Pathogen Genome Data

My life in sequences

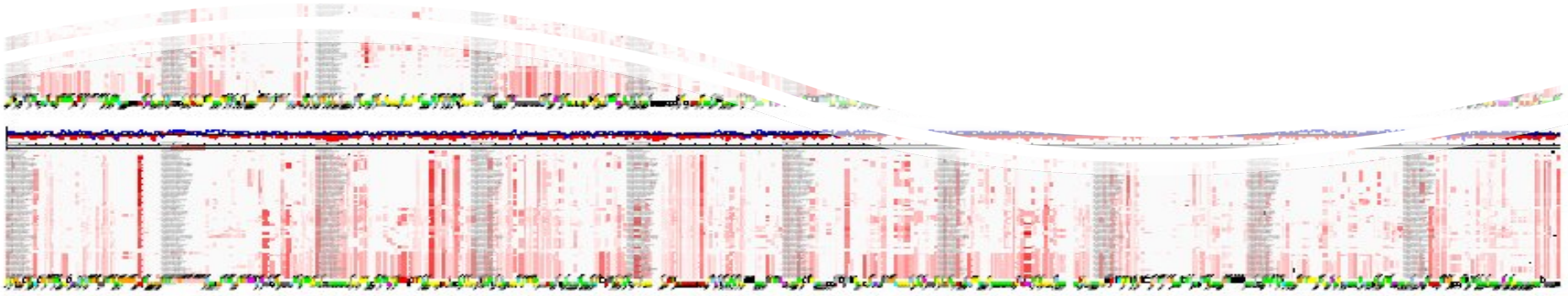
The James Hutton Institute

Leighton Pritchard

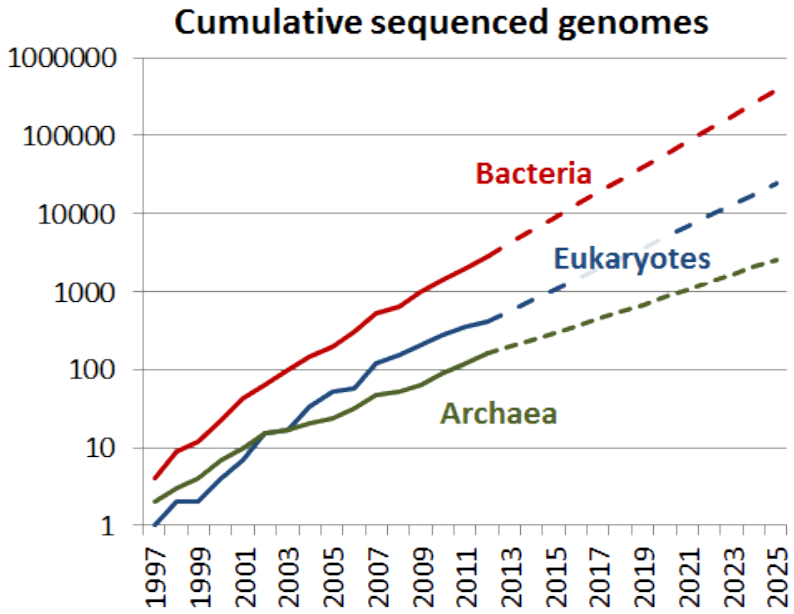
Introduction

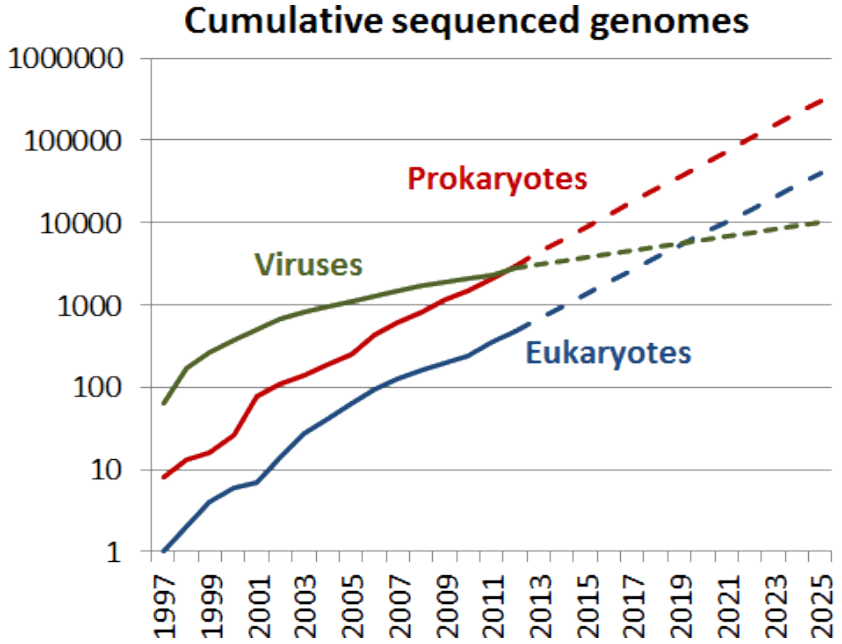
The James Hutton Institute

- 2003: *Erwinia carotovora* subsp. *atroseptica*
- 2009: *Phytophthora infestans*
- 2013: *Dickeya* spp.

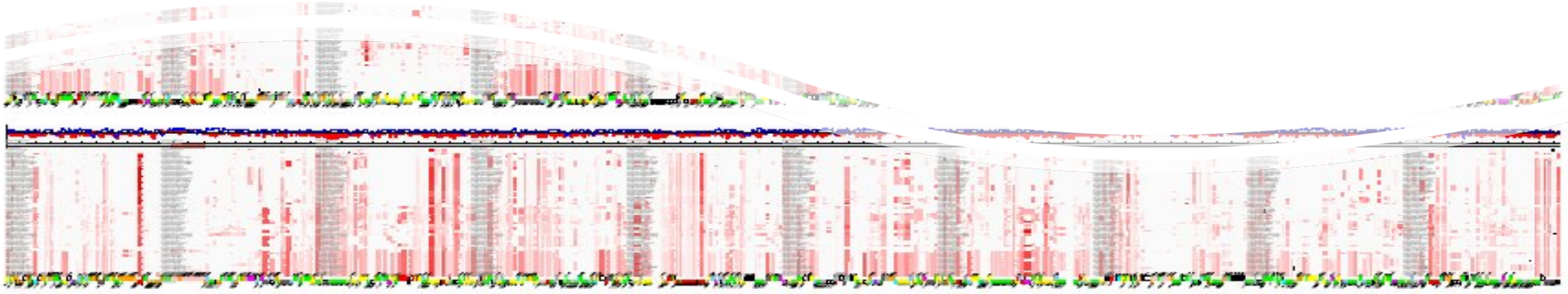


Introduction





Figures and code from: http://sulab.org/2013/06/sequenced-genomes-per-year/

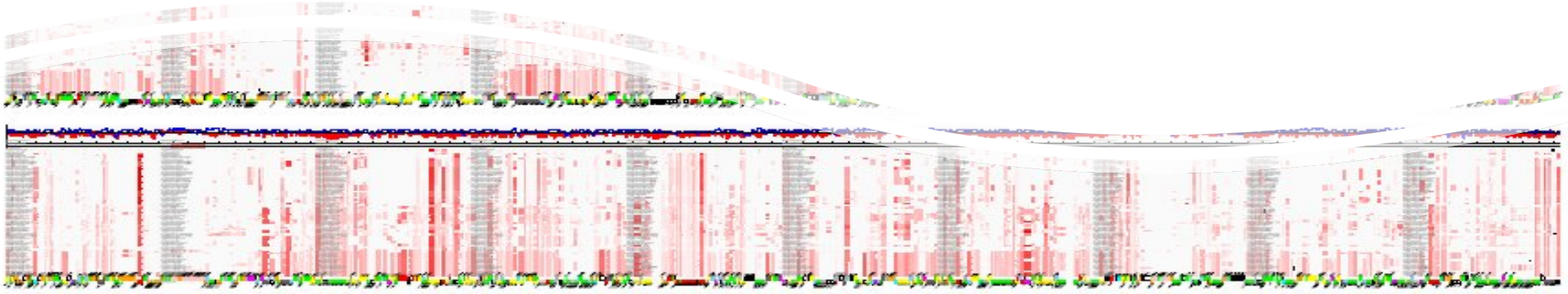


Introduction

The James Hutton Institute

- As the number of sequenced genomes rises…
  - The collective power of genomics to explain biology increases…
  - The impact of a single genome falls…
  - The impact of a single genome paper falls faster…
  - The time available for any scientist to annotate and examine an individual genome in detail falls even faster than that!

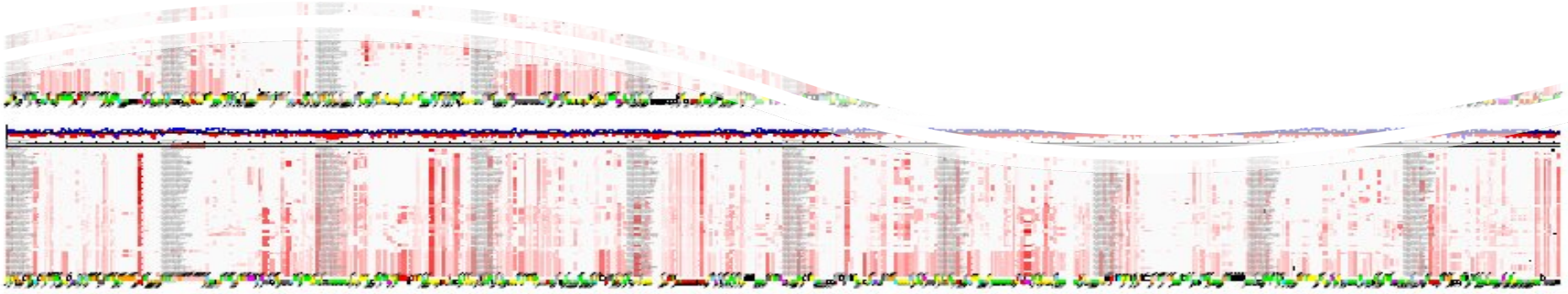
- I discovered this the hard way…



Introduction

The James Hutton Institute

- What can pathogen genome data do for you?
  - Catalogue of genome components
  - Differences in genomic complement: hypotheses for function
  - Characteristic signals for diagnosis and epidemiology
  - Parts lists and wiring diagrams for systems and synthetic biology
- How do we get to biology from genome data?



2003: *Erwinia carotovora* subsp. *atroseptica*

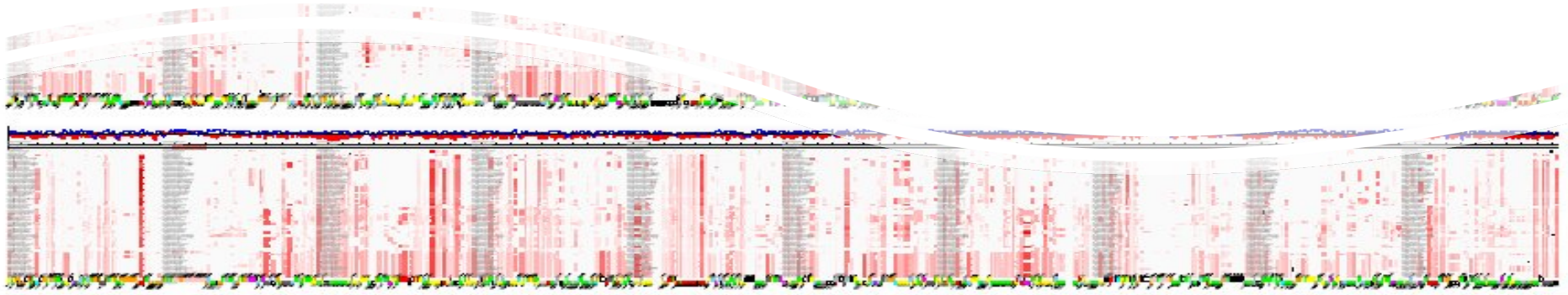
The James Hutton Institute

Genome sequence of the enterobacterial phytopathogen *Erwinia carotovora* subsp. *atroseptica* and characterization of virulence factors

K. S. Bell*†, M. Sebaihia*, L. Pritchard†, M. T. G. Holden*, L. J. Hyman†, M. C. Holeva†, N. R. Thomson*, S. D. Bentley*, L. J. C. Churcher*, K. Mungall*, R. Atkin*, N. Bason*, K. Brooks*, T. Chillingworth*, K. Clark*, J. Doggett*, A. Fraser*, Z. Hance*, H. Hauser*, K. Jagels*, S. Moule*, H. Norbertczak*, D. Ormond*, C. Price*, M. A. Quail*, M. Sanders*, D. Walker*, S. Whitehead*, G. P. C. Salmond‡, P. R. J. Birch†, J. Parkhill*§, and I. K. Toth†§

*The Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SA, United Kingdom; †Plant–Pathogen Interactions Programme, Scottish Crop Research Institute, Invergowrie, Dundee DD2 5DA, United Kingdom; and ‡Department of Biochemistry, Tennis Court Road, Cambridge University, Cambridge CB2 1QW, United Kingdom

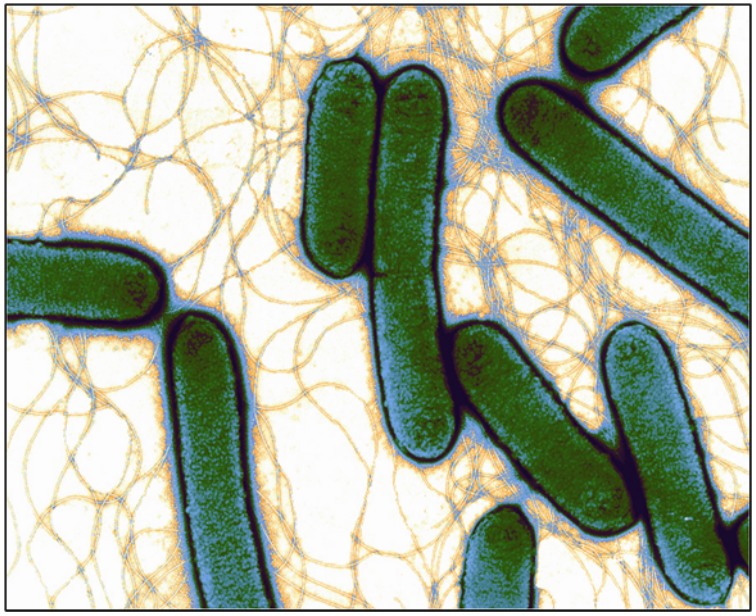
Bell *et al.* (2004) *Proc Natl Acad Sci USA* doi:10.1073/pnas.0402424101.



Pectobacterium atrosepticum

The James Hutton Institute

- Plant pathogenic enterobacterium
- *E. carotovora* subsp. *atroseptica* reclassified as *Pectobacterium atrosepticum*
- Causes blackleg (stem rot), rotting of stored tubers
- Major rot symptoms due to PCWDEs



Pectobacterium atrosepticum

The James Hutton Institute

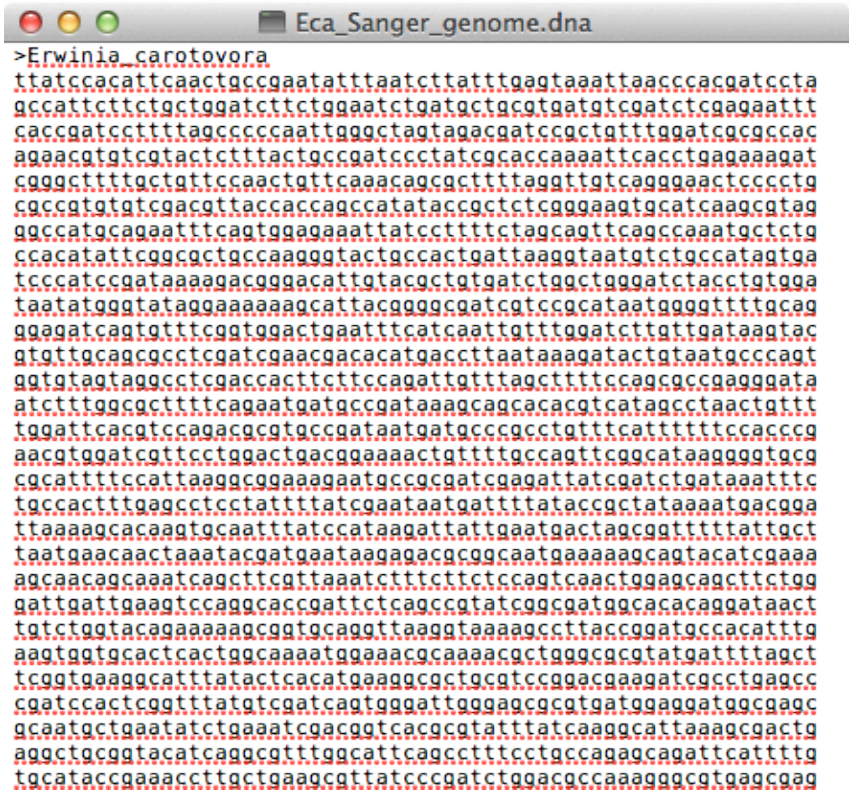
- First plant pathogenic enterobacterium to be sequenced
- £250,000 collaboration between SCRI, University of Cambridge, and WT Sanger Institute
- WGS to 10.2X coverage, 106,500 reads
- Repeats and gaps bridged and sequenced directly
- A single complete high quality 5Mbp circular chromosome
- Published 2004, *PNAS* (32 authors)

Bell *et al.* (2004) *Proc Natl Acad Sci USA* doi:10.1073/pnas.0402424101.

Pectobacterium atrosepticum

The James Hutton Institute

- Having a genome sequence is useful, but…



- Relive the annotation with `Artemis`!
- `Eca_Sanger_genome.dna`
- `Eca_Sanger_annotation.gbk`
- Manual annotation by Sanger Pathogen Sequencing Unit:

  gene ID: `FASTA`, `BLASTN`, `BLASTX`

  genecalls: `ORPHEUS`, `GLIMMER`

  domain ID: `PFAM`, `SIGNALP`, `TMHMM`

  metabolism: `KEGG`

  ncRNA: `RFAM`

  Literature search and comparison
- ≈6 people, ≈6 months

A Word About Annotation

The James Hutton Institute

- Annotation quality is critical to downstream biology!
- Annotation is ***curation***, not cataloguing!

  (but may have no budget on a project)

  (not a new situation: Enzyme Nomenclature Committee never funded...)
- Automation from curated data is the **ONLY** game in town, but you can't propagate information that does not exist!

  (≈30% of metabolic activity has no known gene associated with it)
- Core, well-curated resources are essential

(Dirty little fact: biocurators can spend as much time "de-annotating" literature-based annotations as entering new data)

Chen and Vitkup (2007) *Trends Biotech.* doi:10.1016/j.tibtech.2007.06.001
Bairoch (2009) *Nature Precedings.* doi:10.1038/npre.2009.3092.1

Another Word About Annotation

The James Hutton Institute

- Biocuration (i.e. maintenance of reliable annotation corpus) also requires:
  - software platforms and interfaces (web-based is good…)
  - controlled vocabularies and ontologies (GO, PAMGO, SO, …)
  - computing and network infrastructure, IT and sysadmin support
- **Community annotation:** "Many potential parents, but no-one wants to look after the children"
- Key problem: lack of direct incentives to curate (one's own) data.
- Solution: ELIXIR and centralised resources, e.g. EBI (Ensembl etc.)

Bairoch (2009) *Nature Precedings.* doi:10.1038/npre.2009.3092.1

Comparative Genomics

The James Hutton Institute

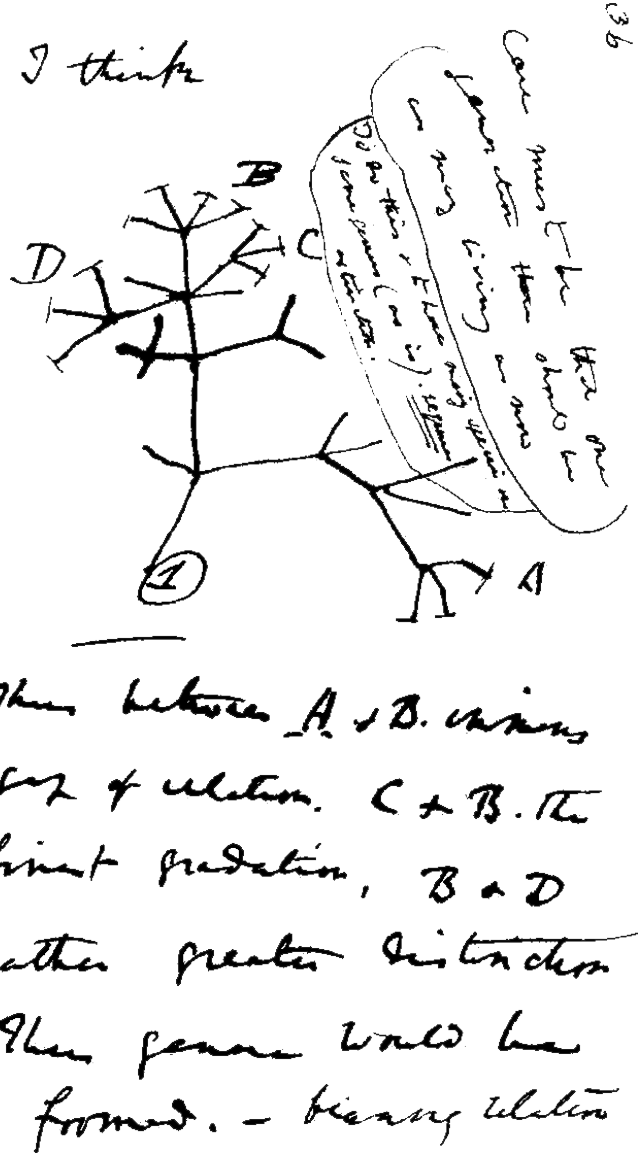
- Having a single genome is useful, but…

…the real power of genomics is ***comparative genomics***

- **The combination of genomic data and comparative and evolutionary biology to address questions of genome structure, evolution and function.**

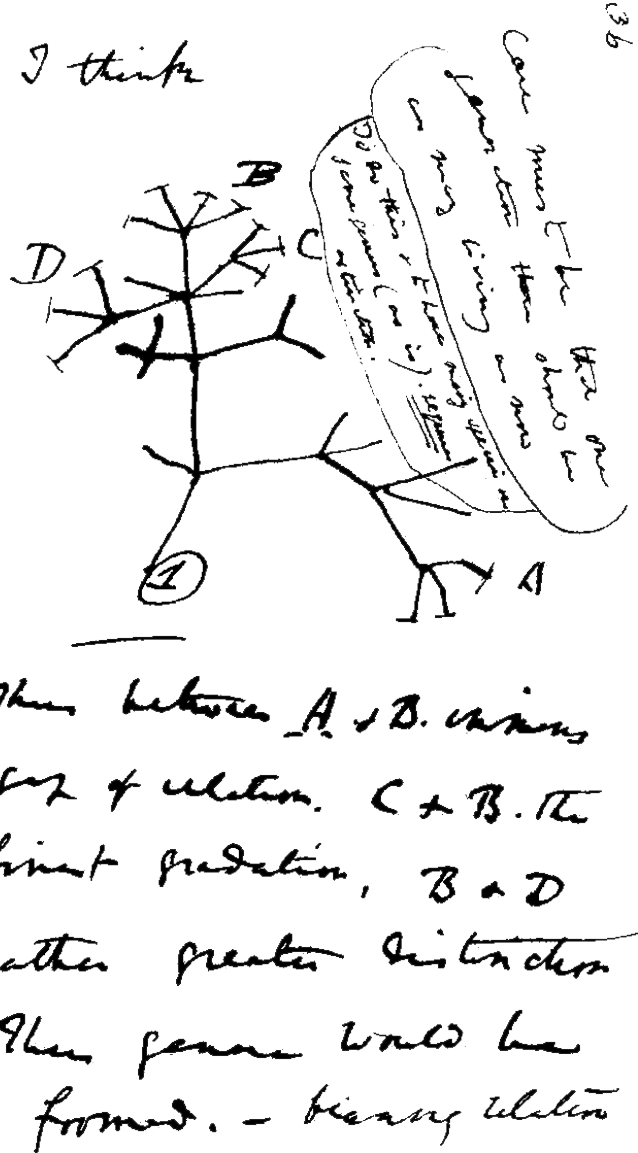
Why Comparative Genomics?

- Genomes describe heritable characteristics
- Related organisms share ancestral genomes
- Functional elements encoded in genomes are common to related organisms
- Functional understanding of model systems (*E. coli*, *A. thaliana*, *D. melanogaster*) can be transferred to non-model systems on the basis of genome comparisons
- Genome comparisons can be informative, even for distantly-related organisms

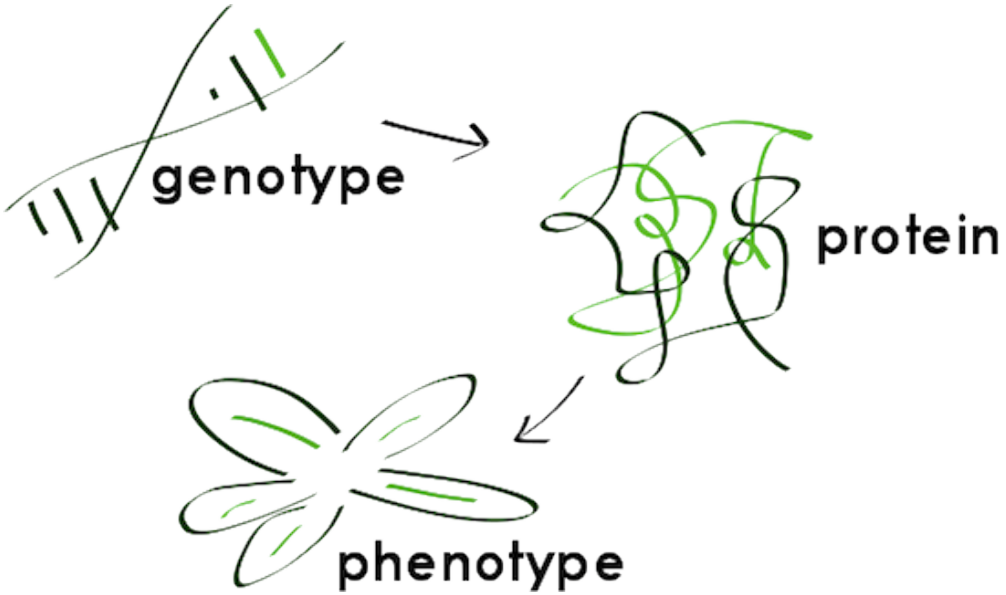


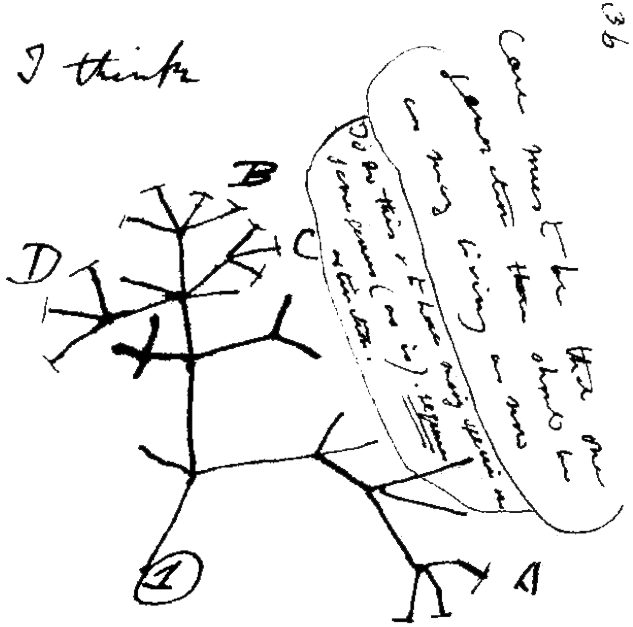
Why Comparative Genomics?

- Genomic differences can underpin phenotypic (morphological, physiological, host range, virulence) differences
- If phenotypes or other organism-level properties are known, comparison of genomes may give mechanistic or functional insight into differences (e.g. environmental persistence).
- Genome comparisons aid identification of functional elements on the genome.
- Studying genomic changes reveals evolutionary processes and constraints.



Caveats of Comparative Genomics





- **BUT:** the genome isn't everything
  - **Context:** epigenetics, tissue differentiation, mesoscale systems, etc.
  - **Phenotypic plasticity:** responses to temperature, stress, environment, etc.

(therefore Systems Biology, of which more later…)

Levels of Genome Comparison

The James Hutton Institute

- Bulk Properties
  - chromosome/plasmid counts and sizes,
  - nucleotide content, *k*-mers, etc.
- Whole Genome Sequence
  - sequence similarity
  - genomic structure and organisation (rearrangements), etc.
- Genome Features/Functional Components
  - numbers and types of features (genes, ncRNA, regulatory elements, etc.)
  - organisation of features (synteny, operons, regulons, etc.)
  - complements of features
  - selection pressure, etc.

Pba Comparative Genomics Questions

The James Hutton Institute

- Just about enough sequenced genomes in 2002/3 to ask:
- What are the gene complement similarities and differences between *Pba* and:
    - other enterobacteria?
    - other plant pathogens?
    - animal pathogens?
    - other environmental bacteria?
- What is genomic evidence for pathogenicity determinants?
- Do pathogenicity determinants show evidence of LGT?
- Are there characteristic features of the *Pba* genome?

Comparing Gene Features

The James Hutton Institute

- Given gene annotations for more than one genome, one can organise and understand relationships in multiple ways
  - Functional similarity (analogy)
  - Evolutionary common origin (homology, orthology, etc.)
  - Evolutionary/functional/family relationships (paralogy)

DISTINGUISHING HOMOLOGOUS FROM ANALOGOUS PROTEINS

WALTER M. FITCH

Abstract

Fitch, W. M. (Dept. Physiological Chem., U. Wisconsin, Madison 53706) 1970. Distinguishing homologous from analogous proteins. Syst. Zool., 19:99–113.—This work provides a means by which it is possible to determine whether two groups of related proteins have a common ancestor or are of independent origin. A set of 16 random

Terms first suggested by Fitch (1970) *Syst. Zool.* doi:10.2307/2412448
Fitch (2000) *Trends Genet*. doi:10.1016/S0168-9525(00)02005-9

Attack of the –logues

The James Hutton Institute

- **Important technical terms describing feature relationships**
- **Homologues:** elements that are similar because they share a common ancestor (**NOTE: There are NOT degrees of homology!**)
- **Analogues:** elements that are (functionally?) similar, maybe through convergent evolution rather than common ancestry
- **Orthologues:** homologues that diverged through speciation
- **Paralogues:** homologues that diverged through duplication within the same genome
- (also co-orthologues, xenologues, etc.)

