

Pathogen Genome Data

EMBL-EBI Bioinformatics of Plants and Plant Pathogens 23rd May 2016



**The James
Hutton
Institute**

Leighton Pritchard^{1,2,3}

¹Information and Computational Sciences,

²Centre for Human and Animal Pathogens in the Environment,

³Dundee Effector Consortium,

The James Hutton Institute, Invergowrie, Dundee, Scotland, DD2 5DA



Acceptable Use Policy

Recording of this talk, taking photos, discussing the content using email, Twitter, blogs, etc. is permitted (and encouraged), providing distraction to others is minimised.

These slides will be made available on SlideShare.

These slides, and supporting material including exercises, are available at <https://github.com/widdowquinn/Teaching-EMBL-Plant-Path-Genomics>



Table of Contents

1 Introduction

- Pathogen Genome Data

2 Public Genome Data Sources

- Online Resources

3 Comparative Genomics

- Why Comparative Genomics?
- Whole Genome Comparisons
- Feature Comparisons

4 Effector Prediction

- Effector Characteristics



Introduction

What can pathogen genome data do for you?

Combining genomic data with comparative and evolutionary biology, addresses questions of pathogen evolution, adaptation and lifestyle.

“NOTHING IN BIOLOGY MAKES SENSE EXCEPT
IN THE LIGHT OF EVOLUTION.”

THEODOSIUS DOBZHANSKY

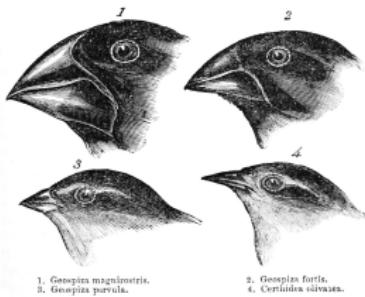




Table of Contents

1 Introduction

- Pathogen Genome Data

2 Public Genome Data Sources

- Online Resources

3 Comparative Genomics

- Why Comparative Genomics?
- Whole Genome Comparisons
- Feature Comparisons

4 Effector Prediction

- Effector Characteristics



<http://www.ncbi.nlm.nih.gov/>

Repository of record for pathogen (and other) genome data

■ Example: *Ralstonia solanacearum*

- Browser interface
- FTP repositories of genome data
 - RefSeq
 - GenBank

Index of /genomes/refseq/bacteria/Ralstonia_solanacearum/latest_assembly_versions

Name	Last modified	Size
GCF_000009125_1_ASM9...>	17-May-2016 16:30	-
GCA_000167955_1_ASM1...>	03-Mar-2015 06:18	-
GCF_000212635_3_ASM9...>	17-May-2016 19:06	-
GCF_000211970_1_ASM2...>	17-May-2016 21:15	-
GCA_000009125_1_ASM9...>	17-May-2016 16:30	-

Index of /genomes/genbank/bacteria/Ralstonia_solanacearum/latest_assembly_versions

Name	Last modified	Size
GCA_000009125_1_ASM9...>	17-May-2016 16:29	-
GCA_000167955_1_ASM1...>	17-May-2016 12:11	-
GCA_000197855_1_ASM1...>	13-Jun-2015 18:15	-
GCA_000212635_2_ASM2...>	17-May-2016 19:06	-
GCA_000009125_1_ASM9...>	17-May-2016 21:15	-



GenBank vs RefSeq

GenBank

- part of International Nucleotide Sequence Database Collaboration (INSDC): EMBL/NCBI/DDBJ
- records 'owned' by submitter
- may include redundant information

RefSeq

- not part of INSDC
- records derived from GenBank, 'owned' by NCBI
- stable non-redundant foundation for functional and diversity studies



<http://www.ensembl.org>

Automated annotation on selected genomes

■ Specialised sub-collections

- Ensembl Protists: <http://protists.ensembl.org/>
- Ensembl Bacteria: <http://bacteria.ensembl.org/>
- Ensembl Fungi: <http://fungi.ensembl.org/>

■ Downloadable resources

- e.g. <ftp://ftp.ensemblgenomes.org/pub/protists/>

■ Ready-made comparative genomics!

- *Phytophthora* genomics alignments (Avr3a)
- Gene trees (Avr3a)



Other Sources

- **Sequencing centres, e.g.**
 - JGI Genome Portals
 - Broad Institute - now retiring their online resources
- **Specialist databases, e.g.**
 - PhytoPath - plant pathogens
 - PHI-Base 4 - curated plant-pathogen interaction data
 - FungiDB - fungi and oomycetes
 - CPGR - fungi and oomycetes (not recently updated)
- **Your friendly local sequencing centre!**
 - Aspera is commonly used to connect to your private data



OPTIONAL WORKSHEET

worksheets/01-downloading_data_biopython.ipynb

Downloading genome data from NCBI with Biopython

- MyBinder link



Table of Contents

- 1 Introduction**
 - Pathogen Genome Data
- 2 Public Genome Data Sources**
 - Online Resources
- 3 Comparative Genomics**
 - Why Comparative Genomics?
 - Whole Genome Comparisons
 - Feature Comparisons
- 4 Effector Prediction**
 - Effector Characteristics

Why comparative genomics?

- Transfer functional information from model systems (*E. coli*, *A. thaliana*, *D. melanogaster*) to non-model systems
- Genome similarity \propto phenotype? (*functional genomics*): virulence and host range
- Genome similarity \propto relatedness? (*phylogenomics*): record of evolutionary processes and constraints

I think



then between *A* & *B*. *cavias*
less of relation. *C* & *B*. the
first generation, *B* & *D*
rather greater distinction
then genera would be
formed. - binary relation



Genomes aren't everything...

Context

- epigenetics
- tissue differentiation/differential expression
- mesoscale systems, etc.

Phenotypic plasticity, responses to

- temperature
- stress
- community, etc.

...and therefore systems biology...

I think



Then between A + B. variation
less of selection. C + D. The
first generation, B and D
rather greater distinction
Then genera would be
formed. - binary selection



Levels of comparison

Bulk Properties

- e.g. k -mer spectra ([MaSH](#), [MetaPalette](#), etc.)

Whole Genome Sequence

- sequence similarity ([BLAST](#), [BLAT](#), [MUMmer](#), etc.)
- structure and organisation ([Mauve](#), [ACT](#), etc.)

Genome Features/Functional Components

- numbers and types of features: genes, ncRNA, regulatory elements, etc.
- organisation of features: synteny, operons, regulons, etc.
- functional complement ([KEGG](#), etc.)



Table of Contents

1 Introduction

- Pathogen Genome Data

2 Public Genome Data Sources

- Online Resources

3 Comparative Genomics

- Why Comparative Genomics?
- Whole Genome Comparisons
- Feature Comparisons

4 Effector Prediction

- Effector Characteristics



Whole genome comparisons

Whole genome comparison

Comparisons of one complete or draft genome with another
(...or many others)

Minimum requirement: **two genomes**

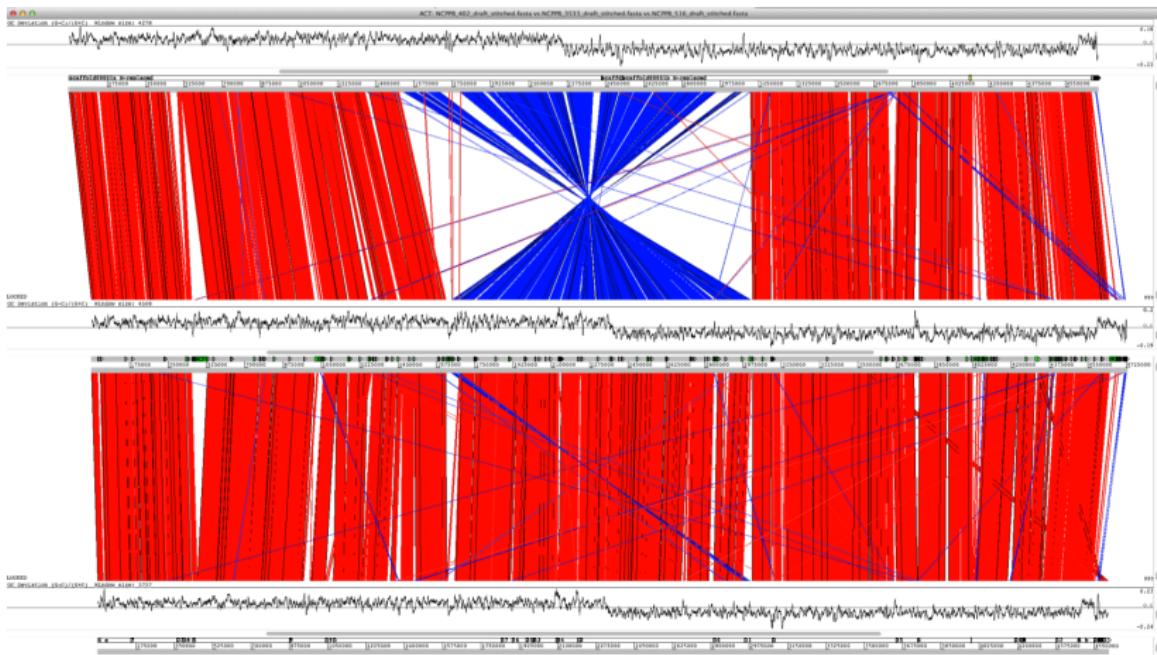
- Reference Genome
- Comparator Genome

The experiment produces a comparative result *that is dependent on the choice of genomes.*



