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

cmd="makeblastdb -in ${genome\_prefix}\*.fa -dbtype nucl"

cmd2="blastn -query ${repository\_path}/Phy\_ITSregions\_all\_20160601.fasta -db ${genome\_prefix}.fa -outfmt 6 -out n.ITS\_vs\_${genome\_prefix}.out"

cmd\_python\_ITS="python ${repository\_path}/generate\_ITS\_GFF.py --blast n.ITS\_vs\_${genome\_prefix}.out --prefix ${genome\_prefix} -o ${genome\_prefix}.ITS.GFF"

### 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.

hmmsearch --cut\_ga --domtblout ${genome\_prefix}.gap4000.refine.hmm.tab ${repository\_path}/ITS\_remove\_redundancy/Phy\_ITSnr100.muscle.gap4000.refine.hmm ${genome\_prefix}.fa

hmmsearch --cut\_ga --domtblout ${genome\_prefix}.Phy\_ITSregions\_all\_20160601\_nr100.fasta.aligned.refine.tab ${repository\_path}/ITS\_remove\_redundancy/Phy\_ITSregions\_all\_20160601\_nr100.fasta.aligned.refine.hmm ${genome\_prefix}.fa

### 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

cmd\_python\_BUSCO="python3 ${HOME}/scratch/Downloads/BUSCO\_v1.1b1/BUSCO\_v1.1b1.py -in ${genome\_prefix}.fa -l ../LINEAGE/eukaryota -o busco -m genome -f -Z 827000000 --long --cpu ${num\_threads}"

### 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!!

cmd\_trimming="java -jar ${trimmomatic\_path}/trimmomatic-0.32.jar PE -threads ${num\_threads} -phred33 ${SRA\_prefix}\_1.fastq.gz ${SRA\_prefix}\_2.fastq.gz R1.fq.gz unpaired\_R1.fq.gz R2.fq.gz unpaired\_R2.fq.gz ILLUMINACLIP:${repository\_path}/TruSeq3-PE.fa:2:30:10 LEADING:3 TRAILING:3 SLIDINGWINDOW:4:15 MINLEN:45"

cmd\_index="bowtie2-build --quiet -f ${genome\_prefix}.fa bowtie\_index\_files"

cmd\_mapping="bowtie2 --very-sensitive --non-deterministic --seed 1 --no-mixed --no-unal -p ${num\_threads} -x bowtie\_index\_files -1 R1.fq.gz -2 R2.fq.gz | samtools view -@ ${num\_threads} -S -b -o $TMP/tmp\_unsorted.bam -"

cmd\_sort="samtools sort -@ ${num\_threads} $TMP/tmp\_unsorted.bam ${genome\_prefix}"

samtools index ${genome\_prefix}.bam

cmd\_Busco\_count="bedtools multicov -bams ${genome\_prefix}.bam -bed ${genome\_prefix}\_BUSCO\_GENES.gene.gff > ${genome\_prefix}\_BUSCO\_GENES.gene.gff.genes.cov"

cmd\_count="bedtools multicov -bams ${genome\_prefix}.bam -bed ${genome\_prefix}.gene.gff > ${genome\_prefix}\_genomic.genes.cov"

cmd\_ITS\_count="bedtools multicov -bams ${genome\_prefix}.bam -bed ${genome\_prefix}.ITS.consensus.GFF > ${genome\_prefix}\_genomic.ITS.cov"

summary stats:

python ${repository\_path}/summary\_stats.py --ITS ${genome\_prefix}\_genomic.ITS.cov.values --GFF ${genome\_prefix}.ITS.consensus.GFF --all\_genes\_cov ${genome\_prefix}\_genomic.genes.cov.values -o ${genome\_prefix}\_stats\_all\_genes\_versus\_ITS.out

python ${repository\_path}/summary\_stats.py --ITS ${genome\_prefix}\_genomic.ITS.cov.values --GFF ${genome\_prefix}.ITS.consensus.GFF --all\_genes\_cov ${genome\_prefix}\_BUSCO\_GENES.gene.gff.genes.cov.values -o ${genome\_prefix}\_stats\_BUSCO\_versus\_ITS.out