Fitch (1970) *Syst. Zool.* doi:10.2307/2412448
Fitch (2000) *Trends Genet*. doi:10.1016/S0168-9525(00)02005-9

Note on "Orthology"

The James Hutton Institute

- Frequently abused/misused as a term
- "Orthology" was defined as an evolutionary relationship, but now often bent into service as a functional descriptor
- Strictly defined only for two species or clades!
  - (cf. cluster definitions OrthoMCL, etc.)
- Orthology is not always transitive
  (A is orthologue of C and B is orthologue of C does not necessarily imply A is an orthologue of B)

Storm & Sonnhammer (2002) *Bioinformatics.* doi:10.1093/bioinformatics/18.1.92
Jensen (2001) *Genome Biol.* doi:10.1186/gb-2001-2-8-interactions1002
Fitch (2000) *Trends Genet*. doi:10.1016/S0168-9525(00)02005-9

Finding “Orthologues”

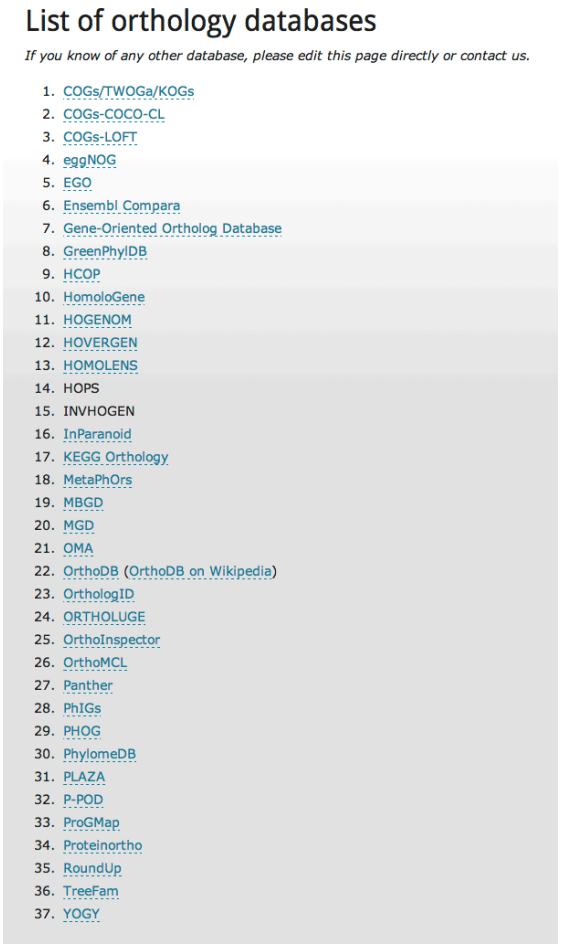
The James Hutton Institute

Or, more pragmatically for hypothesis generation…

- **Finding evolutionary (and/or functional) equivalents of features across two or more organisms’ genomes.**

Why "orthologues"?

- Need to formalise the concept of ***corresponding genes*** across multiple organisms.
  - Evolutionary
  - Functional? (**"The Ortholog Conjecture"**)
- Many (>35) databases attempt to describe orthology relationships
  - http://questfororthologs.org/orthology_databases



List of orthology databases

If you know of any other database, please edit this page directly or contact us.

1. COGs/TWOGs/KOGs
2. COGs-COCO-CL
3. COGs-LOFT
4. eggNOG
5. EGO
6. Ensembl Compara
7. Gene-Oriented Ortholog Database
8. GreenPhylDB
9. HCOP
10. HomoloGene
11. HOGENOM
12. HOVERGEN
13. HOMOLENS
14. HOPS
15. INVHOGEN
16. InParanoid
17. KEGG Orthology
18. MetaPhOrs
19. MBGD
20. MGD
21. OMA
22. OrthoDB (OrthoDB on Wikipedia)
23. OrthologID
24. ORTHOLUGE
25. OrthoInspector
26. OrthoMCL
27. Panther
28. PhIGs
29. PHOG
30. PhylomeDB
31. PLAZA
32. P-POD
33. ProGMap
34. Proteinortho
35. RoundUp
36. TreeFam
37. YOGY

Dessimoz (2011) *Brief. Bioinf.* doi:10.1093/bib/bbr057
Chen & Zhang (2012) *PLoS Comp. Biol*. doi:10.1371/journal.pcbi.1002784
Altenhoff *et* al. (2012) *PLoS Comp.* Biol. doi:10.1371/journal.pcbi.1002514

How to find "orthologues"?

The James Hutton Institute

- Many published methods and databases:
  - Pairwise between two genomes:
    - RBBH (aka BBH, RBH, etc.), RSD, InParanoid, RoundUp
  - Multi-genome:
    - Graph-based: COG, eggNOG, OrthoDB, OrthoMCL, OMA, MultiParanoid
    - Tree-based: TreeFam, Ensembl Compara, PhylomeDB, LOFT
- Methods may apply different - or *refined* - definitions of orthology, paralogy, etc.

Salichos *et al.* (2011) *PLoS One.* doi:10.1371/journal.pone.0018755
Trachana *et al.* (2011) *Bioessays* doi:10.1002/bies.201100062
Kristensen *et al.* (2011) *Brief. Bioinf.* doi:10.1093/bib/bbr030

Evaluating "Orthologue" Prediction

The James Hutton Institute

- Works the same way as for all prediction tools

1. Define a "validation set" (gold standard), unseen by the prediction tool
2. Make predictions with the tool
3. Evaluate confusion matrix and performance statistics
   - Sensitivity
   - Specificity
   - Accuracy

Standard:	+ve	-ve
Predict +ve	TP	FP
Predict -ve	FN	TN

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

"Orthologue" Prediction Performance

The James Hutton Institute

- Performance varies by choice of method and interpretation of "orthology"
- Biggest influence is often annotation quality
- Relative performance varies with benchmark choice
- (clustering) RBH outperforms the more complex algorithms under many circumstances

Wolf *et al.* (2012) *Genome Biol. Evol.* doi:10.1093/gbe/evs100
Salichos *et al.* (2011) *PLoS One.* doi:10.1371/journal.pone.0018755
Altenhoff & Dessimoz (2009) *PLoS Comp. Biol.* doi:10.1371/journal.pcbi.1000262
Trachana *et al.* (2011) *Bioessays.* doi:10.1002/bies.201100062

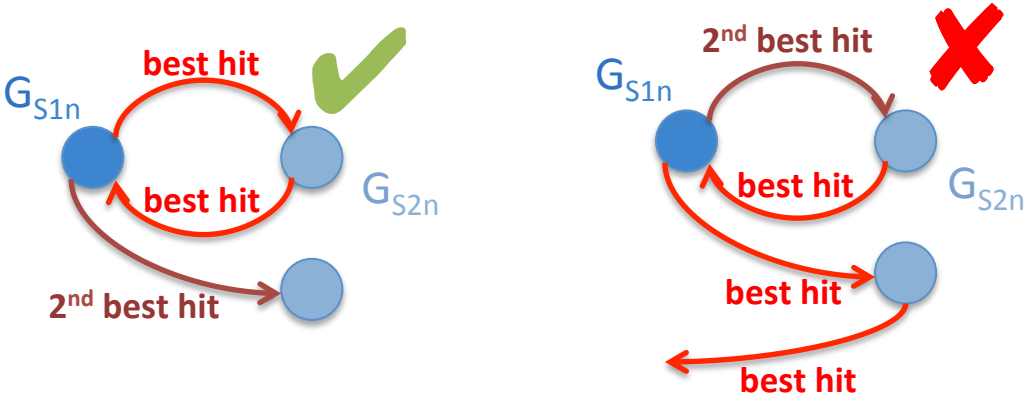
Back To Our *Pba* Questions

The James Hutton Institute

- What are the gene complement similarities and differences between *Pba* and:
    - other enterobacteria?
    - other plant pathogens?
    - animal pathogens?
    - other environmental bacteria?
- What is genomic evidence for pathogenicity determinants?
- Do pathogenicity determinants show evidence of LGT?
- Are there characteristic features of the *Pba* genome?

Pba Reciprocal Best Hits (RBH)

- RBH not necessarily orthologues, but "*equivalent*" features
- Compared to 64 completely sequenced bacterial genomes
  (most of those available at the time, also limited by computing power)
- Limited matches to:
  >30% identity (above "twilight zone")
  >80% coverage (exclude domain-only matches)



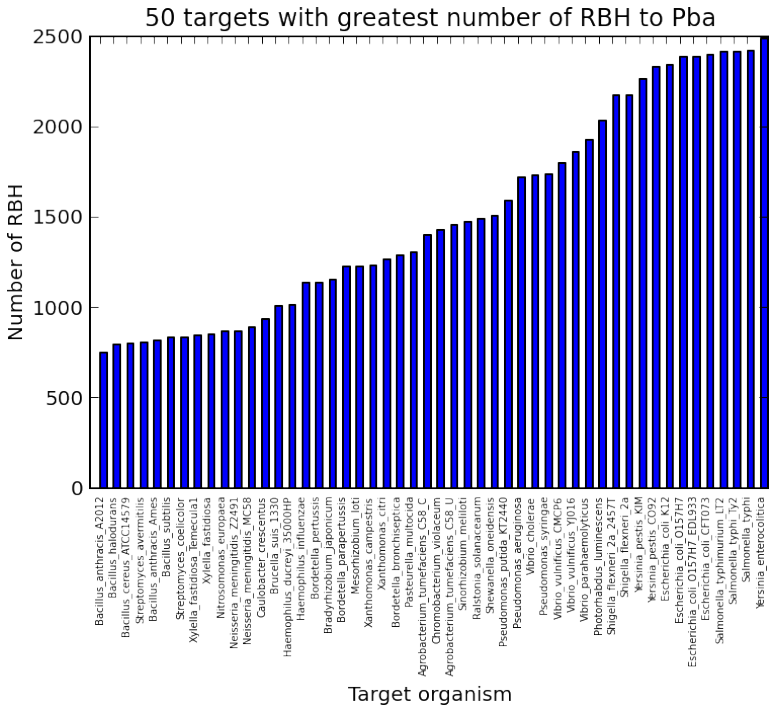
Pba Reciprocal Best Hits (RBH)

The James Hutton Institute

- **Advantages:**
  - quick
  - easy
  - performs surprisingly well
- **Disadvantages:**
  - 1:1 matches miss paralogues
  - not good at identifying gene families or *-to-many relationships without more detailed analysis.
  - no strong theoretical/phylogenetic basis.

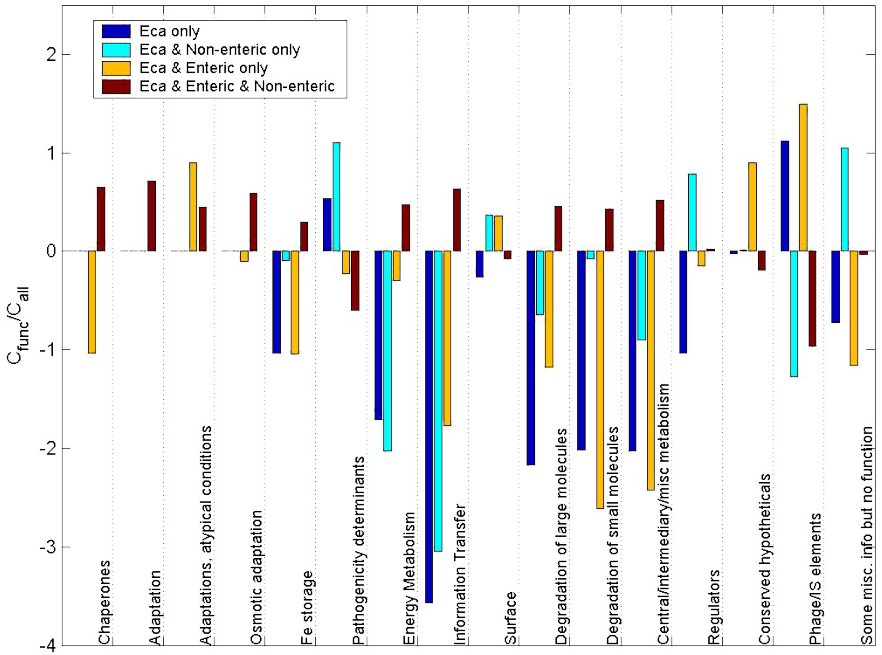
Pba Reciprocal Best Hits (RBH)

- Most RBH with other enterobacteria (see `pba_rbh.ipynb`)
- Pathogenicity factors also shared with plant-associated non-enterobacterial genomes



Functional Classification

- Compare frequency of functional classifications in *Pba* and other organism classes (**NOTE**: annotation-dependent)
- Test statistically for enrichment/overrepresentation
- Many, many methods & programs for this
  (68 tools summarised in 2009 review)
- *Pba*-specific:
  - Pathogenicity determinants
  - Phage/IS elements
- *Pba* & non-enteric (LGT?)
  - Pathogenicity determinants
  - Surface proteins
  - Regulatory proteins

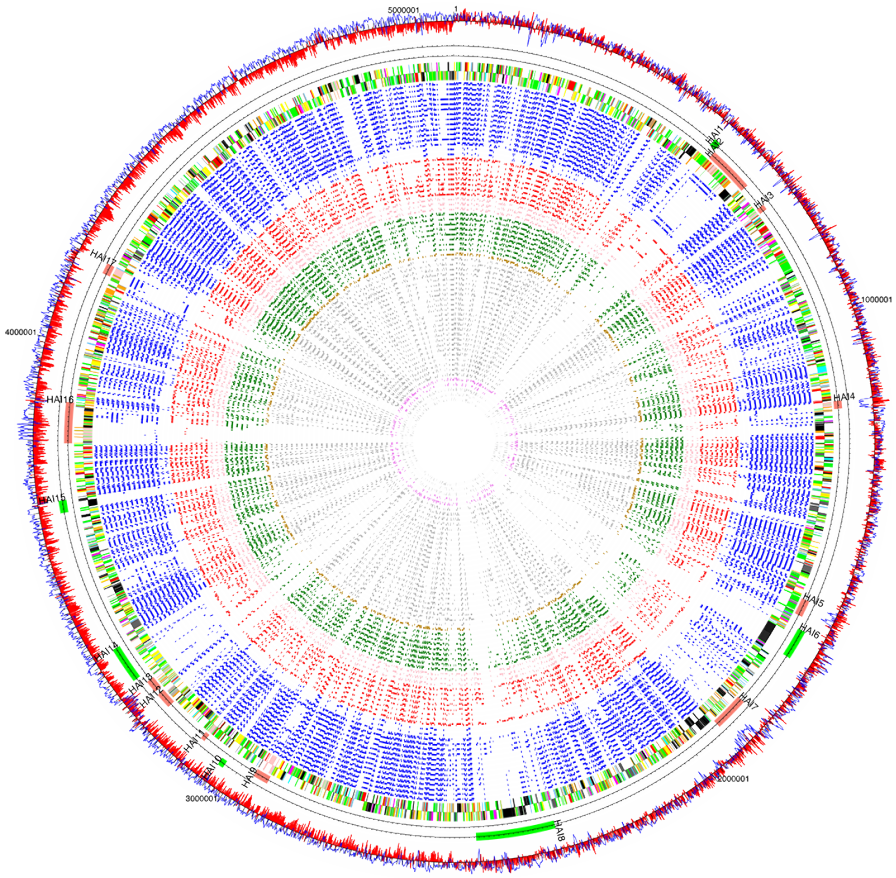


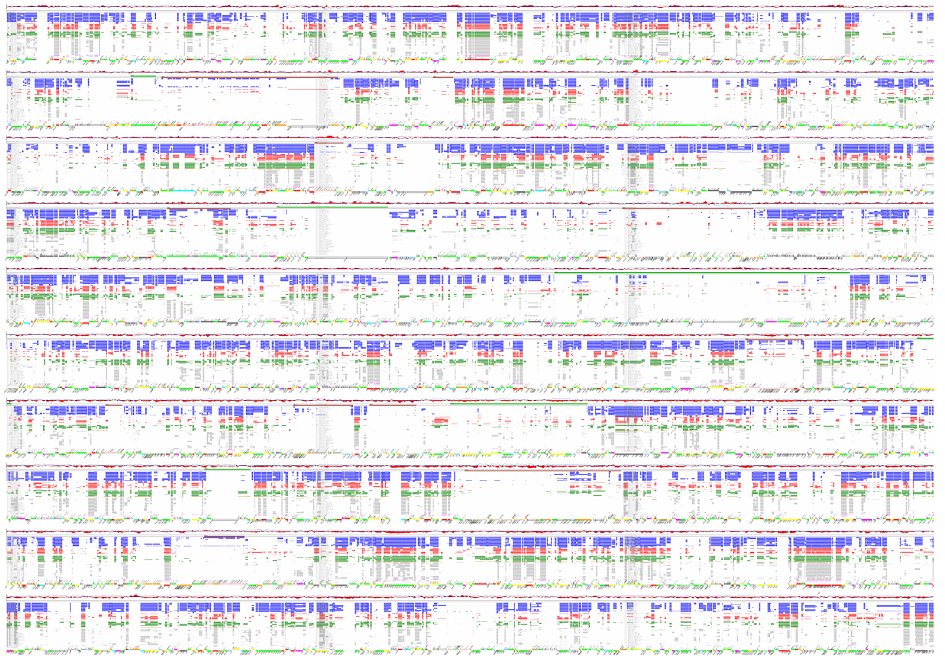
Huang *et al.* (2009) *Nucleic Acids Res.* doi:10.1093/nar/gkn923.

Pba RBH Visualisation

The James Hutton Institute

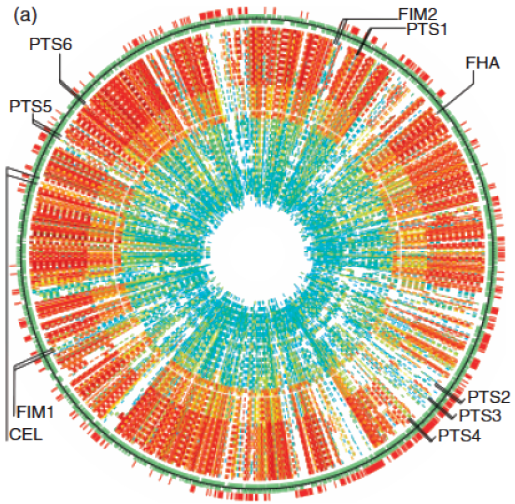
- 114,087 RBH between *Pba* and target bacteria
  too much to go through by hand/eye
- Visualisation tools were essential

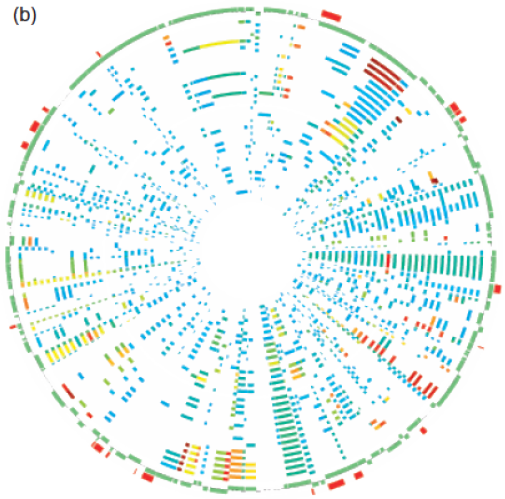


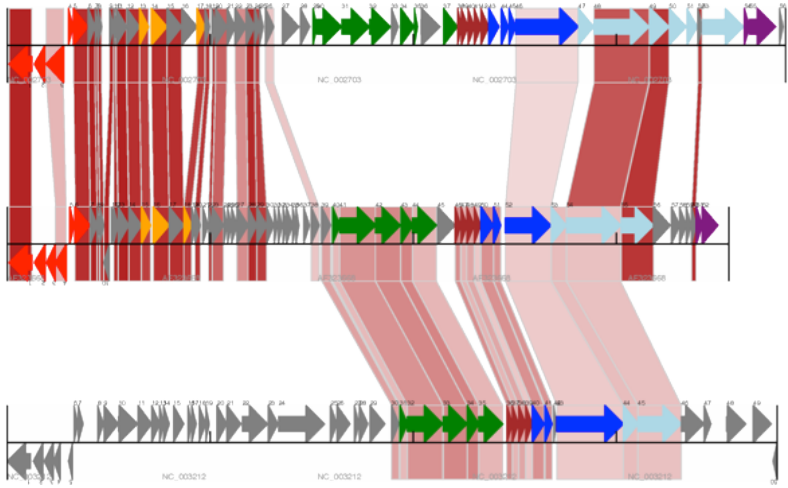


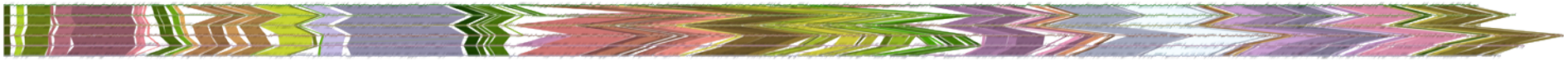
GenomeDiagram

- Comparative genomics visualisation package
- Developed in 2003 for *Pba* sequencing, later incorporated into Biopython









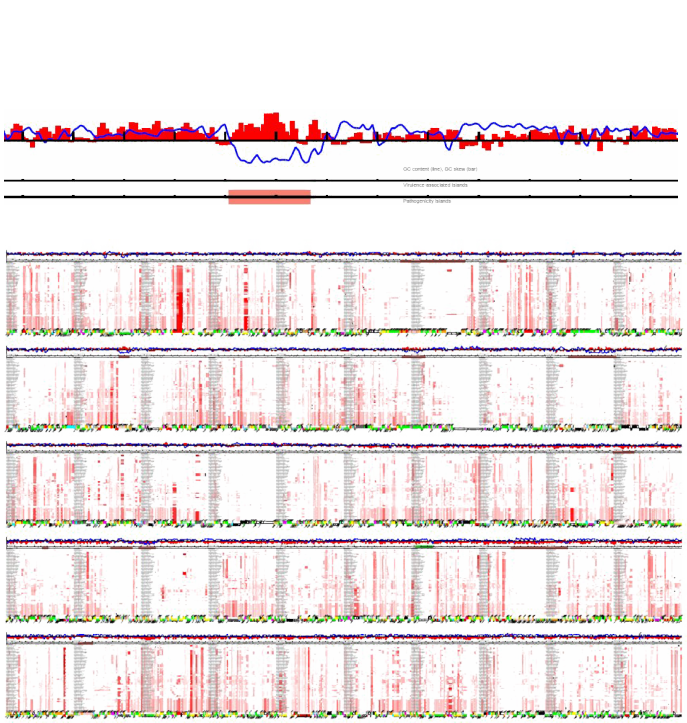
http://www.biopython.org

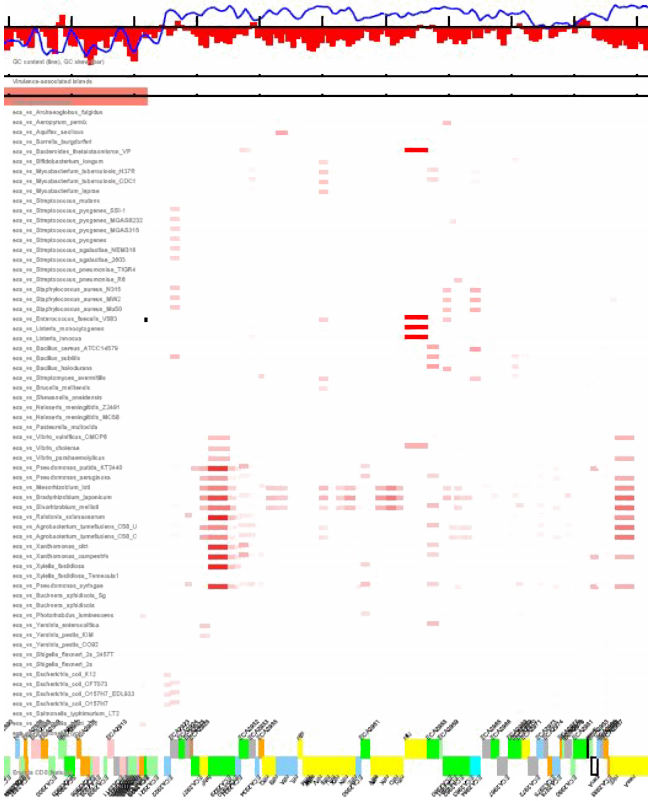
Pritchard *et al.* (2006) *Bioinformatics* doi:10.1093/bioinformatics/btk021.

Visualising LGT in *Pba*

The James Hutton Institute

- Looked for CDS where RBH sequence similarity doesn't follow the taxonomic species tree, and GC% anomalies
- Colour sequence similarity: more red = more similar





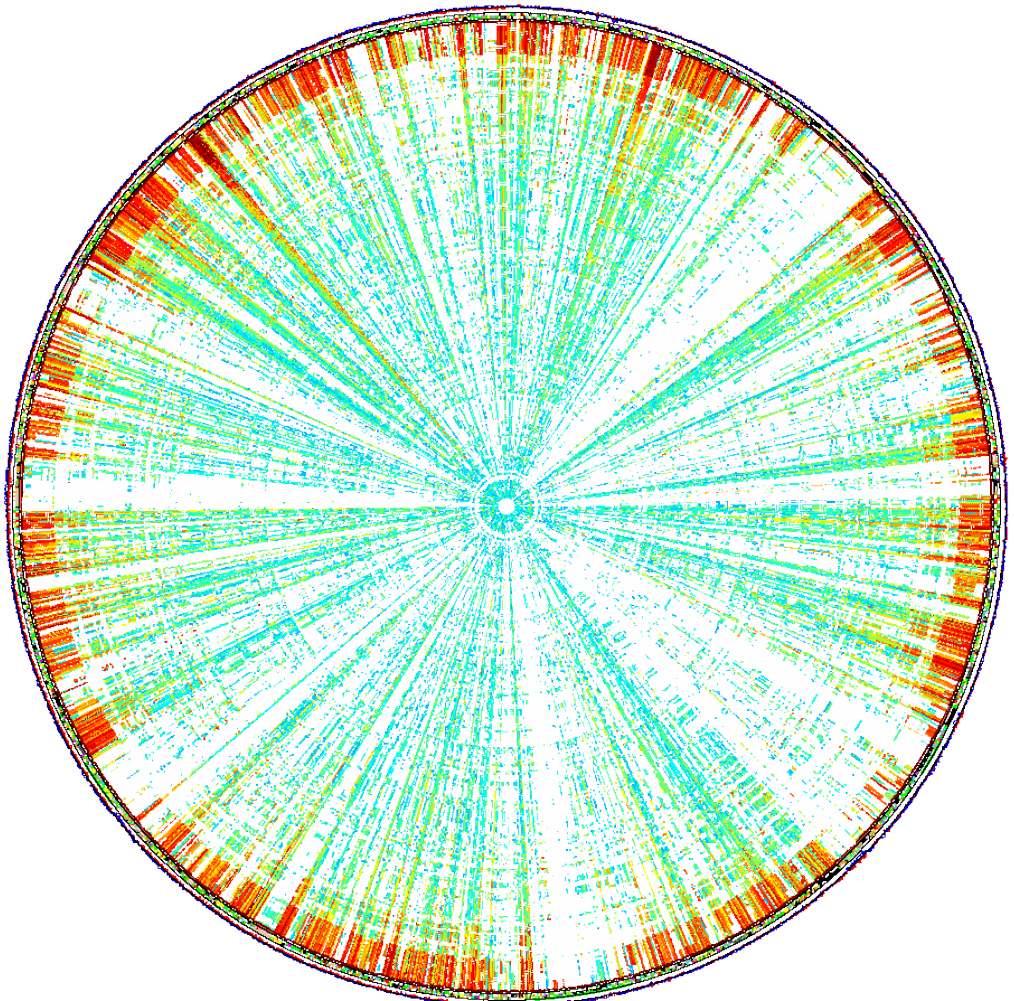
Pba LGT

The James Hutton Institute

- Initially, 11 putative horizontally-acquired islands (HAIs) identified by atypical G+C, proximity of tRNA, etc.
- Six further putative HAIs identified on the basis of RBH (may be more ancient acquisitions)
- **HAI8**: Type III Secretion System, *Pseudomonas syringae*
- **HAI2**: Phytotoxin (*cfa*) synthesis, *P. syringae*
- **HAI15**: Adherence/agglutination, *Vibrio* spp.
- **HAI14**: Nitrogen fixation, *Bradyrhizobium* spp.

