Using Phytophthora ITS1 for identification of field samples

# Introduction

*Phytophthora* species need to be tracked to determine their spread and accordance of new species.

How:

* Water or roots samples containing *Phytophthora* is collected from nurseries/ rivers/ etc …
* DNA extracted from filters or roots.
* Nested PCR to amplify *Phytophthora* ~ specifically.
* PCR products sent to Illumina sample prep.
* Overlapping reads.
* MiSeq sequencing - barcoded samples, spit out to specific files based on barcodes.
* Bioinformatics

## How many ITS1 regions do *Phytophthora* genomes have?

Before any serious analysis is started we need to know a few things.

1. How many ITS1 regions are there within a *Phytophthora* genome

There are shell scripts for each species which runs the pipeline from start to finish for all of them

1. How much diversity is there within a single *Phytophthora* genome.

### Identify ITS regions in each genome by BLAST using a “database” of ITS sequences.

* Using a database of 402 regions from *PhytophthoraDB,* these are BLASTN BLAST+ (version 2.2.30) ([Camacho, et al. 2009](#_ENREF_3)) searched against the genomes.
* A consensus “HIT” is made due to overlapping 5 prime and 3 prime BLAST hits. Python

### HMMsearch to identify ITS regions

* Redundancy was removed from the ITS database using CD-HIT (4.5.4) ([Li and Godzik 2006](#_ENREF_9)) at various thresholds. 100% was used for the final HMM model
* HMMprofile ([Finn, et al. 2011](#_ENREF_6)), was also made from aligned ITS and used to identify ITS regions within a given genome.

### Identify single copy genes using BUSCO (EOG)

* BUSCO (BUSCO\_v1.1b1) ([Simão, et al. 2015](#_ENREF_14)) is a tool to predict core eukaryotic genes.
* The models are not well refined for *Phytophthora,* therefore just used the generic eukaryotic models.
* Convert the output the a GFF for use later

### MAP genomic reads back to the genome

* This massively reduced the number of genomes we can work with, either reads are not available or …. From XXX to 7!
* After the reads have been QC FastQC ([Andrews 2010](#_ENREF_1)) , Trimmomatic ([Bolger and Giorgi](#_ENREF_2)) - trimmed, they were mapped back to the genome. Bowtie 2 ([Langmead 2010](#_ENREF_8))
* Reads WILL have multiple mapping due the nature of the ITS region. Therefore randomly place these. BOWTIE setting.
* Bedtools ([Quinlan and Hall 2010](#_ENREF_12)) Count the number of reads that fall within the identified ITS regions (BLASTN step mentioned above) and compare those to the ~single copy EOG - BUSCO genes (mentioned above) and also the rest of the genes in the genome.
* Compare the ratios of the coverage!!

## Pipeline: A species identification pipeline given Illumina data

A number of pipelines exist that have been developed to identify species based of OTUs, however mostly for fungi: FHiTINGS ([Dannemiller, et al. 2014](#_ENREF_4)), Mycofier ([Delgado-Serrano, et al. 2016](#_ENREF_5)), PIPITS ([Gweon, et al. 2015](#_ENREF_7)). However, to my knowledge only one pipeline exists for use with *Phytophthora* ([Scibetta, et al. 2012](#_ENREF_13)).

Therefore a pipeline needs to be developed for *Phytophthora* species, which utilises a *Phytophthora* specific ITS1 database.

### How

A number of tools exist for each step. The following listed are the current “best” tools for the job, in my opinion. Figure 1 represents the major steps required for the identification of species when given Illumina data. A known ITS1 database is required (OTUs).

