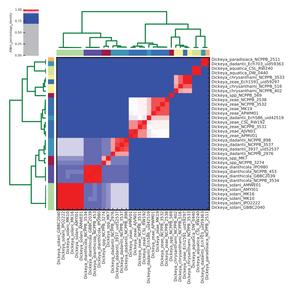
# Whole-genome bacterial classification

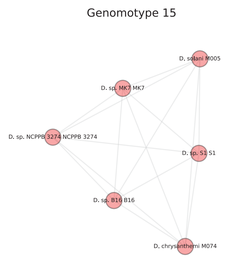
## Whole genome taxonomic reclassification and a universal prokaryotic identifier scheme

Leighton Pritchard[1]

[1]Information and Computational Sciences, The James Hutton Institute, Invergowrie, Dundee, DD2 5DA, Scotland

Email: leighton.pritchard@hutton.ac.uk






The James Hutton Institute

## Introduction

Binomial taxonomic nomenclature is problematic for prokaryotes. Much existing classification has origins in polyphasic and phenotypic classifications that do not reflect relatedness at molecular level, and is under continual active revision that results in widespread confusion in literature, databases, and historical collections.

Taxonomic classification nevertheless remains central to many areas of significant public impact, including development of political policy for legislation, and border control that aims to reduce disease risks to agriculture from pathogenic bacteria. To meet policy goals effectively with diagnostic tools and associate disease risk with identity, historical classifications need to be revisited.

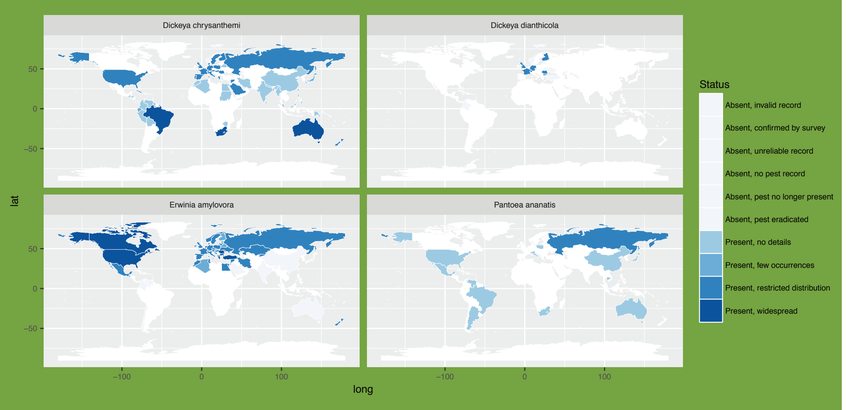



**Figure 1** Global prevalence of enterobacterial plant pathogens

We use measures of genomic relatedness and a graph decomposition approach to subdivide enterobacterial plant pathogens into groupings (*cliques*) based only on inherent properties of their complete genomes. These groups require no arbitrary thresholding and are stable to introduction of new sequences. They are capable of providing a basis for:

- **universal indexing of prokaryotes**
- **probabilistic estimates of risk conditioned on *clique* membership**.

PYANI: Software for average nucleotide identity (ANI)

SlideShare presentation from EAPR 2016

Pritchard et al. (2016)

Pritchard et al. (2013)

**Acknowledgements**

Special thanks go to:

- John Elphinstone
- Rachel Glover
- Sonia Humphris
- Ian Toth

Scottish Government
Riaghaltas na h-Alba
gov.scot

## Methods

Average Nucleotide Identity (ANI) is a sequence comparison method that calculates pairwise nucleotide identity between two genomes. We used PYANI (http://widdowquinn.github.io/pyani/) to calculate and visualise all-vs-all ANI values for 127 enterobacterial plant pathogen genomes from *Erwinia*, *Dickeya* and *Pectobacterium* (Figure 4).

### Graph decomposition

ANI (%identity) values are used to construct a graph in which nodes represent sequenced isolates, and edges are weighted by ANI. The graph is progressively trimmed of edges from lowest to highest ANI. As edges are removed, nodes coalesce into *cliques*, in which all nodes are connected to all other nodes in the clique, with ANI greater than the current threshold. At some thresholds, it is observed that all genomes/nodes are unambiguously assigned to a single *clique* (Figure 2).

The frequency of unambiguous *cliques* is greatest at thresholds corresponding to existing genus-level, species-level and subspecies-level classifications (Figure 3), suggesting that these thresholds correspond to meaningful biological divisions.

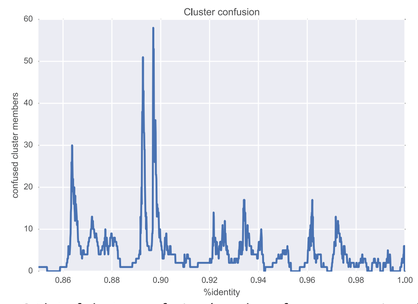



**Figure 2** Plot of cluster confusion (number of genomes assigned to more than one *clique*) against %identity threshold. Zeros occur when every genome is unambiguously assigned to a *clique*.

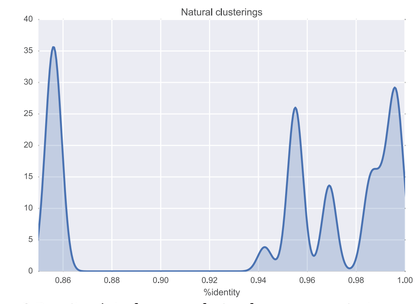



**Figure 3** Density plot of zero confusion frequency against percentage identity edge trimming. Peaks indicate graph decompositions where all genomes are unambiguously assigned to *cliques*.