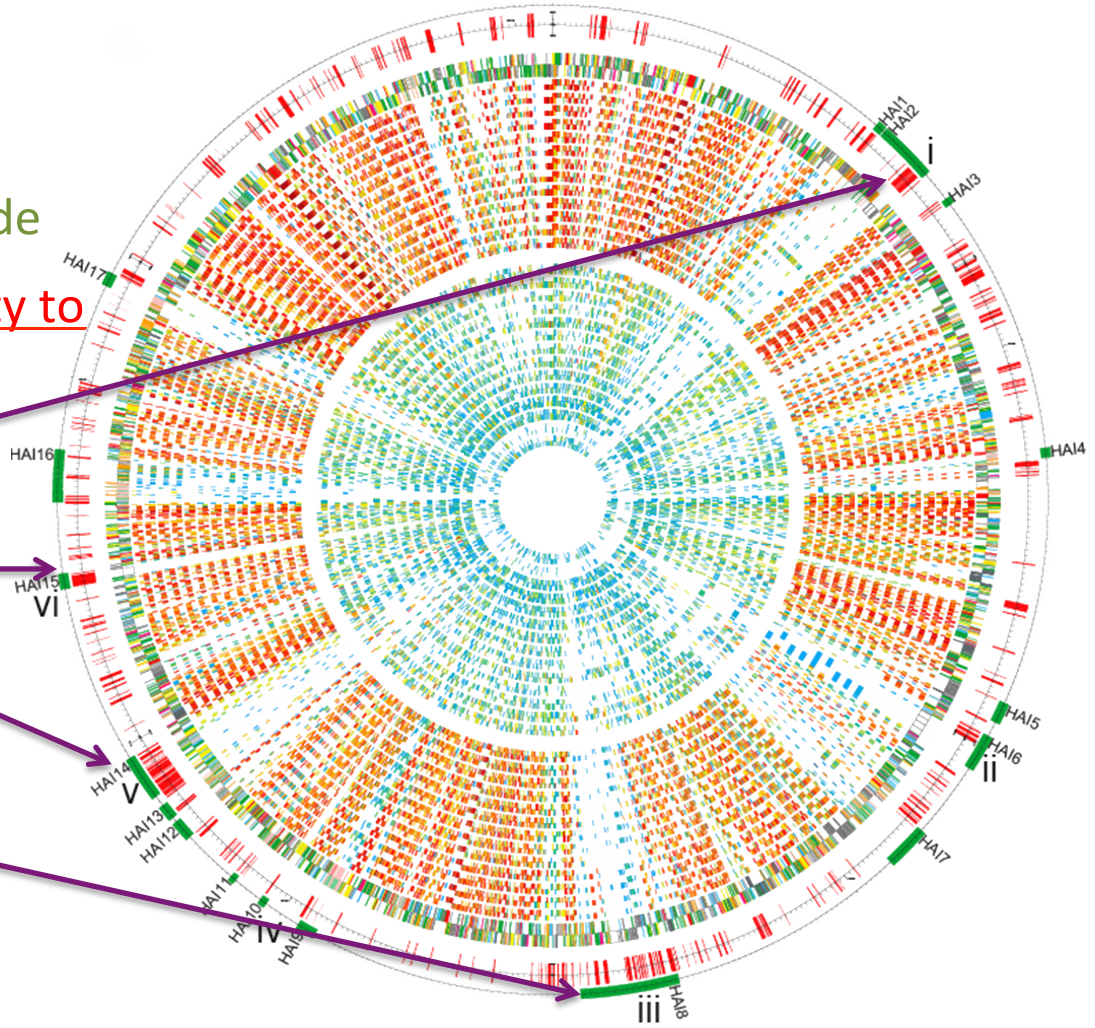
Pba LGT: Plant-associated bacteria

- 2006: Two years later, far more genomes (400) for comparison
- Larger selection of plant- and animal-associated bacteria



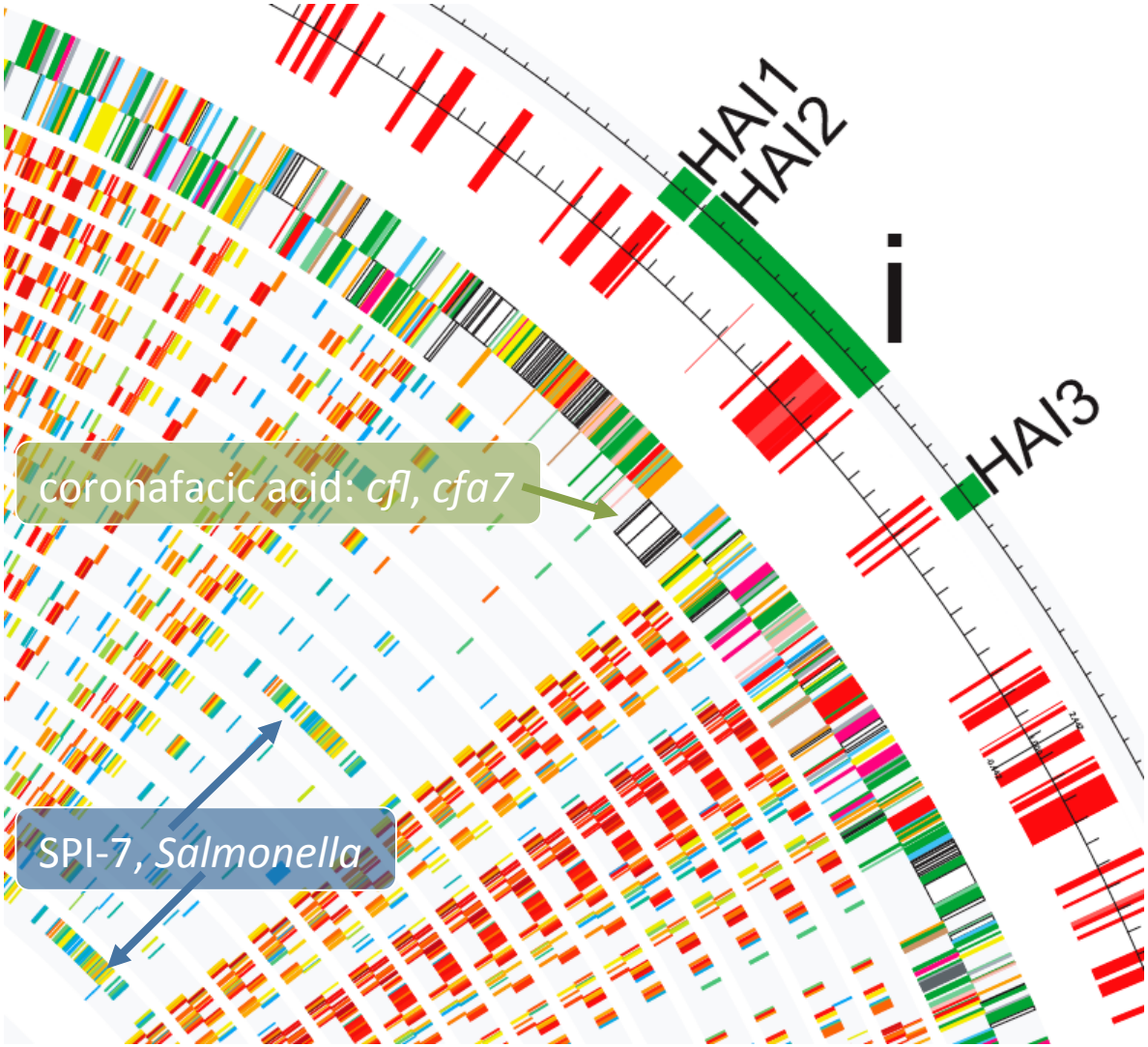
Pba LGT: Plant-associated bacteria

- Comparison against plant- (13) and animal-associated (14) bacteria
- Plant-associated in centre
- Animal-associated on outside
- Red marks: greater similarity to plant-associated bacteria
- **HAI2**: Phytotoxin
- **HAI15**: Adherence
- **HAI14**: Nitrogen fixation
- **HAI8**: T3SS



Toth *et al.* (2006) *Ann. Rev. Phytopathol.* doi:10.1146/annurev.phyto.44.070505.143444

SPI-7 and Coronafacic Acid



Coronatine (*P. syringae*) interferes with jasmonate responses in host, as a jasmonate mimic

Coronafacic acid – *Pseudomonas syringae* phytotoxin precursor (coronatine)
- payload

SPI-7 - *Salmonella* Typhi Pathogenicity island
- delivery system

Block *et al*. (2005) *Mol. Plant Pathol.* doi:10.1111/j.1364-3703.2004.00265.x.
Toth *et al.* (2006) *Ann. Rev. Phytopathol.* doi:10.1146/annurev.phyto.44.070505.143444

Coronafacic Acid Synthesis

- Important for virulence? Testable hypothesis.
- Present in *Pba*, some *Pcc*, virtually no *Dickeya*
- *cfl* and *cfa7* knockouts show reduced *Pba* virulence
    - 200cfu inoculation, measure 17dpi

WT | cfl::mTn5 | cfa7::mTn5

Slawiak & Lojkowska (2009) *Eur. J. Plant Pathol.* doi:10.1007/s10658-008-9418-7.

Conclusions I

The James Hutton Institute

- **Comparative genomics is powerful!**
- Good annotation is essential to understanding genomic data
- Useful, falsifiable hypotheses can be generated from inspection of genomic data
- **Visualisation is extremely useful for large genomic datasets**
- Determining orthology/functional equivalence on large scale data is not yet a solved problem

2009: *Phytophthora infestans*

Vol 461|17 September 2009|doi:10.1038/nature08358 nature

LETTERS

Genome sequence and analysis of the Irish potato famine pathogen *Phytophthora infestans*

Brian J. Haas[1]*, Sophien Kamoun[2,3]*, Michael C. Zody[1,4], Rays H. Y. Jiang[1,5], Robert E. Handsaker[1], Liliana M. Cano[2], Manfred Grabherr[1], Chinnappa D. Kodira[1]†, Sylvain Raffaele[2], Trudy Torto-Alalibo[3]†, Tolga O. Bozkurt[2], Audrey M. V. Ah-Fong[6], Lucia Alvarado[1], Vicky L. Anderson[7], Miles R. Armstrong[8], Anna Avrova[8], Laura Baxter[9], Jim Beynon[9], Petra C. Boevink[8], Stephanie R. Bollmann[10], Jorunn I. B. Bos[3], Vincent Bulone[11], Guohong Cai[12], Cahid Cakir[3], James C. Carrington[13], Megan Chawner[14], Lucio Conti[15], Stefano Costanzo[16], Richard Ewan[15], Noah Fahlgren[13], Michael A. Fischbach[17], Johanna Fugelstad[11], Eleanor M. Gilroy[8], Sante Gnerre[1], Pamela J. Green[18], Laura J. Grenville-Briggs[7], John Griffith[14], Niklaus J. Grünwald[10], Karolyn Horn[14], Neil R. Horner[7], Chia-Hui Hu[19], Edgar Huitema[3], Dong-Hoon Jeong[18], Alexandra M. E. Jones[2], Jonathan D. G. Jones[2], Richard W. Jones[20], Elinor K. Karlsson[1], Sridhara G. Kunjeti[21], Kurt Lamour[22], Zhenyu Liu[3], LiJun Ma[1], Daniel MacLean[2], Marcus C. Chibucos[23], Hayes McDonald[24], Jessica McWalters[14], Harold J. G. Meijer[5], William Morgan[25], Paul F. Morris[26], Carol A. Munro[27], Keith O'Neill[1]†, Manuel Ospina-Giraldo[14], Andrés Pinzón[28], Leighton Pritchard[8], Bernard Ramsahoye[29], Qinghu Ren[30], Silvia Restrepo[28], Sourav Roy[6], Ari Sadanandom[15], Alon Savidor[31], Sebastian Schornack[2], David C. Schwartz[32], Ulrike D. Schumann[7], Ben Schwessinger[2], Lauren Seyer[14], Ted Sharpe[1], Cristina Silvar[2], Jing Song[3], David J. Studholme[2], Sean Sykes[1], Marco Thines[2,33], Peter J. I. van de Vondervoort[5], Vipaporn Phuntumart[26], Stephan Wawra[7], Rob Weide[5], Joe Win[2], Carolyn Young[3], Shiguo Zhou[32], William Fry[12], Blake C. Meyers[18], Pieter van West[7], Jean Ristaino[19], Francine Govers[5], Paul R. J. Birch[34], Stephen C. Whisson[8], Howard S. Judelson[6] & Chad Nusbaum[1]

Haas *et al.* (2009) *Nature* **461**: 393-398. doi:10.1038/nature08358.

Phytophthora infestans

The James Hutton Institute

- The most destructive pathogen of potato ($6.7bn/yr)
- Causes potato late blight
- Irish Potato Famine (1850s)
- Oomycete model organism, hemibiotrophic
- Adapts rapidly to overcome control measures and bred resistance

Phytopthora infestans

The James Hutton Institute

- Major international collaboration, 34 institutions
- WGS at Broad Institute of MIT and Harvard, to 9X coverage
- Four library types, due to highly repetitive genome

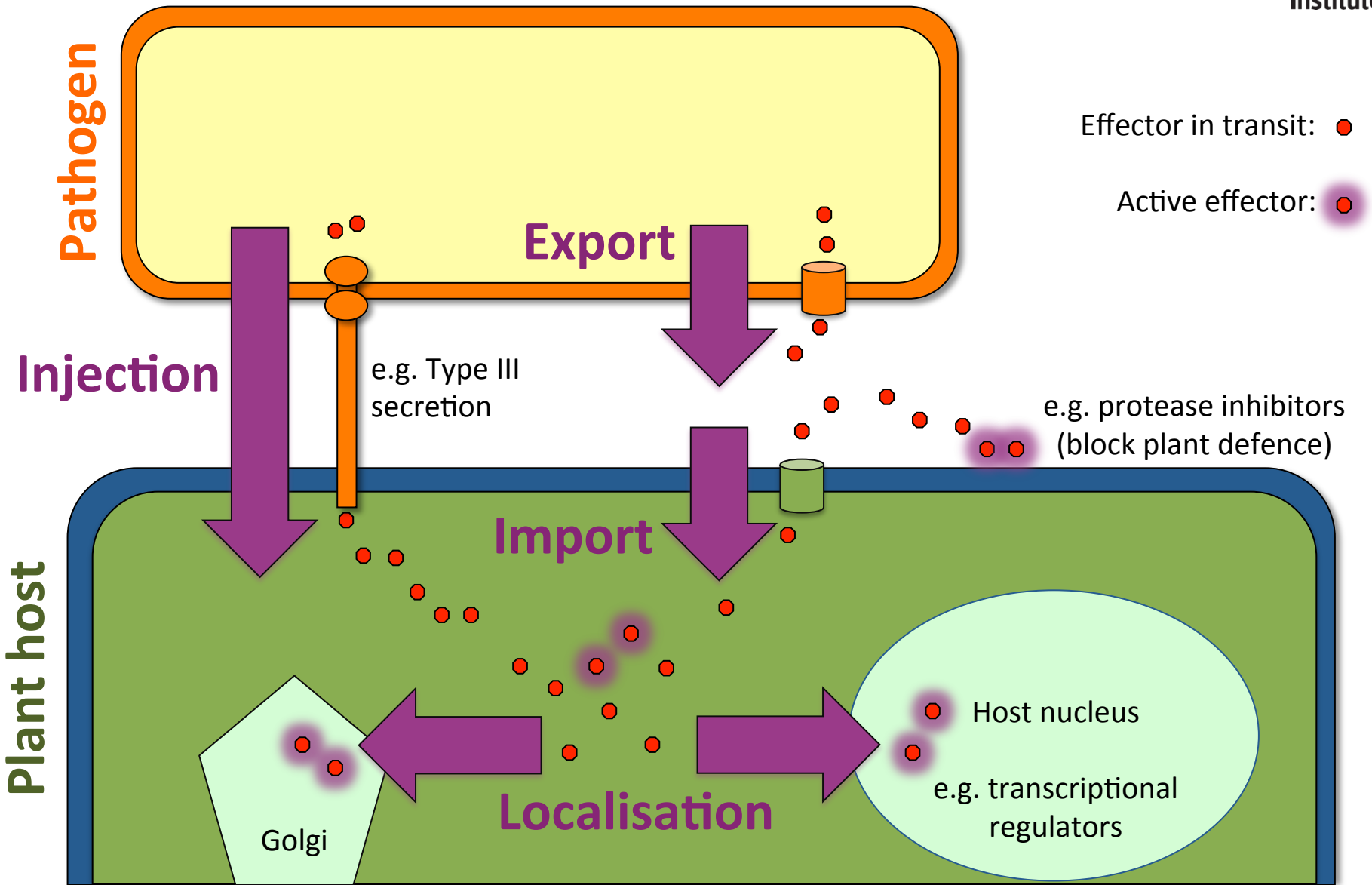
4kb plasmids, 10kb plasmids, 40kb fosmids, 80kb BACs

- Annotation mostly automated; some protein families received manual review by the community

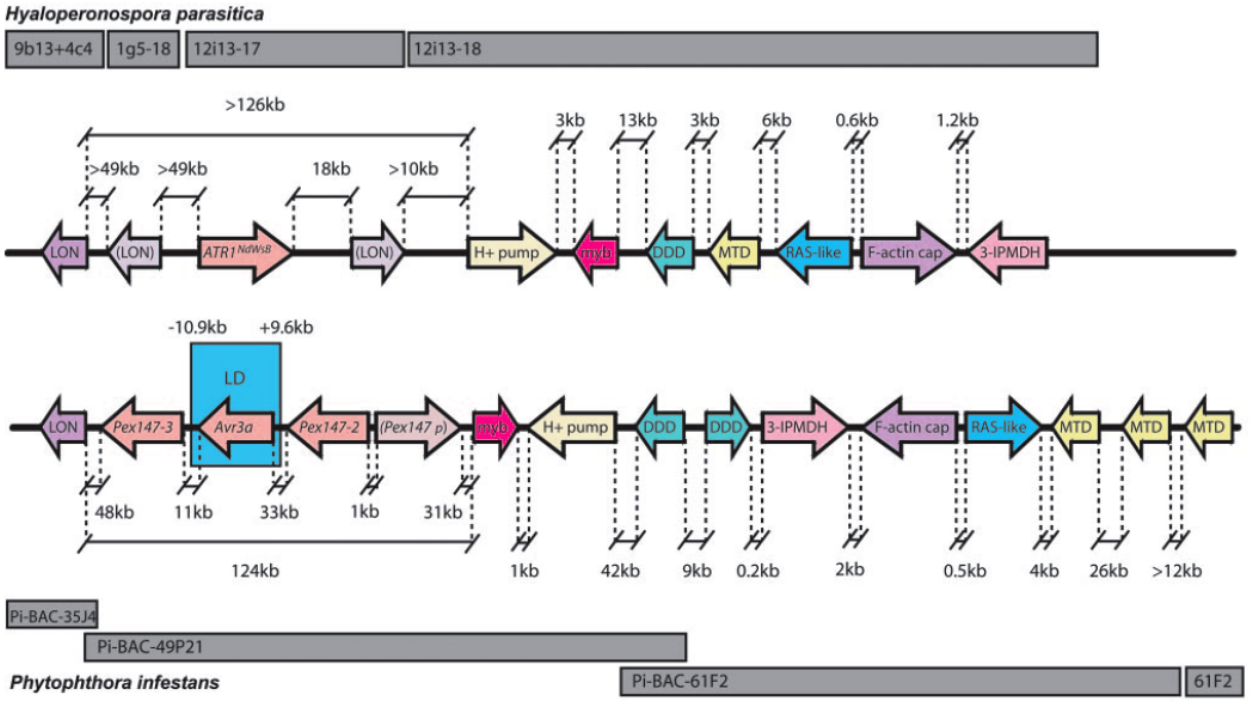
CRN, NPP1-like, elicitin-like, small secreted cysteine-rich, RxLR and transporter

- Published 2009, Nature (96 authors)
  Haas *et al.* (2009) *Nature* **461**: 393-398. doi:10.1038/nature08358.
- Public data: 18,288 contigs (4,921 supercontigs); 250Mbp (possibly triploid)
  http://www.broadinstitute.org/annotation/genome/phytophthora_infestans

My Interest: Effectors



P. infestans Avr3a locus



Contigged, annotated *P. infestans* and *H. peronospora* BAC fragments

Identified microsynteny surrounding the Avr3a and ATR1 avirulence loci

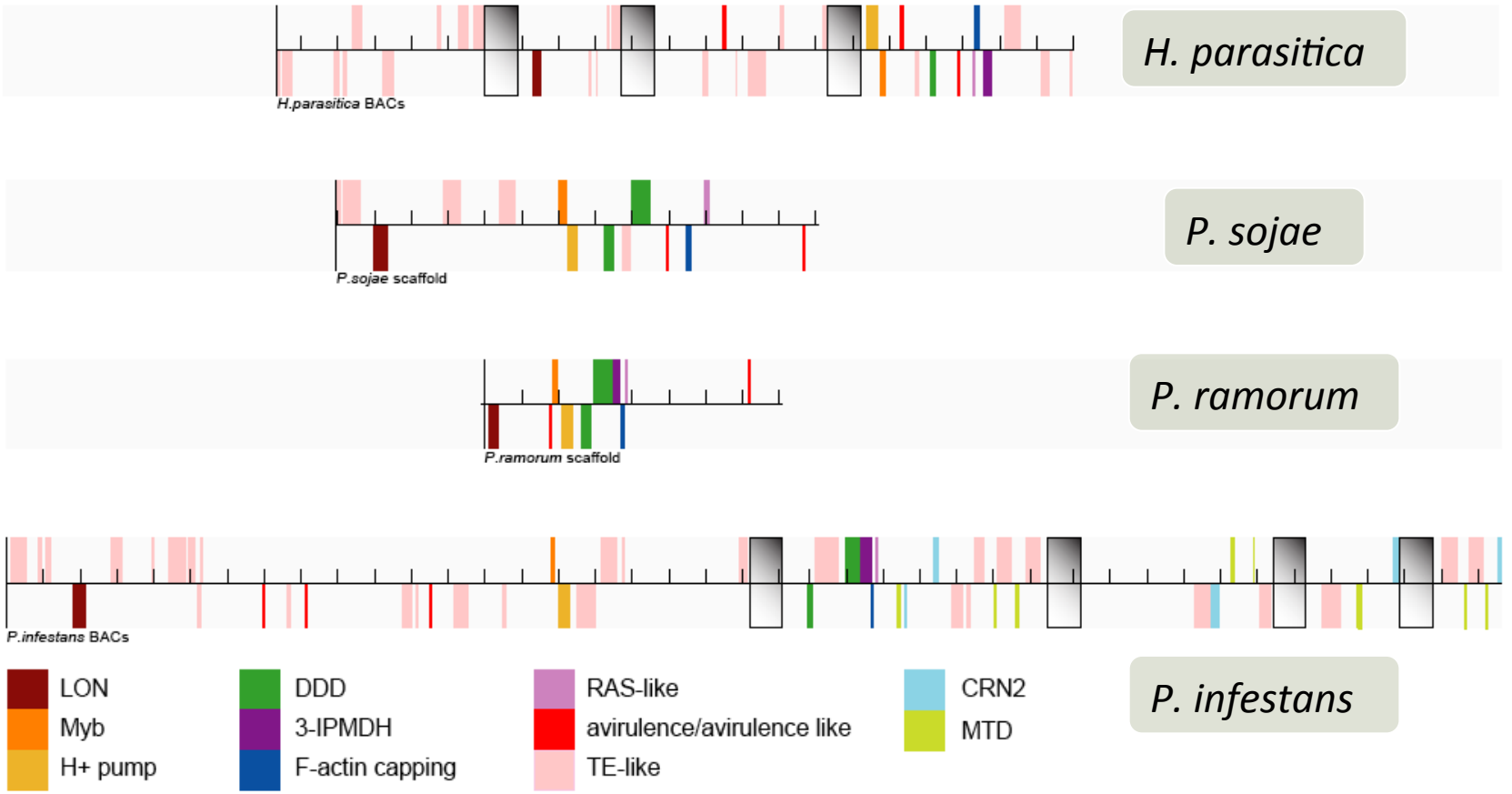
Avr3a, ATR1 belong to RXLR family

- RXLR N-terminal motif observed in a number of *P. infestans* effectors
- Also observed in *Plasmodium falciparum*, and thought associated with translocation
- **Necessary for translocation** in *P. infestans*

Armstrong *et al.* (2005) *Proc. Natl. Acad. Sci. USA* doi:10.1073/pnas.0500113102
Whisson *et al.* (2007) *Nature* doi:10.1038/nature06203

Genome expansion in oomycetes

- Early evidence from BAC assemblies of syntenous loci



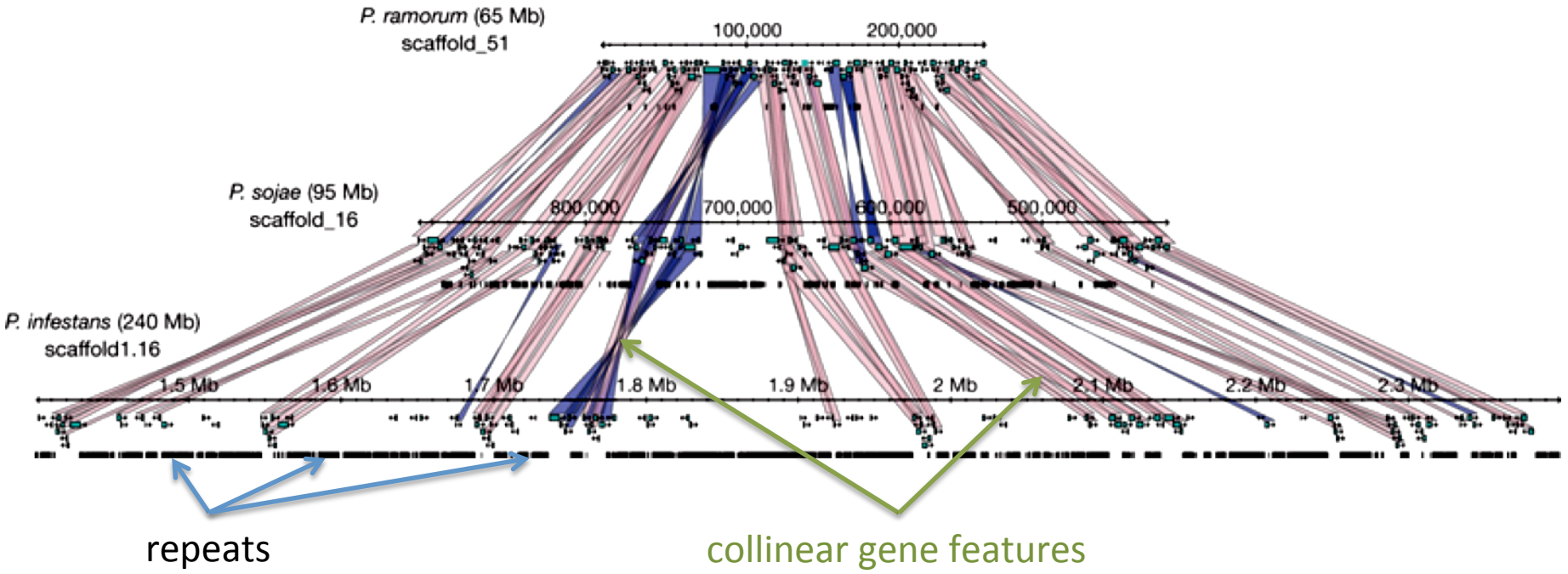
Armstrong *et al.* (2005) *Proc. Natl. Acad. Sci. USA* doi:10.1073/pnas.0500113102

Whisson *et al.* (2007) *Nature* doi:10.1038/nature06203

Repeat-driven Genome Expansion

The James Hutton Institute

- *P. infestans* genome >70% repeat regions (mostly TEs)
  - "Our findings suggest a two-speed genome…" https://www.broadinstitute.org/news/1328
  - High repeat/TE density results in 'experimentation'

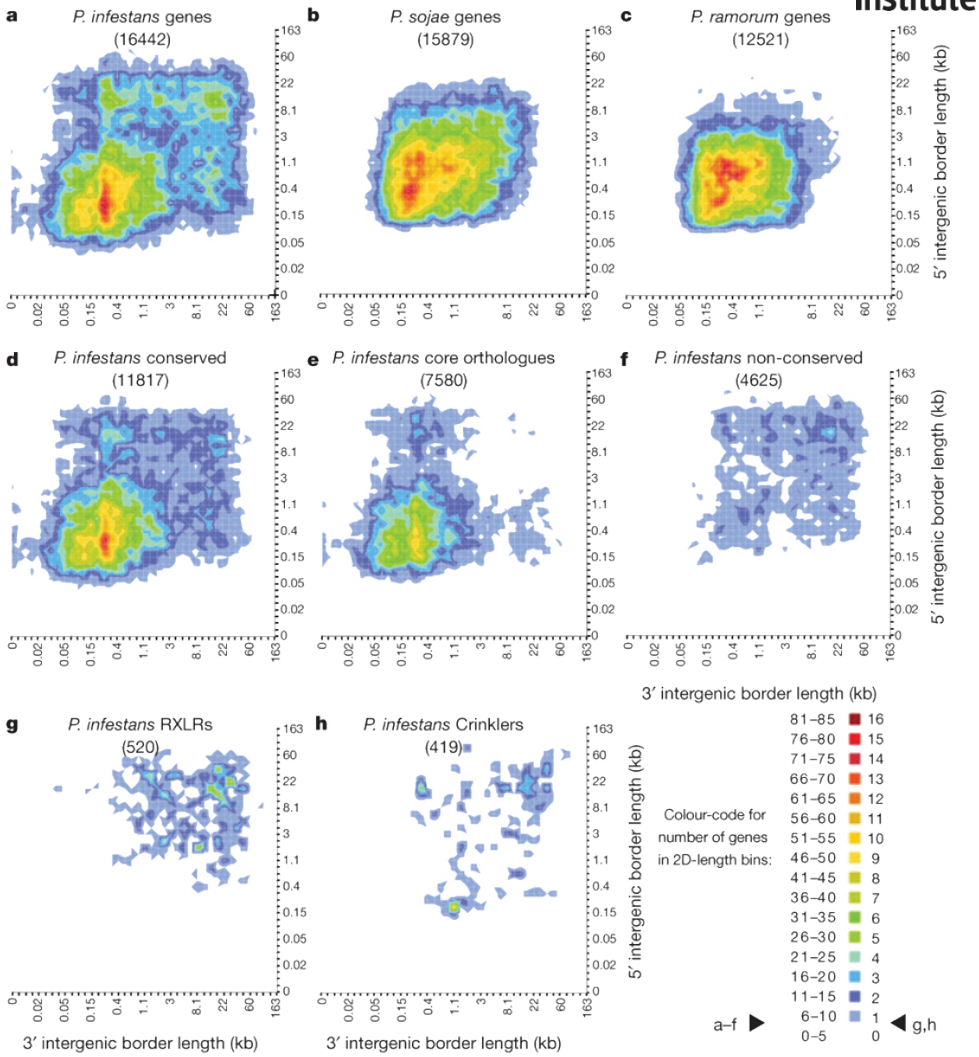


Haas *et al.* (2009) *Nature* doi:10.1038/nature08358

P. infestans two-speed genome

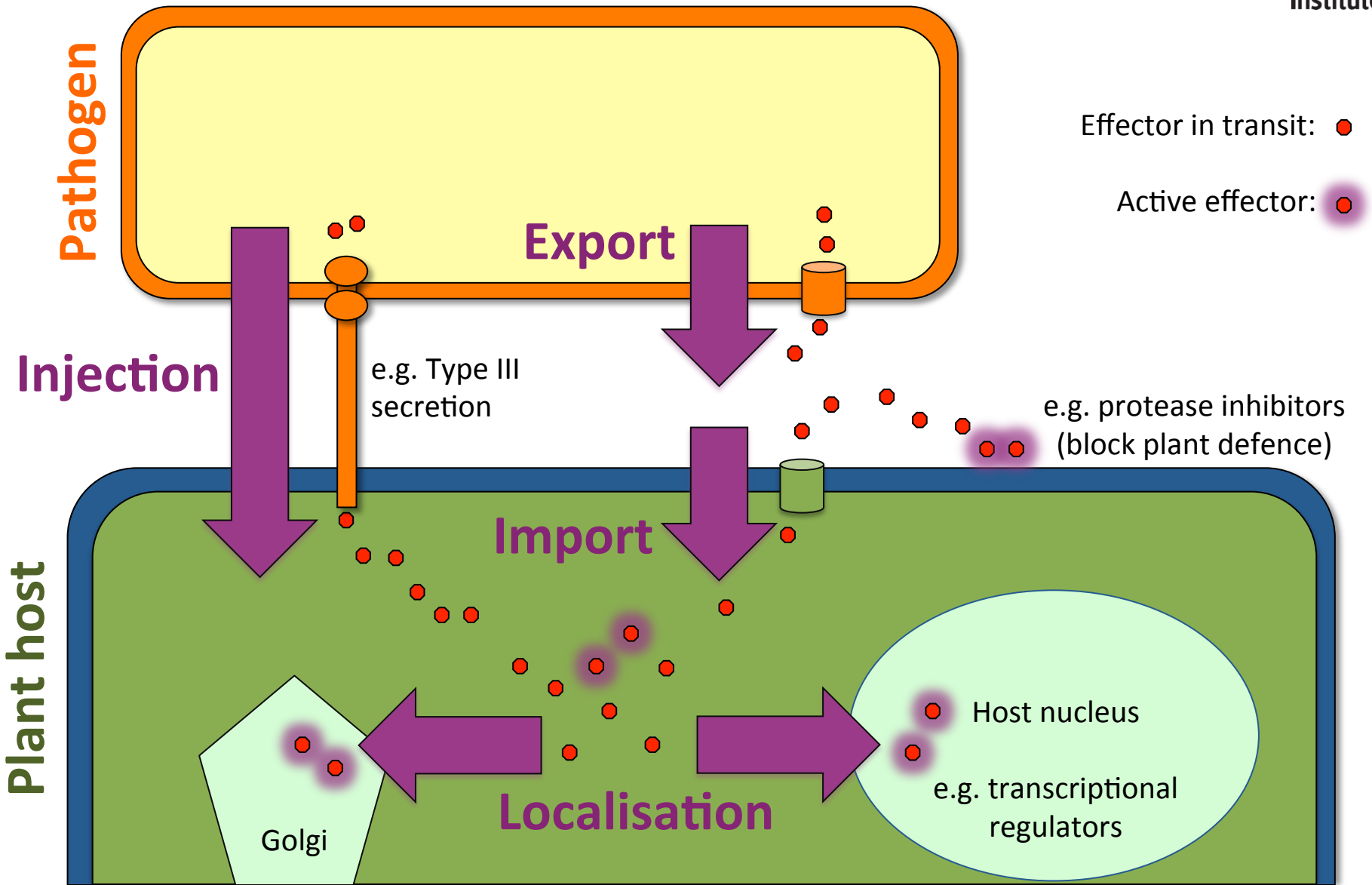
- Diverse population of transposons drives genome expansion specifically in *P. infestans*
- Pathogen effectors disproportionately located in expansion regions
- Expansion drives effector diversity, drives adaptation to overcome novel resistance?