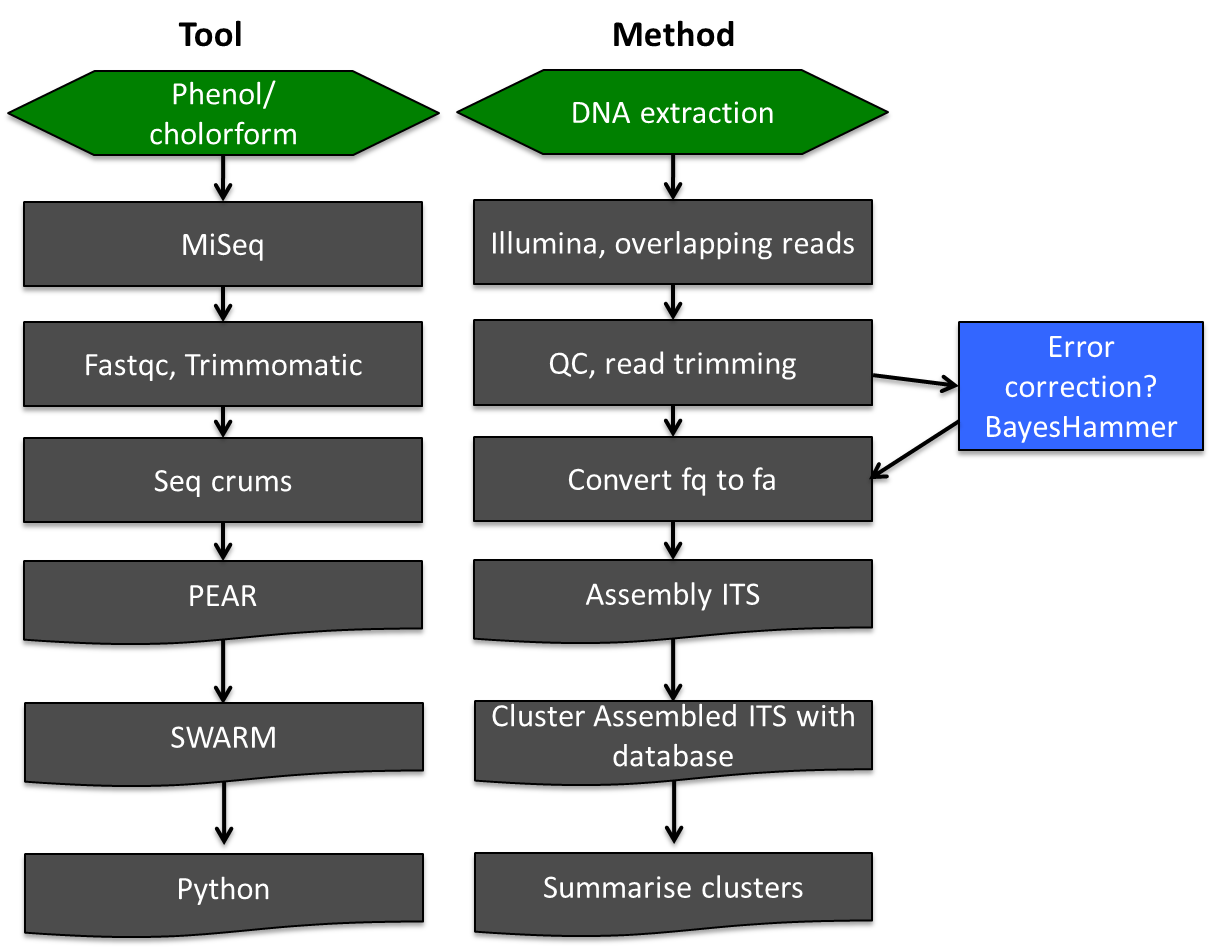


Figure : Diagrammatic representation of the tasks and tools needed in order to perform species identification when given ITS1 sequences.

There are a number of different tools which can be used at each step. Arguably assembly and clustering tools can have the biggest impact. For assembly PEAR ([Zhang, et al. 2014](#_ENREF_15)) was chosen. PEAR is very fast and applies statistical tests to minimise false positive results. For small overlaps it was reported that FLASH ([Magoč and Salzberg 2011](#_ENREF_10)) often fails to merge such reads ([Zhang, et al. 2014](#_ENREF_15)), therefore PEAR was chosen.

For clustering, a number of tools exist. The data analysed in ([Scibetta, et al. 2012](#_ENREF_13)) was clustered with BlastClust and OTUs picked with Qimme. Other clustering tools exist such as CD-HIT-454, DNAclust, Swarm and Usearch. Tools such as CD-HIT-454 and Usearch have inherent problem: They are biased to the order the dataset is presented in. Thus results can alter depending on the order of the dataset. Swarm ([Mahé, et al. 2015](#_ENREF_11)) is an amplicon clustering tool that has been developed which aims to overcome the limitations of other clustering tools. To summarise the clusters, custom Python scripts are made.

### Making and clustering an ITS1 database

The database from Santii was aligned with a known ITS1 region from P. infestans obtained from GenBank. The original database had sequences of around 800nt, which would not be suitable for clustering with swarm. Therefore, the alignment was trimmed to only the ITS1 region for all sequences. Thus yielding a database of ITS1 regions only.

### Test pipeline with known data (PCRMIX\_S96) - previous data:

Below (Table 1) is a table which represents the *Phytophthora* which was used to make the PCRMIX\_S96 dataset. This is a test Illumina MiSeq run composed of a known mix of PCR ITS1 products (originating from those in the table).

|  |  |  |  |
| --- | --- | --- | --- |
| **species** | **Clade** | **pr** | **size** |
| ***P.idaei*** | 1 | 3 | 184 |
| ***P.capsici*** | 2 | 1 | 140 |
| ***P.plurivora*** | 2 | 2 | 152 |
| ***P.palmivora*** | 4 | 4 | 179 |
| ***P.katsurae*** | 5 | 1 | 182 |
| ***P.megasperma*** | 6 | 1 | 193 |
| ***P.rubi*** | 7 | 1 | 199 |
| ***P.cryptogea*** | 8 | 1 | 174 |
| ***P.fallax*** | 9 | 3 | 172 |
| ***P.boehmeriae*** | 10 | 1 | 172 |

Table : Table representing the species used

# Results

## Number of ITS regions within a genome

The number of ITS regions identified by BLAST varied hugely between the genomes. The HMM search identified slightly less, but the trend was consistent with the number identified by BLAST.

Table 2 represents the coverage stats for each class of genes (all genes, EOG and ITS regions). Using these coverage stats the theoretical number of ITS regions was predicted based on the following assumptions: Assuming genome read coverage is normal and extra coverage of ITS regions is directly proportional to the real number of ITS regions. The following formula represent how the theoretical ITS number was calculated.

ITS(theoretical) = ∑ITS\_hits ⋅ (x̅ ITS\_coverage(assembled) / x̅ gene\_coverage)

Or for the EOG genes

ITS(theoretical) = ∑ITS\_hits ⋅ (x̅ ITS\_coverage(assembled) / x̅ EOG\_coverage)

The number of ITS regions identified by BLAST, and are therefore assembled into the genome is an under representation of the total number which is truly present. It is highly likely that genome assemblers collapse such region into a consensus sequence due to the complexity of the assembling highly similar repetitive regions. It may even be possible that some of these regions are so similar that Illumina sequencing technology is not able to resolve such regions (also read length limitations).

