

Phytophthora genomes

05/10/2016

Contents

Public Data	1
Genomes	1
Bioprojects	2
Short reads	3
Annotation	3
Ongoing projects	3
Which genomes	4
ANI: 17 ‘species’ groups.	4
Phylogeny & clades	5
Genomic Points of Interest	6
Fragmented	6
Variable size	7
Expansion	7
Two-speed genome	8
Molecular interactions	9
Host resistance	9
Two-speed genome	9
Modular effectors	11
Functional Differences	11
Metabolic model	11
Map differences	12
Niche association	13

Public Data

Genomes

- 59 assemblies at NCBI: 21 ‘species’

NCBI Resources ▾ How To ▾

Assembly [Create alert](#) [Advanced](#) [Browse by organism](#)

Organism group
 Protists (59)
 Customize ...

Status
☒ Latest (59)
 Latest GenBank (59)
 Latest RefSeq (3)

Assembly level
 Scaffold (52)
 Contig (7)

RefSeq category
 Representative (21)

Display Settings: ▾ Summary, 20 per page, Sorted by Default

Search results
 Items: 1 to 20 of 59

Filters activated: Latest, Exclude anomalous. [Clear all](#) to show 77 items.

☐ [ASM14294v1](#)

1. **Phytophthora infestans** T30-4 (potato late blight agent)
 Intraspecific name: Strain: T30-4
 Submitter: Broad Institute
 Date: 2009/06/17
 Assembly level: Scaffold
 Genome representation: full
 RefSeq category: representative genome
 GenBank assembly accession: GCA_000142945.1 (latest)
 RefSeq assembly accession: GCF_000142945.1 (latest)
 IDs: 165248 [UID] 165248 [GenBank] 1009248 [RefSeq]

Bioprojects

- 133 bioprojects at NCBI: “statements of intent” (isolate and higher)

NCBI Resources ▾ How To ▾

BioProject [Create alert](#) [Advanced](#)

Project Types
 Umbrella (2)
 Primary submission (128)
 RefSeq (3)

Data Types
 Epigenomics (1)
 Genome sequencing (58)
 Other (11)
 Transcriptome (26)

Project Data
 Nucleotide (63)
 Protein (16)
 Assembly (61)
 SRA (63)
 GEO DataSets (18)

Display Settings: ▾ Summary, 20 per page, Sorted by Default order

[See also 21 genomes matching your organism search](#)

Search results
 Items: 1 to 20 of 133

☐ [Phytophthora capsici](#)

1. **Phytophthora capsici** Genome sequencing
 Taxonomy: [Phytophthora capsici](#)
 Project data type: Genome sequencing
 Scope: Monoisolate
 DOE Joint Genome Institute
 Accession: PRJNA340629 ID: 340629

☐ [Phytophthora sojae](#)

2. **Phytophthora sojae** Genome sequencing
 Taxonomy: [Phytophthora sojae](#)
 Project data type: Genome sequencing

Short reads

- 630 short read sets at NCBI

The screenshot shows the NCBI SRA search interface. The search term is "Phytophthora[Organism]". The results show 630 public accessions. The first result is "Genome sequencing of *Phytophthora ramorum* lineage EU1 isolate CC1048" with 1 ILLUMINA (Illumina HiSeq 2000) run: 19.2M spots, 2.8G bases, 1.5Gb downloads. The accession is SRX2200946.

Annotation

- 16 assemblies at Ensembl Protists
- Comparative genomics; ($9 \times P. parasitica$)

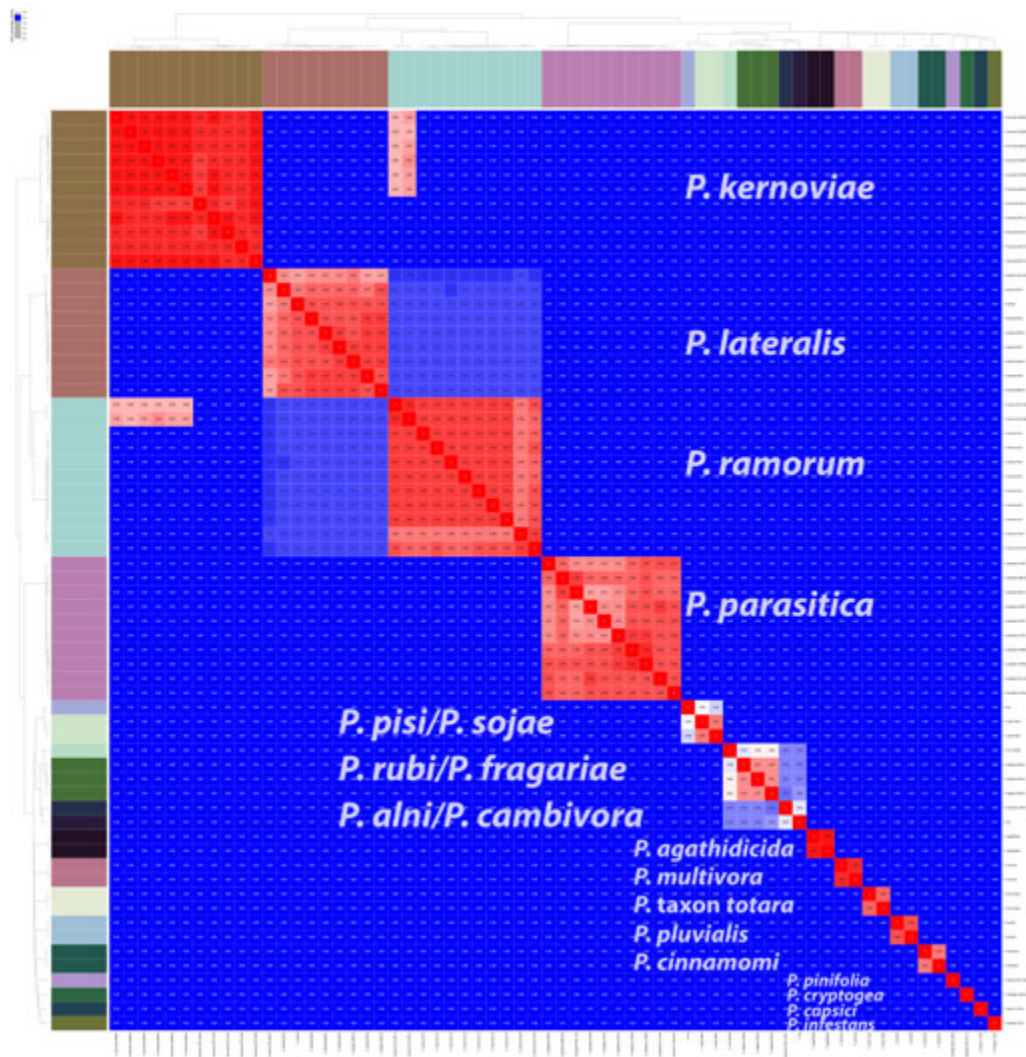
Name	Classification	Taxon ID	Assembly	Accession	Variation database	Regulation database	Whole genome alignments	Other alignments	In peptide compare
Phytophthora infestans	Stramenopiles	60677	AGP1425v1	GCA_00112548.1	✓	✓	✓	✓	✓
Phytophthora kernoviae	Stramenopiles	284330	GCA_000303075.1	GCA_000303075.1	✓	✓	✓	✓	✓
Phytophthora lateralis	Stramenopiles	284330	GCA_000318485.1	GCA_000318485.1	✓	✓	✓	✓	✓
Phytophthora nicotianae	Stramenopiles	179	AGP14001v1	GCA_00140013.1	✓	✓	✓	✓	✓
Phytophthora nicotianae (GCA_001402985.1)	Stramenopiles	179	AGP14001v1	GCA_001402985.1	✓	✓	✓	✓	✓
Phytophthora parasitica	Stramenopiles	317000	GCA_000303075.1	GCA_000303075.1	✓	✓	✓	✓	✓
Phytophthora parasitica (GCA_000509466.1)	Stramenopiles	179	Phyt_pam_GCA000509466.1	GCA_000509466.1	✓	✓	✓	✓	✓