## 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).

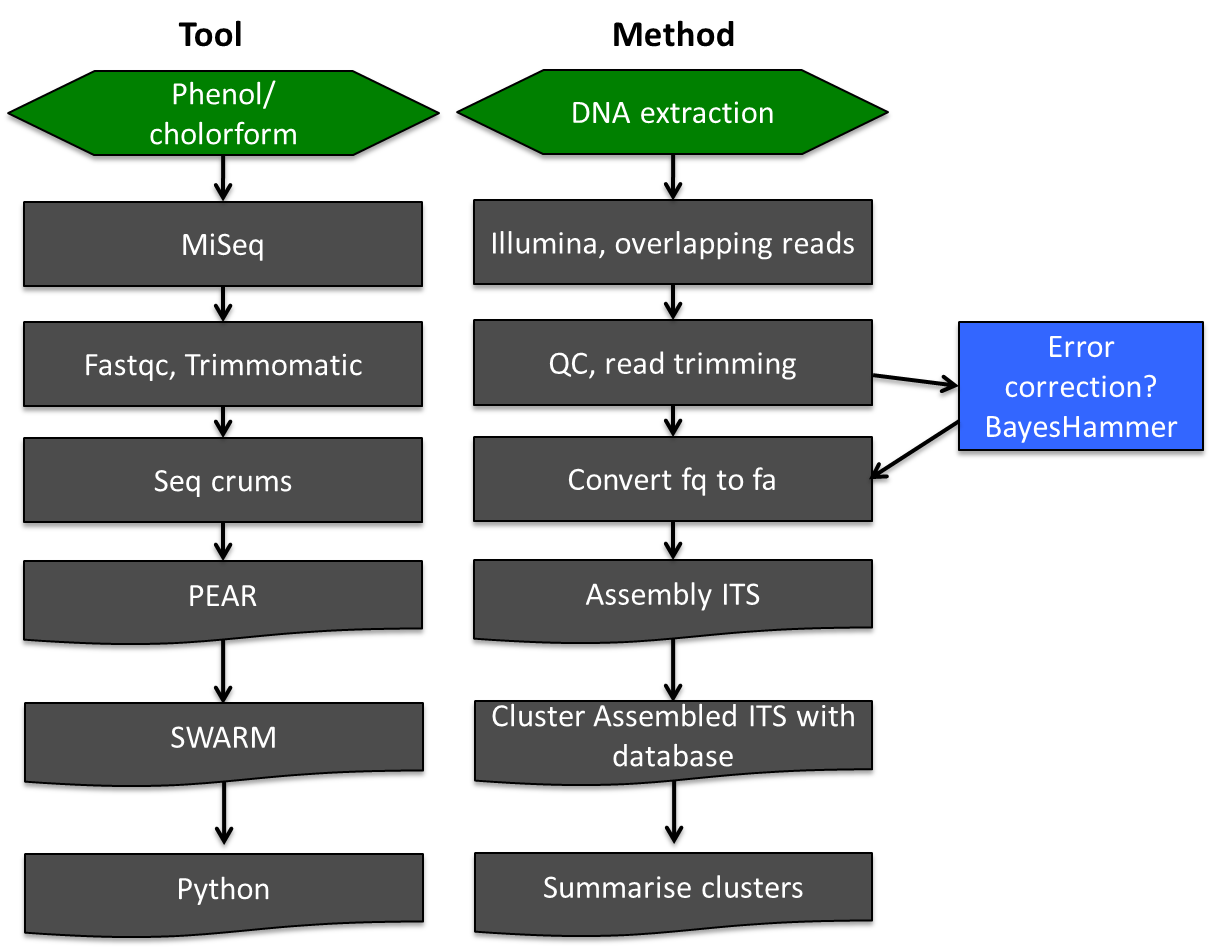


Figure : Diagrammatic representation of the tasks and tools needed in order to perform species identification when given ITS1 sequences.

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.

# 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 1 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 2).

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | ***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.

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 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.