Pairwise genome alignments

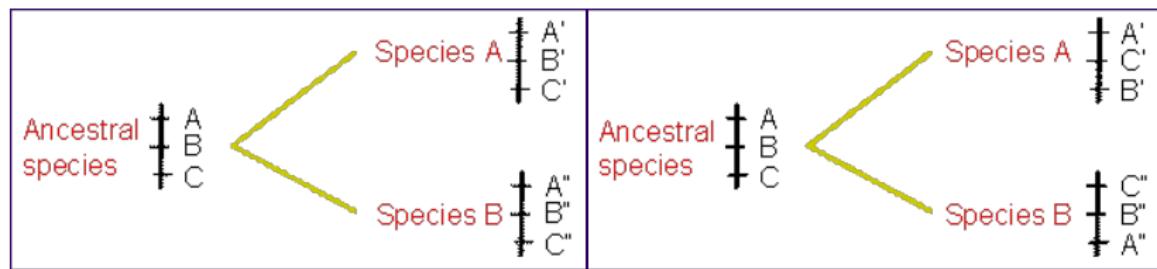
Pairwise comparisons produce alignments of similar regions.





Synteny and Collinearity

Genome rearrangements may occur post-species divergence
Sequence similarity, and order of similar regions, may be conserved



- *collinear* conserved elements lie in the same linear sequence
- *syntenous* (or *syntenic*) elements:
 - (orig.) lie on the same chromosome
 - (mod.) are collinear

Evolutionary constraint (e.g. indicated by synteny) may indicate functional constraint (and help determine *orthology*)



Vibrio mimicus ^a

^a Hasan et al. (2010) Proc. Natl. Acad. Sci. USA 107:21134-21139 doi:10.1073/pnas.1013825107

Chromosomes

- C-I: virulence genes.
- C-II: environmental adaptation

C-II has undergone extensive rearrangement; C-I has not.

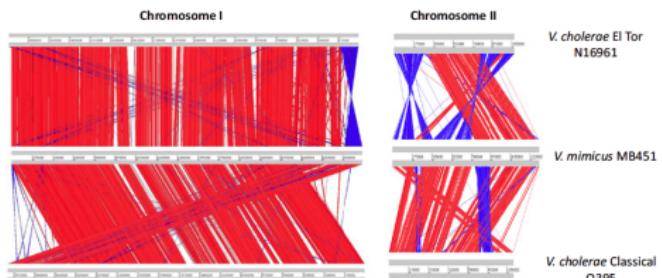


Fig. 2. Linear pairwise comparison of the *Vibrio mimicus* genome by Artemis Comparison Toll. Regions with similarity are highlighted by connecting red or blue lines between the genomes; red lines indicate homologous blocks of sequence, and blue lines indicate inversions. Gaps indicate unique DNA. The gray bars represent forward and reverse strands.

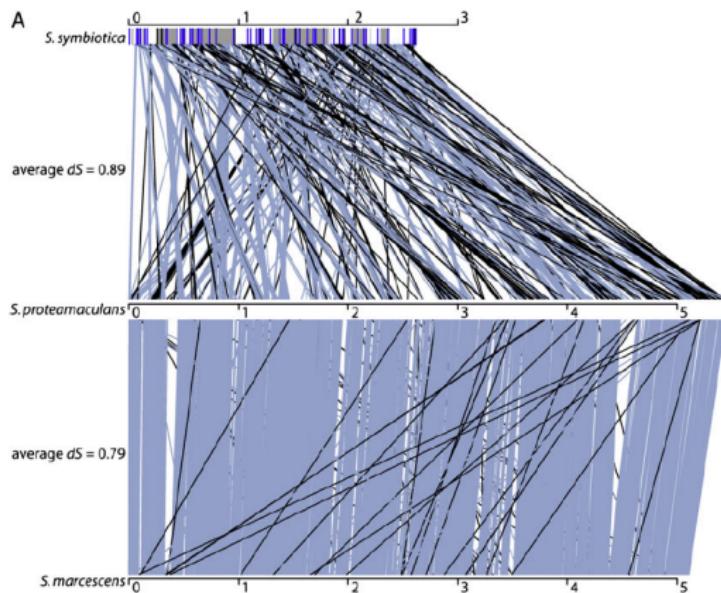
Suggests modularity of genome organisation, as a mechanism for adaptation (HGT, two-speed genome).



Serratia symbiotica ^a

^a Burke and Moran (2011) *Genome Biol. Evol.* 3:195-208 doi:10.1093/gbe/evr002

S. symbiotica is a recently adapted symbiont of aphids
Massive genomic decay: consequence of adaptation





Whole genome classification ^{a b}

^a Baltrus (2016) *Trends Microbiol.* doi:10.1016/j.tim.2016.02.004

^b Pritchard *et al.* (2016) *Anal. Methods* doi:10.1039/c5ay02550h

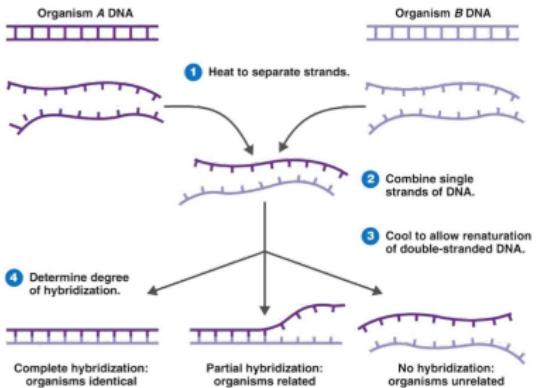
- Widespread confusion about bacterial strain classification and nomenclature
 - Taxonomies contradicted by bioinformatic classification
 - Databases populated by non-taxonomists
- Philosophy and practice of taxonomy are in conflict
- Classification can be independent of existing nomenclature
 - The route from genotype to phenotype is complicated
 - Time to abandon traditional bacterial species concepts?
- An unambiguous sequence-based classification scheme is possible



DNA-DNA hybridisation^a

^a Morello-Mora and Amann (2001) *FEMS Micro. Rev.* doi:10.1016/S0168-6445(00)00040-1

- “Gold Standard” for prokaryotic taxonomy, since 1960s. “70% identity ≈ same species.”
- Denature DNA from two organisms.
- Allow to anneal. Reassociation ≈ similarity, measured as ΔT of denaturation curves.