Based of genomic reads coverage, the theoretical ITS regions increases substantially compared to the assembled value for all genomes interrogated. Figure 2 is a boxplot representing the data obtained for *Phytophthora infestans* T30-4 genomic reads coverage for each class of genes. BUSCO- EOG genes do not help in generating a theoretical ITS gene count number, as the theoretical ITS gene count number predicted when using EOG coverage ratio is less than the total found by BLASTN, which is clearly incorrect. Experimental evidence validates the hypothesis of a greater theoretical ITS gene count number compared to assembled ITS gene count number. QPCR data using actin as a base line for single gene expression profile implies there could be between 48 and 508 ITS regions in a given genome (Table 3).

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | ***P. infestans*** | ***P. lateralis*** | ***P. sojae*** | ***P. cinnamomi*** | ***P. kernoviae*** | ***P. ramorum*** | ***P. cambivora*** | ***notes*** |
| **ITS** |  |  |  |  |  |  |  | p capsici no read data |
| Total BLAST hits | 133 | 2 | 7 | 2 | 1 | 1 | 12 |  |
| min/ max read coverage | 0/3153 | 386/6801 | 3/706285 | 113/318066 | 201002/201002 | 29824/29824 | 0/513 | # see shell scripts for exact genomes and reads used. |
| Average read coverage | 1440 | 3593 | 101100 | 159089 | 201002 | 29824 | 73 |  |
| Median read coverage | 1436 | 3593 | 291 | 159089 | 201002 | 29824 | 2 |  |
|  |  |  |  |  |  |  |  |  |
| **BUSCO, EOG** |  |  |  |  |  |  |  |  |
| Total EOG genes identified | 1029 | 1184 | 1135 | 1184 | 1221 | 1142 | 1222 |  |
| min/ max read coverage | 0/568537 | 78/50607 | 0/1955219 | 6/3828473 | 966/133767 | 140/21272 | 0/530 |  |
| Average read coverage | 2823 | 11151 | 19615 | 14557 | 10819 | 3568 | 19 |  |
| Median read coverage | 1977 | 11746 | 17477 | 5302 | 8596 | 3709 | 10.5 |  |
|  |  |  |  |  |  |  |  |  |
| **GENES** |  |  |  |  |  |  |  |  |
| Total no. of annotated genes | 17791 | 12222 | 18196 | NA | 10855 | 15608 | NA |  |
| min/ max read coverage | 0/61450 | 6/86516 | 0/1497842 | NA | 22/79779 | 0/20345 | NA |  |
| Average read coverage | 275 | 2123 | 3181 | NA | 1691 | 653 | NA |  |
| Median read coverage | 190 | 1492 | 2181 | NA | 1216 | 476 | NA |  |
|  |  |  |  |  |  |  |  |  |
| **RESULTS BASED ON MEAN** |  |  |  |  |  |  |  | \* mean is bad for extreme values |
| ITS - BUSCO ratio | 0.5 | 0.3 | 5.2 | 10.9 | 18.6 | 8.4 | 3.8 |  |
| **predicted ITS** | **68** | **0.6** | **36** | **21.9** | **18.6** | **8.4** | **46.1** |  |
|  |  |  |  |  |  |  |  |  |
| ITS - gene ratio | 5.2 | 1.7 | 31.8 | NA | 118.8 | 45.6 | NA |  |
| **predicted ITS** | **698** | **3.4** | **222.5** | NA | **118.8** | **45.6** | NA |  |
|  |  |  |  |  |  |  |  |  |
| **RESULTS BASED ON MEDIAN** |  |  |  |  |  |  |  |  |
| ITS - BUSCO ratio | 0.7 | 0.3 | 0 | 30 | 23.4 | 8 | 0.2 |  |
| **predicted ITS** | **96.6** | **0.6** | **0.1** | **60** | **23.4** | **8** | **2.3** |  |
|  |  |  |  |  |  |  |  |  |
| ITS - gene ratio | 7.6 | 2.4 | 0.1 | NA | 165.3 | 62.7 | NA |  |
| **predicted ITS** | **1005** | **4.8** | **0.9** | **NA** | **165.3** | **62.7** | **NA** |  |
|  |  |  |  |  |  |  |  |  |
| **ITS** |  |  |  |  |  |  |  |  |
| Total HMM hits | 102 | 0 | 3 , 4 loose | 1 | 2 loose 1 | 1 | 5, 8loose | # p. lateralis ITS region is really short for unknown reasons. |
| min/ max read coverage |  |  |  |  |  |  |  |  |
| Average read coverage |  |  |  |  |  |  |  |  |
| Median read coverage |  |  |  |  |  |  |  |  |

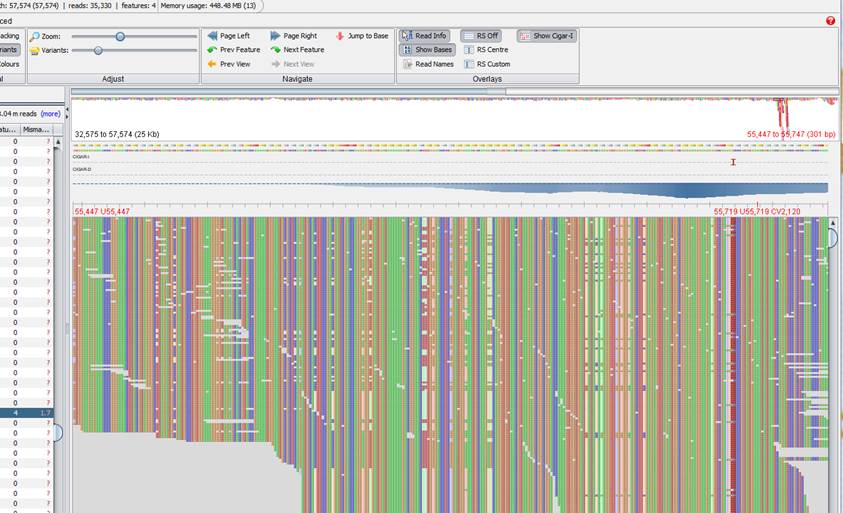
Table : Table representing coverage statistics for each class of genes. These statistics are then used to estimate the number of ITS regions within a genome.