(see `pi_two_speed.ipynb`)



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

Effector Properties: action and localisation



What is an effector?

The James Hutton Institute

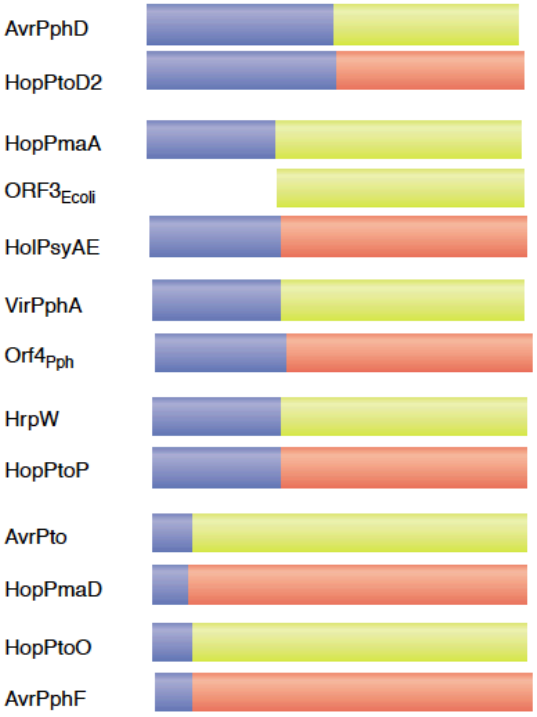
- Molecule produced by pathogen that (directly?) modifies host molecular/biochemical 'behaviour', e.g.
  - Inhibits enzyme action (*Cladosporium fulvum* AVR2, AVR4; *Phytophthora infestans* EPIC1, EPIC2B; *P. sojae* glucanase inhibitors)
  - Cleaves protein target (*Pseudomonas syringae* AvrRpt2)
  - (De-)phosphorylates protein target (*Pseudomonas syringae* AvrRPM1, AvrB)
  - Additional component in/retargeting host system, e.g. E3 ligase activity (*P. syringae* AvrPtoB; *P. infestans* Avr3a)
  - Regulatory control (*Xanthomonas campestris* AvrBs3, TAL effectors)

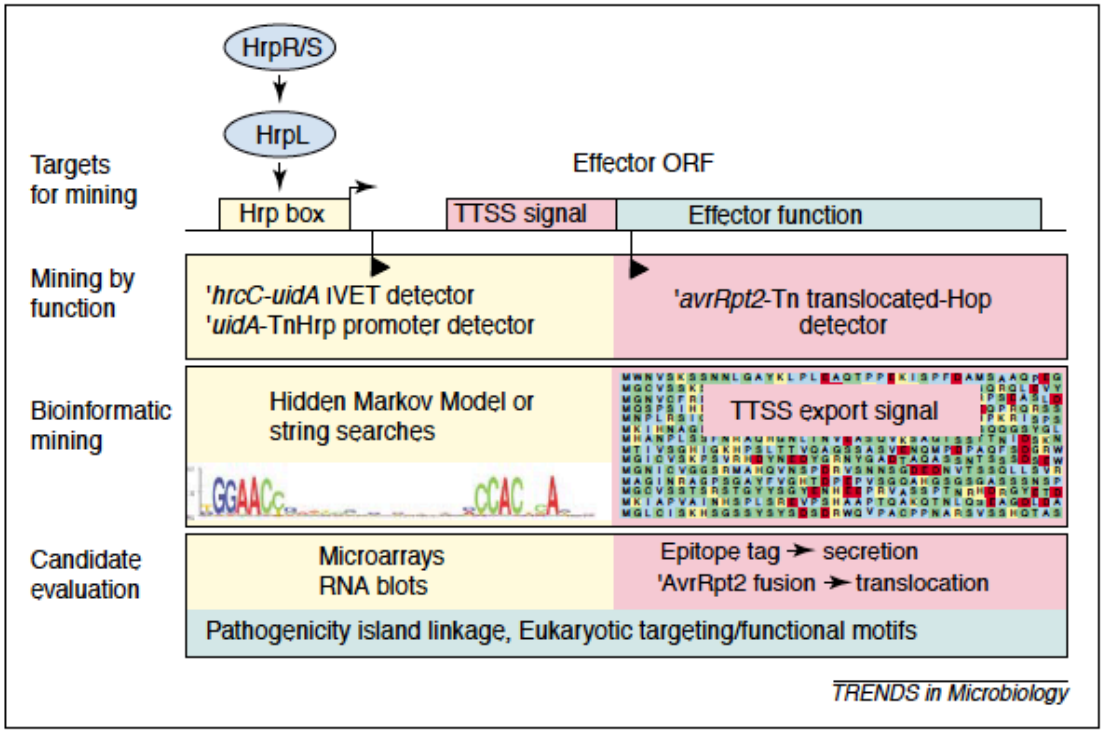
What is an effector?

- No unifying biochemical mechanism; may act inside or outwith host cell, many functions and final locations
- **No single test** for 'candidate effectors'
  - Tests are for protein family membership and/or evidence of 'effector-like behaviour'
  - A general sequence classification problem (functional annotation)
  - Many possible bioinformatic/computational approaches
  - No big red button

Characteristics of known effectors

- Modularity
  - **Delivery**: localisation/translocation domain(s)
  - **Activity**: functional/interaction domain(s)

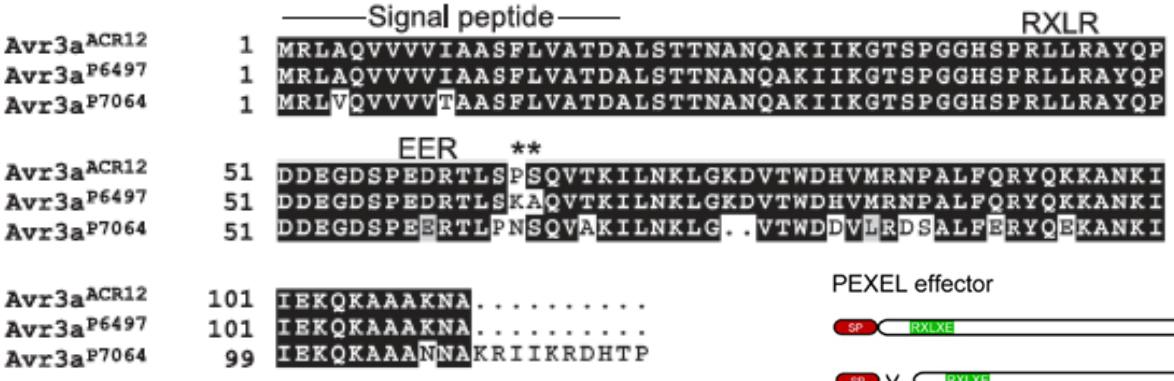




Greenberg & Vinatzer (2003) *Curr Opin Microbiol* doi:10.1016/S1369-5274(02)00004-8
Collmer *et al.* (2002) *Trends in Microbiology* doi:10.1016/S0966-842X(02)02451-4

Characteristics of known effectors

- Modularity
  - **Delivery**: localisation/translocation domain(s)
  - **Activity**: functional/interaction domain(s)

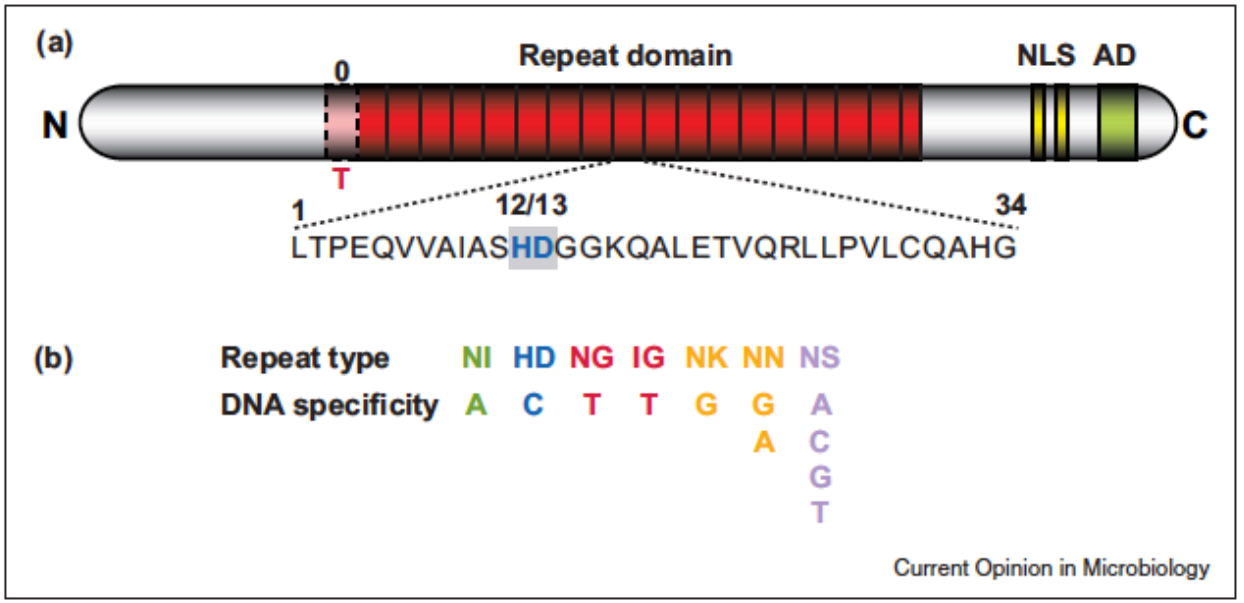


Dong *et al.* (2011) *PLoS One* doi:10.1371/journal.pone.0020172.

Bouwmeester *et al.* (2011) *Front. Plant-Microbe Int.* doi:10.3389/fpls.2011.00075

Characteristics of known effectors

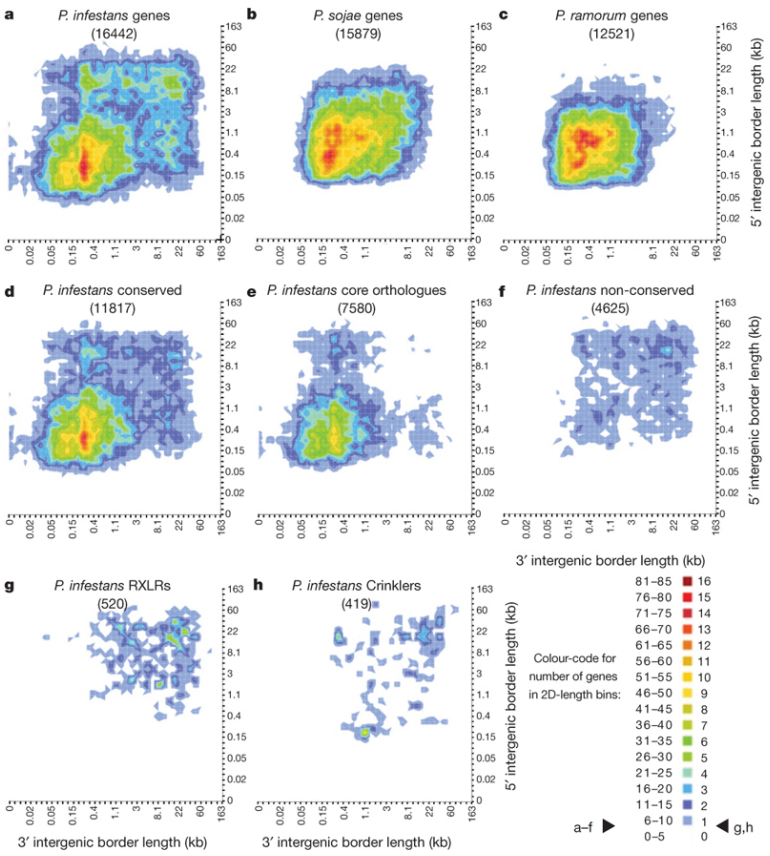
- Sequence motifs
  - Localisation/translocation domain(s) typically common to effector class (e.g. RxLR, T3E, CHxC)
  - Functional domain(s) may be common to effector class (e.g. TAL), or divergent (e.g. RxLR, T3E in general)



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

Characteristics of known effectors

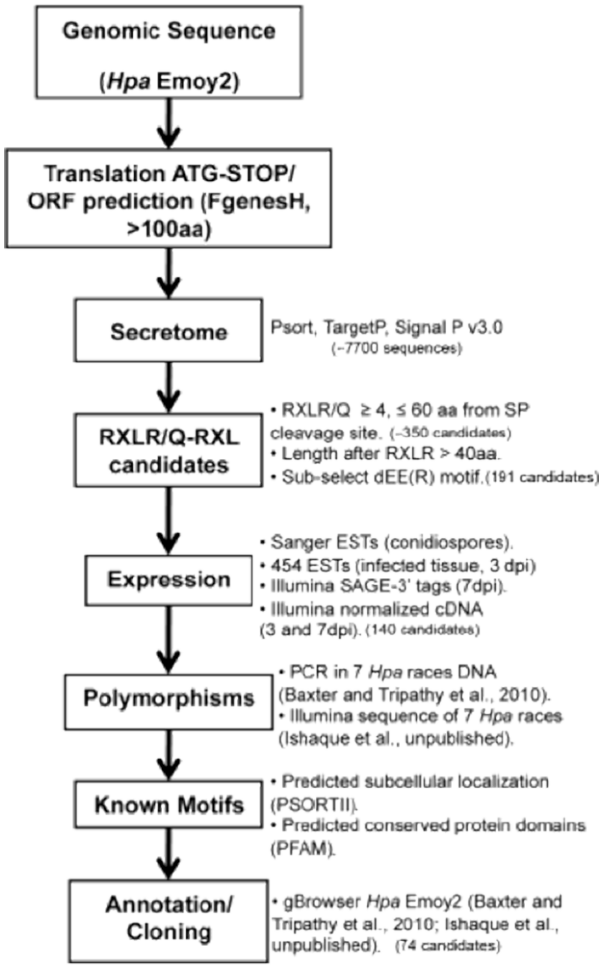
- "Arms Races" occur:
  - Host defences track effector evolution
  - Effectors evade host defences
- Divergence of effectors under selection pressure
  - Diversifying selection; divergence may result from evasion of detection, rather than change of biochemical 'function'
- Effectors may be found preferentially in characteristic locations
  - *P. infestans* 'gene sparse' regions



Raffaele *et al.* (2010) *BMC Genomics* doi:10.1186/1471-2164-11-637.

Characteristics of known effectors

- Application of 'filters': reduce the number of sequences to check (similar to stratification)
  - Presence/absence filters:
    - SignalP (export signal)
    - RxLR/T3SS (translocation signal)
    - Expression (used by pathogen)
    - Positive selection (suggests arms race)
    - Location in repeat-rich regions
    - etc…
- Workflows (e.g. Galaxy, Taverna) can be useful here

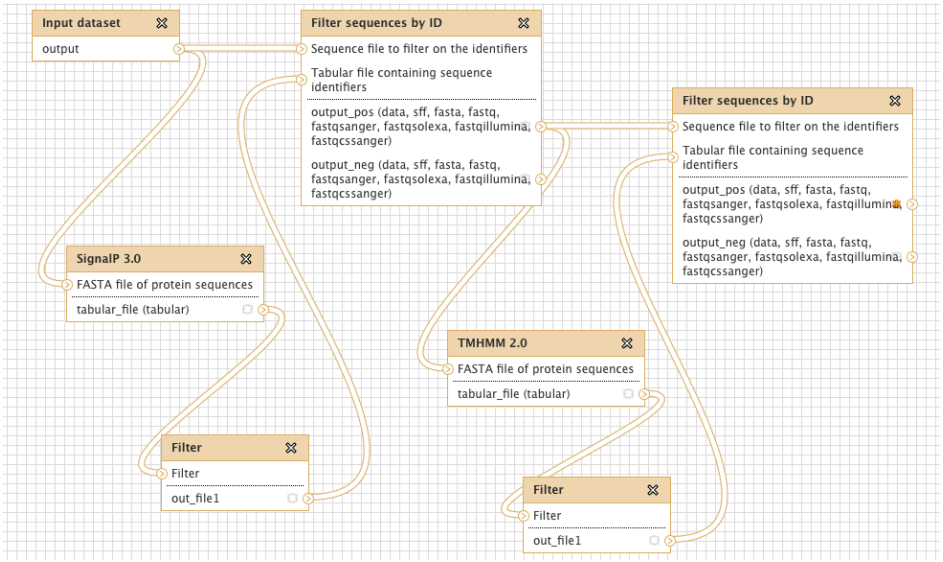


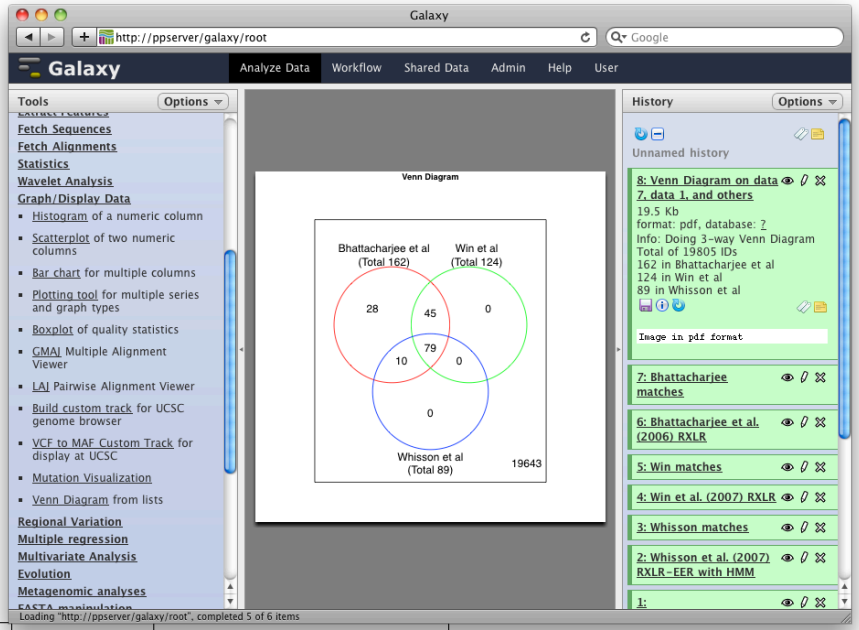
Fabro *et al*. (2011) *PLoS Pathog*. doi:10.1371/journal.ppat.1002348.
Cock *et al.* (2013) *PeerJ* doi:10.7717/peerj.167/.
Cock & Pritchard (2014) *Methods Mol. Biol.* doi:10.1007/978-1-62703-986-4_1
Pritchard & Broadhurst (2014) *Methods Mol. Biol.* doi:10.1007/978-1-62703-986-4_4

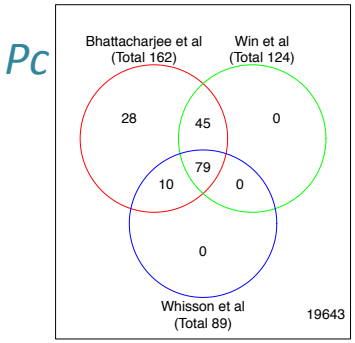
Galaxy Workflow – Effector finding

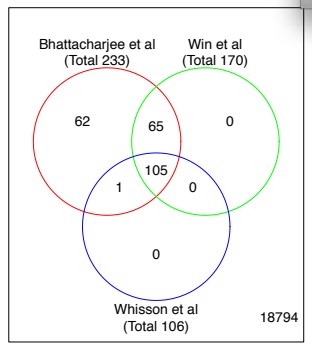
- Public Galaxy workflow for RxLR-finding

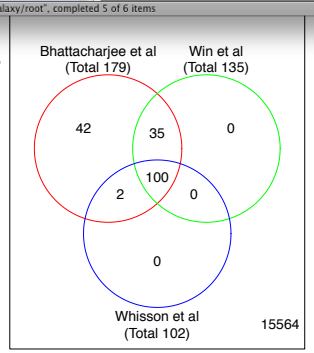
http://toolshed.g2.bx.psu.edu/repository?repository_id=0984925f74bd671c







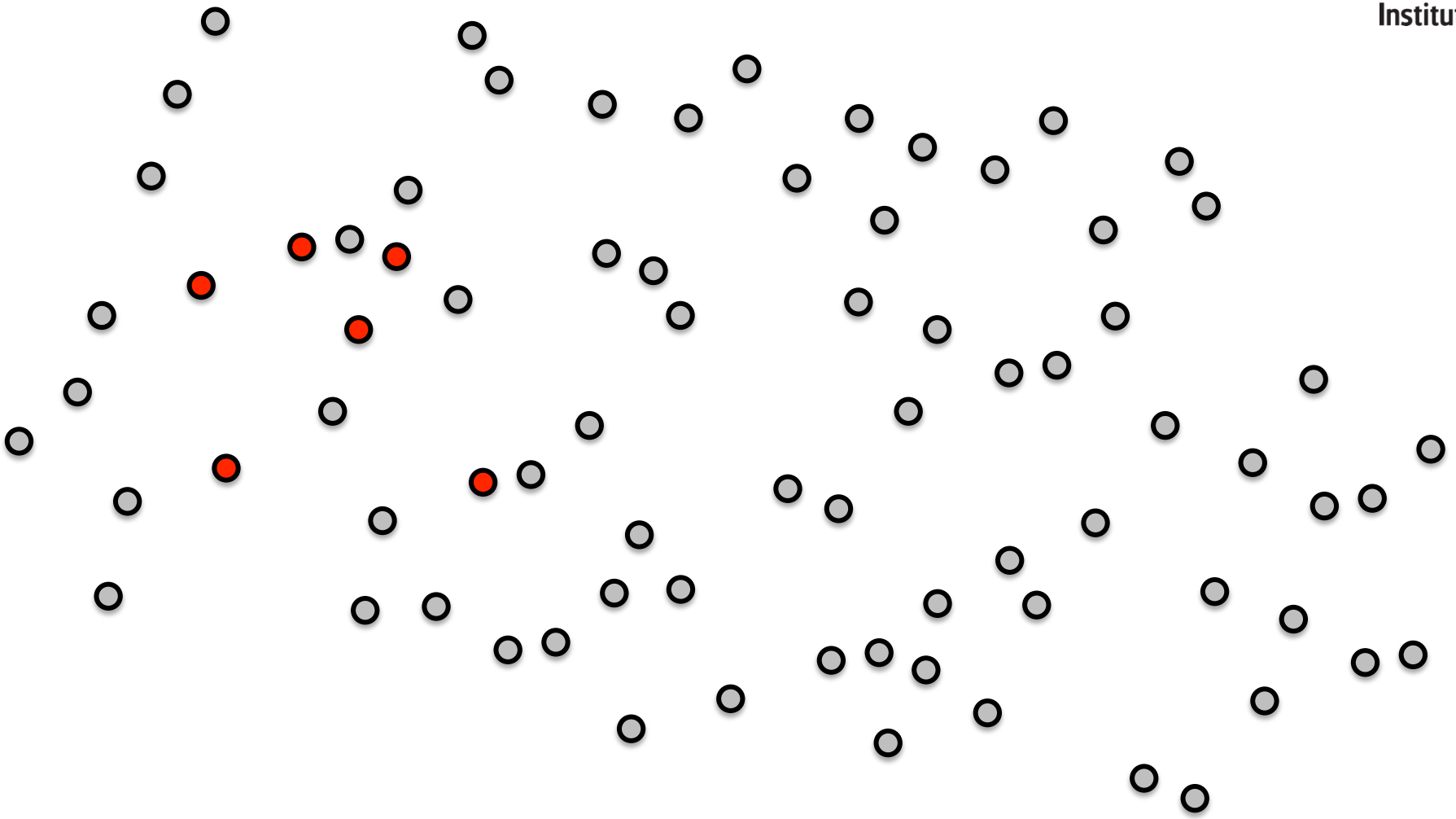




Cock *et al.* (2013) *PeerJ* doi:10.7717/peerj.167/.