## Results

We applied ANI to sequenced genomes of 34 *Dickeya*, 55 *Pectobacterium*, and 38 *Erwinia* isolates to obtain an initial classification. This indicated that several novel species-level groups for *Dickeya* (2), *Pectobacterium* (4) and *Erwinia* (2) can be proposed.

We decomposed the ANI graph into natural groupings at genus (≈86% id) and species (≈95%) identity levels. This indicated that the *Pectobacterium* genus grouping is stable and consistent with sequenced isolates. However the *Dickeya* genus can be subdivided into three (splitting off *D. aquatica* and *D. paradisiaca*) genus-level groups, and the *Erwinia* genus divided into 13 (thirteen) genus-level groups.

The decomposition supports several reclassifications and novel species groupings for each genus (Figure 5).

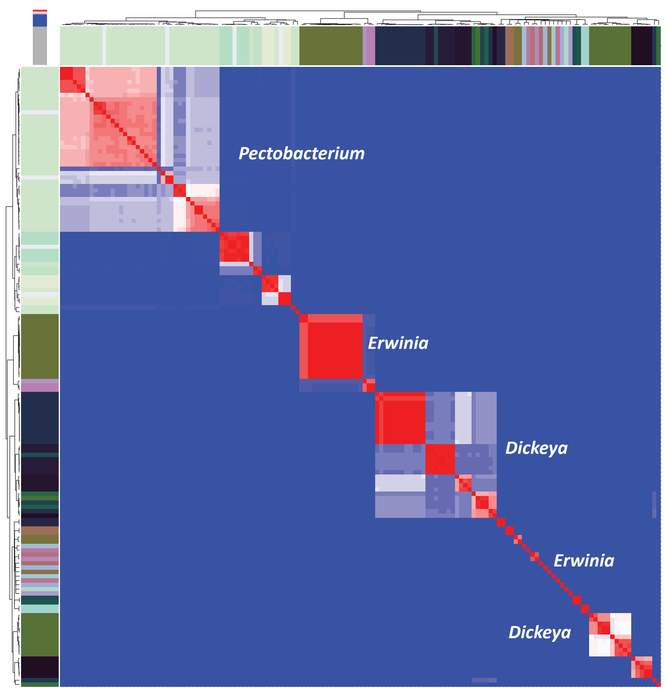



**Figure 4** Average nucleotide identity plot for 127 enterobacterial plant pathogen genomes produced by PYANI (http://widdowquinn.github.io/pyani/). Red squares indicate that genomes share >95% nucleotide identity; blue squares indicate that genomes share <95% identity. Row and column dendrograms cluster genomes by how similar their identity profiles are; dendrogram colours indicate isolate species assignment before ANI analysis. Large red blocks on the diagonal correspond to species (at >95% id).

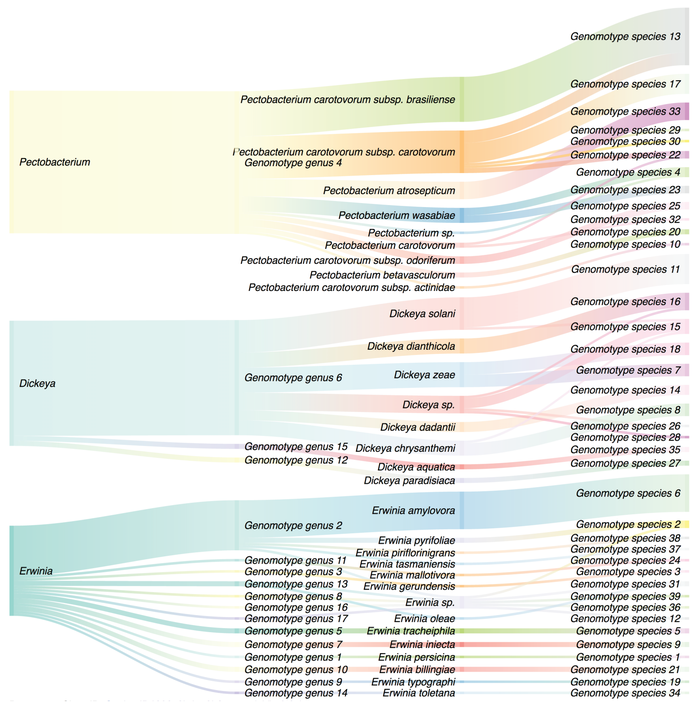



**Figure 5** Sankey (river) diagram showing proposed reclassifications of enterobacterial plant pathogen isolates at genus and species level, after graph decomposition (L to R: isolate classification; genus by decomposition; isolate classification; species by decomposition). The *Pectobacterium* genus is stably assigned, but reassortment of isolates at species level is indicated. Decomposition suggests *D. aquatica* and *D. paradisiaca* should be reassigned at genus level, and several *Dickeya* species level reclassifications are proposed. Graph decomposition suggests much reorganisation of *Erwinia* is necessary

## Conclusions

- Bacterial taxonomic and nomenclature assignments in historical collections and public databases do not always agree with molecular and genomic evidence
- ANI is a global measure of genome similarity useful for bacterial classification
- Decomposition of graphs constructed from all-vs-all ANI of bacterial genomes produces classifications that broadly agree with, but refine existing classifications at biologically-meaningful levels
- Classifications produced by this approach suggest widespread reclassification of some genera (e.g. *Erwinia*) but confirm existing groupings in other clades

**References**

- Pritchard *et al*. (2016) "Genomics and taxonomy in diagnostics for food security: soft-rotting enterobacterial plant pathogens" *Anal. Methods*, 2016,8, 12-24 DOI: 10.1039/C5AY02550H
- Pritchard *et al.* (2013) "Detection of phytopathogens of the genus Dickeya using a PCR primer prediction pipeline for draft bacterial genome sequences" *Plant Pathology* 62:587–596.