Proxy for sequence similarity - replace with genome analysis^{1?}

¹ Chan et al (2012) *BMC Microbiol.* doi:10.1186/1471-2180-12-302

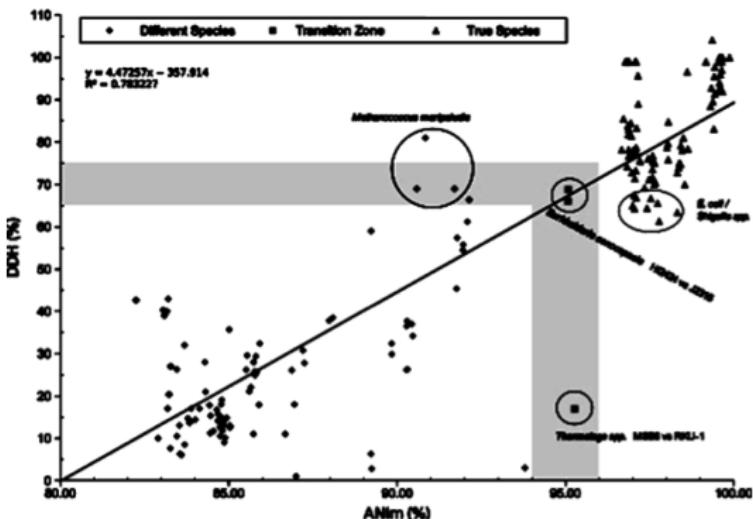


Average Nucleotide Identity (ANIm)^a

^a Richter and Rossello-Mora (2009) *Proc. Natl. Acad. Sci. USA* doi:10.1073/pnas.0906412106

1. Align genomes
(MUMmer)
2. **ANIm:** Mean
% identity of all
matches

- DDH:ANIm
linear
- 70%ID ≈
95%ANib



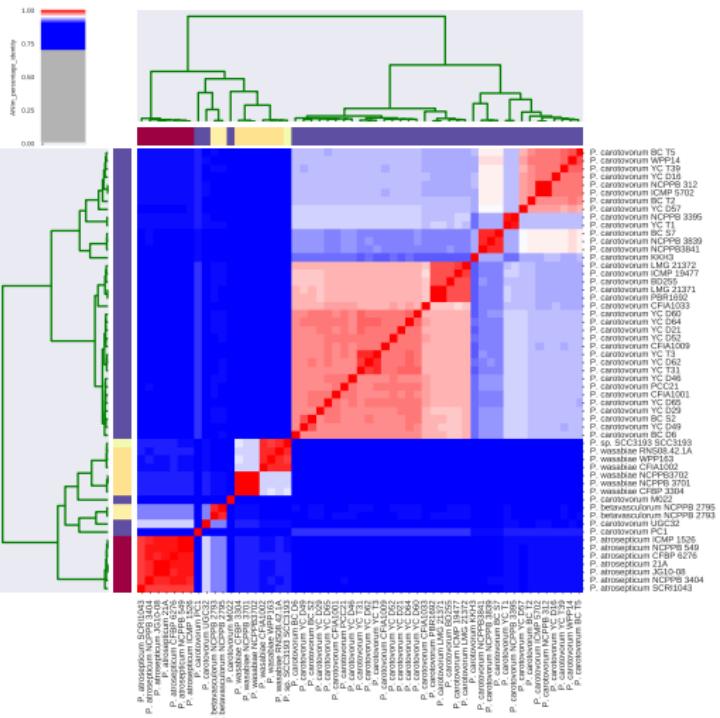
55 *Pectobacterium* spp. ANIm ^a



^aPritchard et al. (2016) *Anal. Methods* doi:10.1039/c5ay02550h



- Ten species-level groups (four novel)
 - *P. carotovorum* split: several species
 - *P. wasabiae* split: two species

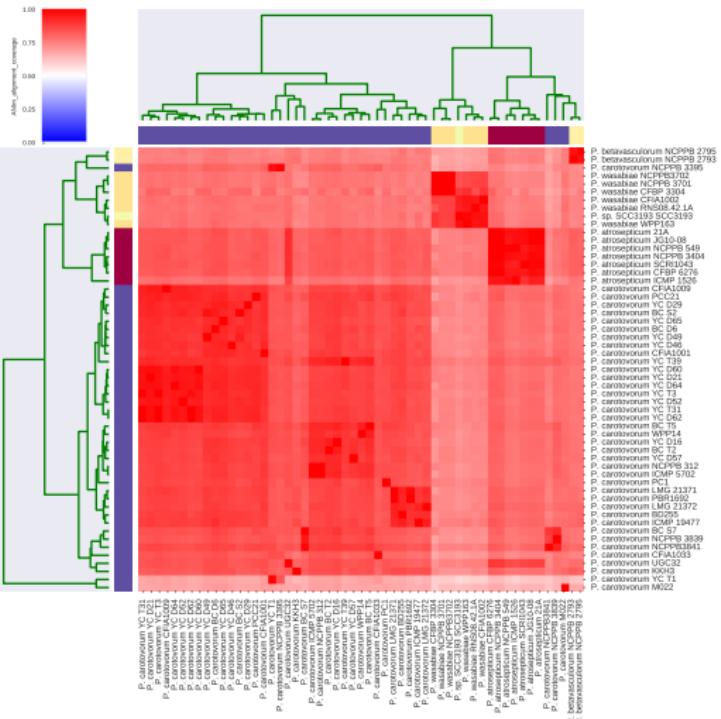


55 *Pectobacterium* spp. ANIm^a



^aPritchard et al. (2016) *Anal. Methods* doi:10.1039/c5ay02550h

- All isolates align over >50% of whole genome



Advantages

- Average identity of all 'homologous' regions
- Approximates limiting case of MLST/MLSA/multigene comparisons
- Adaptable to variable thresholding (LINS) classifications

Criticisms

- Thresholds 'arbitrary', based on homologous regions only
- Taxonomic classification, not phylogenetic reconstruction
- No functional (or gene-based) interpretation; still need pangenome classification and analysis



EXERCISE

`exercises/01-whole_genome_comparisons.ipynb`

- Pairwise comparison of *Pseudomonas* genomes
- ANIm classification of *Pseudomonas* isolates

- MyBinder link



Chromosome painting ^a

^aYahara et al. (2013) *Mol. Biol. Evol.* 30:1454-1464 doi:10.1093/molbev/mst055

- “Chromosome painting” (**FINESTRUCTURE**) infers recombination-derived ‘chunks’
- Genome’s haplotype constructed in terms of recombination events from a ‘donor’ to a ‘recipient’ genome



Fig. 1. Chromosome painting in silico. Each lane indicates the chromosome of a strain shown on the right. The strains are classified by *findSTRUCTURE* into subgroups labeled by colors (table 1 and fig. 2) on the left. A color along the chromosome indicates the subgroup that donated a chunk of SNPs through homologous recombination. All genomic positions are transformed to those of a reference strain (26695).



Chromosome painting ^a

^aYahara et al. (2013) Mol. Biol. Evol. 30:1454-1464 doi:10.1093/molbev/mst055

- Recombination events summarised in a *coancestry matrix*.
- *H. pylori*: most within geographical bounds, but asymmetrical donation from Amerind/East Asian to European isolates.

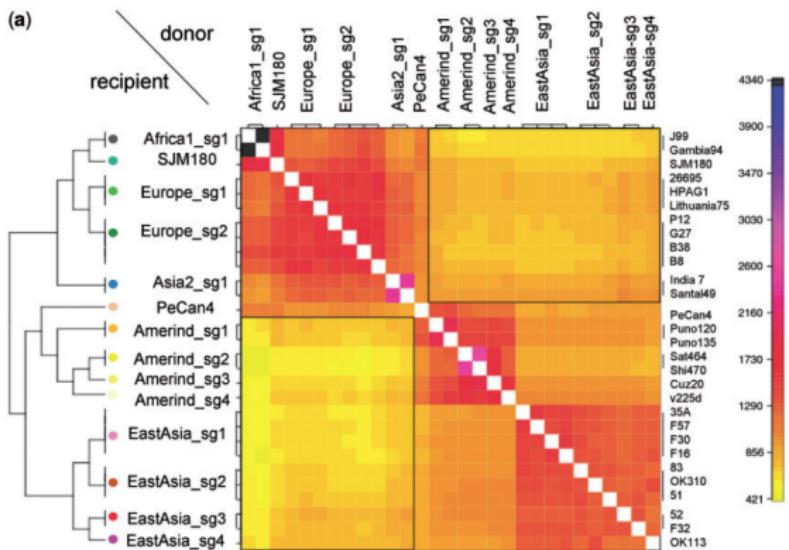




Table of Contents

1 Introduction

- Pathogen Genome Data

2 Public Genome Data Sources

- Online Resources

3 Comparative Genomics

- Why Comparative Genomics?
- Whole Genome Comparisons
- Feature Comparisons

4 Effector Prediction

- Effector Characteristics



Feature comparisons

Feature comparisons

Comparisons of the annotated features of one genome with another
(...or many others)

- gene features
- RNA features
- regulatory features



Equivalent features

The power of genomics is comparative genomics!