***P. lateralis* is weird:**

*Phy Lateralis* ITS blast region is that area with huge amount of coverage indicated by the arrow.

The low coverage around it is not representative of other areas in my opinion. There is not obvious reason

Why the ITS would be predicted to be short either…



|  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Species | Clade | Primer | Size (bp) | CT Actin (1/10) | | CT ITS (1/10) | | Difference | Ctdiff/3.3 | Copies/genome | | | Copies/genome |
| ***P. idaei*** | 1 | 3 | 184 | 23.5 | | 17.89 | | 5.61 | 1.7 | 50.11872336 | | | **48.84029469** |
| ***P. capsici*** | 2 | 1 | 140 | 25.84 | | 20.27 | | 5.57 | 1.687878788 | 48.73924393 | | | **47.50475426** |
| ***P. plurivora*** | 2 | 2 | 152 | 19.47 | | 10.7 | | 8.77 | 2.657575758 | 454.5438201 | | | **436.5490646** |
| ***P. palmivora*** | 4 | 4 | 179 | 17.18 | | 11.31 | | 5.87 | 1.778787879 | 60.08801796 | | | **58.48521281** |
| ***P. katsurae*** | 5 | 1 | 182 | 22.38 | | 14.82 | | 7.56 | 2.290909091 | 195.3930405 | | | **188.7064598** |
| ***P. megasperma*** | 6 | 1 | 193 | 25.54 | | 18.11 | | 7.43 | 2.251515152 | 178.4494244 | | | **172.4458978** |
| ***P. rubi*** | 7 | 1 | 199 | 25.64 | | 18.19 | | 7.45 | 2.257575758 | 180.9571541 | | | **174.8531529** |
| ***P. cryptogea*** | 8 | 1 | 174 | 19.49 | 10.5 | | 8.99 | | 2.724242424 | | 529.9591859 | **508.4633577** | |
| ***P. fallax*** | 9 | 3 | 172 | 23.43 | 16.65 | | 6.78 | | 2.054545455 | | 113.3823501 | **109.8963759** | |
| ***P. boehmeriae*** | 10 | 1 | 172 | 21.65 | 14.17 | | 7.48 | | 2.266666667 | | 184.7849797 | **178.5271893** | |

Table : Table representing QPCR data (from SANTI??? DAVID???) The QPCR amplification curves were compared to Actin in order to estimate the copy number of the amplified ITS regions. The number of ITS regions estimated by QPCR was between 48 and 508.

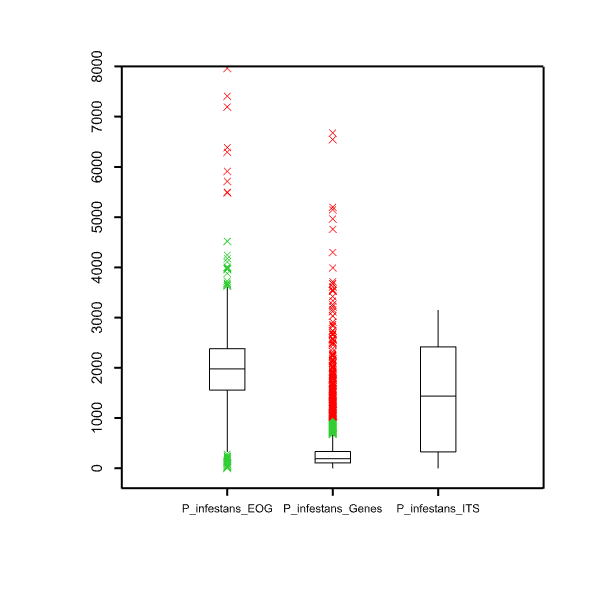


Figure : Boxplot representing genomic reads coverage for each class of genes from the *Phytophthora infestans* T30-4 genome. The ITS genes compared to all gene have greater coverage, implying greater copy number.

## How much diversity is there within a single *Phytophthora* genome

Using the full database as a BLAST search against the *P. infestans* genome we identify a range of similarity from 80% 100%. This was also, and more importantly observed when using a single *P. infestans* ITS sequence was BLASTN searched against the *P. infestans* genome. This implies that within a single genome, the ITS regions can vary from 80% - 100% similar to a given ITS query (limit of a BLAST search).

Of the ITS BLAST hits from *P. infestans* T30-4, these were extracted and trimmed to the ITS1 region. Variation and indels were observed. The resulting 47 sequences (which were not removed due to missing regions etc …) were then clustered with the database for barcoding.