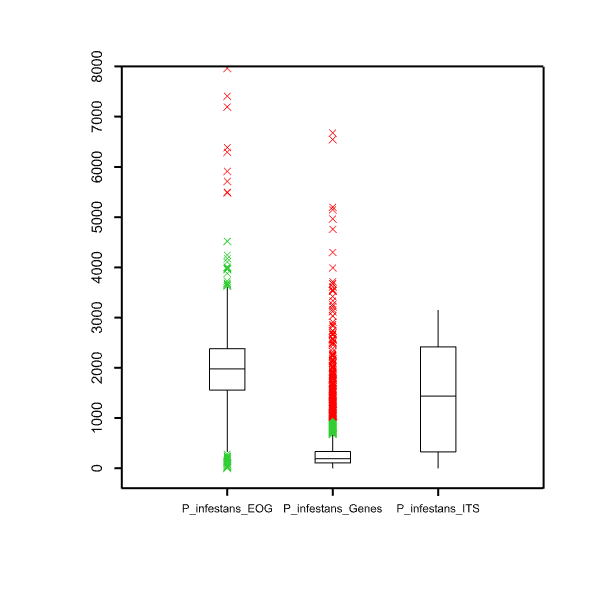


Figure : Boxplot representing genomic reads coverage for each class of genes from the *Phytophthora infestans* T30-4 genome. The ITS genes compared to all gene have greater coverage, implying greater copy number.

## 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.

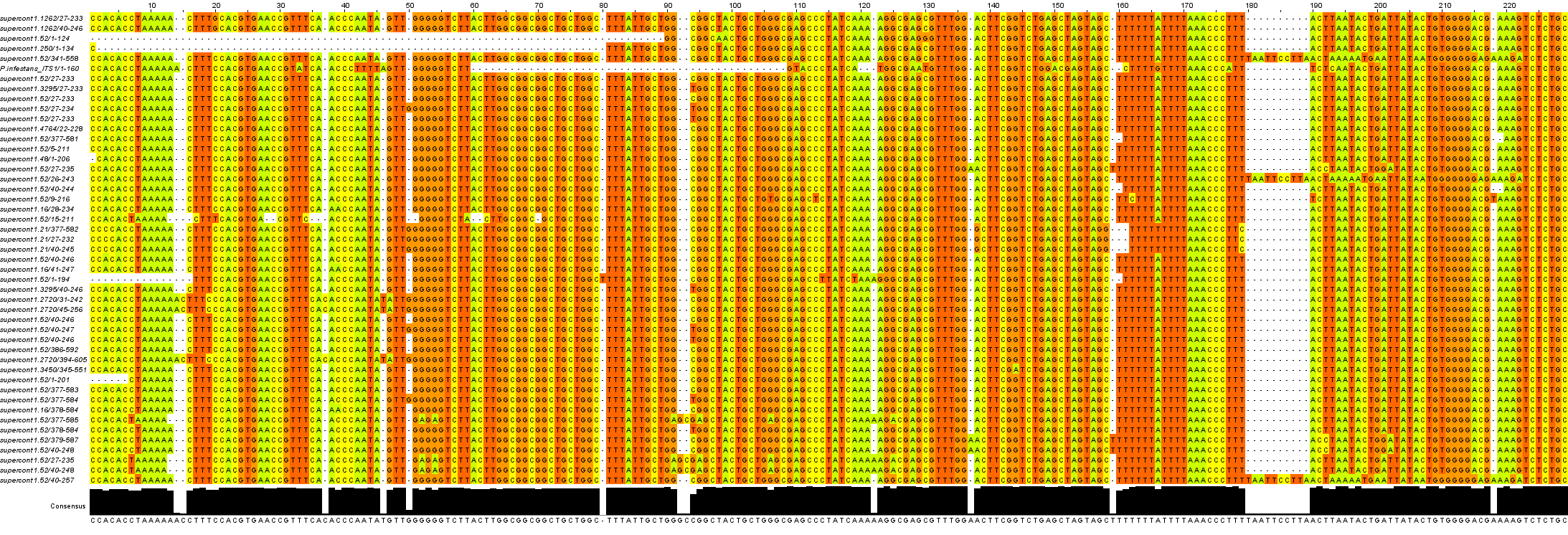


Figure : Alignment of ITS1 regions found with a single genome, *P. infestans* T30-4. Variation is found, including indels.

**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. 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

# 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.