Ongoing projects

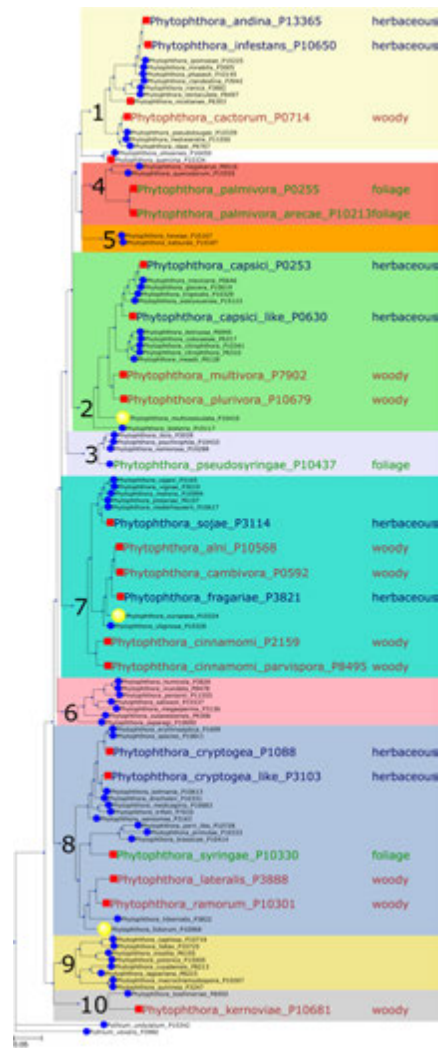
- We don't have full scope, but...
- Grunwald lab: *P. syringae*, *P. ramorum*
- Mayr-Stihl-Stiftung: *P. plurivora*, *P. quercina*
- and no doubt others...

Which genomes

ANI: 17 'species' groups.