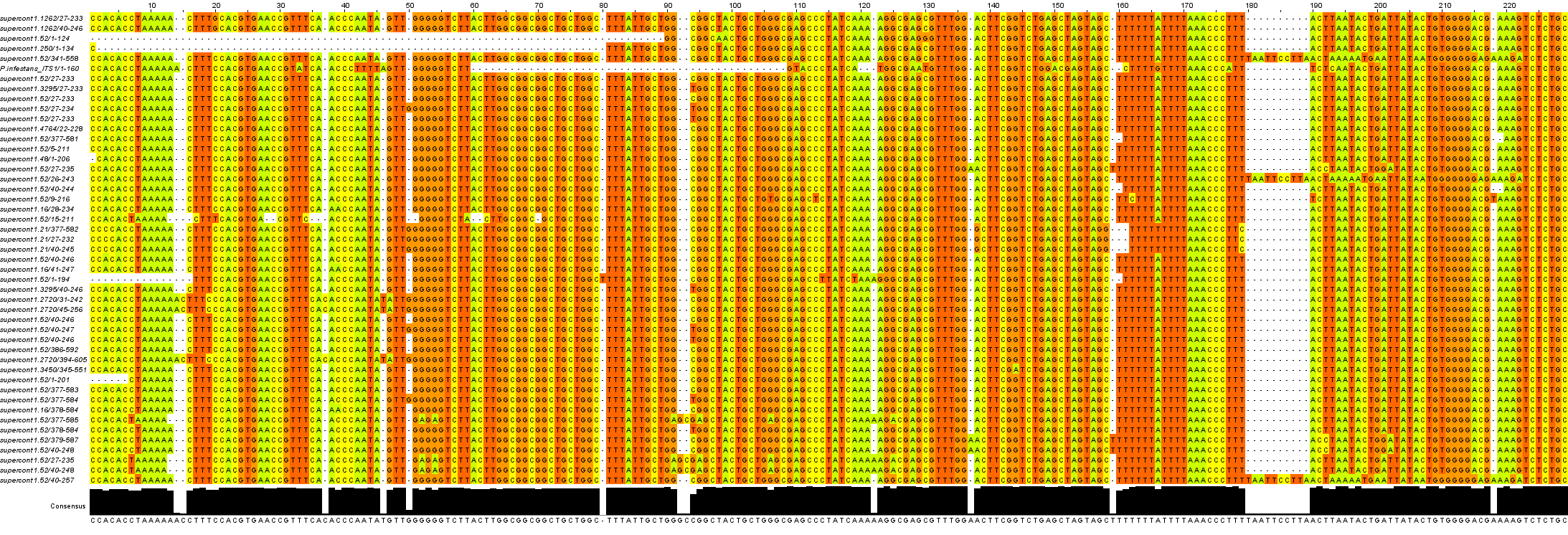


Figure : Alignment of ITS1 regions found with a single genome, *P. infestans* T30-4. Variation is found, including indels.

**THE RESULT:** When clustered with the “database” using Swarm, all *P. infestans* T30-4 genomic ITS1 sequences either cluster with CLADE1C, or reside as singletons. Therefore, these DO NOT cluster with false positive species. Although the other species are only represented by a single ITS and not the several hundred to may exist. Conclusion, based on these data, clustering of ITS1 regions obtained from individual species should be sufficient to identify species, or at the very least, the clade of which the species belongs to.

## Pipeline: A species identification pipeline given Illumina data

### Making and clustering an ITS1 database

In order to determine if the sequences in this database alone are sufficient to identify sequences when clustered together, it was necessary to subject the database to clustering analysis. Just to clarify: Do the sequences in the database cluster with one another, if so, then these could never resolve sequences originating from these.

122 clusters were produced out of 154 sequences. Therefore some of these ITS1 sequences DID cluster with another ITS1 sequences. Using Swarm with a setting of d1. However, only sequences from a certain clade will cluster with a same clade. Therefore ITS1 sequences from clade 2 will only cluster with others from clade 2 and not 3, 4 etc…

The sequences which clustered with other ITS sequences are (Table 4):

6Phytophthora\_sp\_nov\_CBS114338 Phytophthora\_taxon\_PgChlamydo\_VHS3753\_EU301160

Phytophthora\_gibbosa\_CBS127951\_HQ012933 Phytophthora\_gregata\_CBS127952\_HQ012942 Phytophthora\_taxon\_raspberry\_P1050\_AF541905

P.\_taxon\_kwongon\_TCH009 P.\_taxon\_kwongonlike\_CLJO100

P.\_aff.\_rosacearum\_DDS2909 P.\_rosacearum\_P292

Phytophthora\_amnicola\_CBS131652\_JQ029956 P.\_amnicola\_CBS131652

8Phytophthora\_aff\_primulae\_P6817 8Phytophthora\_primulae\_CBS110162

8Phytophthora\_erythroseptica\_CBS111343 8Phytophthora\_himalayensis\_CBS35759

7Phytophthora\_europaea\_BR1072 7Phytophthora\_uliginosa\_P10413

7Phytophthora\_cinnamomi\_var\_robiniae\_P16351 7Phytophthora\_sojae\_CBS38261

7Phytophthora\_melonis\_CBS58269 7Phytophthora\_sinensis\_CBS55788

7Phytophthora\_fragariae\_CBS20946 7Phytophthora\_rubi\_CBS109892

5Phytophthora\_katsurae\_CBS58785 5Phytophthora\_sp\_novaeguinea\_P1256

4Phytophthora\_arecae\_CBS30562 4Phytophthora\_palmivora\_CBS17926

3Phytophthora\_quercina\_CBS78195 3Phytophthora\_sp\_ohioensis\_P16050

3Phytophthora\_pseudosyringae\_391716 3Phytophthora\_nemorosa\_CBS114870

2Phytophthora\_capsici\_CBS12823 2Phytophthora\_sp\_glovera\_P10618 2Phytophthora\_mexicana\_P0646

2Phytophthora\_sp.\_FFM22.2 2Phytophthora\_sp.\_UK92615 2Phytophthora\_citricola\_CBS22188 2Phytophthora\_acerina\_BO57 2Phytophthora\_pini\_CBS18125

2Phytophthora\_multivora\_DDS3480 2Phytophthora\_multivora\_VHS21011

2Phytophthora\_colocasiae\_P6290 2Phytophthora\_meadii\_CBS21988 2Phytophthora\_aff\_meadii\_CBS23530

2Phytophthora\_capensis\_P1819 2\_Phytophthora\_emanzi\_CMW35510

1Phytophthora\_aff\_infestans\_P13660 1Phytophthora\_infestans\_CBS36651 1Phytophthora\_ipomoeae\_P10225 1Phytophthora\_mirabilis\_CBS67885 1Phytophthora\_sp\_andina\_EC3163 1Phytophthora\_phaseoli\_CBS55688

1Phytophthora\_cactorum\_CBS113344 1Phytophthora\_idaei\_P6767

Table : List of *Phytophthora* ITS1 sequences that cluster with another from the database. More explicitly, these do not exist as singletons and are not able to resolve clustering beyond these clusters. These will lead to false positive identification and thus are a limitation to the method/ sequence used for analysis.