- Makes catalogues of genome components comparable between organisms
- Differences, e.g. presence/absence of equivalents may support hypotheses for functional or phenotypic difference
- Can identify characteristic signals for diagnosis/epidemiology
- Can build parts lists and wiring diagrams for systems and synthetic biology





Orthologues ^{a b}

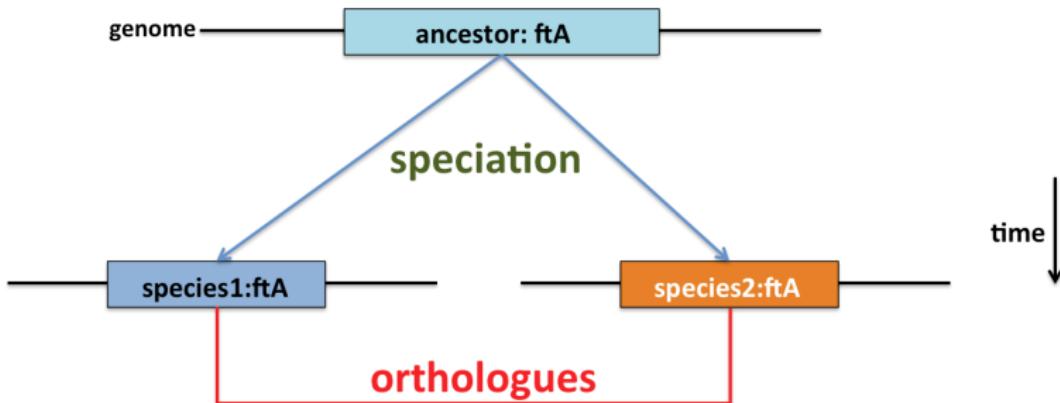
^a Nehrt et al. (2011) *PLoS Comp. Biol.* doi:10.1371/journal.pcbi.1002073

^b Chen et al. (2012) *PLoS Comp. Biol.* doi:10.1371/journal.pcbi.1002784

Orthologs/Orthologues

"Homologs that diverged through speciation" (orig.)

"Genes/products we think are probably the same thing" (mod. inform.)





Why orthologues? ^a ^b ^c

^aChen and Zhang (2012) *PLoS Comp. Biol.* doi:10.1371/journal.pcbi.1002784

^bDessimoz (2011) *Brief. Bioinf.* doi:10.1093/bib/bbr057

^cAltenhoff and Dessimoz (2009) *PLoS Comp. Biol.* 5:e1000262 doi:10.1371/journal.pcbi.1000262

- Formalise the idea of *corresponding genes* in different organisms.
- Suggest two relationships:
 - Evolutionary equivalence
 - Functional equivalence ("The Ortholog Conjecture")

The Ortholog Conjecture

Without duplication, a gene product is unlikely to change its basic function, because this would lead to loss of the original function, and this would be harmful.



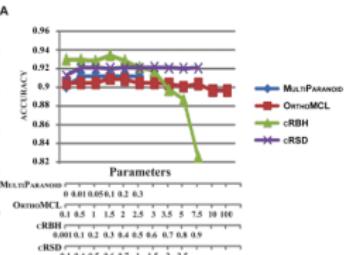
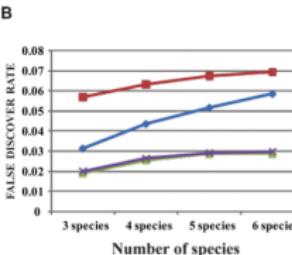
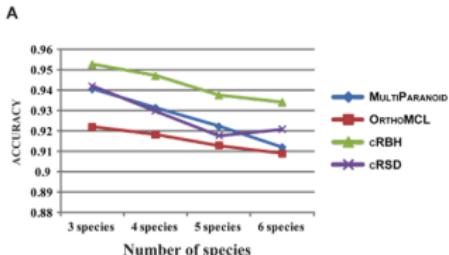
Finding orthologues ^a

^aSalichos and Rokas (2011) *PLoS One* 6:e18755 doi:10.1371/journal.pone.0018755.g006

Which discovery method performs best?

- Four methods tested against 2,723 curated orthologues from six *Saccharomycetes*:
RBBH (and cRBH); RSD (and cRSD); MultiParanoid;
OrthoMCL
- Rated by statistical performance metrics: sensitivity,
specificity, accuracy, FDR

cRBH most accurate and specific, with lowest FDR.





EXERCISE

exercises/02-cds_feature_comparisons.ipynb

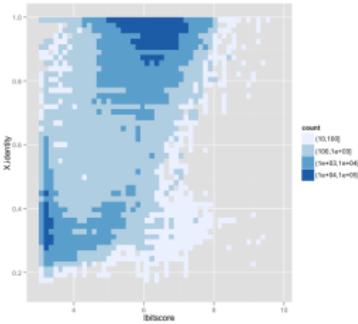
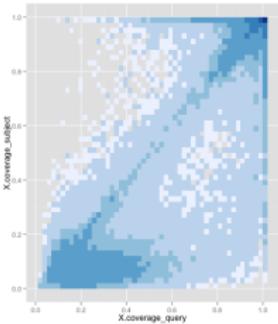
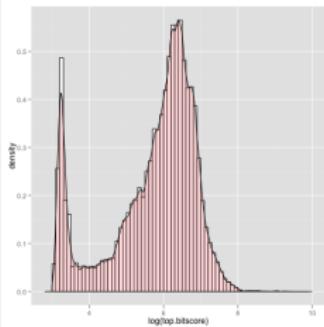
- RBBH analysis of *Pseudomonas* CDS feature annotations

- MyBinder link



One-way BLAST vs RBBH

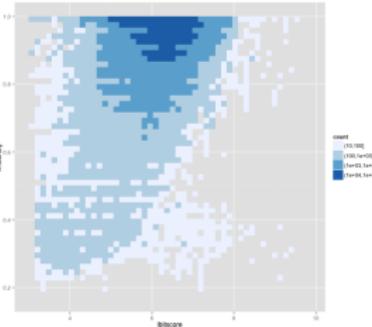
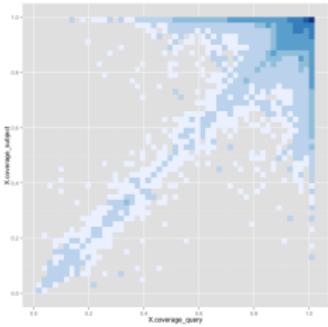
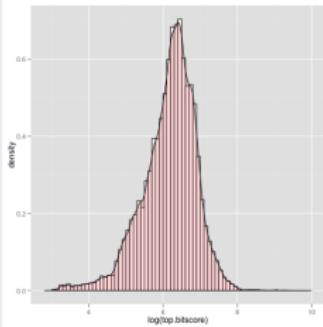
One-way BLAST includes many low-quality hits





One-way BLAST vs RBBH

Reciprocal best BLAST removes many low-quality matches





The Pangenome

The Core Genome Hypothesis

“The core genome is the primary cohesive unit defining a bacterial species”

- Once equivalent genes have been identified, those present in all related isolates can be identified: **the core genome**.
- The remaining genes are **the accessory genome**, and are expected to mediate function that distinguishes between isolates.

Roary: Rapid large-scale prokaryote pan-genome analysis - works on a desktop machine.

Accessory genome ^{a b}

^a Croll and McDonald (2012) *PLoS Path.* 8:e1002608 doi:10.1371/journal.ppat.1002608

^b Baltrus et al. (2011) *PLoS Path.* 7:e1002132 doi:10.1371/journal.ppat.1002132

Accessory genomes

A cradle for adaptive evolution, particularly for bacterial pathogens, such as *Pseudomonas* spp.