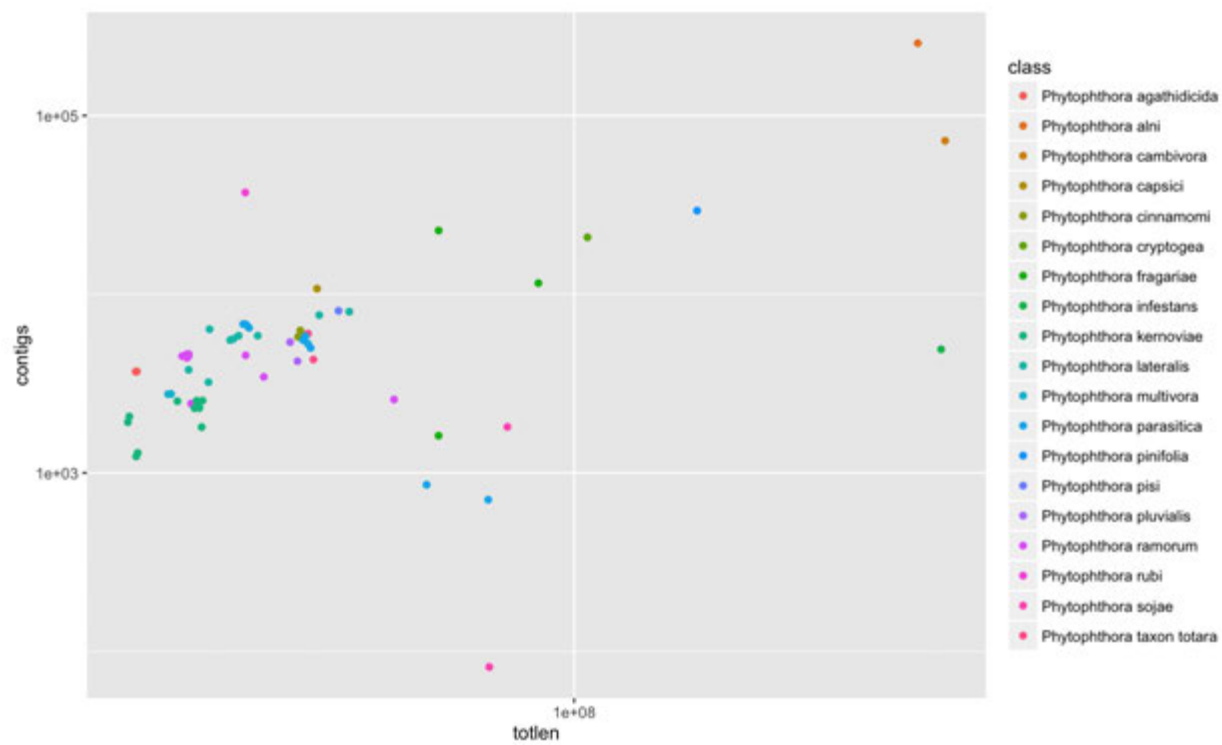


Phylogeny & clades

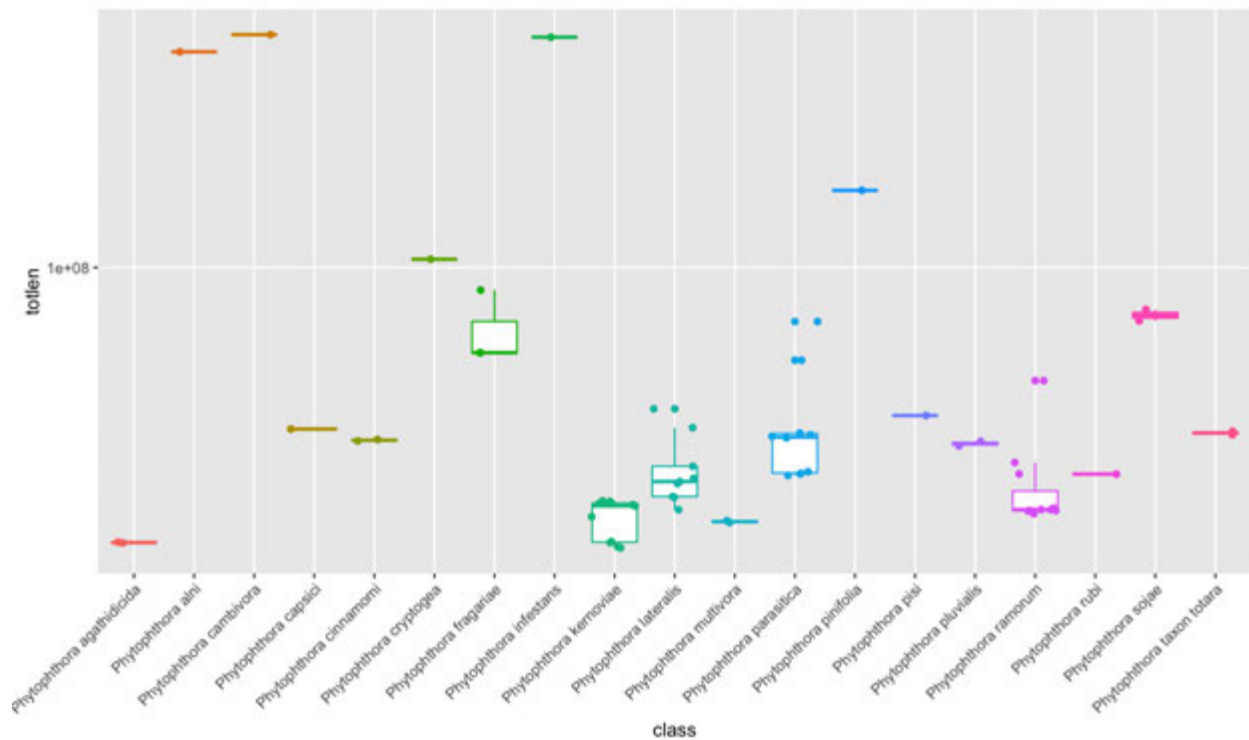


Genomic Points of Interest

Fragmented



Variable size



Expansion

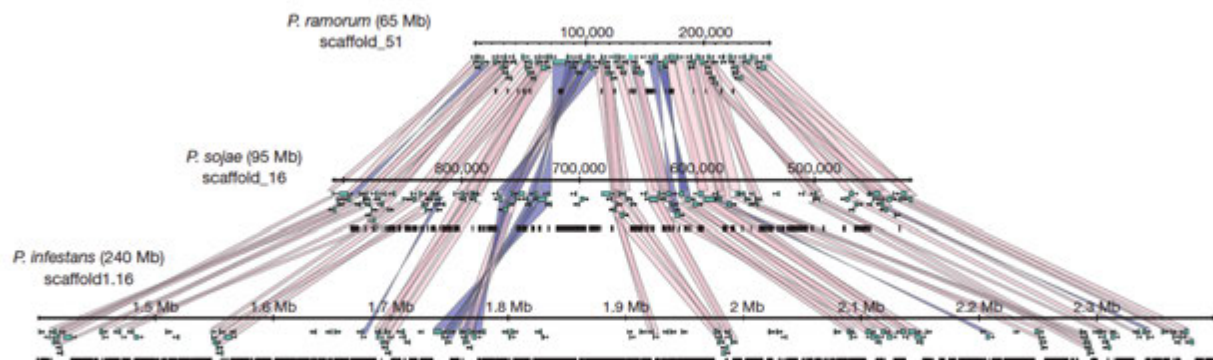
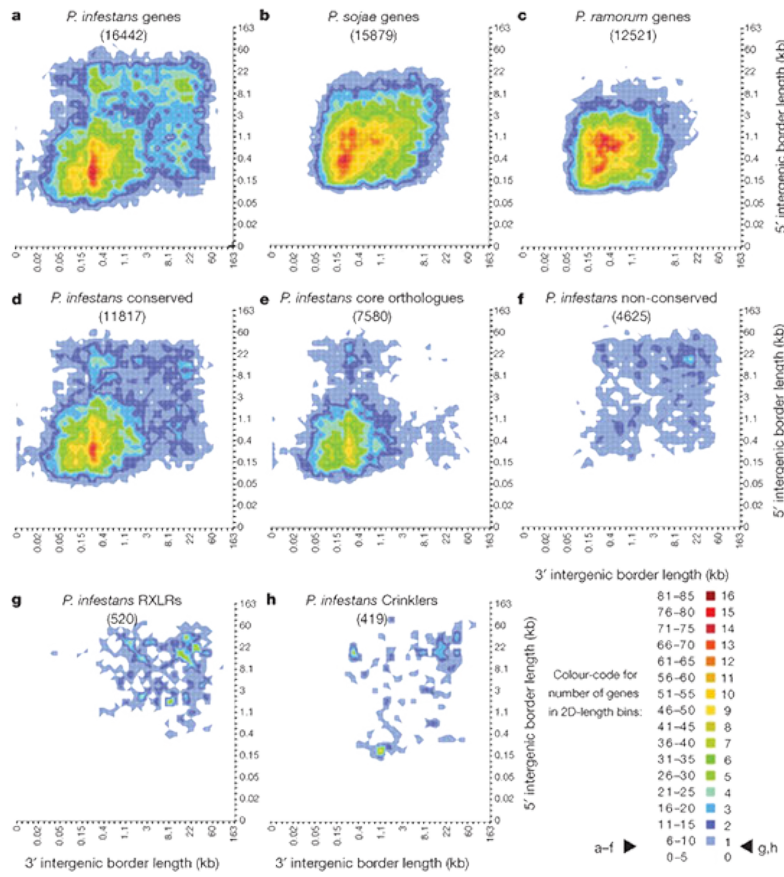


Figure 1 | Repeat-driven genome expansion in *Phytophthora infestans*. Conserved gene order across three homologous *Phytophthora* scaffolds. Genome expansion is evident in regions of conserved gene order, a

consequence of repeat expansion in intergenic regions. Genes are shown as turquoise boxes, repeats as black boxes. Collinear orthologous gene pairs are connected by pink (direct) or blue (inverted) bands.