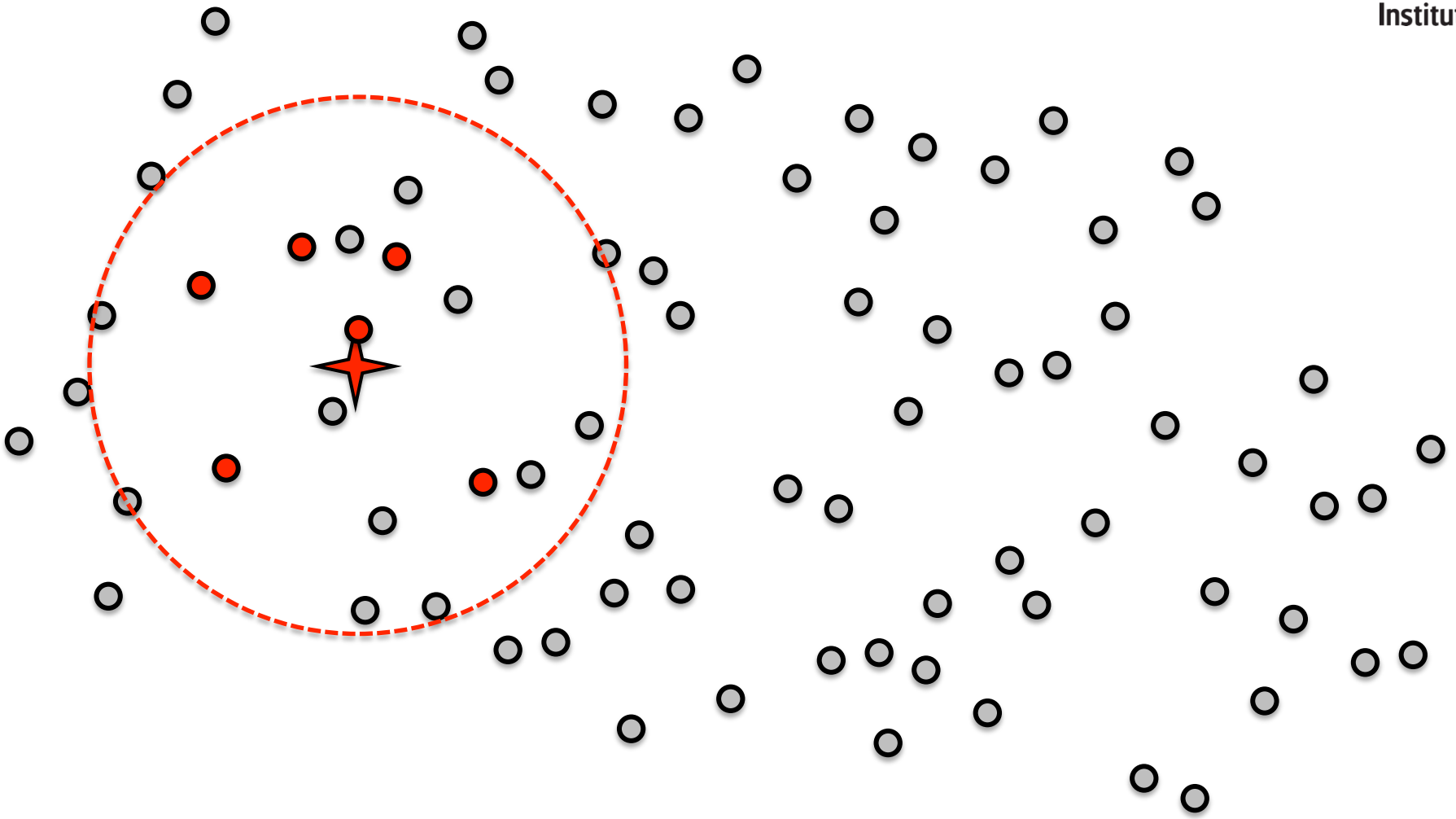
Cock & Pritchard (2014) *Methods Mol. Biol.* doi:10.1007/978-1-62703-986-4_1

Sequence space



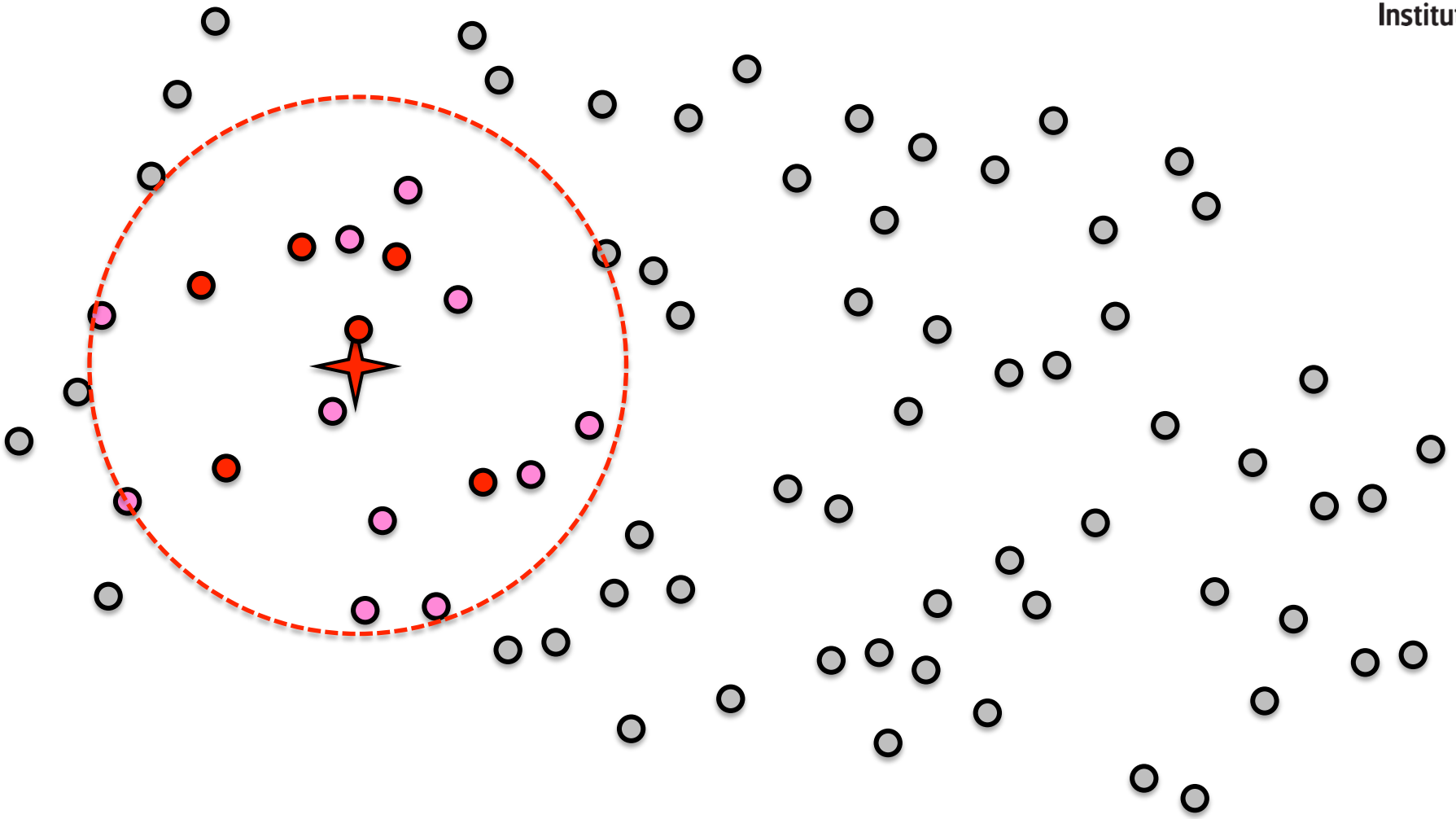
Known exemplars: red

Sequence space



Define a centre, and a distance that includes the examples

Sequence space



Classify ‘similar’ sequences

Finding effectors

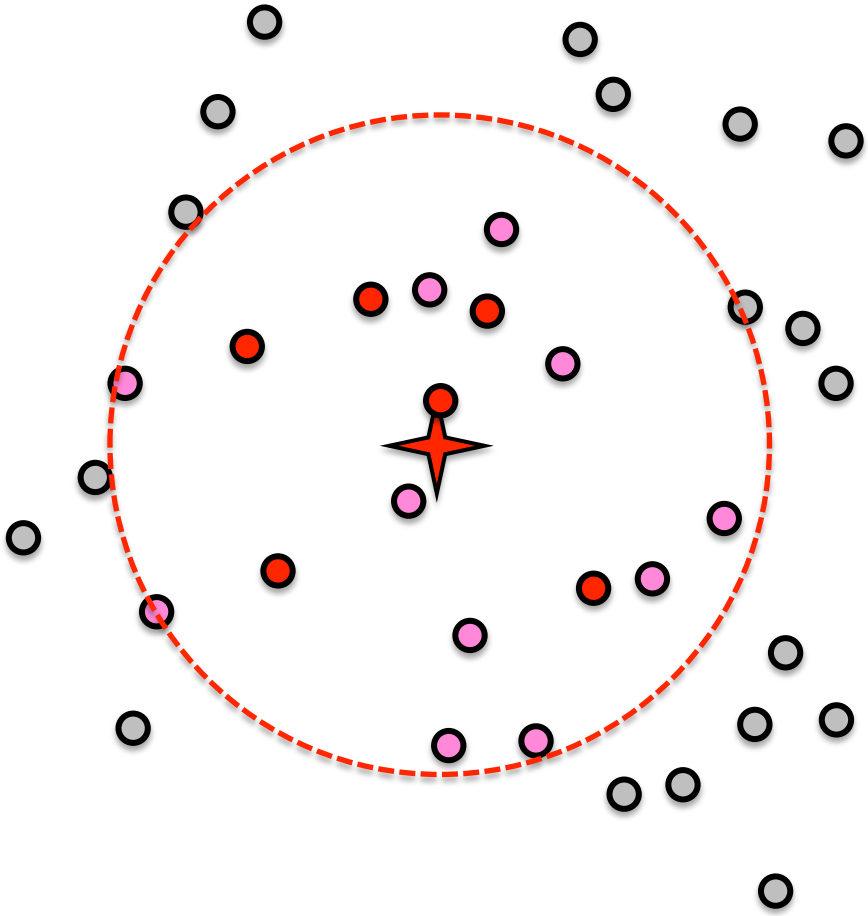
The James Hutton Institute

- Finding effectors is simple(-ish):

1. Have one or more examples of your effector (class)
2. Define some kind of appropriate threshold of similarity
3. Check all the gene/gene product sequences in the genome against that threshold

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

It's not *that* simple



- How do we define 'distance'?
- How large a 'distance' do we take?
- How do we know we've chosen a sensible 'distance'?

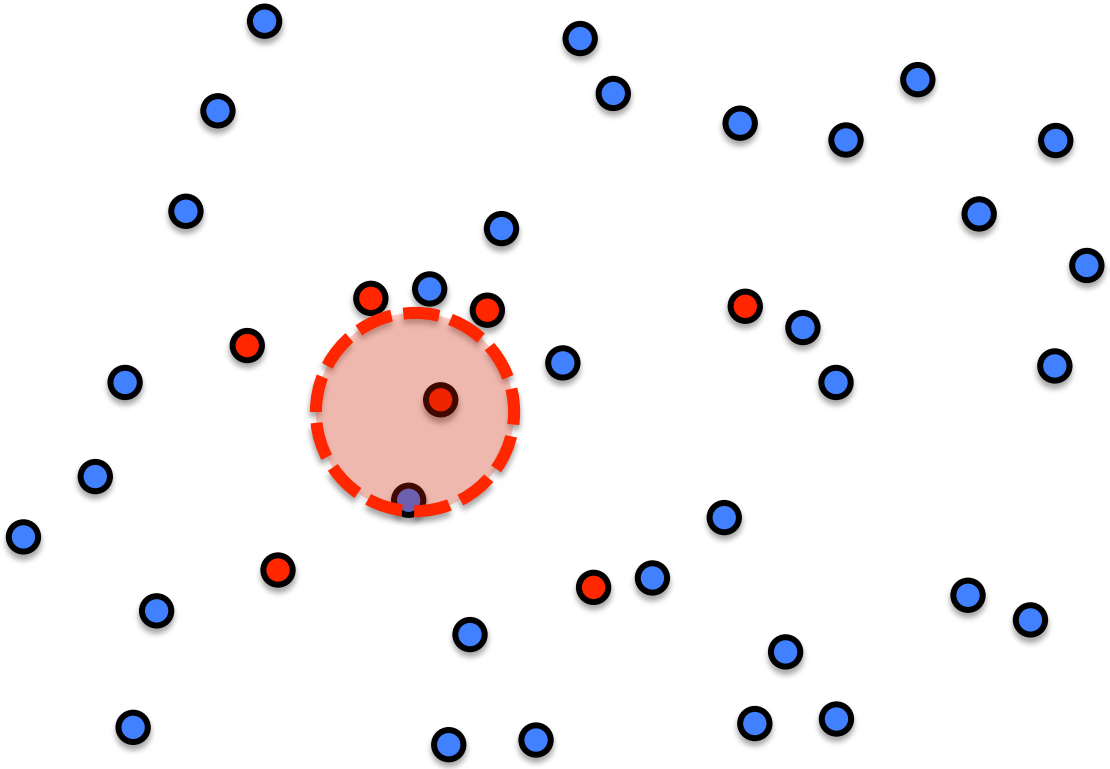
Pritchard & Broadhurst (2014) *Methods Mol. Biol.* doi:10.1007/978-1-62703-986-4_4

How do we define 'distance'

- Sequence identity
- Derived score (based on sequence identity/ alignment)
  - Bit score in BLAST
  - E-value in BLAST
- Derived score (based on other measures)
  - Bit score in HMMer
- Clustering (not strictly a distance)
  - Sequence identity (e.g. CD-HIT)
  - MCL

(we're really assessing criteria for class membership)

How large a distance do we take?

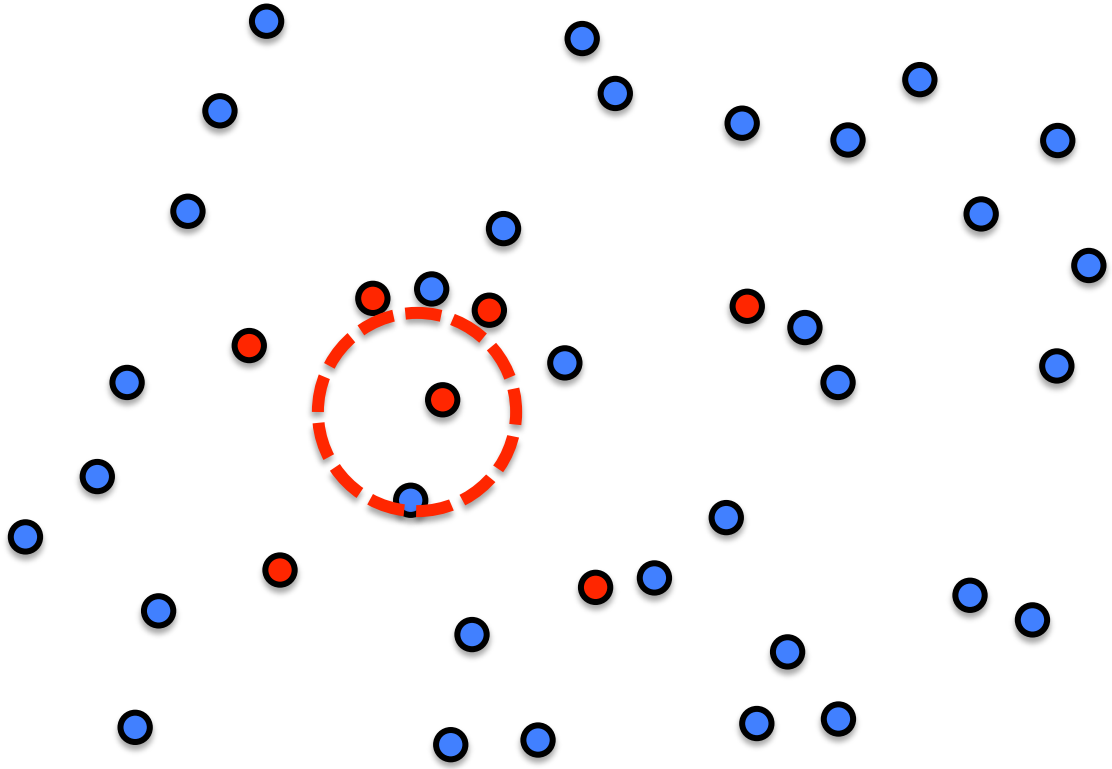


Confusion matrix:

	IN	OUT
Red	1	5
Blue	1	36

- Our distance/boundary classifies sequences as ‘in’ or ‘out’
  - ‘red’ or ‘blue’
- Changing distance/bound results in various degrees of success…

How large a distance do we take?



True positive

False negative

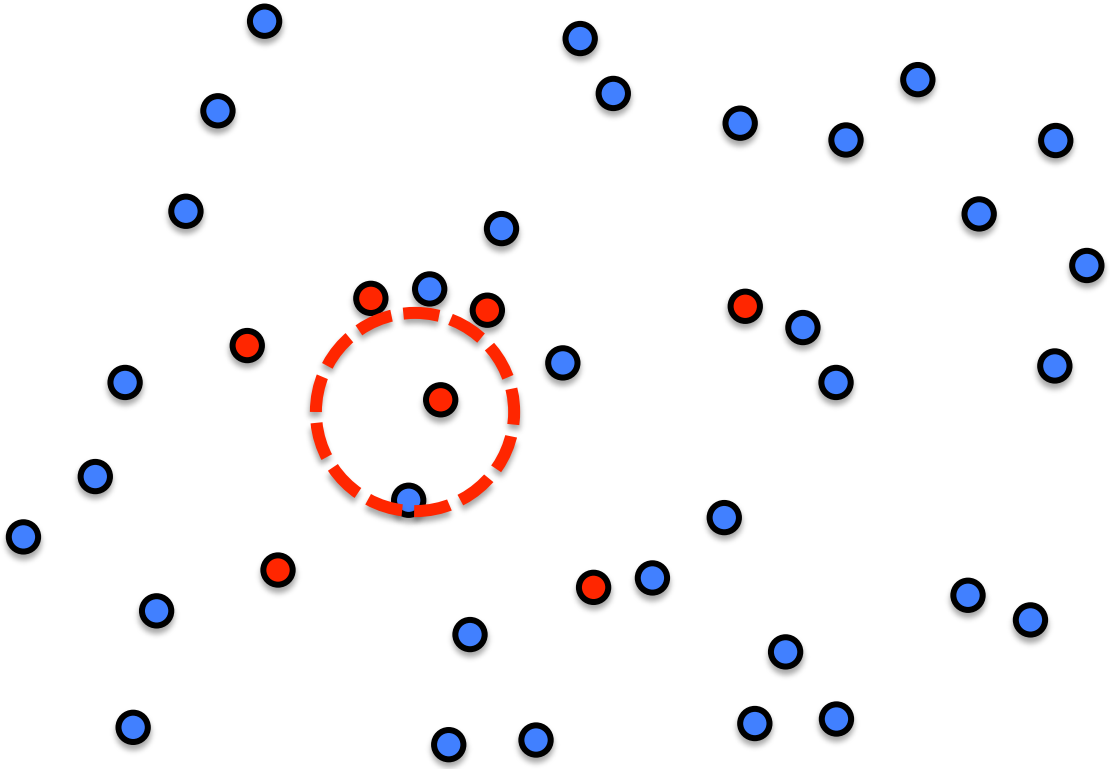
	IN	OUT
Red	1	5
Blue	1	36

False positive

True negative

- Our distance/boundary classifies sequences as ‘in’ or ‘out’
  - ‘red’ or ‘blue’
- Changing distance/bound results in various degrees of success…

How large a distance do we take?

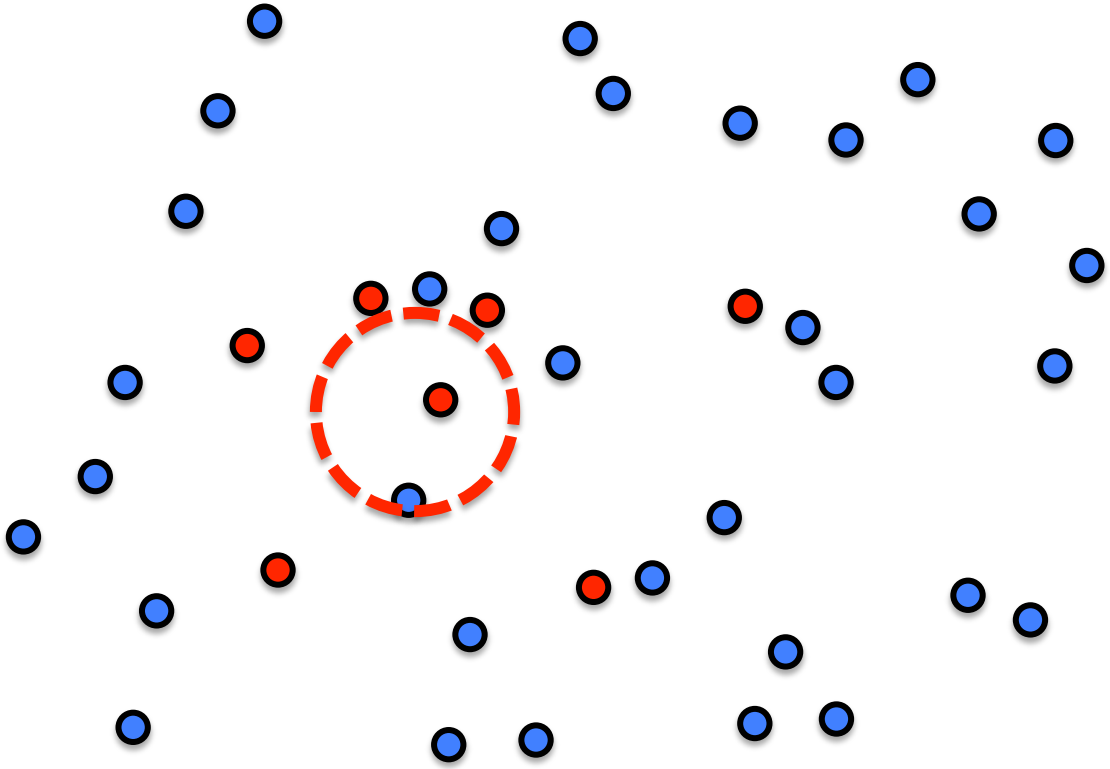


	IN	OUT
Red	1	5
Blue	1	36

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)

- Our distance/boundary classifies sequences as ‘in’ or ‘out’
  - ‘red’ or ‘blue’
- Changing distance/bound results in various degrees of success…

How large a distance do we take?

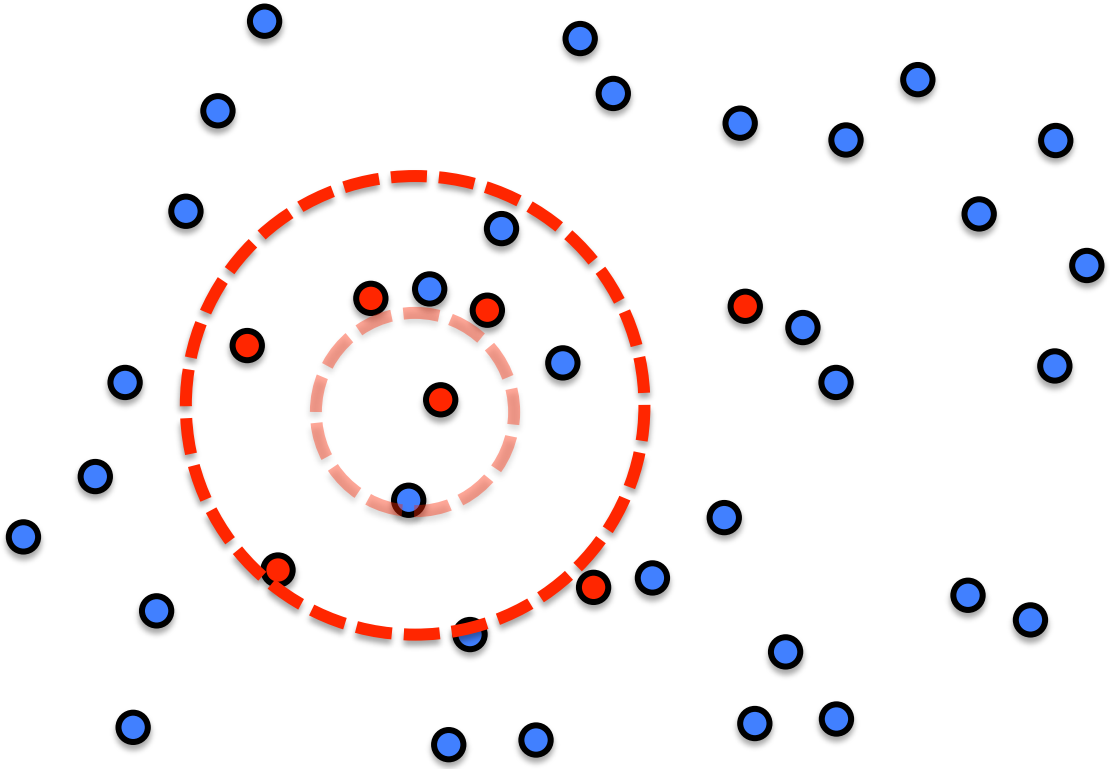


	IN	OUT
Red	1	5
Blue	1	36

False positive rate	1/37 = 0.03
False negative rate	5/6 = 0.83
Sensitivity	1/6 = 0.17
Specificity	36/37 = 0.97

- Our distance/boundary classifies sequences as ‘in’ or ‘out’
  - ‘red’ or ‘blue’
- Changing distance/bound results in various degrees of success…

How large a distance do we take?

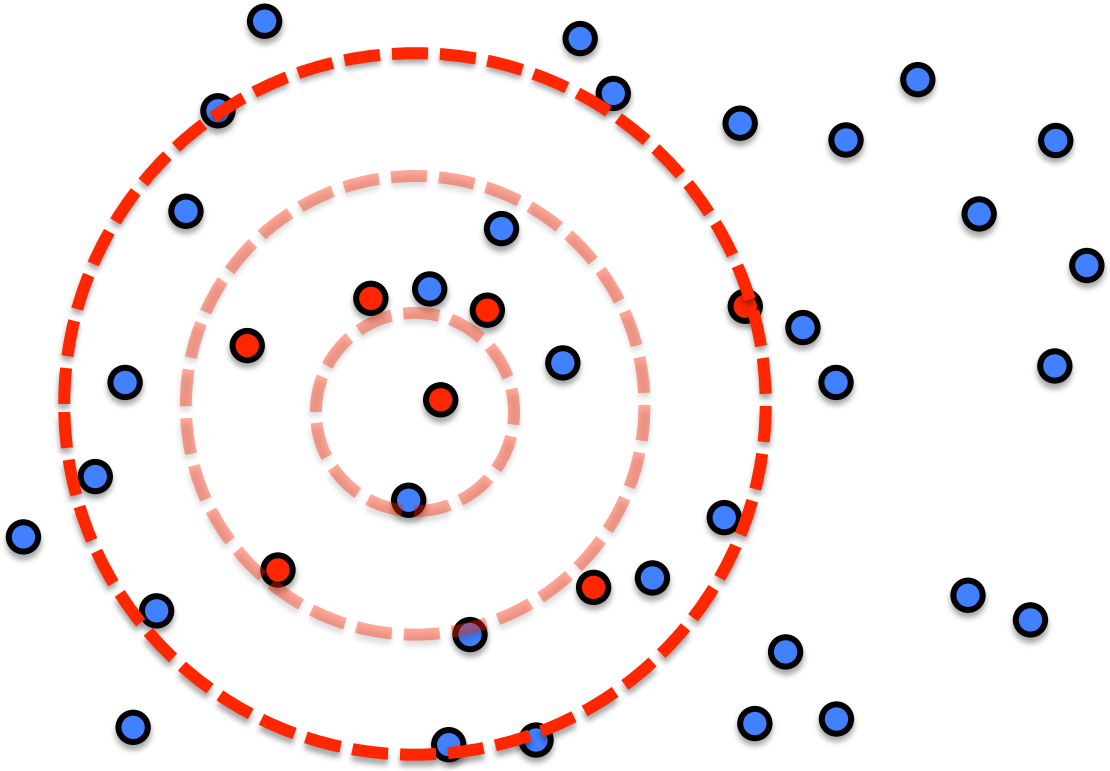


	IN	OUT
Red	5	2
Blue	4	33

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

- Our distance/boundary classifies sequences as ‘in’ or ‘out’
  - ‘red’ or ‘blue’
- Changing distance/bound results in various degrees of success…

How large a distance do we take?

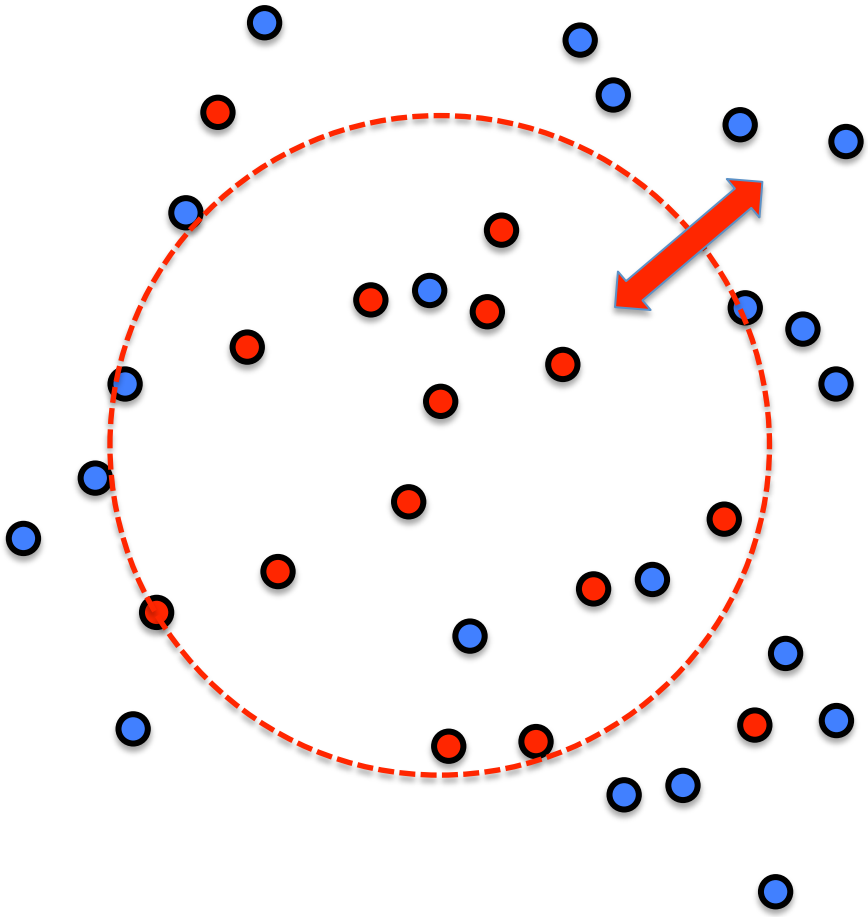


	IN	OUT
Red	7	0
Blue	14	23

False positive rate	0.38
False negative rate	0
Sensitivity	1
Specificity	0.62

- Our distance/boundary classifies sequences as ‘in’ or ‘out’
  - ‘red’ or ‘blue’
- Changing distance/bound results in various degrees of success…

How large a distance do we allow?