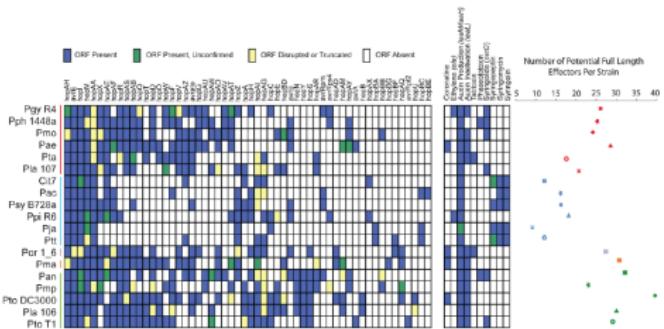


Figure 3. *P. syringae* isolates harbor extensive diversity in virulence gene repertoires. TTE, toxin, and plant hormone biosynthesis genes are listed across the top. *P. syringae* genomes, color-coded by phylogenetic group as in Figure 1. At the left, a blue box indicates presence of full-length ORFs or complete pathways within each genome. Green boxes indicate that genes or pathways are present by similarity searches, but the presence of full-length genes could not be verified by PCR, or the pathways are potentially incomplete. Yellow boxes indicate that genes are either significantly truncated or are disrupted by insertion sequence elements. White boxes indicate absence of genes or pathways from the strains based on homology searches. At the far right, the total number of potentially functional TTE proteins is shown for each genome and displayed according to doi:10.1371/journal.ppat.1002132.g003



OPTIONAL WORKSHEET

worksheets/02-prokka_roary.ipynb

- Annotation of pathogen genomes with Prokka
 - Calculation of the *Pantoea agglomerans* pangenome with Roary
-
- MyBinder link



Table of Contents

1 Introduction

- Pathogen Genome Data

2 Public Genome Data Sources

- Online Resources

3 Comparative Genomics

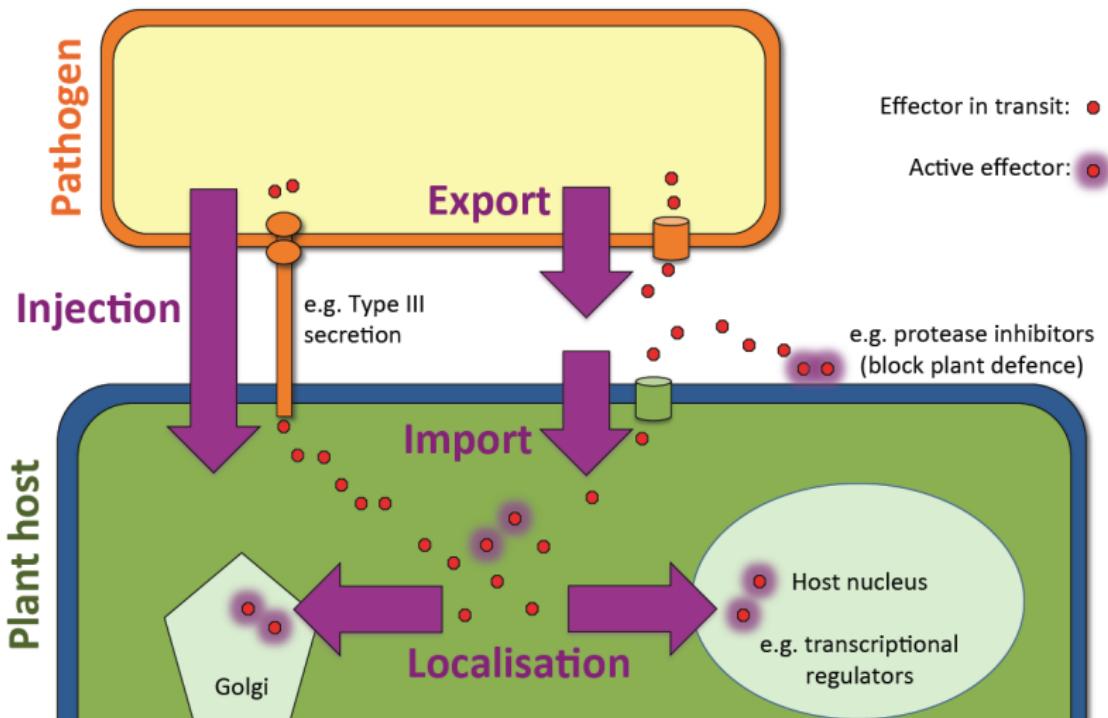
- Why Comparative Genomics?
- Whole Genome Comparisons
- Feature Comparisons

4 Effector Prediction

- Effector Characteristics



What is an effector?





What is an effector?

Effector

A molecule produced by pathogen that (directly?) modifies host molecular/biochemical 'behaviour'

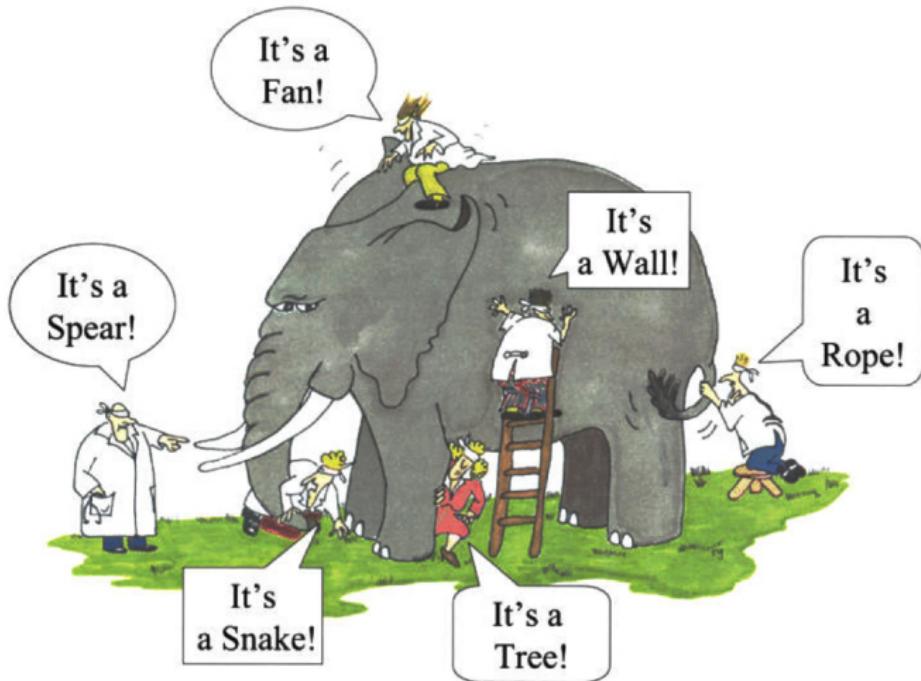
- Inhibits enzyme action (e.g. *Cladosporium fulvum* AVR2, AVR4; *Phytophthora infestans* EPIC1, EPIC2B; *P. sojae* glucanase inhibitors)
- Cleaves a protein target (e.g. *Pseudomonas syringae* AvrRpt2)
- (De-)phosphorylates a protein target (e.g. *P. syringae* AvrRPM1, AvrB)
- Retargeting host system such as E3 ligase (e.g. *P. syringae* AvrPtoB; *P. infestans* Avr3a)
- Regulatory control (e.g. *Xanthomonas campestris* AvrBs3)



What is an effector?

No unifying biochemical mechanism

No single test for 'candidate effectors', even in one organism



Effectors are modular ^{a b}

^aGreenberg & Vinatzer (2003) *Curr. Opin. Microbiol.* doi:10.1016/S1369-5274(02)00004-8

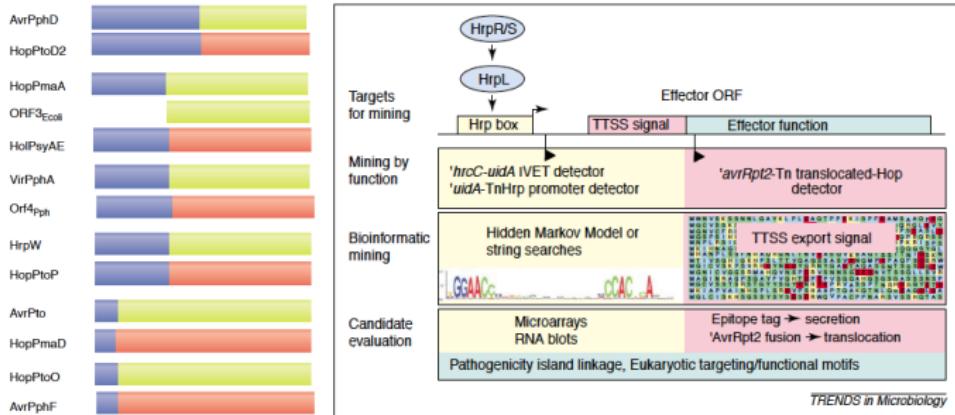
^bCollmer *et al.* (2002) *Trends Microbiol.* doi:10.1016/S0966-842X(02)02451-4

