Stern, A.R.; Tedeschi, R.; Alma, A.; Bianco, P.A. Multiple gene analyses reveal extensive genetic diversity among 'Candidatus *Phytoplasma mali'* populations. *Ann. Appl. Biol.,* **2011**, *158*(3), 257-266.

[17] Saccardo, F.; Martini, M.; Palmano, S.; Ermacora, P.; Scortichini, M.; Loi, N.; Firrao, G. Genome drafts of four phytoplasma strains of the ribosomal group 16SrIII. *Microbiology*, **2012**, *158*(Pt 11), 2805-2814.

[18] Chung, W.-C.; Chen, L.-L.; Lo, W.-S.; Lin, C.-P.; Kuo, C.-H. Comparative analysis of the peanut witches'-broom phytoplasma genome reveals horizontal transfer of potential mobile units and effectors. *PLoS One*, **2013**, *8*(4), e62770. Available from: journals.plos.org/plosone/article?id=10.1371/journal.pone.0062770

[19] Davis, R.E.; Zhao, Y.; Dally, E.L.; Lee, I.-M.; Jomantiene, R.; Douglas, S.M. 'Candidatus *Phytoplasma pruni'*, a novel taxon associated with X-disease of stone fruits, Prunus spp.: Multilocus characterization based on 16S rRNA, secY, and ribosomal protein genes. *Int. J. Syst. Evol. Microbiol.,* **2013**, *63*(Pt 2), 766-776.

[20] Quaglino, F.; Zhao, Y.; Casati, P.; Bulgari, D.; Bianco, P.A.; Wei, W.; Davis, R.E. 'Candidatus *Phytoplasma solani'*, a novel taxon associated with stolbur- and bois noir-related diseases of plants. *Int. J. Syst. Evol. Microbiol.,* **2013**, *63*(Pt 8), 2879-2894.

[21] Chen, W.; Li, Y.; Wang, Q.; Wang, N.; Wu, Y. Comparative genome analysis of wheat blue dwarf phytoplasma, an obligate pathogen that causes wheat blue dwarf disease in China. *PLoS One*, **2014**, *9*(5), e96436. Available from: journals.plos.org/plosone/article?id=10.1371/journal.pone.0096436

[22] Quaglino, F.; Kube, M.; Jawhari, M.; Abou-Jawdah, Y.; Siewert, C.; Choueiri, E.; Sobh, H.; Casati, P.; Tedeschi, R.; Lova, M.M.; Alma, A.; Bianco, P.A. 'Candidatus *Phytoplasma phoenicium*' associated with almond witches'-broom disease: from draft genome to genetic diversity among strain populations. *BMC Microbiol.*,

2015, *15*(1), 148. Available from: https://bmcmicrobiol.biomedcentral.com/articles/10.1186/s12866-015-0487-4

[23] Chitsaz, H.; Yee-Greenbaum, J.L.; Tesler, G.; Lombardo, M.; Dupont, C.L.; Badger, J.H.; Novotny, M.; Rusch, D.B.; Fraser, L.J.; Gormley, N.A.; Schulz-Trieglaff, O.; Smith, G.P.; Evers, D.J.; Pevzner, P.A.; Lasken, R.S. Efficient *de novo* assembly of single-cell bacterial genomes from short-read data sets. *Nat. Biotechnol.,* **2011**, *29*(10), 915-921.

[24] Liu, S.; Trapnell, C. Single-cell transcriptome sequencing: Recent advances and remaining challenges. *F1000Res.*, **2016**, 5, F1000 Faculty Rev-182. Available from: https://f1000research.com/articles/5-182/v1

[25] Dierckxsens, N.; Mardulyn, P.; Smits, G. NOVOPlasty: *De novo* assembly of organelle genomes from whole genome data. *Nucleic Acids Res.,* **2016**, *45*(4), e18.

[26] Soorni, A.; Haak, D.; Zaitlin, D.; Bombarely, A. Organelle_PBA, a pipeline for assembling chloroplast and mitochondrial genomes from PacBio DNA sequencing data. *BMC Genom.,* **2017**, *18*(1), 49. Available from: https://bmcgenomics.biomedcentral.com/articles/10.1186/s12864-016-3412-9

[27] Zonneveld, B.J.M.; Leitch, I.J.; Bennett, M.D. First nuclear DNA amounts in more than 300 angiosperms. *Ann. Bot.,* **2005**, *96*(2), 229-244.

[28] Tritt, A.; Eisen, J.A.; Facciotti, M.T.; Darling, A.E. An integrated pipeline for *de novo* assembly of microbial genomes. *PLoS One*, **2012**, *7*(9), e42304. Available from: journals.plos.org/plosone/article?id=10.1371/journal.pone.0042304

[29] Li, H.; Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, **2009**, *25*(14), 1754-1760.

[30] Simão, F.A.; Waterhouse, R.M.; Ioannidis, P.; Kriventseva, E.V; Zdobnov, E.M. BUSCO: Assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics*, **2015**, *31*(19), 3210-3212.

[31] Delcher, A.L.; Phillippy, A.; Carlton, J.; Salzberg, S.L. Fast algorithms for large-scale genome alignment and comparison. *Nucleic Acids Res.,* **2002**, *30*(11), 2478-2483.

[32] Altenhoff, A.M.; Kunca, N.; Glover, N.; Train, C.-M.; Sueki, A.; Piliota, I.; Gori, K.; Tomiczek, B.; Muller, S.; Redestig, H.; Gonnet, G.H.; Dessimoz, C. The OMA orthology database in 2015: Function predictions, better plant support, synteny view and other improvements. *Nucleic Acids Res.,* **2015**, *43*(D1), D240-D249.

[33] Firrao, G.; Martini, M.; Ermacora, P.; Loi, N.; Torelli, E.; Foissac, X.; Carle, P.; Kirkpatrick, B.C.; Liefting, L.; Schneider, B.; Marzachì, C.; Palmano, S. Genome wide sequence analysis grants unbiased definition of species boundaries in 'Candidatus Phytoplasma'. *Syst. Appl. Microbiol.,* **2013**, *36*(8), 539-548.

[34] Aziz, R.K.; Bartels, D.; Best, A.A.; DeJongh, M.; Disz, T.; Edwards, R.A.; Formsma, K.; Gerdes, S.; Glass, E.M.; Kubal, M.; Meyer, F.; Olsen, G.J.; Olson, R.; Osterman, A.L.; Overbeek, R.A.; McNeil, L.K.; Paarmann, D.; Paczian, T.; Parrello, B.; Pusch, G.D.; Reich, C.; Stevens, R.; Vassieva, O.; Vonstein, V.; Wilke, A.; Zagnitko, O. The RAST server: Rapid annotations using subsystems technology. *BMC Genomics*, **2008**, *9*(1), 75. Available from: https://bmcgenomics.biomedcentral.com/articles/10.1186/1471-2164-9-75