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

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

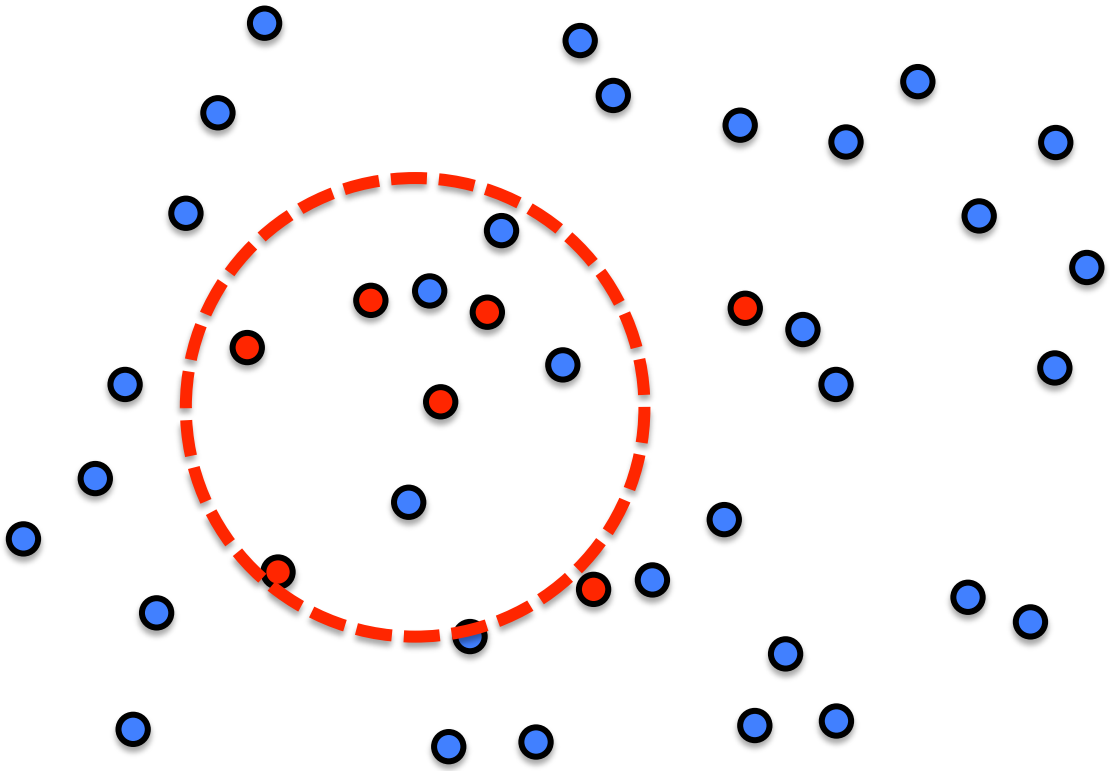
Cross-validation

The James Hutton Institute

- Estimation of classifier performance depends on
  - distance measure
  - composition of training set ('positives' and 'negatives')
- **Cross-validation gives objective estimate of performance**
- Many strategies available, including:
  - leave-one-out (LOO)
  - *k*-fold crossvalidation
  - repeated (random) subsampling
- Essentially: always keep a **hold-out set** (not used to train)

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

After Cross-validation



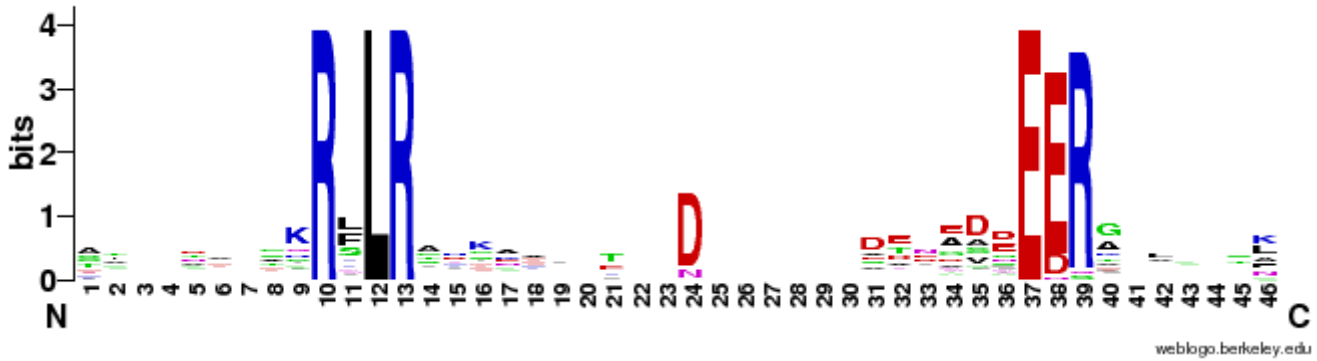
False positive rate	0.11
False negative rate	0.29
Sensitivity	0.81
Specificity	0.89
Precision	0.56

- Used crossvalidation to find ‘best’ method & parameters
- This gives estimated performance metrics on unseen data
- Apply ‘best’ method to complete dataset for prediction

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

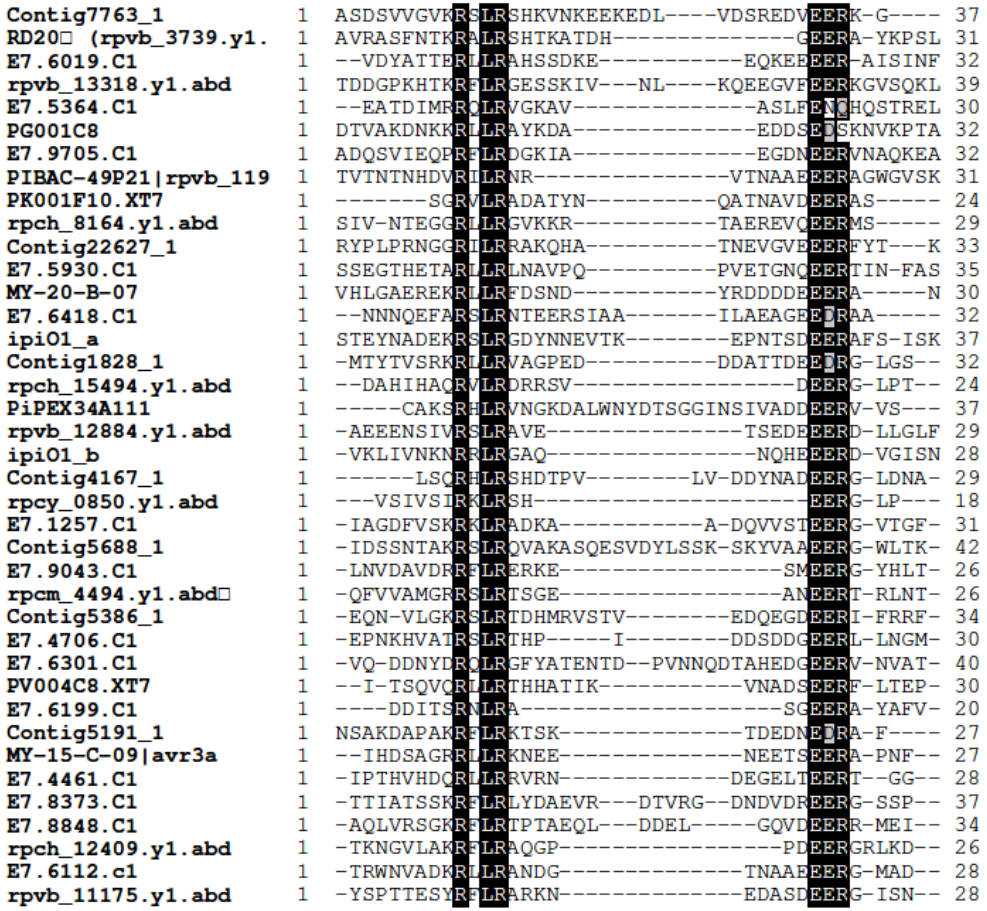
Building an RxLR training set

- Starting point: 49 candidate sequences (reference set)
- Known:
  - Contain (putatively) RxLR-EER motif
  - All but one transcribed (i.e. not bad gene calls)
- Assumed:
  - Presence of signal peptide and RxLR-EER categorises effectors



Building a classifier

- We have a recognisable motif, with substantial local variation and indels
  - Therefore chose profile HMM
  - Use HMMer software
- Profile HMMs sensitive to quality of alignment
- Therefore treat alignment as a parameter of the HMM (much difference between alignments – several tested!)



Whisson *et al.* (2007) *Nature* doi:10.1038/nature06203

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

Building a classifier

The James Hutton Institute

- Parameters modified for HMM
  - Alignment package (no alignment, anchored, Clustal, DiAlign, T-Coffee) on default settings
  - Full-length and truncated (no signal peptide) alignments to test for influence of signal peptide region on classifier
    - **Plus one alignment of RxLR-EER plus flanking region only (‘cropped’)**
- HMM built for each of eleven alignments
  - Default HMMer parameters
- Once built, the HMM was the classifier.
  Stratification by presence of signal peptide (SignalP)

```
hmmbuild --amino <output> <alignment>
```

Whisson *et al.* (2007) *Nature* doi:10.1038/nature06203
Haas *et al.* (2009) *Nature* doi:10.1038/nature08358

Testing the classifiers

The James Hutton Institute

- Eleven classifiers to test
- **Step 1**: Consistency test
  - Does the classifier correctly call as positive the sequences used to train it?
  - Estimates recovery of the information in the training set
- **Step 2**: Recovery of full sequences
  - Estimates performance of classifier on complete sequence data

Alignment	Consistency	Reference	Minimum bit score
core_rxlr_aa	38	38	10.0
core_aa_anchored_rxlr_eer	36	36	18.6
core_default_clustal	33	38	12.9
core_default_dialign	2	38	6.8
core_default_t_coffee	14	38	5.8
truncated_core_aa	33	33	3.6
truncated_core_aa_anchored_rxlr_eer	34	34	13.8
truncated_default_clustal	31	35	4.4
truncated_default_dialign	5	34	1.9
truncated_default_t_coffee	17	37	2.0
cropped	36	39	5.6

SigP-RxLR-Cterm

RxLR-Cterm

RxLR

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

Testing the classifiers

- **Step 3:** Leave-One-Out Crossvalidation
  - But only have positive examples!
  - **Removes possibility that classifier matches on basis of having 'seen' a sequence before**

Better match to classifier than to control

Alignment	Bit score								
	-20	-15	-10	-5	0	5	10	15	20
core	22	19	17	15	11	7	6	3	2
anchored_rxlr_eer	23	21	17	14	11	6	5	3	2
default_clustal	31	27	25	19	14	12	7	2	1
default_dialign	37	36	27	24	19	10	2	2	1
default_t_coffee	36	35	34	31	28	26	20	12	10
truncated_core	18	11	4	2	2	2	2	2	2
truncated_anchored_rxlr_eer	22	20	16	13	8	4	3	1	1
truncated_default_clustal	27	25	18	12	8	4	1	0	0
truncated_default_dialign	39	39	39	27	10	5	0	0	0
truncated_default_t_coffee	36	35	33	31	24	16	7	2	0
cropped	39	39	39	39	39	39	37	35	25

SigP-RxLR-Cterm

RxLR-Cterm

RxLR

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

Testing the classifiers

The James Hutton Institute

- **Step 4:** Tests on negative samples
  - Completely shuffled sequences
  - Shuffled downstream of the signal peptide only
  - Replace RxLR-EER with AAAA-AAA

No classifier identifies a false positive

(no classifier matches on sequence composition alone)

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

Testing the classifiers

The James Hutton Institute

- **Step 4:** Tests on negative samples
  - Completely shuffled sequences
  - Shuffled downstream of the signal peptide only
  - Replace RxLR-EER with AAAA-AAA

(some recognition on basis of signal peptide)

Alignment	Recovered	Reference	Rate	
core_rxlr_aa	3	38	8%	SigP-RxLR-Cterm
core_aa_anchored_rxlr_eer	0	36	0%	
core_default_clustal	9	38	24%	
core_default_dialign	2	38	5%	
core_default_t_coffee	0	38	0%	
truncated_core_aa	0	33	0%	RxLR-Cterm
truncated_core_aa_anchored_rxlr_eer	0	34	0%	
truncated_default_clustal	0	35	0%	
truncated_default_dialign	0	34	0%	
truncated_default_t_coffee	0	37	0%	
cropped	0	39	0%	RxLR

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

Testing the classifiers

The James Hutton Institute

- **Step 4:** Tests on negative samples
  - Completely shuffled sequences
  - Shuffled downstream of the signal peptide only
  - Replace RxLR-EER with AAAA-AAA

(some recognition on sequence other than motif)

Alignment	Recovered	Reference	Rate	
core_rxlr_aa	39	38	102%	SigP-RxLR-Cterm
core_aa_anchored_rxlr_eer	27	36	75%	
core_default_clustal	35	38	92%	
core_default_dialign	38	38	100%	
core_default_t_coffee	32	38	84%	
truncated_core_aa	20	33	61%	RxLR-Cterm
truncated_core_aa_anchored_rxlr_eer	16	35	46%	
truncated_default_clustal	33	35	94%	
truncated_default_dialign	12	34	35%	
truncated_default_t_coffee	16	37	43%	
cropped	0	39	0%	RxLR

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

Choosing a classifier

The James Hutton Institute

- **The 'cropped' classifier has:**
  - 100% recovery of positive training sequences
  - 0% recovery of negative test sequences
- Some variation in classifier performance on whole genome:
  284 candidates from 'cropped' (≈80-200 common to other models)

		HMM										
		full-length					truncated					
Method	total	core	anchored	clustal	dialign	t-coffee	core	anchored	clustal	dialign	t-coffee	cropped
core_rxlr_aa	1999	X										
core_aa_anchored_rxlr_eer	1994	1890	X									
core_default_clustal	541	478	465	X								
core_default_dialign	196	89	90	107	X							
core_default_t_coffee	292	106	111	141	166	X						
truncated_core_aa	929	923	925	402	60	63	X					
truncated_core_aa_anchored_rxlr_eer	195	170	184	149	75	87	163	X				
truncated_default_clustal	81	65	70	64	68	77	62	69	X			
truncated_default_dialign	131	44	50	52	63	75	35	44	39	X		
truncated_default_t_coffee	143	70	87	79	103	118	60	77	69	53	X	
cropped	284	84	98	114	133	191	60	86	71	80	120	X

Whisson *et al.* (2007) *Nature* doi:10.1038/nature06203
Haas *et al.* (2009) *Nature* doi:10.1038/nature08358

A Trip To The Doctor

The James Hutton Institute

- Routine medical checkup
- Test for disease X (horrible, unpleasant, potentially suppurating)
- Test has ***sensitivity*** (i.e. predicts disease where there is disease) of **95%**
- Test has ***false positive rate*** (i.e. predicts disease where there is no disease) of **1%**
- Your test is **positive** (bad luck, there)
- What is the probability that you have disease X?

0.01 0.05 0.50 0.95 0.99

A trip to the doctor, part II

The James Hutton Institute

- Test for disease X (horrible, unpleasant, potentially suppurating)
- Test has ***sensitivity*** (i.e. predicts disease where there is disease) of **95%**
- Test has ***false positive rate*** (i.e. predicts disease where there is no disease) of **1%**
- Your test is positive
- To calculate the probability that the test correctly determines whether you have the disease, you need to know the **baseline occurrence**.

Baseline occurrence: 1% ⇒ P(disease|+ve)=0.490

Baseline occurrence: 80% ⇒ P(disease|+ve)=0.997

What is the baseline occurrence for effectors?

The James Hutton Institute

- **Usually rely on predictions for expected baseline**
- **Bacterial genomes:** ≈4500 genes
  - Type III effectors: 1-10% (Arnold *et al.* 2009); 1-2% (Collmer *et al.* 2002); 1% (Boch and Bonas, 2010)
- **Oomycete/fungal genomes**: ≈20000 genes
  - RxLRs: 120-460 (1-2%; Whisson *et al.* 2007); ≤563 (≲2% Haas *et al.* 2009);
  - CRNs: 19-196 (≲1%; Haas *et al.* 2009)
  - CHxC: ≈30 (<1%; Kemen *et al*. 2011)
- **We need to take care over result interpretation:**
  - Prediction method with 5% false negative rate and 1% false positive rate, with 1% baseline, predicting 500 effectors:
    - ***P*(effector|positive test)≈0.5**

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

A lesson from the literature?

The James Hutton Institute

- "The resulting computational model revealed a strong type III secretion signal in the N-terminus that can be used to detect effectors with **sensitivity of 71%** and **[specificity] of 85%**."
  - Sensitivity [*P*(+ve|T3E)] = 0.71; FPR [1-Specificity; *P*(+ve|not T3E)] = 0.15
  - Base rate [*P*(T3E)] ≈ 3%; Genes = 4500

$$P(\mathrm{T3E}|+\mathrm{ve}) = \frac{P(+\mathrm{ve}|\mathrm{T3E})P(\mathrm{T3E})}{P(+\mathrm{ve}|\mathrm{T3E})P(\mathrm{T3E}) + P(+\mathrm{ve}|\overline{\mathrm{T3E}})P(\overline{\mathrm{T3E}})}$$

  - **We expect *P*(T3E|+ve) ≈ 0.13**
  - **(and a significant number, up to 15% of the genome, of false positives… see `pred_acc.ipynb`)**

A lesson from the literature?

The James Hutton Institute

- **"The surprisingly high number of (false) positives in genomes without TTSS exceeds the expected false positive rate (Table 1)"**

TTSS status	G+C content	Number of proteins	Positives	Z-Score
-	56.3%	1841	3.7%	0.6
-	43.1%	1963	1.6%	-2.3
-	[illegible]	[illegible]	[illegible]	8.3
-	[illegible]	[illegible]	[illegible]	6.8
+	57.9%	5169	3.8%	7.2
+	[illegible]	[illegible]	[illegible]	[illegible]
+	58.3%	5607	3.7%	[illegible]
-	42.8%	2120	10.8%	[illegible]
-	44.8%	2385	12.2%	[illegible]
-	40.1%	3541	3.3%	[illegible]
-	45.8%	4008	3.5%	[illegible]
+	51.4%	4510	2.8%	[illegible]
+	52.0%	4614	2.7%	[illegible]
+	52.1%	4659	2.8%	[illegible]
+	52.1%	4617	2.8%	[illegible]
+	52.2%	4205	2.8%	[illegible]
+	52.2%	3964	2.6%	[illegible]
+	52.1%	4779	2.5%	[illegible]
+	52.2%	4805	2.6%	[illegible]
+	52.2%	4091	2.7%	[illegible]
+	52.1%	5601	3.1%	[illegible]
+	52.2%	4627	2.8%	[illegible]
+	51.9%	4753	2.9%	1.9
+	52.1%	4312	2.7%	0.8
+	52.2%	4523	2.8%	1.8
-	53.6%	3645	3.0%	5.0
+	46.2%	4489	4.7%	7.6
-	46.3%	4394	4.7%	7.2
+	46.2%	4687	4.8%	7.7
-	46.3%	4440	4.8%	7.3
-	45.1%	3754	4.9%	5.0

196 Type III effectors?

TTSS status	G+C content	Number of proteins	Positives	Z-Score
-	58.4%	4777	3.3%	-1.0
-	66.9%	2157	3.6%	1.1
-	62.4%	4587	7.7%	18.9
-	65.4%	4506	7.2%	20.7
-	59.2%	1631	6.1%	6.9
-	60.2%	1998	6.2%	8.8
-	60.1%	1728	6.4%	6.2
-	59.9%	2416	5.3%	7.3
-	72.5%	3079	5.0%	12.6
-	72.4%	2940	4.0%	8.0
-	53.5%	2272	5.8%	8.5
-	63.1%	2950	9.3%	19.4
-	53.8%	3056	5.9%	9.5
-	[illegible]	[illegible]	[illegible]	[illegible]
-	54.1%	3076	6.4%	11.9
-	61.4%	2119	10.3%	21.1

218 Type III effectors, no T3SS?

0.038 x 5169 x 0.13 ≈ 26

[No. +ve x *P*(T3E|+ve)]

1-2% (Collmer *et al.* 2002); 1% (Boch and Bonas, 2010)

Conclusions II

The James Hutton Institute

- Genomic context is almost as interesting and useful as genome features
- **Functional prediction is largely sequence classification**
- **The statistics of large datasets can be unintuitive: always try to think statistically**
- Workflows (e.g. Galaxy) are great tools

2013: *Dickeya* spp.

The James Hutton Institute

genomeA Journals.ASM.org

Draft Genome Sequences of 17 Isolates of the Plant Pathogenic Bacterium *Dickeya*

Leighton Pritchard,[a] Sonia Humphris,[b] Gerry S. Saddler,[c] John G. Elphinstone,[d] Minna Pirhonen,[e] Ian K. Toth[b]

genomeA Journals.ASM.org

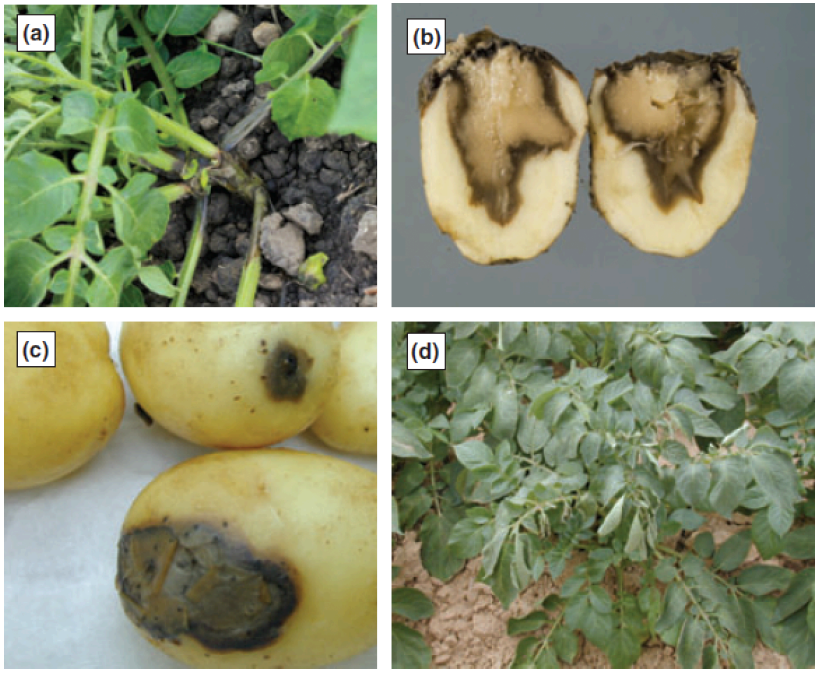
Draft Genome Sequences of Four *Dickeya dianthicola* and Four *Dickeya solani* Strains

Leighton Pritchard,[a] Sonia Humphris,[b] Steve Baeyen,[c] Martine Maes,[c] Johan Van Vaerenbergh,[c] John Elphinstone,[e] Gerry Saddler,[d] Ian Toth[b]

Pritchard *et al.* (2013) *Genome Announc*. doi:10.1128/genomeA.00978-13.
Pritchard *et al.* (2013) *Genome Announc*. doi:10.1128/genomeA.00087-12.

Dickeya spp.

- Soft rot pathogen of crop and ornamental plants
- *Dickeya* species have distinct host ranges
- Expanding and diversifying across Europe
- ***D. solani* a particular emerging threat**
- Promising SynthBio/BioFuels target



Toth *et al.* (2011) *Plant Path*. **60**: 385-399
doi:10.1111/j.1365-3059.2011.02427.x.

Host: chrysanthemum (*Chrysanthemum* spp.)
Landesanst. f. Pflanzenbau und Pflanzenschutz, Mainz Archive

Dickeya spp.

The James Hutton Institute

- 25 new bacterial genomes, spanning six species of genus
- Several short read methods, several sequencing services

454 (Liverpool), Illumina (Glasgow, Belgium), SE and PE

- 6-84X coverage, 170k-4m reads
- Genomes in draft form: 27-273 contigs
- Annotation mostly automated
- Published 2013, *Genome Announcements* (6, 8 authors)

Pritchard *et al.* (2013) *Genome Announc*. doi:10.1128/genomeA.00978-13.
Pritchard *et al.* (2013) *Genome Announc*. doi:10.1128/genomeA.00087-12.

Dickeya classification is in flux

- a.k.a. *Pectobacterium chrysanthemi* or *Erwinia chrysanthemi*
- *Erwinia chrysanthemi* complex (from 1980)
  - Pathovars *chrysanthemi*, *dianthicola*, *dieffenbachiae*, *parthenii*, *zeae*, *paradisiaca*
  - (*E.chrysanthemi* pv *paradisiaca* later *Brenneria paradisiaca*)
- Classification by phenotype

Characteristic	Phenon 1	Phenon 2	Phenon 4	Phenon 3	Phenon 5	Phenon 6	*P. cypripedii*
(−)-D-Arabinose	+	−	−	+	−	+	+
(−)-D-Tartrate	−	d (25)	−	−	+	+	+
Inulin	−	−	+	−	d (88)	−	−
Lactose	+	d (75)	d (20)	−	−	d (17)	+
Growth at 39 °C	+	+	+	+	−	d (83)	−
cis-Aconitate	+	−	d (20)	d (80)	−	−	+
(+)-D-Melibiose, (+)-D-raffinose	+	+	+	−	d (44)	d (83)	−
5-Keto-D-gluconate	−	−	−	d (20)	−	+	+
Mannitol	+	+	+	+	+	−	+
Lecithin	+	+	+	+	+	−	−
ADH Moeller	d (15)	−	+	−	d (69)	−	−
meso-Tartrate	+	d (75)	−	+	+	+	+
myo-Inositol	+	+	d (80)	+	+	−	+
Casein	+	d (75)	+	d (80)	d (75)	−	−
Novel species	*D. dadantii* +*D. zeae*	*D. chrysanthemi* bv. *parthenii*	*D. chrysanthemi* bv. *chrysanthemi*	*D. dieffenbachiae*	*D. dianthicola*	*D. paradisiaca*	

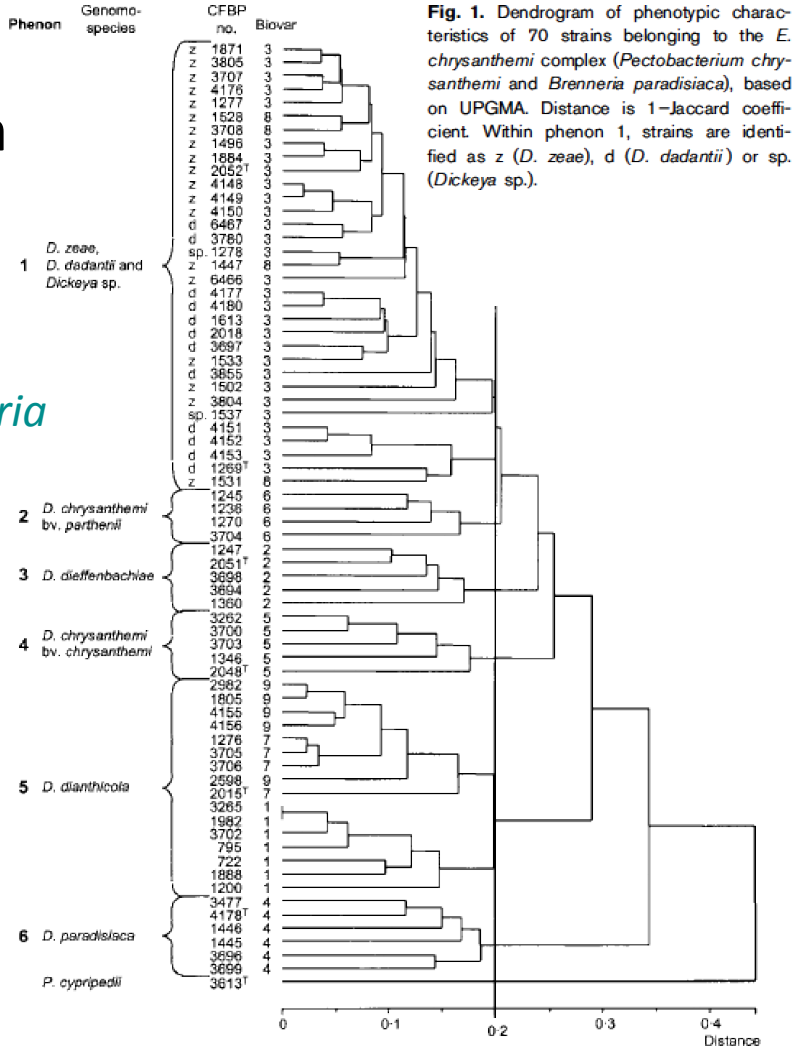


Fig. 1. Dendrogram of phenotypic characteristics of 70 strains belonging to the *E. chrysanthemi* complex (*Pectobacterium chrysanthemi* and *Brenneria paradisiaca*), based on UPGMA. Distance is 1−Jaccard coefficient. Within phenon 1, strains are identified as z (*D. zeae*), d (*D. dadantii*) or sp. (*Dickeya* sp.).

Samson *et al.* (2005) *Int J Syst Evol Micr*. doi:10.1099/ijs.0.02791-0.

Current classification of *Dickeya*

- Complex transferred to *Dickeya* gen. nov. with six assigned species:
  - *Dickeya chrysanthemi*
  - *Dickeya paradisiaca*
  - *Dickeya dadantii*
  - *Dickeya dianthicola*
  - *Dickeya dieffenbachiae*
  - *Dickeya zeae*

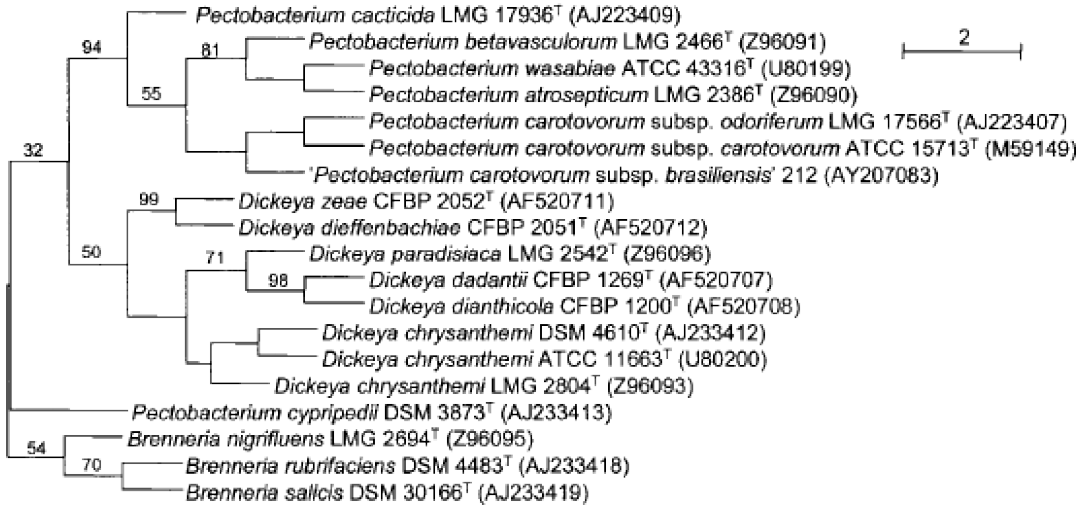
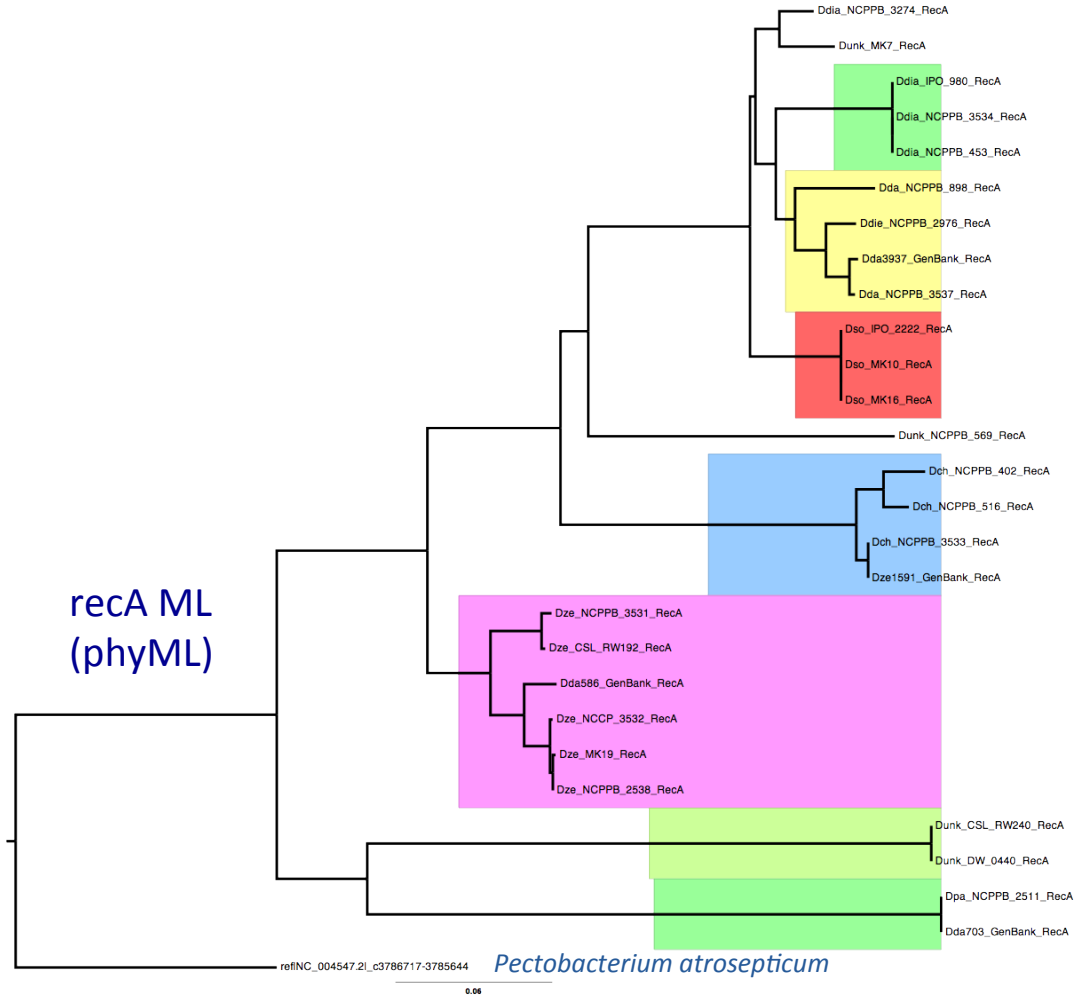


Fig. 2. Unrooted tree, the result of a phylogenetic analysis (parsimony shown) of 16S rRNA gene sequences of type strains of *Dickeya*, *Pectobacterium* and *Brenneria* species. Bootstrap values (expressed as percentages of 1000 replications, BIONJ+Kimura two-parameter) are indicated only for branches also retrieved by BIONJ and maximum-likelihood analyses ($P<0{\cdot}01$).