- Haas *et al.* (2009) doi: 10.1038/nature08358

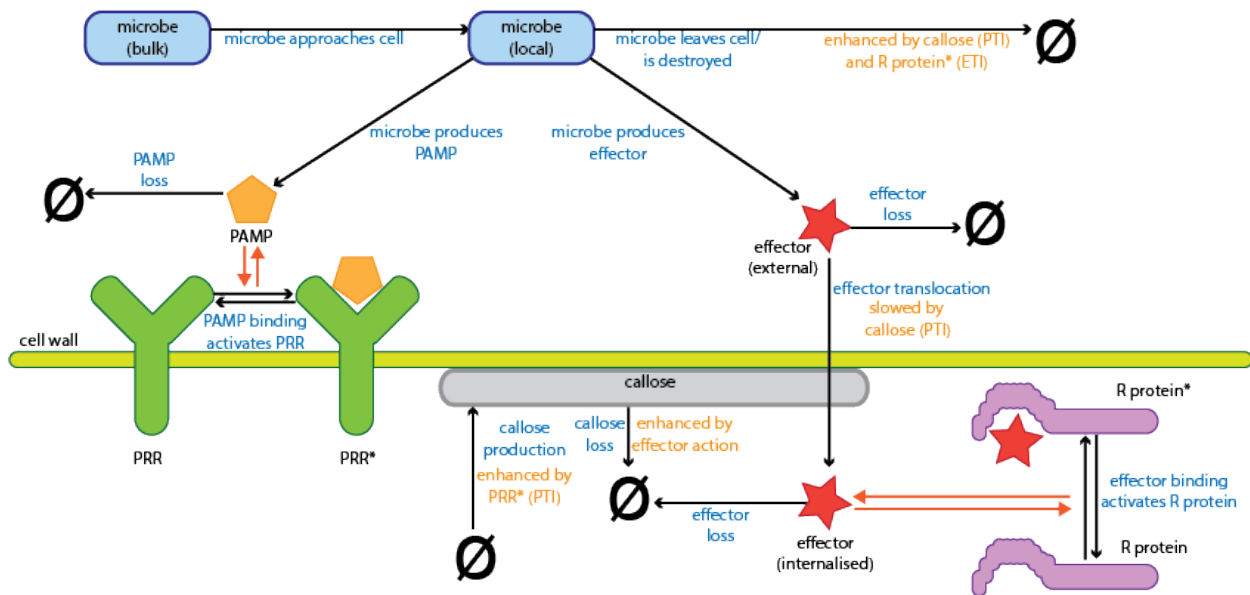
Two-speed genome



- Haas *et al.* (2009) doi: 10.1038/nature08358

Molecular interactions

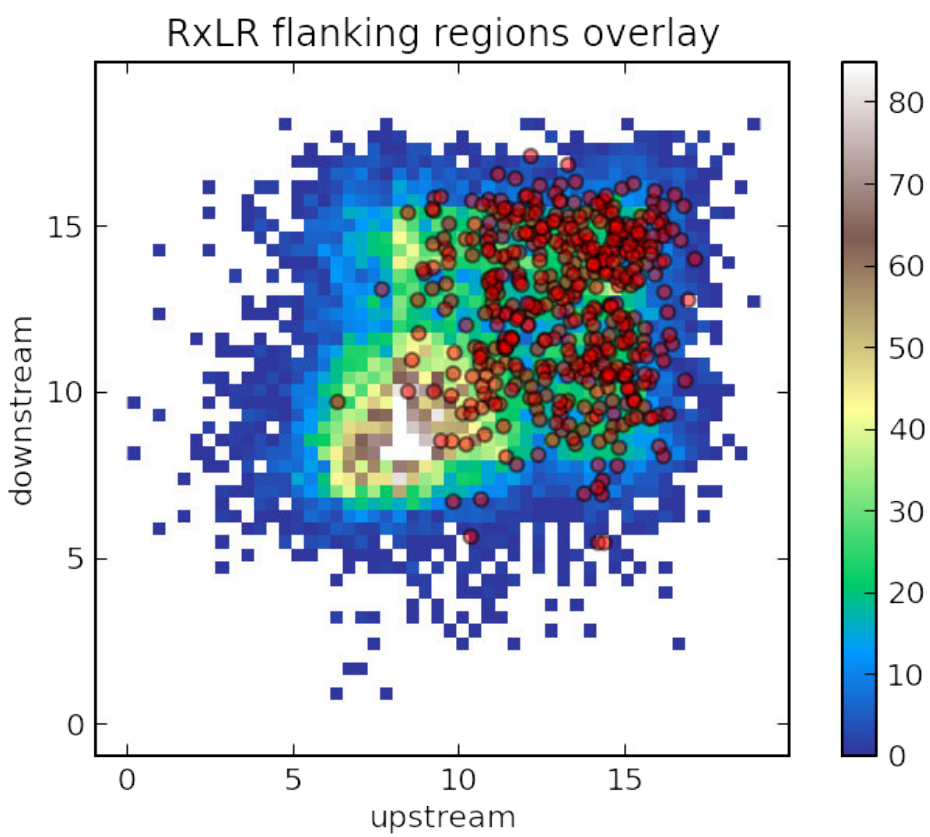
Host resistance



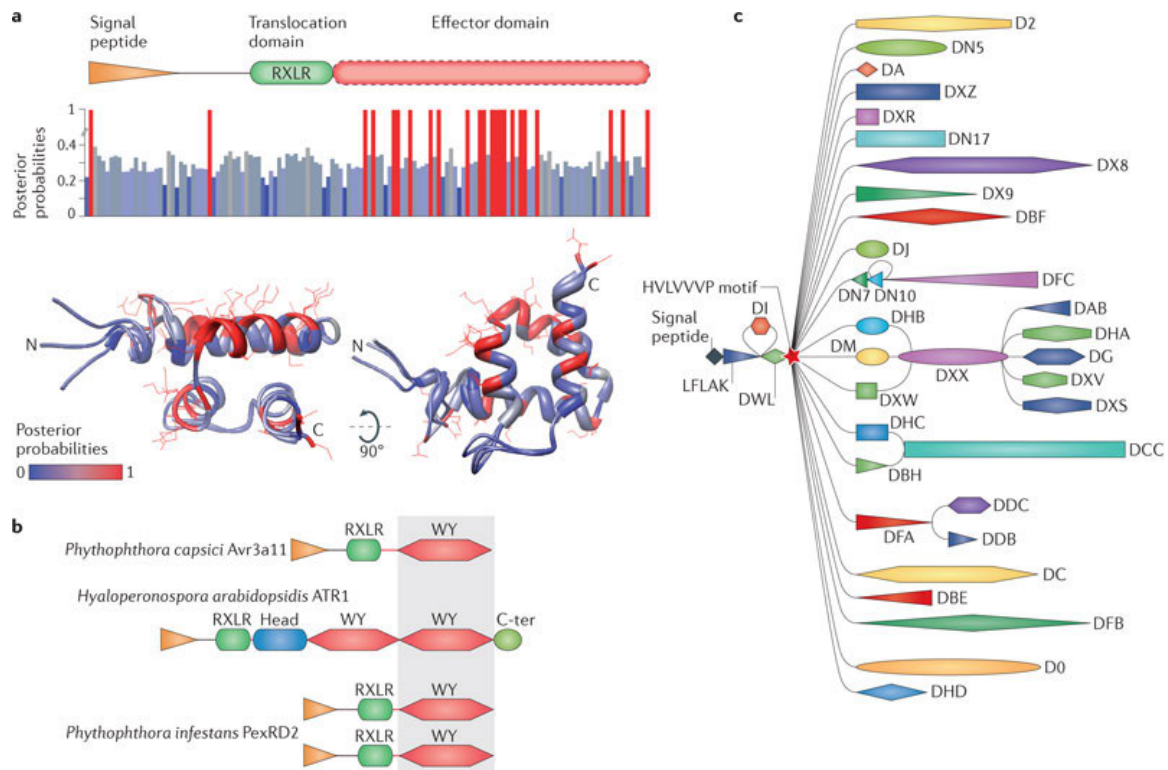
- Pritchard & Birch (2014) doi: 10.1111/mpp.12210