### Unresolved clusters?

Can we refine the clustering by making the clustering threshold more stringent? If we run Swarm with a d0 setting, which is the most strict. We get 138 clusters out of the 154 sequences. The following species cannot be resolved by Swarm ITS1 clustering (Table 5). In conclusion, some can be further resolved, but clade 1C cannot be sub-divided beyond what is stated below:

6Phytophthora\_sp\_nov\_CBS114338 Phytophthora\_taxon\_PgChlamydo\_VHS3753\_EU301160

Phytophthora\_gibbosa\_CBS127951\_HQ012933 Phytophthora\_gregata\_CBS127952\_HQ012942

P.\_aff.\_rosacearum\_DDS2909 P.\_rosacearum\_P292

8Phytophthora\_aff\_primulae\_P6817 8Phytophthora\_primulae\_CBS110162

8Phytophthora\_erythroseptica\_CBS111343 8Phytophthora\_himalayensis\_CBS35759

7Phytophthora\_melonis\_CBS58269 7Phytophthora\_sinensis\_CBS55788

4Phytophthora\_arecae\_CBS30562 4Phytophthora\_palmivora\_CBS17926

3Phytophthora\_quercina\_CBS78195 3Phytophthora\_sp\_ohioensis\_P16050

2Phytophthora\_capsici\_CBS12823 2Phytophthora\_sp\_glovera\_P10618

2Phytophthora\_sp.\_FFM22.2 2Phytophthora\_sp.\_UK92615

2Phytophthora\_multivora\_DDS3480 2Phytophthora\_multivora\_VHS21011

2Phytophthora\_capensis\_P1819 2\_Phytophthora\_emanzi\_CMW35510

1Phytophthora\_aff\_infestans\_P13660 1Phytophthora\_infestans\_CBS36651 1Phytophthora\_ipomoeae\_P10225 1Phytophthora\_mirabilis\_CBS67885 1Phytophthora\_sp\_andina\_EC3163

Table : Names of species which cluster together with the most strict setting with Swarm clustering. Therefore, these cannot be resolved any further than those stated above using Swarm.

Now we have looked into what happens when we cluster to database together with no extra data. What happens when we include know data?

### Test pipeline with known data (PCRMIX\_S96) - previous data:

PCRmix is roughly normalised for the quantity of starting sample.

Using Swarm with a setting of d1.

#### Without error correction:

For a sequencing sample made up of 10 species, the resulting pipeline clustered the ITS sequences with 17 species. This was partly due to the extra clustering of same clade species, e.g 7Phytophthora\_fragariae\_CBS20946 7Phytophthora\_rubi\_CBS109892**. (**Table 5**)**

Of the 10 species put into the same 9 were correctly identified: set(['katsurae', 'palmivora', 'rubi', 'fallax', 'boehmeriae', 'capsici', 'plurivora', 'idaei', 'megasperma']). The species which was not identified was : cryptogea (this does have an ITS represented in the ITS database used for clustering). The following are false positives identified by clustering: set(['mexicana', 'fragariae', 'glovera', 'cactorum', 'citricola', 'novaeguinea', 'kelmania', 'arecae'])

The identification of *P. mexicana, P. novaeguinea, P. kelmania* are true false positives as these are singleton database clusters (did not cluster with other species from the dataset). However, the following were identified due to clusters containing multiple members from the database: *P. fragariae* clusters with *P. rubi, P. glovera* clusters with *P. capsici, P. citricola* clusters with *P. pluruvora, P. arecae* clusters with *P. palmivora, P. cactorum* clusters with *P. idaei*. This can be explained by several reasons: 1) contamination, possible but unlikely. 2) The ITS regions of the species overlap in sequence space. Possible, but very hard to prove. 3) Illumina errors? Let’s try Illumina sequence error correction to see if this alters the false positive rate.

|  |  |
| --- | --- |
| **#species** | **number\_of\_reads\_hitting\_species** |
| 9Phytophthora\_fallax\_P10722 | 241 |
| 10Phytophthora\_boehmeriae\_CBS29129 | 409 |
| 7Phytophthora\_fragariae\_CBS20946 7Phytophthora\_rubi\_CBS109892 | 455 |
| 8Phytophthora\_sp\_kelmania\_CBS30762 | 537 |
| Phytophthora\_megasperma\_IMI133317\_AF266794 | 480 |
| 2Phytophthora\_citricola\_CBS22188 2Phytophthora\_plurivora\_CBS37961 | 318 |
| 2Phytophthora\_capsici\_CBS12823 2Phytophthora\_sp\_glovera\_P10618 2Phytophthora\_mexicana\_P0646 | 452 |
| 4Phytophthora\_arecae\_CBS30562 4Phytophthora\_palmivora\_CBS17926 | 449 |
| 5Phytophthora\_katsurae\_CBS58785 5Phytophthora\_sp\_novaeguinea\_P1256 | 383 |
| 1Phytophthora\_cactorum\_CBS113344 1Phytophthora\_idaei\_P6767 | 464 |
| #Fasta file assembly summary: #min\_contig = 158 max\_contig = 241 avg\_contig = 195 | |
| #Total number of assemblerd sequences = 4595 | |
| #number of reads clustering with Phy = 4188 | |
| #number of starting reads = 5354 | |
| #percent of reads clustering with Phyto = 91.14 | |

Table : Table representing the species and number of reads which clustered with the species when using a known-spiked sample.