Delivery

N-terminal localisation/translocation domain

Activity

C-terminal functional/interaction domain





Effectors are modular ^{a b}

^aDong et al. (2011) *PLoS One* doi:10.1371/journal.pone.0020172.t004

^bBoch et al. (2009) *Science* doi:10.1126/science.1178811

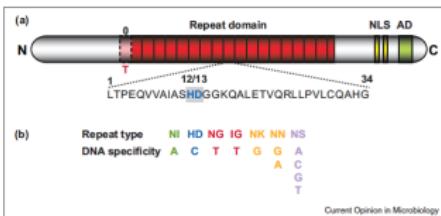
Delivery

Typically common to effector class: RxLR, T3E, CHxC

Activity

May be common (TAL) or divergent within effector class (RxLR, T3E)

	Signal peptide	RXLR
Avr3a ^{ACR12}	1 MRLAQVVVVIAASFLVATDALSTTNANQAKI	KIGTSPGGHSFRLRAYQP
Avr3a ^{P6497}	1 MRLAQVVVVIAASFLVATDALSTTNANQAKI	IIGTSPGGHSFRLRAYQP
Avr3a ^{P7064}	1 MRLAQVVVVIAASFLVATDALSTTNANQAKI	IIGTSPGGHSFRLRAYQP
	EEER **	
Avr3a ^{ACR12}	51 DDEGDSFEDRTLSPSQVTKILNKLGDVTVNDHVMRNPALFQRYQKKANKI	
Avr3a ^{P6497}	51 DDEGDSFEDRTLKAQVTKILNKLGDVTVNDHVMRNPALFQRYQKKANKI	
Avr3a ^{P7064}	51 DDEGDSFEEHTLPNSQVAKILNKLG . . VTWWDVFRDSDALEERYQEKKANKI	
Avr3a ^{ACR12}	101 IEKQKAAAKNA.....	
Avr3a ^{P6497}	101 IEKQKAAAKNA.....	
Avr3a ^{P7064}	99 IEKQKAAAANNAKRIIKERDHTP	





Effector prediction tools (online) ^{a b}



The James
Hutton
Institute

^aSperschneider et al. (2015) *PLoS Pathogens* doi:10.1371/journal.ppat.1004806

^bSonah et al. (2016) *Front. Plant Sci.* doi:10.3389/fpls.2016.00126

Bacterial Type III Effectors

- EffectiveT3
- modlab
- T3SEdb

Fungal/Oomycete Effectors

- EffectorP
- Galaxy Toolshed RxLR predictor



What do we look for? ^a

^aPritchard & Broadhurst (2014) *Methods Mol. Biol.* doi:10.1007/978-1-62703-986-4_4

What if someone hasn't built a classifier for your protein family?

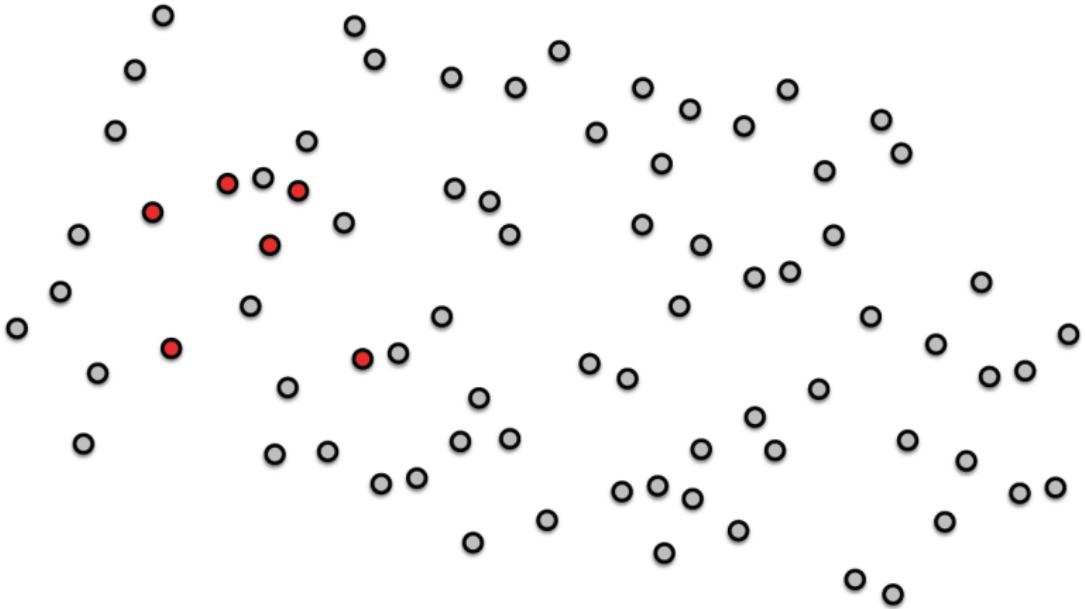
- Tests are for protein family membership and/or 'effector-like' functional signal
- The same as any sequence classification problem (functional annotation)
- Many possible approaches
- (Supervised) machine learning problem:
 - train
 - test
 - validate



Sequence space ^a

^aPritchard & Broadhurst (2014) *Methods Mol. Biol.* doi:10.1007/978-1-62703-986-4_4

Known members of our effector class are in red

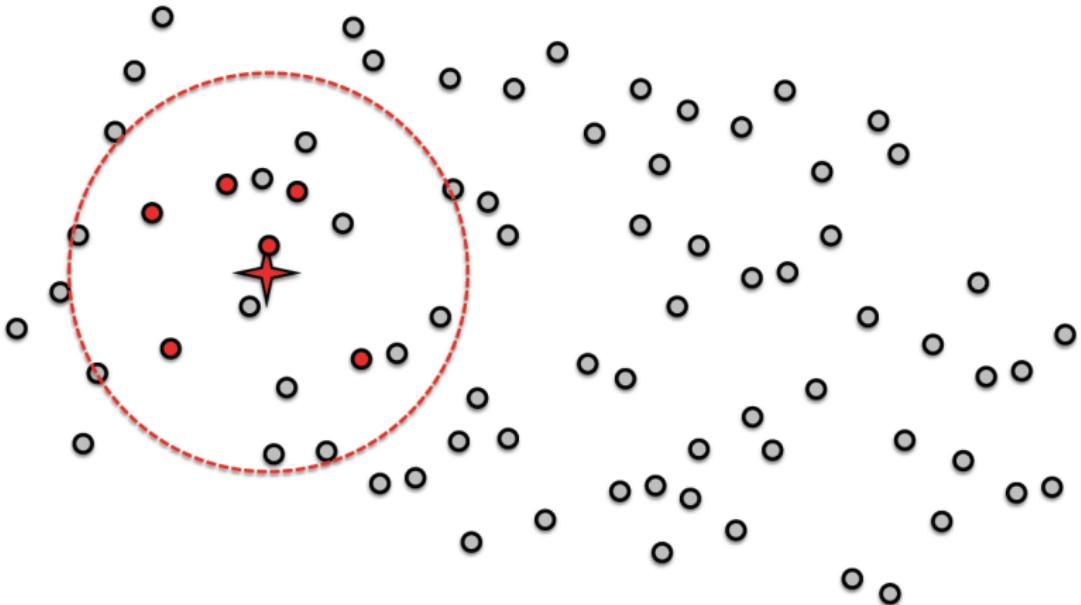




Similarity distance ^a

^aPritchard & Broadhurst (2014) *Methods Mol. Biol.* doi:10.1007/978-1-62703-986-4_4

Define a representative *centre*, and a *distance* from it that includes known effectors