Two-speed genome

Effectors (RXLR) preferentially in 'fast' regions: diversity



Modular effectors



Nature Reviews | Microbiology

- Raffaele & Kamoun (2014) doi: 10.1038/nrmicro2790

Functional Differences

Metabolic model

- Whole-genome reconstruction: BioModels

EMBL-EBI

BioModels Database

BioModels Home Models Submit Support About BioModels Contact us

Whole Genome Metabolism - *Phytophthora infestans* (strain T30-4)

Download SBML Send feedback

Model information

Identifier: BMID000000140564
Format: SBML L2 V4

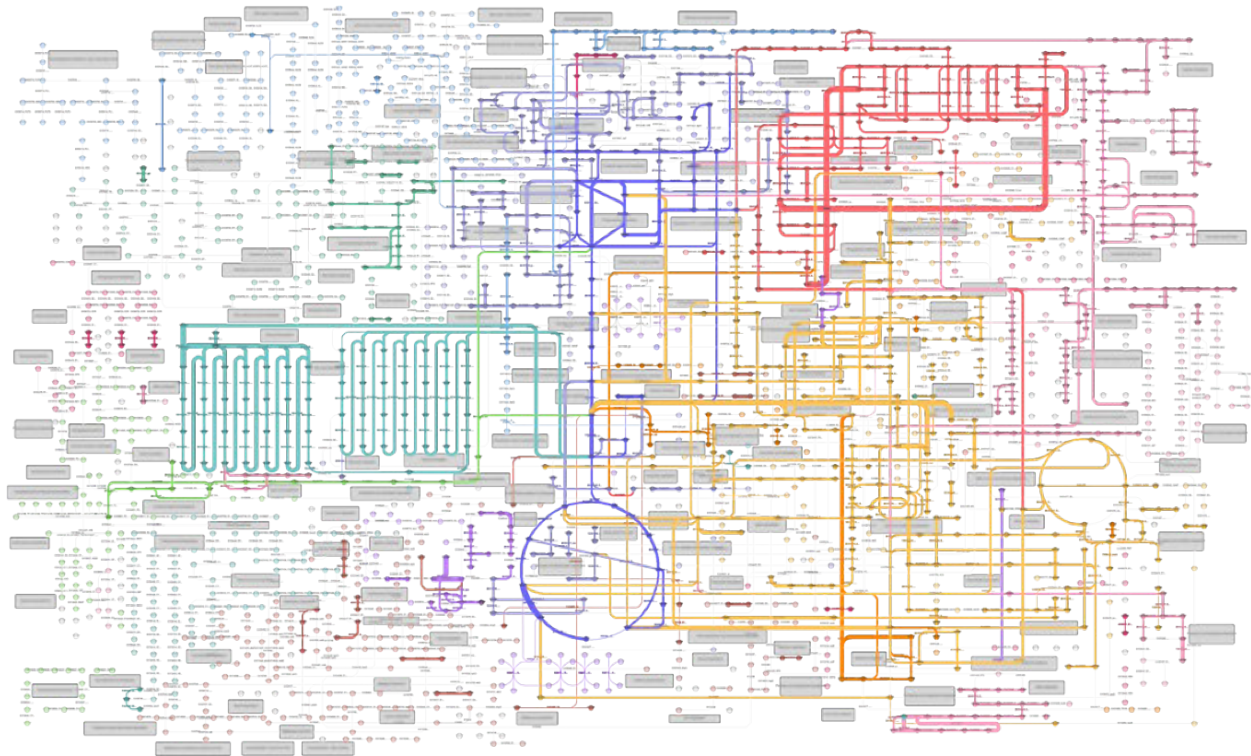
Project: [path2models](#)
Categories: [genome-scale](#)

Submission: 19 May 2012 15:24:58 UTC
Last modified: 04 Oct 2013 01:37:19 UTC
Published: 19 May 2012 23:49:21 UTC

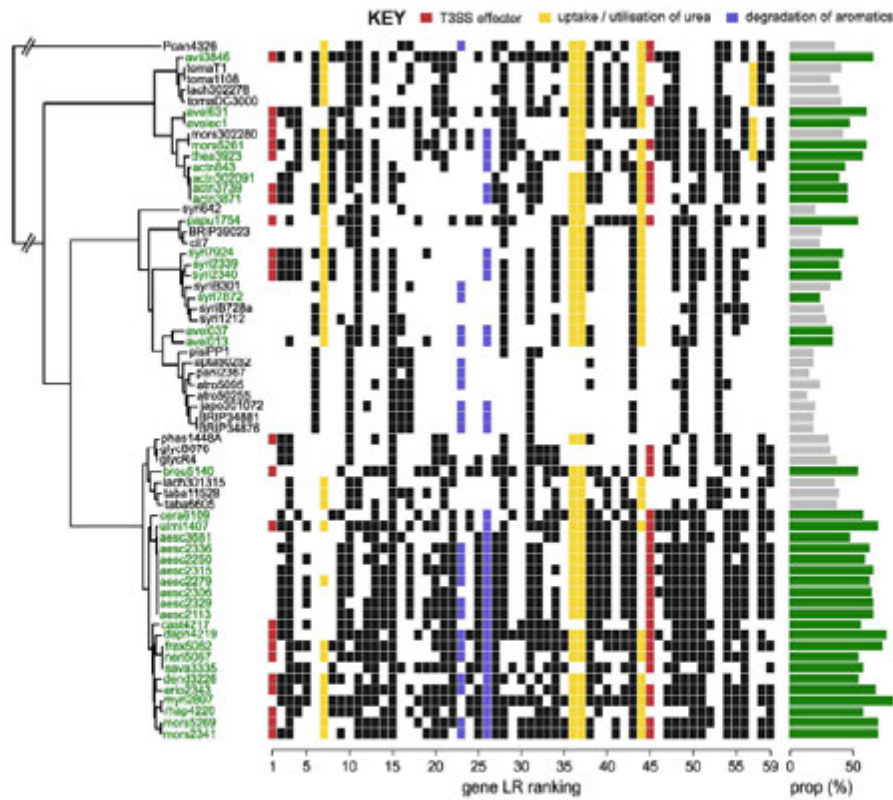
Annotations

occursIn [Phytophthora infestans \(strain T30-4\)](#)
isDerivedFrom [Phytophthora infestans](#)
isDescribedBy [Genome sequence and analysis of the Irish potato famine pathogen *Phytophthora infestans*](#)

Map differences



Niche association



Occurrence profile for 59 genes significantly associated with the woody niche. Genes of particular interest are highlighted in colour (see key). Genes are ordered from 1 to 59 corresponding to the magnitude of the LR statistic (decreasing significance); the order of genes is not indicative of physical proximity on the chromosome. Strains are ordered according to the core genome phylogeny; the bar chart at the right shows the proportion of genes (out of 59) present in woody (green) versus non-woody (grey) host pathogens.