#### With error correction

As mentioned above, could error correction alter the false positive rate we observed with this known dataset when clustering with a known ITS1 database (Table 6).

First thing to note is the increase in reads which cluster with the database. Using the same datasets, error correction has increased the number of reads which cluster with the ITS1 database from 91.14% to 96.72%. Has this increase false positive, or decreased false positive clustering?

Error correction increased the number of species identified from 17 to 18. Of which 9 of the 10 “spiked” species where identified. set(['katsurae', 'palmivora', 'rubi', 'fallax', 'boehmeriae', 'megasperm', 'capsici', 'plurivora', 'idaei'])

The species not identified was again: set(['cryptogea']). Therefore error correction did not solve this false negative. A possible explanation for this could be the ITS1 databse, or the amplified ITS1 from this species could have been one of the hundred that is possesses.

**False positive rate:** The identification of *P. mexicana, P. novaeguinea, P. kelmania* are true false positives as these are singleton database clusters (did not cluster with other species from the dataset). However, the following were identified due to clusters containing multiple members from the database: *P. fragariae* clusters with *P. rubi, P. glovera* clusters with *P. capsici, P. citricola* clusters with *P. pluruvora, P. arecae* clusters with *P. palmivora*, *P. cactorum* clusters with *P. idaei*.

There was **one extra false positive** compared to the non-error corrected dataset: *P. tabaci.* This had one read which clustered with it. It would be down to the user to use initiative on whether or not to treat this as a true positive. How many reads does a species have to have in order for it to truly be positive?

|  |  |
| --- | --- |
| #species | number\_of\_reads\_hitting\_species |
| 9Phytophthora\_fallax\_P10722 | 260 |
| 10Phytophthora\_boehmeriae\_CBS29129 | 443 |
| 7Phytophthora\_fragariae\_CBS20946 7Phytophthora\_rubi\_CBS109892 | 487 |
| 8Phytophthora\_sp\_kelmania\_CBS30762 | 563 |
| Phytophthora\_megasperma\_IMI133317\_AF266794 | 509 |
| 2Phytophthora\_citricola\_CBS22188 2Phytophthora\_plurivora\_CBS37961 | 322 |
| 2Phytophthora\_capsici\_CBS12823 2Phytophthora\_sp\_glovera\_P10618 2Phytophthora\_mexicana\_P0646 | 461 |
| 4Phytophthora\_arecae\_CBS30562 4Phytophthora\_palmivora\_CBS17926 | 496 |
| 5Phytophthora\_katsurae\_CBS58785 5Phytophthora\_sp\_novaeguinea\_P1256 | 406 |
| **1Phytophthora\_tabaci\_CBS30529** | **1** |
| 1Phytophthora\_cactorum\_CBS113344 1Phytophthora\_idaei\_P6767 | 503 |
| #Fasta file assembly summary: #min\_contig = 158 max\_contig = 223 avg\_contig = 195 | |
| #Total number of assemblerd sequences = 4602 | |
| #number of reads clustering with Phy = 4451 | |
| #number of starting reads = 5354 | |
| #percent of reads clustering with Phyto = 96.72 | |

Table : Table representing the species and number of reads which clustered with the species when using a error corrected known-spiked sample

### Test pipeline with known data (DNAMIX\_S96) - previous data:

DNAmix is NOT normalised for the quantity of starting sample (unlike PCRmix96). It does however contain the same “spike” species.

### Testing on an unknown data set: IGB\_300913\_S42

Using Swarm with a setting of d1.

#### With and without Error correction:

Table 7 summarise the results for IGB\_300913\_S42 with error correction, Table 8 without error correction. Error correction increases the number of reads that cluster with Phytophthora ITS1 sequences. There was no difference in the species identified in the error corrected and non-error corrected dataset. The only difference was the number of reads.

##### With error correction:

|  |  |
| --- | --- |
| #species | number\_of\_reads\_hitting\_species |
| 7Phytophthora\_cinnamomi\_var\_robiniae\_P16351 7Phytophthora\_sojae\_CBS38261 | 55 |
| 7Phytophthora\_fragariae\_CBS20946 7Phytophthora\_rubi\_CBS109892 | 96 |
| 7Phytophthora\_cambivora\_P0592 | 132 |
| 3Phytophthora\_quercina\_CBS78195 3Phytophthora\_sp\_ohioensis\_P16050 | 288 |
| 8Phytophthora\_cryptogea\_CBS11319 | 11 |
| 8Phytophthora\_aff\_primulae\_P6817 8Phytophthora\_primulae\_CBS110162 | 33 |
| Phytophthora\_gonapodyides\_H\_14\_02\_JF912516 | 946 |
| 6Phytophthora\_sp\_nov\_CBS114338 Phytophthora\_taxon\_PgChlamydo\_VHS3753\_EU301160 | 211 |
| P.\_riparia\_VI\_3100B9F\_HM004225 Phytophthora\_lacustris\_P245\_AF266793 | 1551 |
| Phytophthora\_bilorbang\_CBS161653\_JQ256377 | 4 |
| 8Phytophthora\_syringae\_CBS110161 | 483 |
| 8Phytophthora\_hibernalis\_CBS119904 | 23 |
| 2Phytophthora\_citricola\_CBS22188 2Phytophthora\_plurivora\_CBS37961 | 192 |
| 1Phytophthora\_aff\_infestans\_P13660 1Phytophthora\_infestans\_CBS36651 1Phytophthora\_ipomoeae\_P10225 1Phytophthora\_mirabilis\_CBS67885 1Phytophthora\_sp\_andina\_EC3163 1Phytophthora\_phaseoli\_CBS55688 | 14 |
| 1Phytophthora\_cactorum\_CBS113344 1Phytophthora\_idaei\_P6767 | 32 |
| #Fasta file assembly summary: #min\_contig = 136 max\_contig = 324 avg\_contig = 216 |  |
| #Total number of assemblerd sequences = 6369 |  |
| #number of reads clustering with Phy = 4071 |  |
| #number of starting reads = 7265 |  |
| #percent of reads clustering with Phyto = 63.92 |  |