Samson *et al.* (2005) *Int J Syst Evol Micr*. doi:10.1099/ijs.0.02791-0.

Classification of sequenced *Dickeya*

The James Hutton Institute



recA ML
(phyML)

D. dianthicola

D. dadantii
(*D. d.* subsp. *dieffenbachiae*)

D. solani

D. chrysanthemi

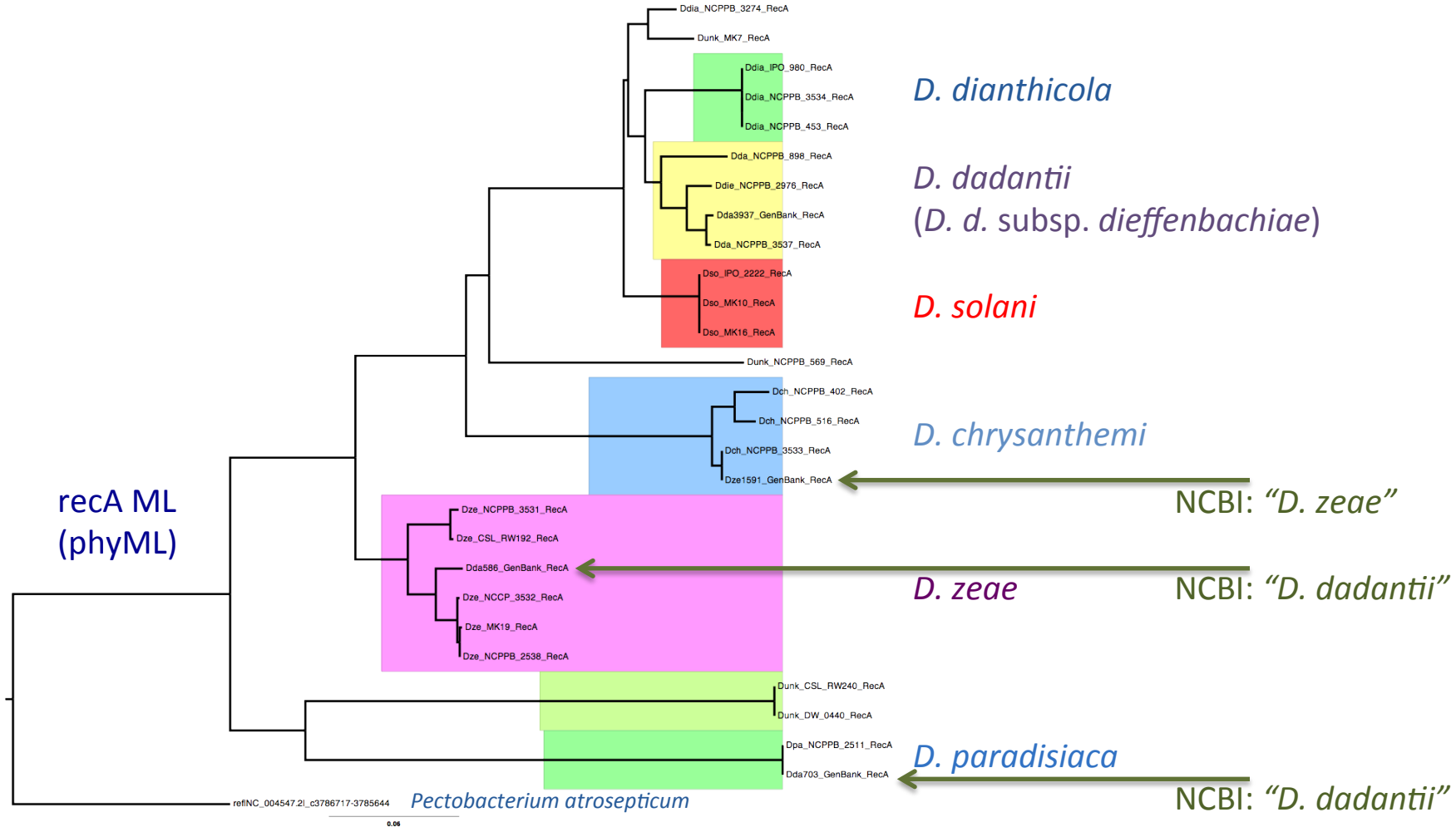
D. zeae

D. paradisiaca

Pectobacterium atrosepticum

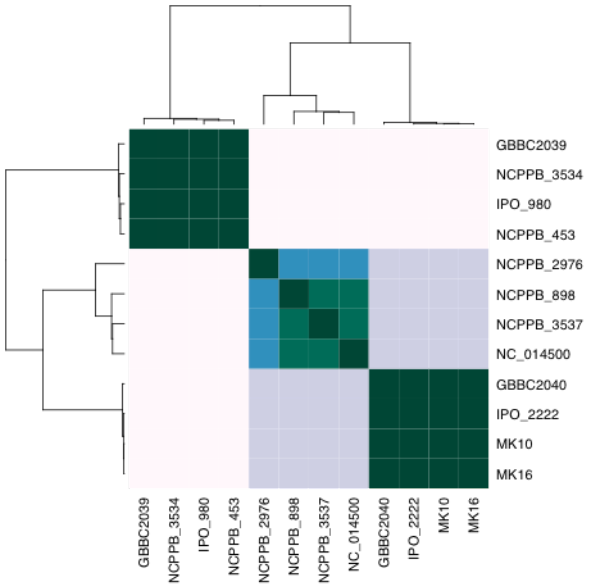
GenBank *Dickeya* genomes misclassified

- 3/4 *Dickeya* genomes in GenBank are misassigned species



D. solani recently described

- There was some question over whether *D. solani* is a species, or subspecies of *D. dadantii*.
- Species boundaries: 70% DNA-DNA hybridisation (DDH)
  - equivalent to 95% Average Nucleotide Identity (ANI), determined by whole-genome alignment



		D.dianthicola				*D.dadantii*				*D.solani*			
		GBBC2039	NCPPB_453	NCPPB_3534	IPO_980	NCPPB_2976	NCPPB_3537	NCPPB_898	NC_014500	GBBC2040	IPO_2222	MK10	MK16
D.dianthicola	GBBC2039	1	0.9936	0.9927	0.993	0.9222	0.9219	0.9221	0.9223	0.9227	0.9229	0.923	0.9231
	NCPPB_453	0.9936	1	0.9947	0.9952	0.9221	0.9216	0.922	0.9219	0.9226	0.9226	0.9229	0.9231
	NCPPB_3534	0.9927	0.9947	1	0.9937	0.922	0.9219	0.922	0.9221	0.9222	0.9223	0.9225	0.9227
	IPO_980	0.993	0.9952	0.9937	1	0.9217	0.9214	0.9218	0.9217	0.9223	0.9224	0.9227	0.9227
D.dadantii	NCPPB_2976	0.9222	0.9221	0.922	0.9217	1	0.9684	0.9669	0.968	0.9411	0.9413	0.9415	0.9416
	NCPPB_3537	0.9219	0.9216	0.9219	0.9214	0.9684	1	0.985	0.985	0.9406	0.941	0.9412	0.9414
	NCPPB_898	0.9221	0.922	0.922	0.9218	0.9669	0.985	1	0.9838	0.94	0.9403	0.9405	0.9407
	NC_014500	0.9223	0.9219	0.9221	0.9217	0.968	0.985	0.9838	1	0.9408	0.9413	0.9415	0.9417
D.solani	GBBC2040	0.9227	0.9226	0.9222	0.9223	0.9411	0.9406	0.94	0.9408	1	0.9987	0.9983	0.9983
	IPO_2222	0.9229	0.9226	0.9223	0.9224	0.9413	0.941	0.9403	0.9413	0.9987	1	0.9994	0.9995
	MK10	0.923	0.9229	0.9225	0.9227	0.9415	0.9412	0.9405	0.9415	0.9983	0.9994	1	0.9997
	MK16	0.9231	0.9231	0.9227	0.9227	0.9416	0.9414	0.9407	0.9417	0.9983	0.9995	0.9997	1

International Journal of Systematic and Evolutionary Microbiology (2014), **64**, 768–774 DOI 10.1099/ijs.0.052944-0

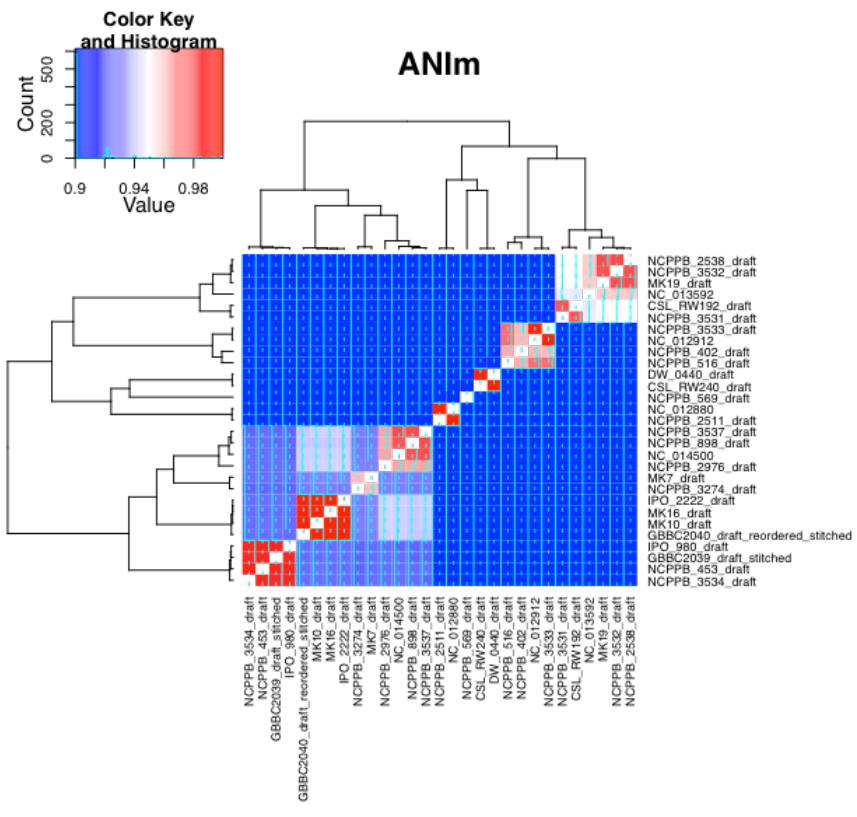
Dickeya solani sp. nov., a pectinolytic plant-pathogenic bacterium isolated from potato (*Solanum tuberosum*)

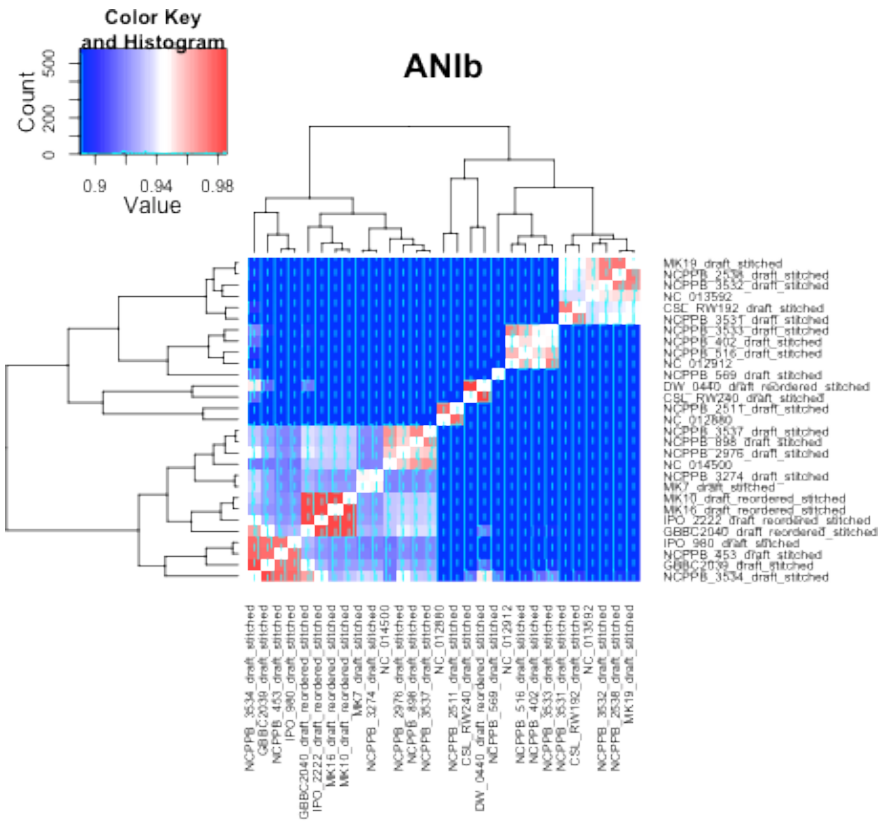
Jan M. van der Wolf,[1] Els H. Nijhuis,[1] Malgorzata J. Kowalewska,[2] Gerry S. Saddler,[2] Neil Parkinson,[3] John G. Elphinstone,[3] Leighton Pritchard,[4] Ian K. Toth,[4] Ewa Lojkowska,[5] Marta Potrykus,[5] Malgorzata Waleron,[5] Paul de Vos,[6] Ilse Cleenwerck,[6] Minna Pirhonen,[7] Linda Garlant,[7] Valérie Hélias,[8] Joël F. Pothier,[9,10] Valentin Pflüger,[11] Brion Duffy,[10] Leah Tsror[12] and Shula Manulis[12]

https://github.com/widdowquinn/scripts/tree/master/bioinformatics

van der Wolf *et al.* (2014) *Int J. Syst. Evol. Micro*. doi:10.1099/ijs.0.052944-0.

Nine *Dickeya* species?





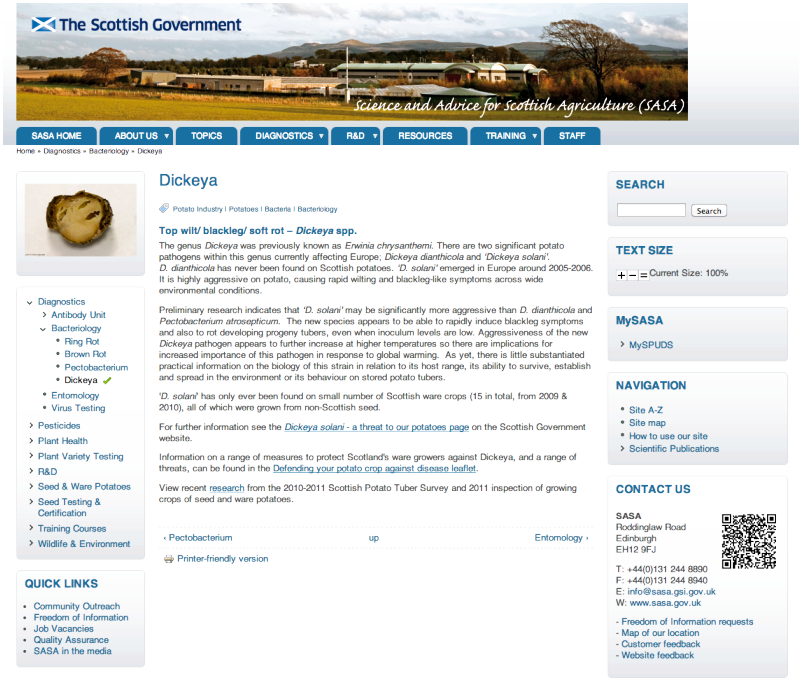
- Availablility of a genome for your pathogen may reveal some taxonomic surprises…
  (*Rhodococcus…*)

https://github.com/widdowquinn/scripts/tree/master/bioinformatics

Dickeya Genomic Diagnostics

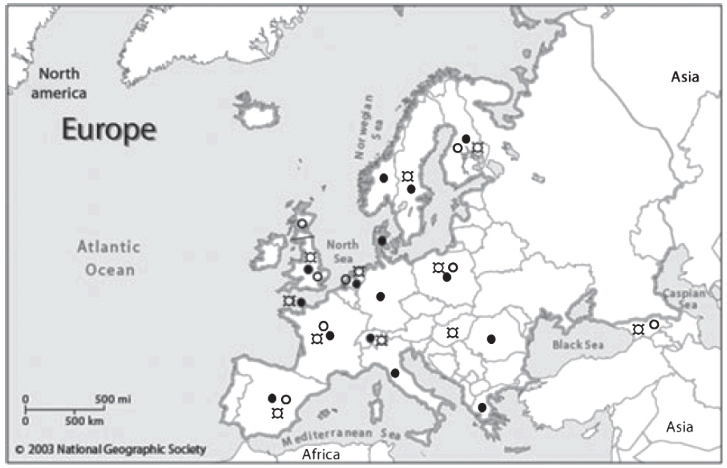
The James Hutton Institute

- **To legislate with respect to, and quarantine materials infected with, a pathogen we must be able to identify it accurately and precisely.**



Dickeya spp. are a threat in Europe

- *D. dianthicola* an established pathogen across Europe
- ***D. solani* an emerging threat**
  - Increased virulence
  - Different temperature preference
- Must be able to distinguish and identify, to legislate and/or quarantine
  - Zero-tolerance in Scotland



- No qPCR primers existed to distinguish among *Dickeya* spp.
- Was asked to find a solution, starting from our draft quality genome sequences

Toth *et al.* (2011) *Plant Path*. doi:10.1111/j.1365-3059.2011.02427.x.

qPCR primer design

The James Hutton Institute

- Typical approaches use ***rational design***
  - Select region that is (thought to be):
    - Similar in target organisms
    - Divergent or absent in off-target organisms
    - Conserved enough to be amplified by single primer set
  - Design primers to amplify that region
  - Frequently-used regions: intergenic transcribed spacers (ITS), ribosomal DNA, "housekeeping" genes, "virulence" genes.
- **Alignment-Based Process:**
  1. Identify common region(s) ("alignment")
  2. Design primers

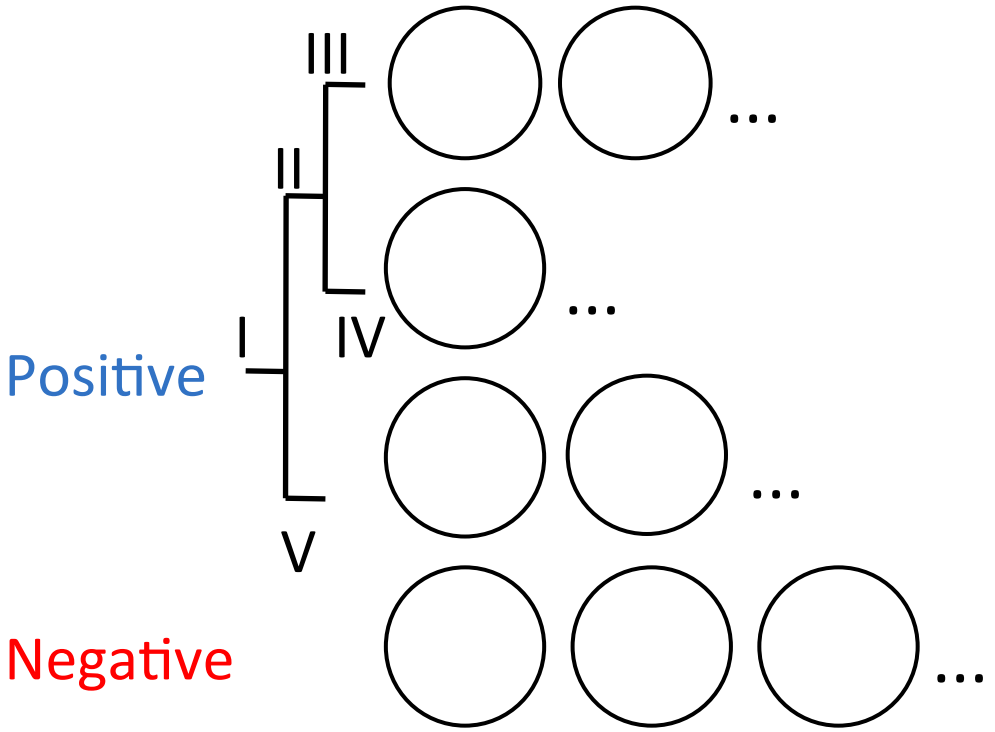
Availability of whole pathogen genomes gives new options…

“Alignment-free” Genomic PCR Design

1. Define positive, negative sets, and classes

2. Convert to single (pseudo)chromosomes:

3. Genome feature locations from GBK file or predicted



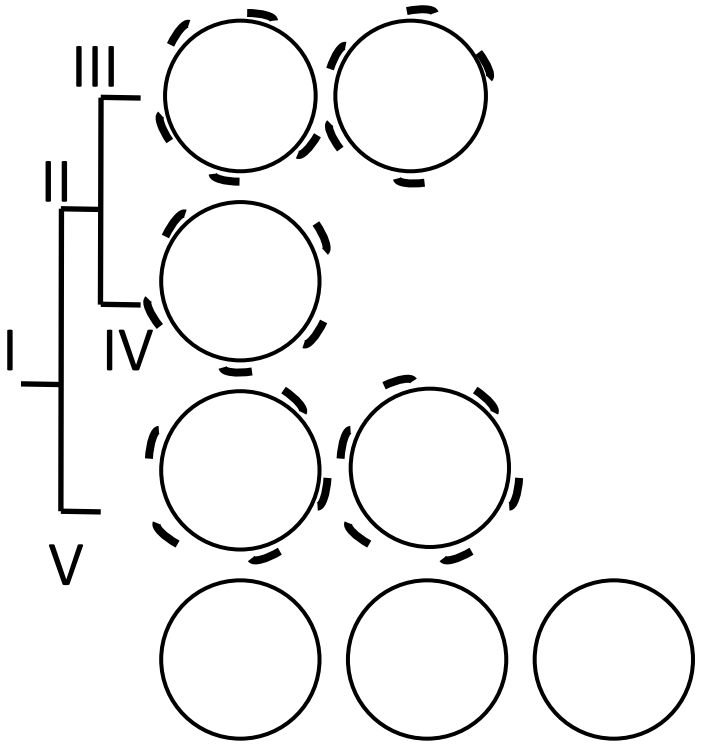
Pritchard *et al.* (2012) *PLoS One*. doi:10.1371/journal.pone.0034498.
Pritchard *et al.* (2013) *Plant Path*. doi:10.1111/j.1365-3059.2012.02678.x.

Primer prediction (on positive set)

4. Predict **> 1000** thermodynamically plausible primer sets on each (pseudo)chromosome

Positive

Negative

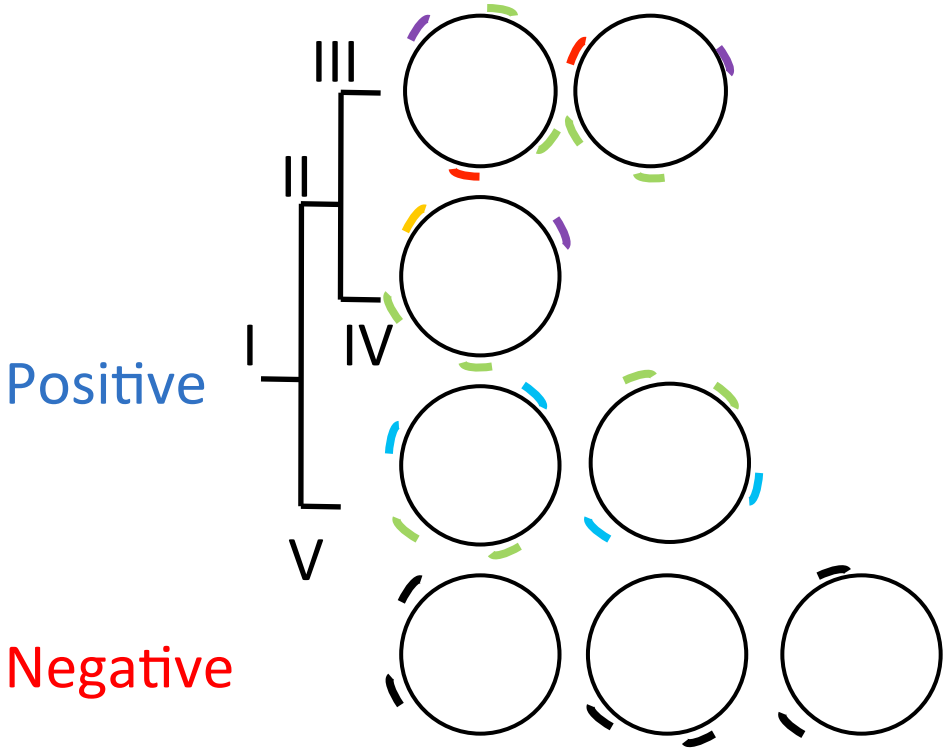


Pritchard *et al.* (2012) *PLoS One*. doi:10.1371/journal.pone.0034498.
Pritchard *et al.* (2013) *Plant Path*. doi:10.1111/j.1365-3059.2012.02678.x.

Test cross-amplification *in silico*

5. Check cross-amplification: All primer sets tested against other organisms, colour by amplified class(es).

6. BLAST screen: All primers screened for off-target sequences with BLAST

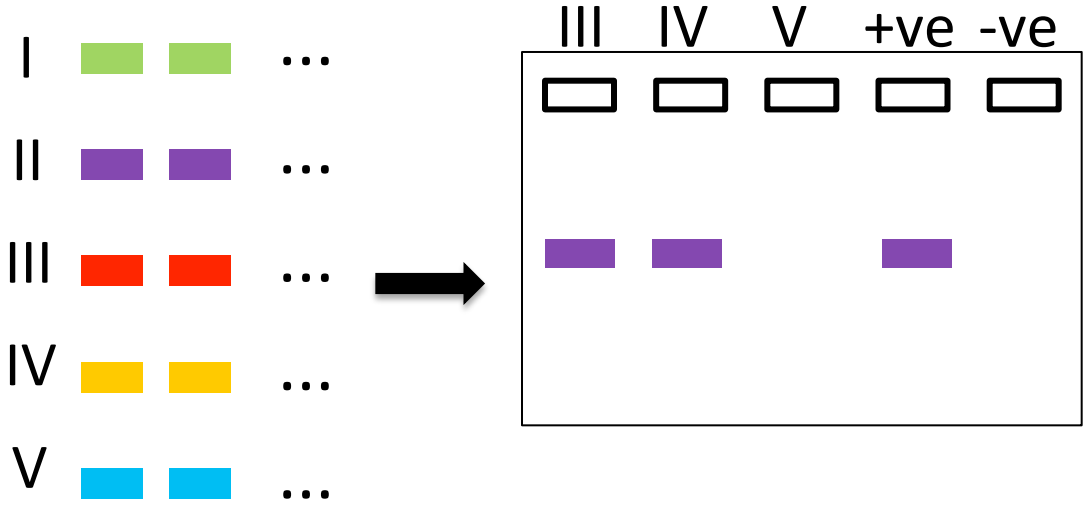


Pritchard *et al.* (2012) *PLoS One*. doi:10.1371/journal.pone.0034498.
Pritchard *et al.* (2013) *Plant Path*. doi:10.1111/j.1365-3059.2012.02678.x.