- Sharp *et al.* (2016) doi: 10.1111/mpp.12423

ITS1 classification

Leighton Pritchard, Peter Thorpe

05/10/2016

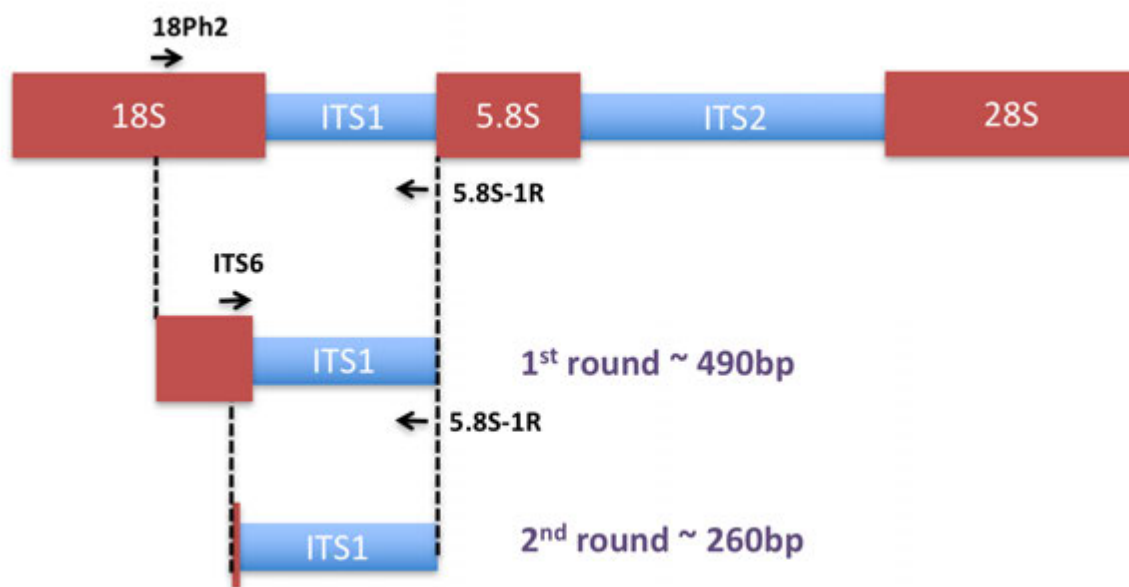
What is the goal?

Broadly-speaking...

Identify and profile *Phytophthora* present in a nursery

1. Sequence sample from nursery
2. Put reads into “black box”
3. Be told which *Phytophthora* spp. are present (and in what relative quantity)

Amplifying ITS1



So what's in the box?

Given a set of reads...

- Assemble reads into complete ITS1 regions
- Compare to known ITS1 sequences (supervised) to identify originating species. Needs reliable database.
- Apply clustering (unsupervised) to identify originating species/novel OTUs
- Use abundance to estimate relative quantity of originating species

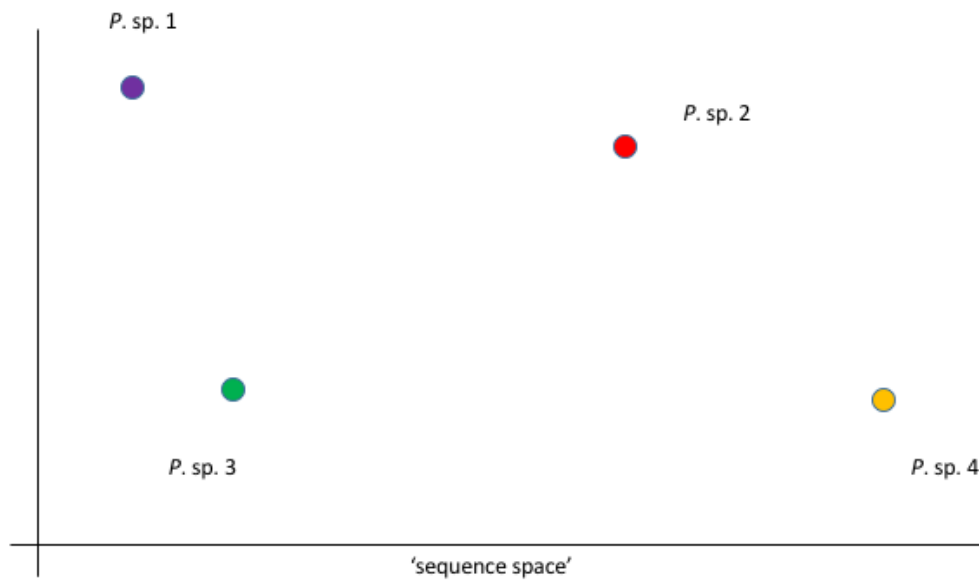
Clustering ITS1

Sequence survey 1

Are either of these statements true? (not accounting for primer choice...)

- There is only one ITS1 sequence per species?
- There is only one ITS1 sequence per isolate?

One ITS1 per species

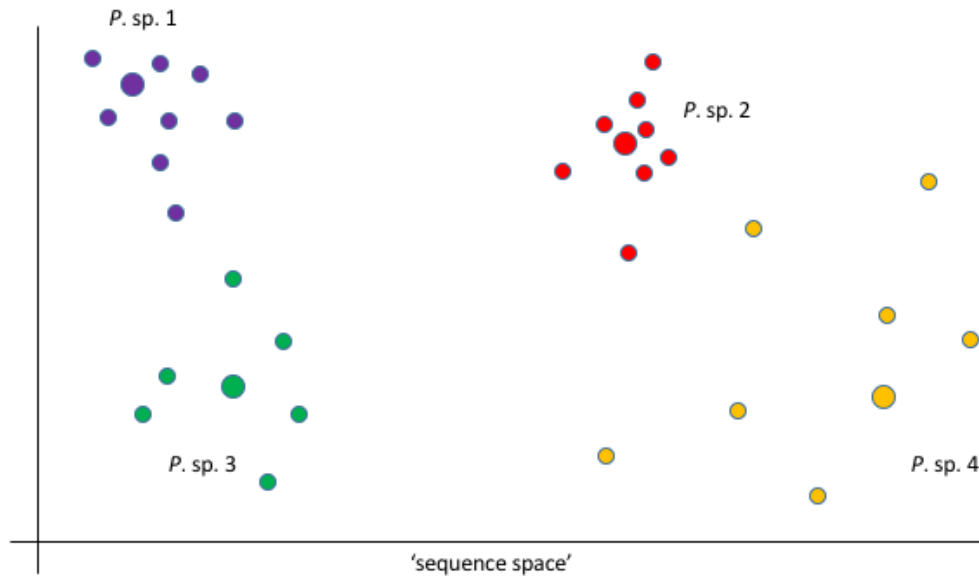


>1 ITS1 per species



NJ tree of ITS1 regions from cocoa-associated *Phytophthora* spp. Appiah *et al.* (2004) *Plant Path.* doi: 10.1111/j.0032-0862.2004.00980.x

One ITS1 per isolate



How many ITS1 per isolate?