Table : Resulting obtained for clustering IGB\_300913\_S42 with the known ITS1 database. With error correction

##### Without error correction:

|  |  |
| --- | --- |
| #species | number\_of\_reads\_hitting\_species |
| 7Phytophthora\_cinnamomi\_var\_robiniae\_P16351 7Phytophthora\_sojae\_CBS38261 | 49 |
| 7Phytophthora\_fragariae\_CBS20946 7Phytophthora\_rubi\_CBS109892 | 91 |
| 7Phytophthora\_cambivora\_P0592 | 120 |
| 3Phytophthora\_quercina\_CBS78195 3Phytophthora\_sp\_ohioensis\_P16050 | 260 |
| 8Phytophthora\_cryptogea\_CBS11319 | 9 |
| 8Phytophthora\_aff\_primulae\_P6817 8Phytophthora\_primulae\_CBS110162 | 33 |
| Phytophthora\_gonapodyides\_H\_14\_02\_JF912516 | 891 |
| 6Phytophthora\_sp\_nov\_CBS114338 Phytophthora\_taxon\_PgChlamydo\_VHS3753\_EU301160 | 200 |
| P.\_riparia\_VI\_3100B9F\_HM004225 | 8 |
| Phytophthora\_lacustris\_P245\_AF266793 | 1467 |
| Phytophthora\_bilorbang\_CBS161653\_JQ256377 | 2 |
| 8Phytophthora\_syringae\_CBS110161 | 464 |
| 8Phytophthora\_hibernalis\_CBS119904 | 23 |
| 2Phytophthora\_citricola\_CBS22188 2Phytophthora\_plurivora\_CBS37961 | 187 |
| 1Phytophthora\_aff\_infestans\_P13660 1Phytophthora\_infestans\_CBS36651 1Phytophthora\_ipomoeae\_P10225 1Phytophthora\_mirabilis\_CBS67885 1Phytophthora\_sp\_andina\_EC3163 1Phytophthora\_phaseoli\_CBS55688 | 14 |
| 1Phytophthora\_cactorum\_CBS113344 1Phytophthora\_idaei\_P6767 | 28 |
| #Fasta file assembly summary: #min\_contig = 136 max\_contig = 324 avg\_contig = 216 | |
| #Total number of assemblerd sequences = 6369 |  |
| #number of reads clustering with Phy = 3846 |  |
| #number of starting reads = 7265 |  |
| #percent of reads clustering with Phyto = 60.39 |  |

Table : Resulting obtained for clustering IGB\_300913\_S42 with the known ITS1 database. Without error correction

# References

Andrews S 2010. FastQC: A quality control tool for high throughput sequence data. Reference Source.

Bolger A, Giorgi F Trimmomatic: A Flexible Read Trimming Tool for Illumina NGS Data. URL <http://www>. usadellab. org/cms/index. php.

Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL 2009. BLAST+: architecture and applications. BMC bioinformatics 10: 421.

Dannemiller KC, Reeves D, Bibby K, Yamamoto N, Peccia J 2014. Fungal High‐throughput Taxonomic Identification tool for use with Next‐Generation Sequencing (FHiTINGS). Journal of basic microbiology 54: 315-321.

Delgado-Serrano L, Restrepo S, Bustos JR, Zambrano MM, Anzola JM 2016. Mycofier: a new machine learning-based classifier for fungal ITS sequences. BMC Research Notes 9: 402.

Finn RD, Clements J, Eddy SR 2011. HMMER web server: interactive sequence similarity searching. Nucleic acids research 39: W29-W37.

Gweon HS, Oliver A, Taylor J, Booth T, Gibbs M, Read DS, Griffiths RI, Schonrogge K 2015. PIPITS: an automated pipeline for analyses of fungal internal transcribed spacer sequences from the Illumina sequencing platform. Methods in Ecology and Evolution 6: 973-980.

Langmead B 2010. Aligning short sequencing reads with Bowtie. Current protocols in bioinformatics: 11.17. 11-11.17. 14.

Li W, Godzik A 2006. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics 22: 1658-1659.

Magoč T, Salzberg SL 2011. FLASH: fast length adjustment of short reads to improve genome assemblies. Bioinformatics 27: 2957-2963.

Mahé F, Rognes T, Quince C, De Vargas C, Dunthorn M 2015. Swarm v2: highly-scalable and high-resolution amplicon clustering. PeerJ 3: e1420.

Quinlan AR, Hall IM 2010. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics 26: 841-842.

Scibetta S, Schena L, Chimento A, Cacciola SO, Cooke DE 2012. A molecular method to assess Phytophthora diversity in environmental samples. Journal of microbiological methods 88: 356-368.

Simão FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM 2015. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics: btv351.

Zhang J, Kobert K, Flouri T, Stamatakis A 2014. PEAR: a fast and accurate Illumina Paired-End reAd mergeR. Bioinformatics 30: 614-620.