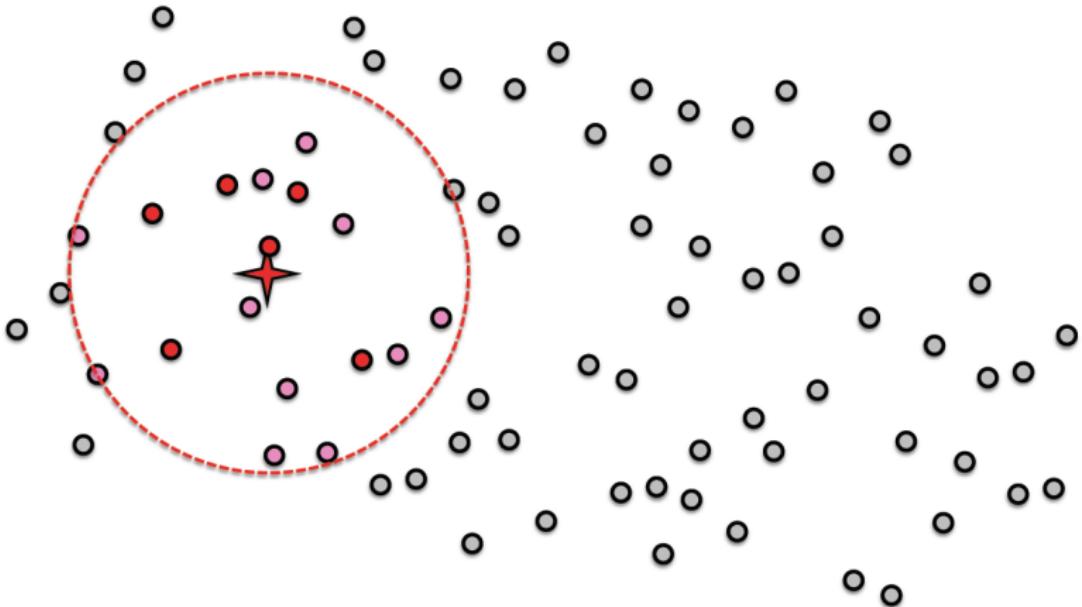




Classify candidates ^a

^aPritchard & Broadhurst (2014) *Methods Mol. Biol.* doi:10.1007/978-1-62703-986-4_4

Classify sequences **within** the distance as **similar**





EXERCISE

```
exercises/03-effector_finding.ipynb
```

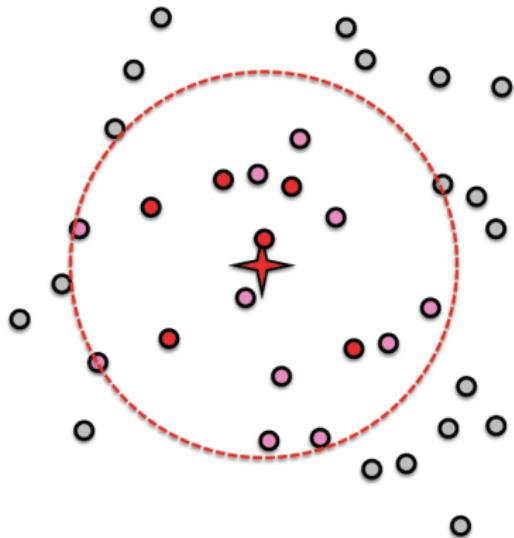
- Downloading annotated *Pseudomonas* AvrPto1 effectors from a public sequence repository
 - Building a (HMM) model from this training set
 - Searching public genome annotations with the model
-
- MyBinder link



Choosing a distance ^a

^aPritchard & Broadhurst (2014) *Methods Mol. Biol.* doi:10.1007/978-1-62703-986-4_4

- How do we define distance?
- How large a distance should we take?
- How do we know if we chose well?

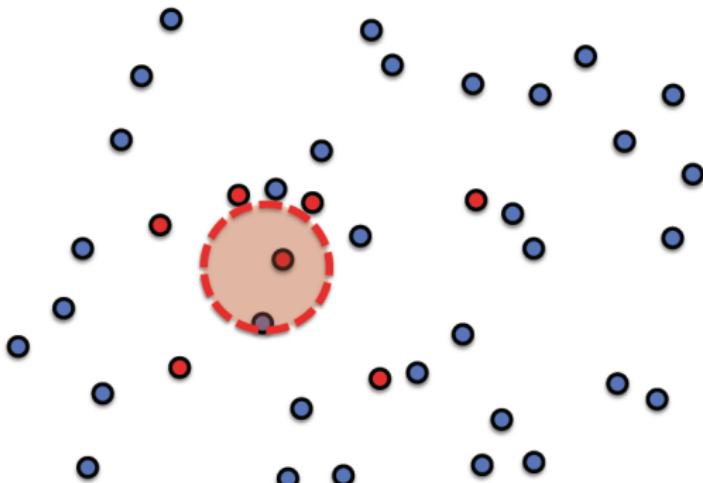




Are you in or out? ^a

^aPritchard & Broadhurst (2014) *Methods Mol. Biol.* doi:10.1007/978-1-62703-986-4_4

- The boundary (distance) classifies sequences as 'in' or 'out'
- Sequences are predicted to be *either* in the class or not in the class
- Changing distance/boundary changes classification



Confusion matrix:

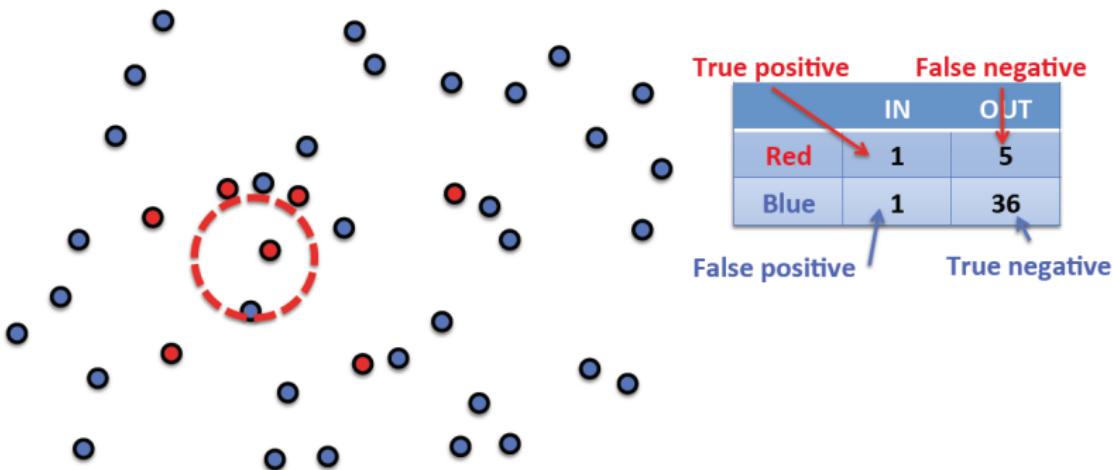
	IN	OUT
Red	1	5
Blue	1	36



TP/TN/FP/FN ^a

^a Pritchard & Broadhurst (2014) *Methods Mol. Biol.* doi:10.1007/978-1-62703-986-4_4

- The boundary (distance) classifies sequences as 'in' or 'out'
- Sequences are predicted to be either **in the class** or **not in the class**
- Changing distance/boundary changes classification

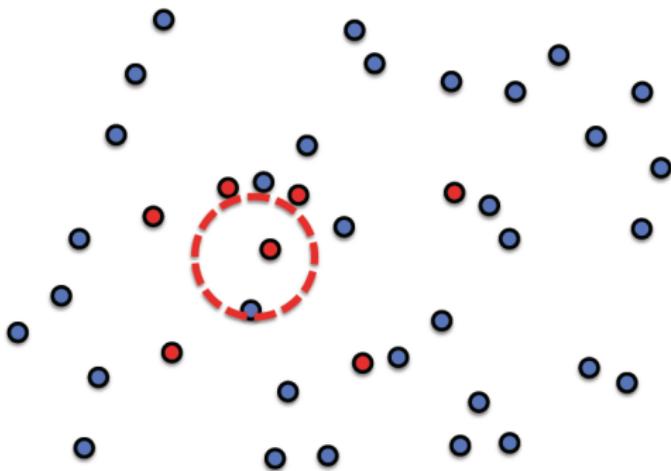




FPR/FNR/Sn/Sp/FDR ^a

^a Pritchard & Broadhurst (2014) *Methods Mol. Biol.* doi:10.1007/978-1-62703-986-4_4

- The boundary (distance) classifies sequences as 'in' or 'out'
- Sequences are predicted to be *either* in the class or not in the class
- Changing distance/boundary changes classification