1. Identify 'canonical' ITS1 sequence
2. Query (BLAST/HMMer) against sequenced genomes
3. Count matches

Six sequenced genomes

- Distinct BLAST hits with ITS1 query
 - *P. infestans*: 133
 - *P. sojae*: 7
 - *P. cinnamomi*: 2
 - *P. kernoviae*: 1
 - *P. ramorum*: 1
 - *P. cambivora*: 12

But...

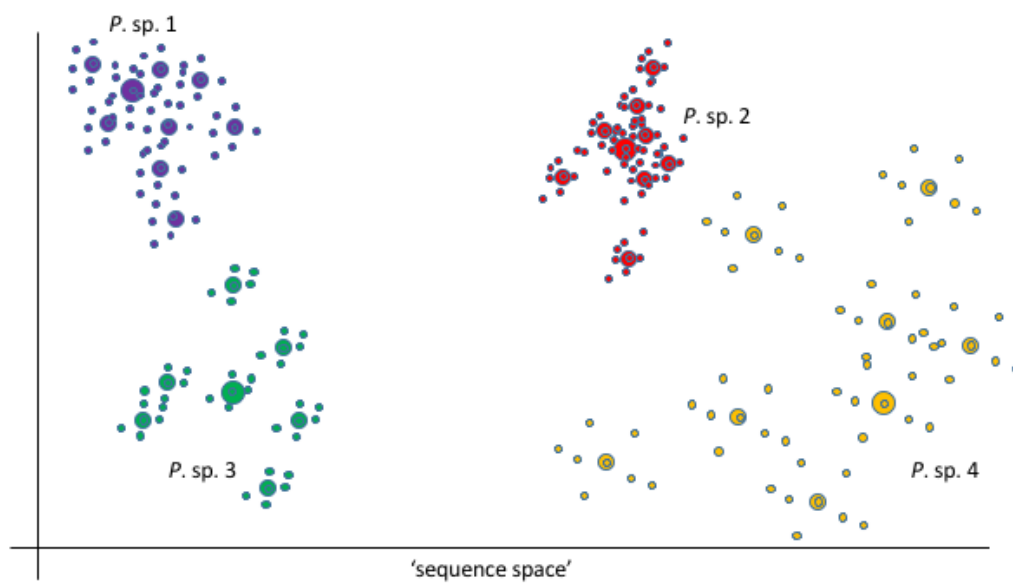
Assembly questions...

- ITS1/rRNA regions are repetitive
- Assemblies collapse repetitive regions
- Is there evidence of assembly collapse?
- Excessive read coverage to genomic ITS1 is evidence of assembly collapse
- Compare coverage to 'conserved, single-copy genes' (BUSCO)

Six sequenced genomes

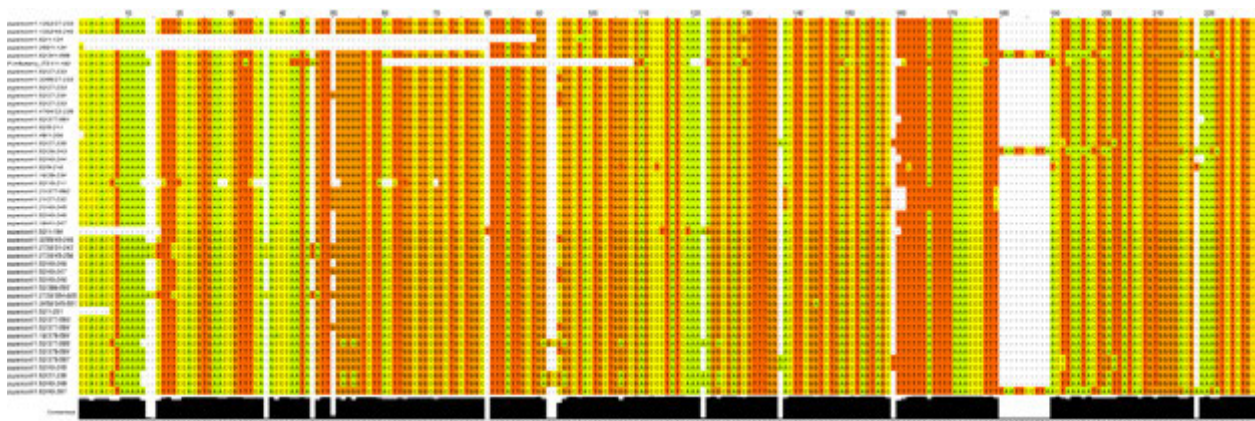
- Ratio ITS1:BUSCO gene coverage = estimated copy number
 - *P. infestans*: 97
 - *P. sojae*: 40
 - *P. cinnamomi*: 60
 - *P. kernoviae*: 23
 - *P. ramorum*: 8
 - *P. cambivora*: 84

>1 ITS1 per isolate



ITS1 diversity

47 sequences from *P. infestans* T30-4



Estimating isolate ITS1 diversity

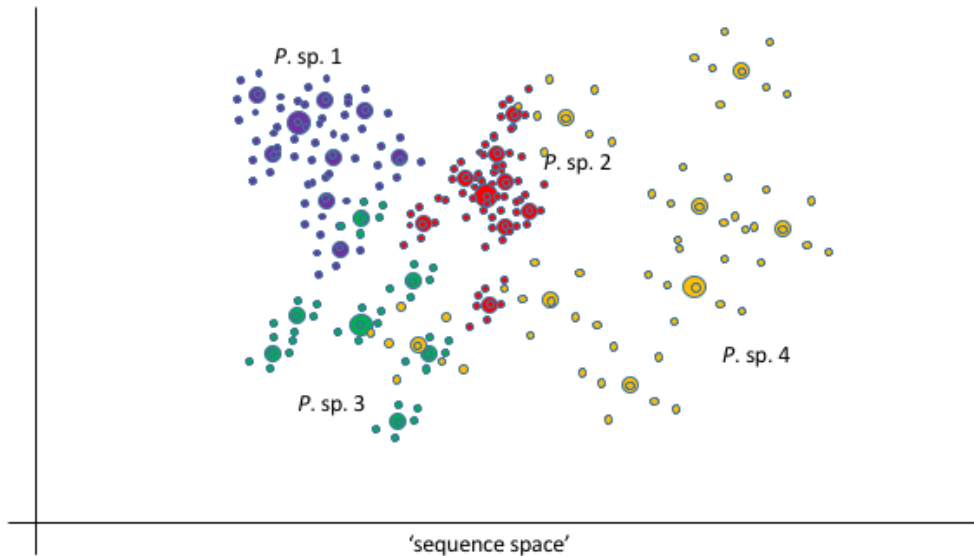
- collapsed regions not yet assembled
- assembly options include: GRAbB, MITObim, and riboSEED
- without assembly, estimate SNPs/phasing/profile

Sequence survey 2

If there is more than one ITS1 sequence per isolate...

- How diverse are ITS1 sequences?
 - <80%-100% sequence identity
- Is there possibility for confusion in classification, or in a representative database?

Confusion?



Confusion

- 154 'representative' ITS1 sequences
- Clustering using SWARM
- 122-138 clusters resulted (dependent on stringency)
- Potential confusion *but* so far only within a clade (*sensu* Martin *et al.*)

Algorithms

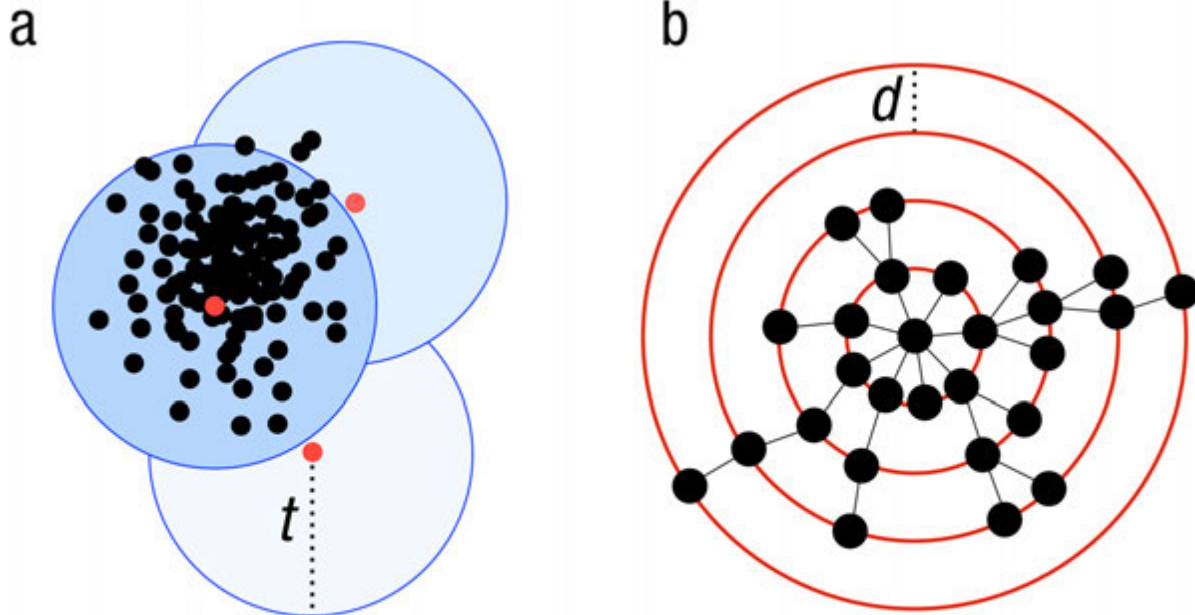


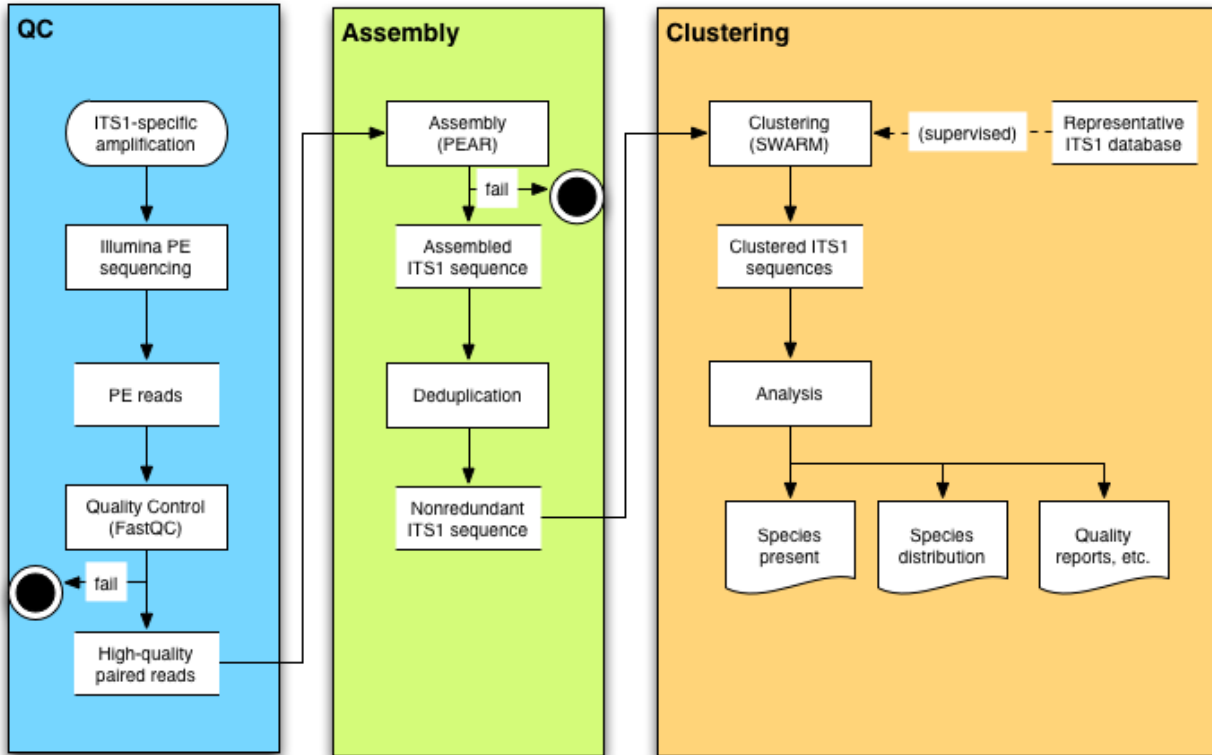
Figure 1. Visualization of the widely used greedy clustering approach based on centroid selection and a global clustering threshold, t , where closely related amplicons can be placed into different OTUs. (b) By contrast, Swarm clusters iteratively by using a small local clustering threshold, d , allowing OTUs to reach their natural limits.

Classification

Three stages

1. **QC:** FastQC to clean reads
2. **Assembly:** PEAR to assemble reads
3. **Clustering:** SWARM to cluster reads (unsupervised) for OTUs

Flowchart



Working prototype

- Still much to do, but groundwork in place
- PCRMIX96 testing: 10 known species
 - 9 species present identified
 - 8 false positives identified
 - 1 false negative
- Some issues due to ITS1 sequence ‘confusion’

Validation

- Influence of primers (filter)?
- We need to establish objectively that the clustering/classification works:
 - samples of known spp. composition to test accuracy of classification
 - FPR, FNR, Sp, Sn, etc.
- We need to build a representative database of ITS1
 - genome analysis to catalogue ITS1 variation
- Comparison against, e.g. PIPITS (requires database development)