	IN	OUT
Red	1	5
Blue	1	36

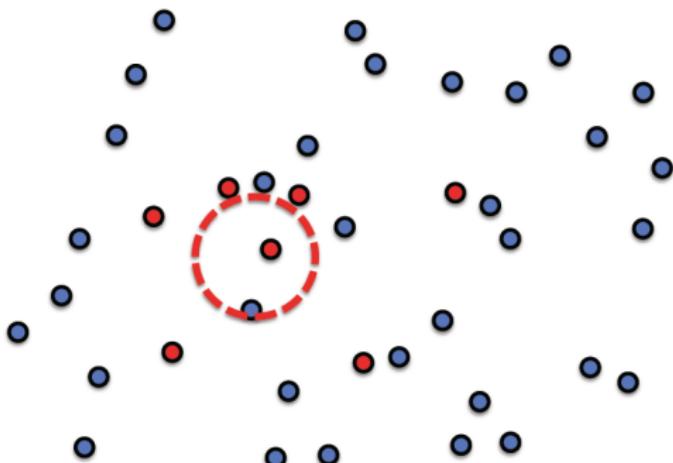
False positive rate	$FP/(FP+TN)$
False negative rate	$FN/(TP+FN)$
Sensitivity	$TP/(TP+FN)$
Specificity	$TN/(FP+TN)$
False discovery rate	$FP/(FP+TP)$



Small Boundary ^a

^aPritchard & Broadhurst (2014) *Methods Mol. Biol.* doi:10.1007/978-1-62703-986-4_4

- The boundary (distance) classifies sequences as 'in' or 'out'
- Sequences are predicted to be *either* **in the class** or **not in the class**
- Changing distance/boundary changes classification



	IN	OUT
Red	1	5
Blue	1	36

False positive rate $1/37 = 0.03$

False negative rate $5/6 = 0.83$

Sensitivity $1/6 = 0.17$

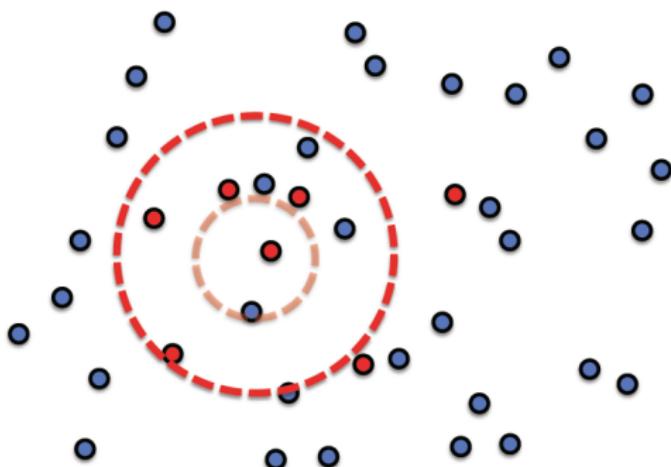
Specificity $36/37 = 0.97$



Medium boundary ^a

^aPritchard & Broadhurst (2014) *Methods Mol. Biol.* doi:10.1007/978-1-62703-986-4_4

- The boundary (distance) classifies sequences as 'in' or 'out'
- Sequences are predicted to be *either* **in the class** or **not in the class**
- Changing distance/boundary changes classification



	IN	OUT
Red	5	2
Blue	4	33

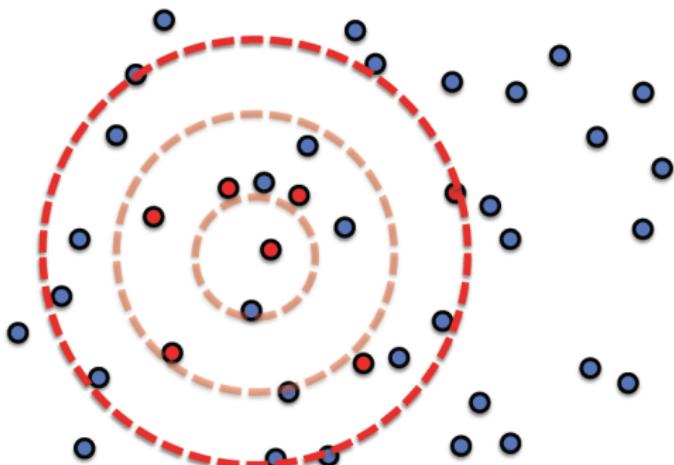
False positive rate	0.11
False negative rate	0.29
Sensitivity	0.81
Specificity	0.89



Large boundary ^a

^aPritchard & Broadhurst (2014) *Methods Mol. Biol.* doi:10.1007/978-1-62703-986-4_4

- The boundary (distance) classifies sequences as 'in' or 'out'
- Sequences are predicted to be *either* **in the class** or **not in the class**
- Changing distance/boundary changes classification



	IN	OUT
Red	7	0
Blue	14	23

False positive rate 0.38

False negative rate 0

Sensitivity 1

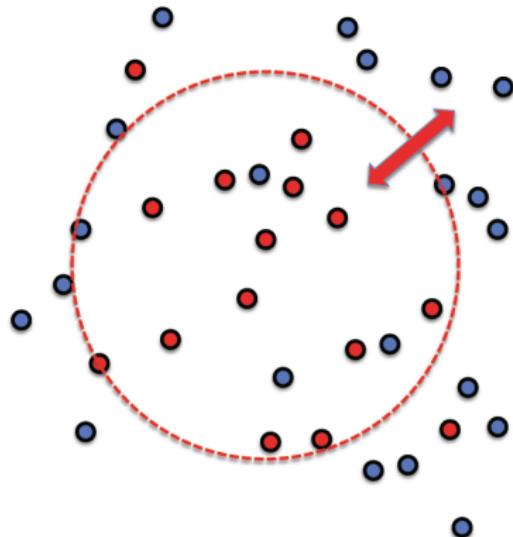
Specificity 0.62



Choosing a boundary ^a

^aPritchard & Broadhurst (2014) *Methods Mol. Biol.* doi:10.1007/978-1-62703-986-4_4

- Assign known ‘positive’ and ‘negative’ examples
- Vary ‘distance’ and measure predictive performance (F-measure, AUC, ...)
- Choose the distance that gives the ‘best’ performance





Crossvalidation ^a

^aPritchard & Broadhurst (2014) *Methods Mol. Biol.* doi:10.1007/978-1-62703-986-4_4

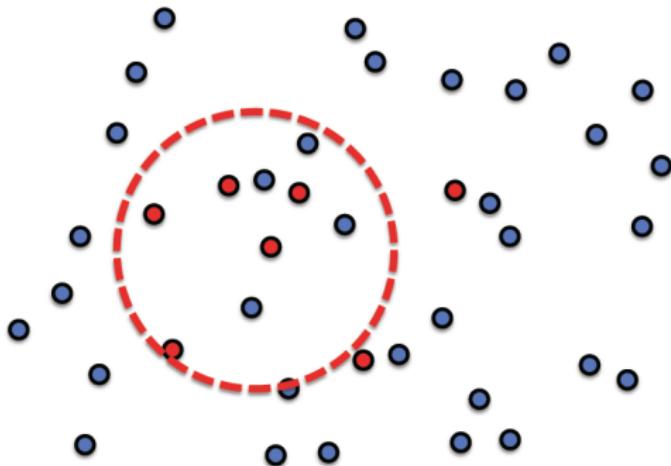
- Estimation of classifier performance depends on:
 - Boundary choice/distance measure
 - Composition of training set ('positives' and 'negatives')
- Cross-validation gives objective estimate of performance
- Many strategies (beyond today's scope), including:
 - leave-one-out (LOO)
 - k -fold crossvalidation
 - repeated (random) subsampling
- Always validate against a *hold-out set* (not used to train the classifier)



Post-crossvalidation ^a

^aPritchard & Broadhurst (2014) *Methods Mol. Biol.* doi:10.1007/978-1-62703-986-4_4

- Crossvalidation gives 'best' method & parameters
- Apply 'best' method to complete dataset for prediction



False positive rate	0.11
False negative rate	0.29
Sensitivity	0.81
Specificity	0.89
Precision	0.56

BEWARE THE BASE RATE FALLACY!



OPTIONAL WORKSHEET

worksheets/03-effector-finding.ipynb

The Baserate Fallacy in effector prediction and classification.

- MyBinder link



Licence: CC-BY-SA

By: Leighton Pritchard

This presentation is licensed under the Creative Commons Attribution ShareAlike license

<https://creativecommons.org/licenses/by-sa/4.0/>