Classify primers and validation

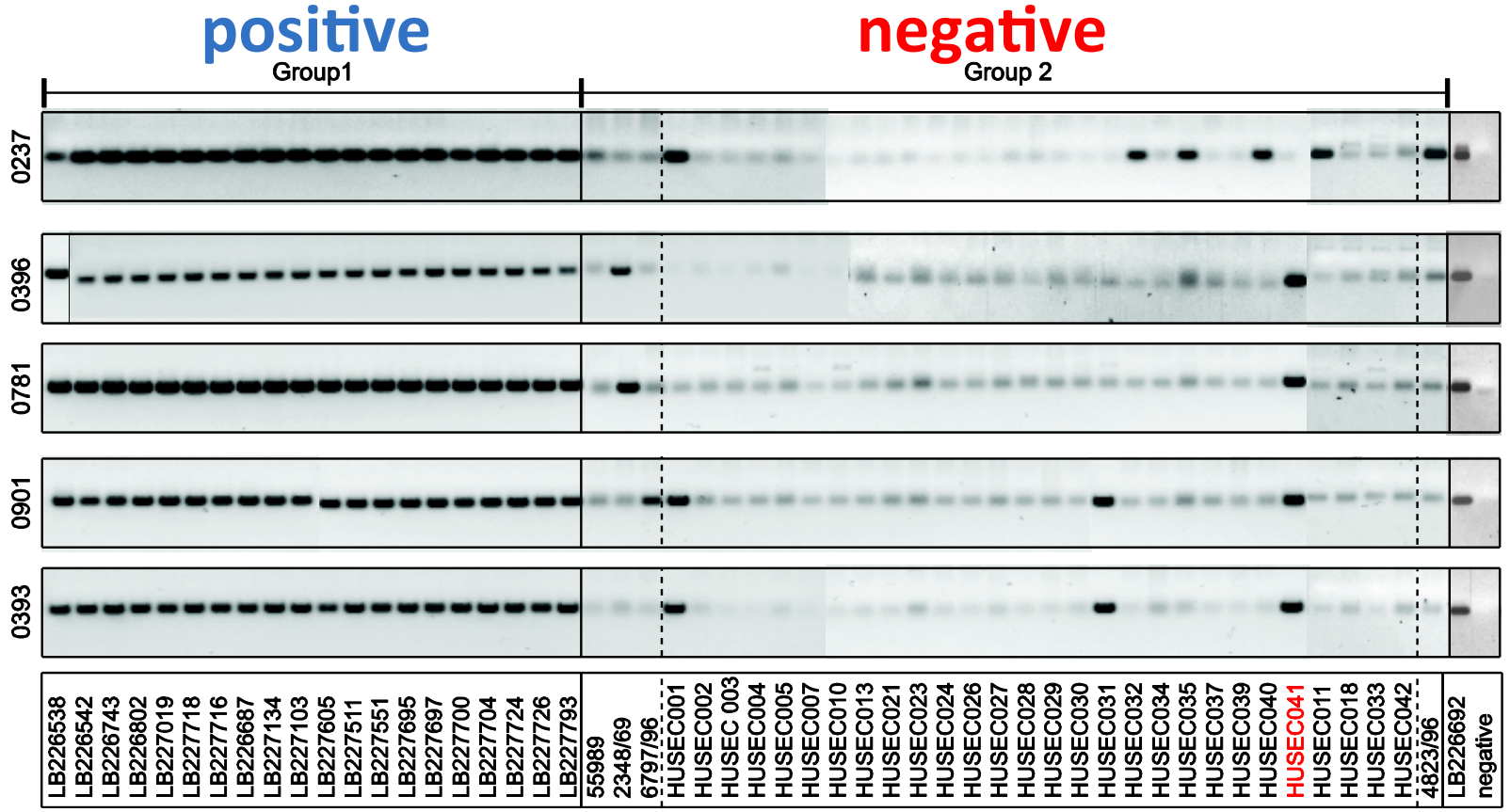
7. Classify primers: according to the ability to amplify specific classes of input sequence.

8. Validate primers: Primer set validated on positive and negative targets *in vitro*.



Pritchard *et al.* (2012) *PLoS One*. doi:10.1371/journal.pone.0034498.
Pritchard *et al.* (2013) *Plant Path*. doi:10.1111/j.1365-3059.2012.02678.x.

Primer design: validated *in vitro* (*E.coli*)



Sub-serotype specificity:
impB (β-lactamase plasmid),
prophage gp20 transfer protein

OPEN ACCESS Freely available online

PLoS one

Alignment-Free Design of Highly Discriminatory Diagnostic Primer Sets for *Escherichia coli* O104:H4 Outbreak Strains

Leighton Pritchard[1*], Nicola J. Holden[2], Martina Bielaszewska[3], Helge Karch[3], Ian K. Toth[2]

Pritchard *et al.* (2012) *PLoS One*. doi:10.1371/journal.pone.0034498.

Primer design: validated *in vitro* (*Dickeya*)

The James Hutton Institute

Table 5 Specificity of real-time qPCR assays predicted by the pipeline compared with existing qPCR and conventional PCR assays

		Number of isolates detected, by assay								
Test species	Isolates tested	DIC	DIA-A	DIA-C	SOL-C	SOL-D	PEC	ECH	ECA	ADE
D. dianthicola	7	7	7	7	0	0	7	7	0	7
'*D. solani*' (DUC-1)	16	16	0	0	16	16	16	16	0	16
DUC-2	5	5	0	0	0	0	5	5	0	5
DUC-3	1	1	0	0	0	0	1	1	0	1
D. dadantii	11	11	0	0	1	1	11	11	0	11
D. dieffenbachiae	6	6	0	0	0	0	6	6	0	6
D. chrysanthemi bv. *chrysanthemi*	7	7	0	0	0	0	7	7	0	7
D. chrysanthemi bv. *parthenii*	3	3	0	0	0	0	3	3	0	3
D. paradisiaca	1	1	0	0	0	0	1	1	0	1
D. zeae	11	11	0	0	0	0	11	11	0	11
New *Dickeya* species level clade (I)	1	1	0	0	0	0	1	1	0	1
New *Dickeya* species level clade (II)	1	1	0	0	0	0	1	1	0	1
Pectobacterium atrosepticum	1	1	0	0	0	0	1	0	1	0
P. carotovorum subsp. *carotovorum*	1	1	0	0	0	0	1	0	0	0
P. betavasculorum	1	1	0	0	0	0	1	0	0	0
P. carotovorum subsp. *odoriferum*	1	1	0	0	0	0	1	0	0	0
P. wasabiei	1	1	0	0	0	0	1	0	0	0
Pantoea agglomerans	1	1	0	0	0	0	0	0	0	0
Brenneria quercina	1	1	0	0	0	0	0	0	0	0
Erwinia amylovora	1	1	0	0	0	0	0	0	0	0

Primers tested against 70 isolates

Plant Pathology (2012) Doi: 10.1111/j.1365-3059.2012.02678.x

Detection of phytopathogens of the genus *Dickeya* using a PCR primer prediction pipeline for draft bacterial genome sequences

L. Pritchard[a], S. Humphris[a], G. S. Saddler[b], N. M. Parkinson[c], V. Bertrand[c], J. G. Elphinstone[c] and I. K. Toth[a]*

Pritchard *et al.* (2013) *Plant Path*. doi:10.1111/j.1365-3059.2012.02678.x.

Dickeya genecalling

- **No free lunch for gene-calling**
- Tested several automated methods against "gold standard" annotations
- **Systematic errors:** *rpmDHJ*, *prfB*, *outE*, *secD*, *hrpD*, *virB7*, etc.

Manual annotation, *Pectobacterium*

Genecaller:	Glimmer3	GeneMark	Prodigal (SD)	Prodigal (motif)	RAST
Prediction count	4679	4606	4467	4468	4648
Absolutely missed CDS	112	133	156	153	202
strict assessment (start and end base match)					
Sensitivity	0.73	0.75	0.86	0.85	0.79
False discovery rate (FDR)	0.31	0.27	0.14	0.14	0.24
Positive predictive value (PPV)	0.69	0.73	0.86	0.86	0.76
lenient assessment (end base matches)					
Sensitivity	0.97	0.97	0.97	0.97	0.95
False discovery rate (FDR)	0.07	0.06	0.03	0.03	0.08
Positive predictive value (PPV)	0.93	0.94	0.97	0.97	0.92

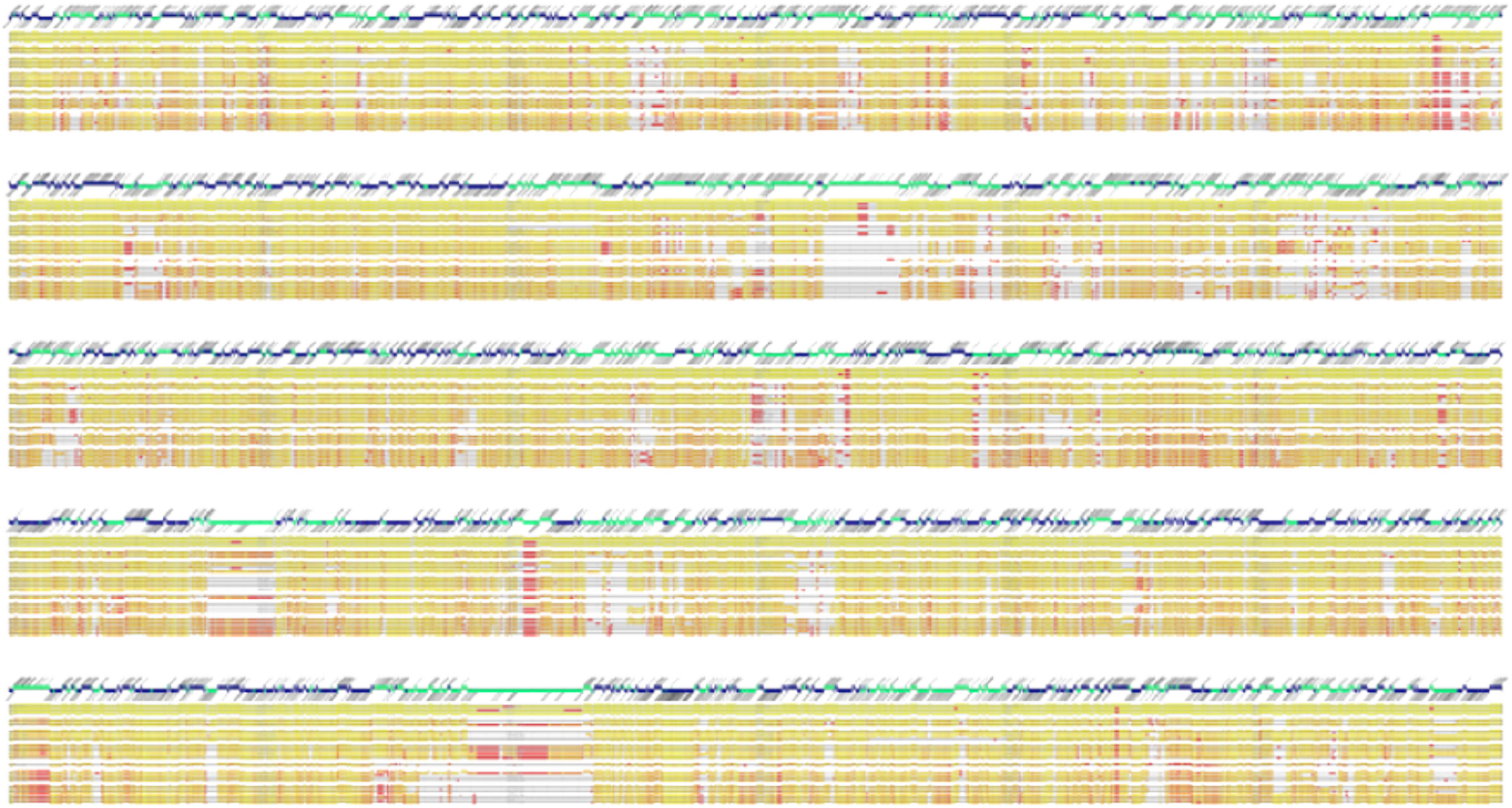
Community annotation, *Dickeya*

Genecaller:	Glimmer3	GeneMark	Prodigal (SD)	Prodigal (motif)	RAST
Prediction count	4752	4440	4287	4291	4604
Absolutely missed CDS	284	393	407	408	461
strict assessment (start and end base match)					
Sensitivity	0.62	0.59	0.71	0.71	0.64
False discovery rate (FDR)	0.41	0.39	0.25	0.25	0.37
Positive predictive value (PPV)	0.59	0.61	0.75	0.75	0.63
lenient assessment (end base matches)					
Sensitivity	0.94	0.91	0.91	0.91	0.90
False discovery rate (FDR)	0.10	0.06	0.03	0.03	0.09
Positive predictive value (PPV)	0.90	0.94	0.97	0.97	0.91

Dickeya RBH

The James Hutton Institute

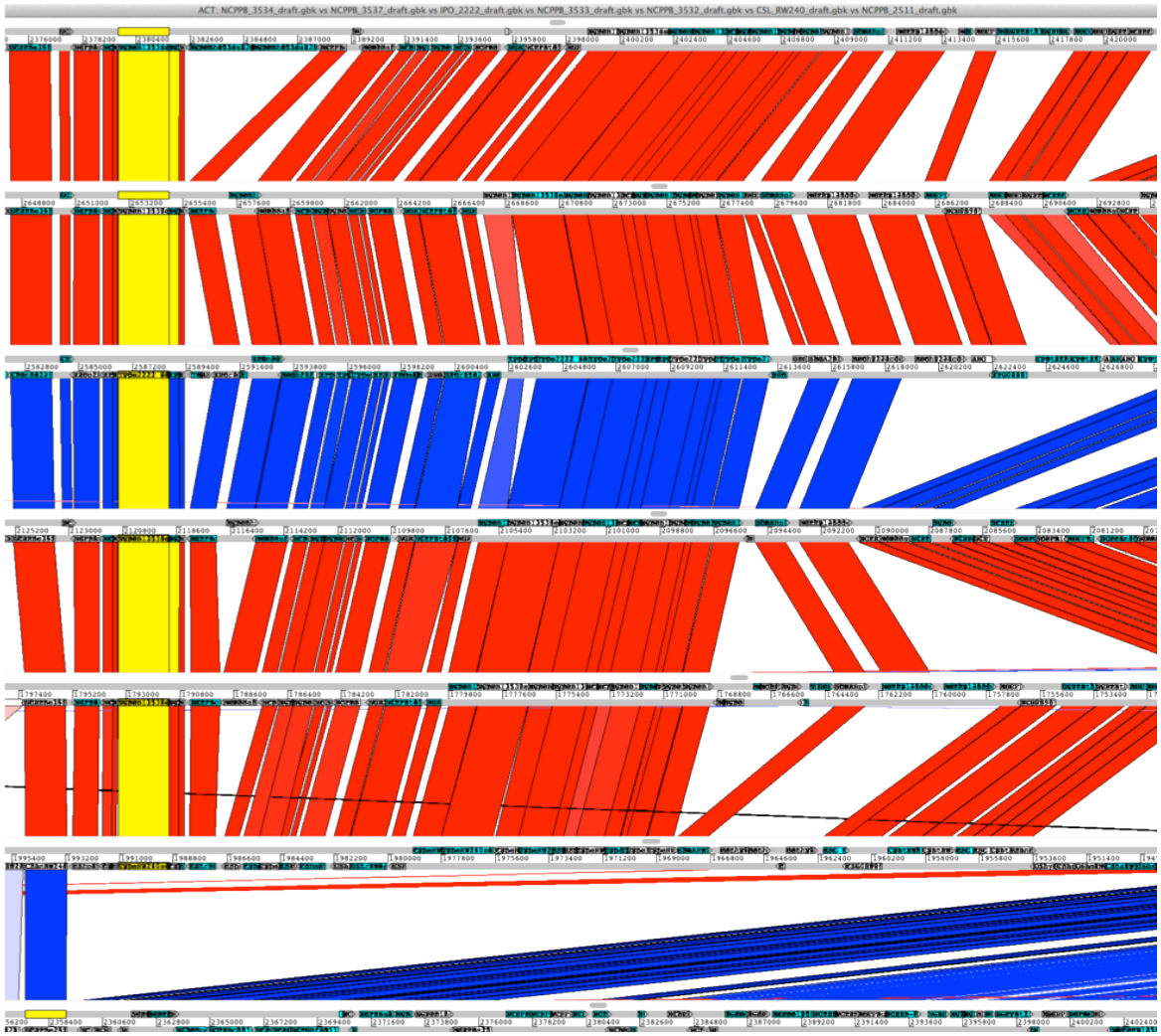
- Comparison of predicted gene complements
- Identifies sequence-similar ‘equivalence’, ***not orthology***



http://armchairbiology.blogspot.co.uk/2012/07/on-reciprocal-best-blast-hits.html

Type III Secretion System

The James Hutton Institute



D. dianthicola NCPPB 3534

D. dadantii NCPPB 3537

D. solani IPO 2222

D. chrysanthemi NCPPB 3533

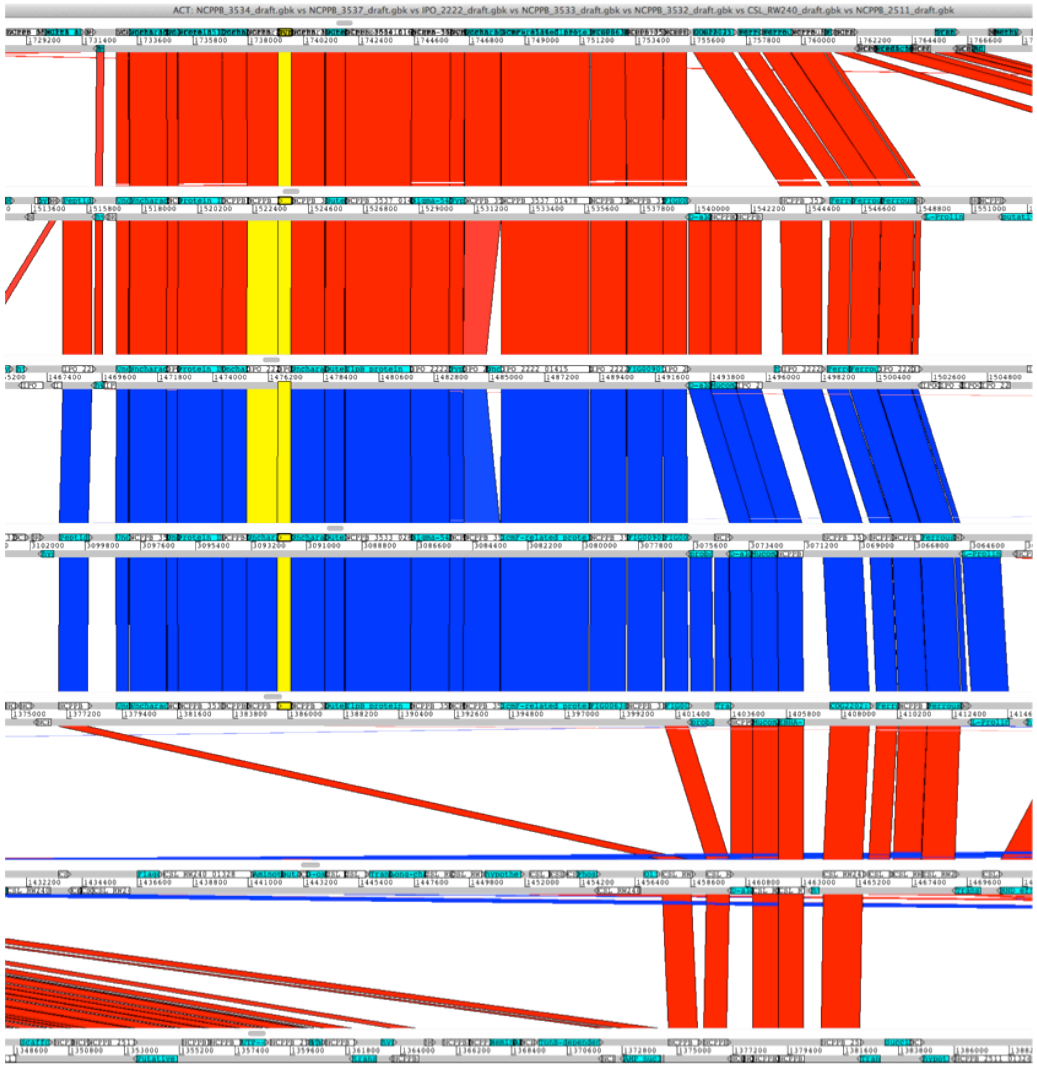
D. zeae NCPPB 3532

"*D. aquatica*" CSL RW240/1

D. paradisiaca NCPPB 2511

Type VI Secretion System

The James Hutton Institute



D. dianthicola NCPPB 3534

D. dadantii NCPPB 3537

D. solani IPO 2222

D. chrysanthemi NCPPB 3533

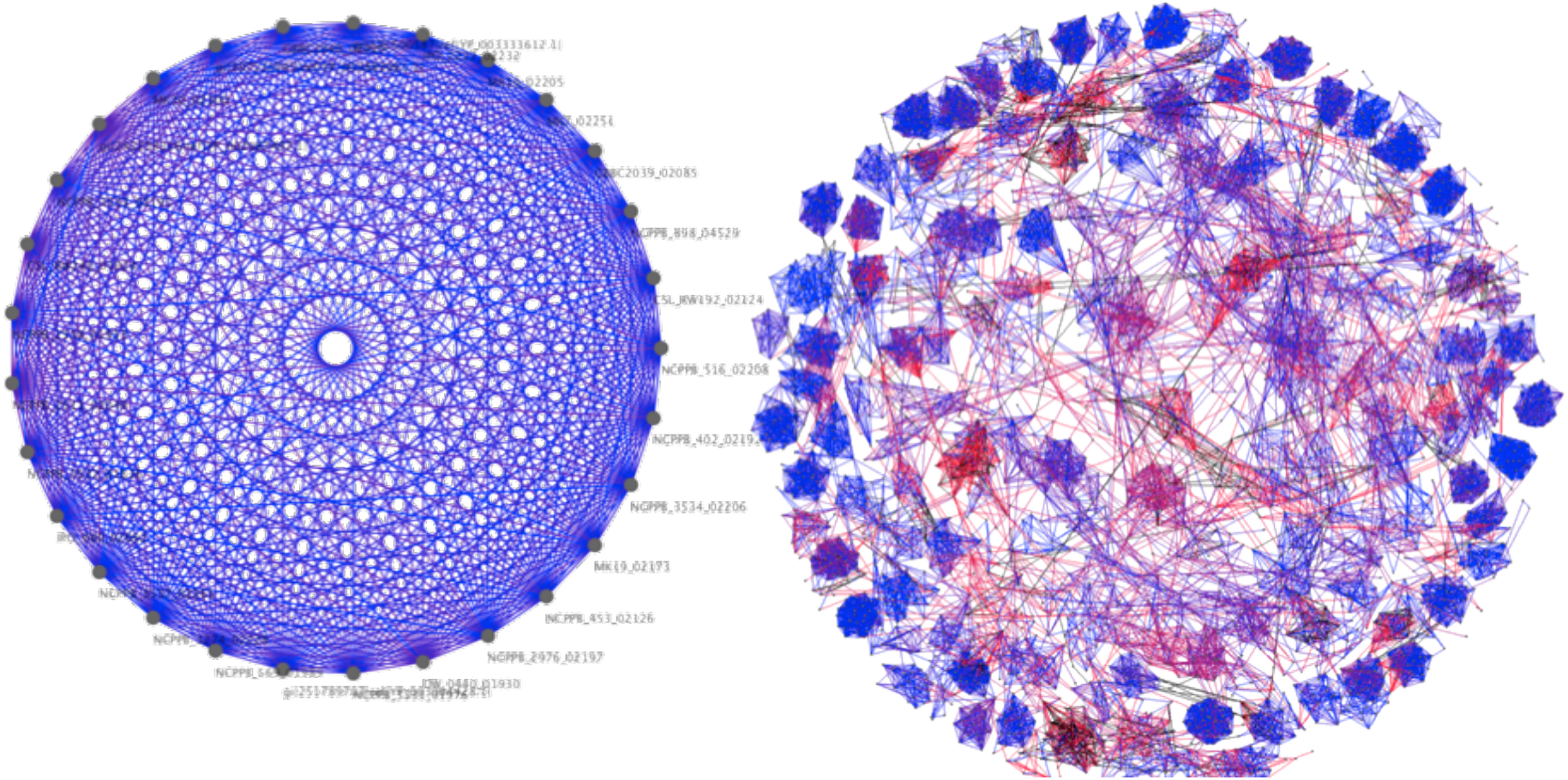
D. zeae NCPPB 3532

"D. aquatica" CSL RW240/1

D. paradisiaca NCPPB 2511

Pangenome identification

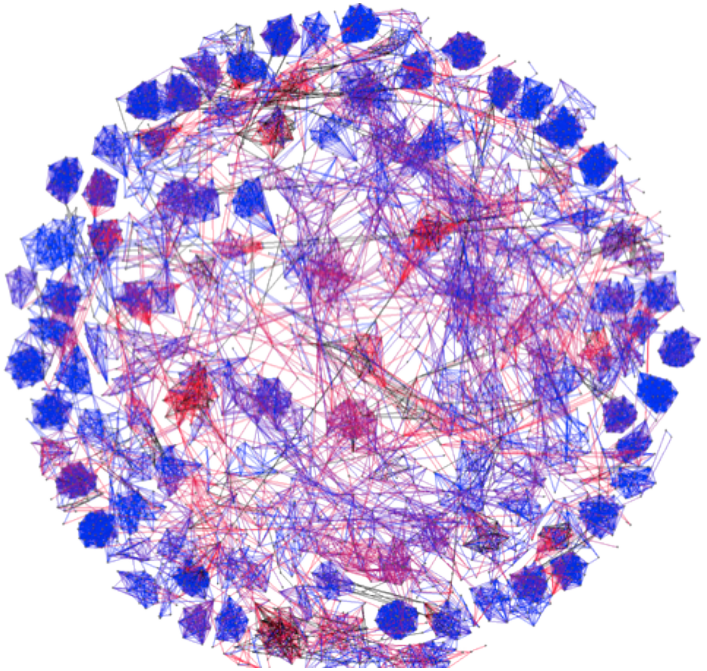
- **Clique:** each CDS makes RBH with every other member of the same *clique*



Pangenome identification

- Prune non-clique graphs to identify accessory genomes for each species

Species	Weak Pruning	Full Pruning
Core Genome	2201	2201
D. chrysanthemi	32	36
D. dadantii	11	14
D. dianthicola	102	127
D. paradisiaca	404	441
D. solani	120	157
D. zeae	33	40



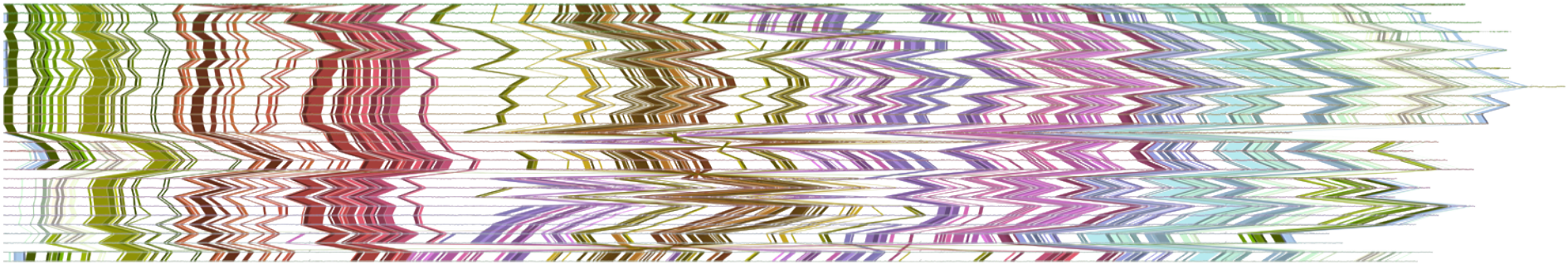
- Accessory: RBBH with all other members of same species, but no other *Dickeya*
- Weak pruning: remove all RBBH <80% identity, <40% coverage
- Full pruning: trim graph (by Mahalanobis distance) until minimal cliques found

Collinearity

The James Hutton Institute

- Using conserved order of RBH pairs to identify collinear regions
  - i-ADHoRe (http://bioinformatics.psb.ugent.be/software/details/i--ADHoRe)

29 *Dickeya* genomes: core genome



Dickeya zeae: core and accessory genome

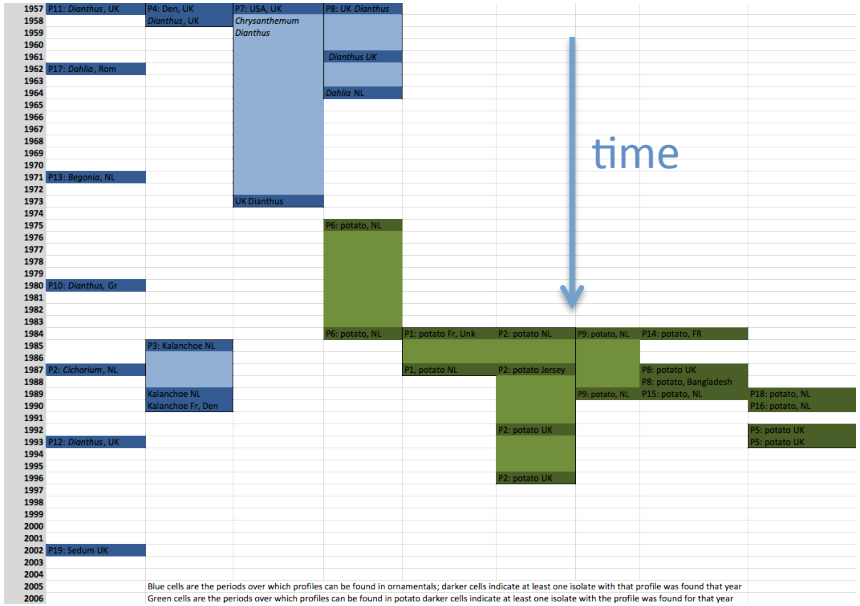
Simillion *et al.* (2008) *Bioinformatics* doi:10.1093/bioinformatics/btm449
http://armchairbiology.blogspot.co.uk/2012/09/the-colours-man-colours.html
https://github.com/widdowquinn/pyADHoRe

Epidemiology

- *Dickeya* infections are well-controlled on potato imports. But *Dickeya* has a broad host range. How did/does it get into Europe?

Variable nucleotide tandem repeats (VNTR)

- Identified in sequenced *D. dianthicola* genomes
- variable repeat numbers in historical isolates



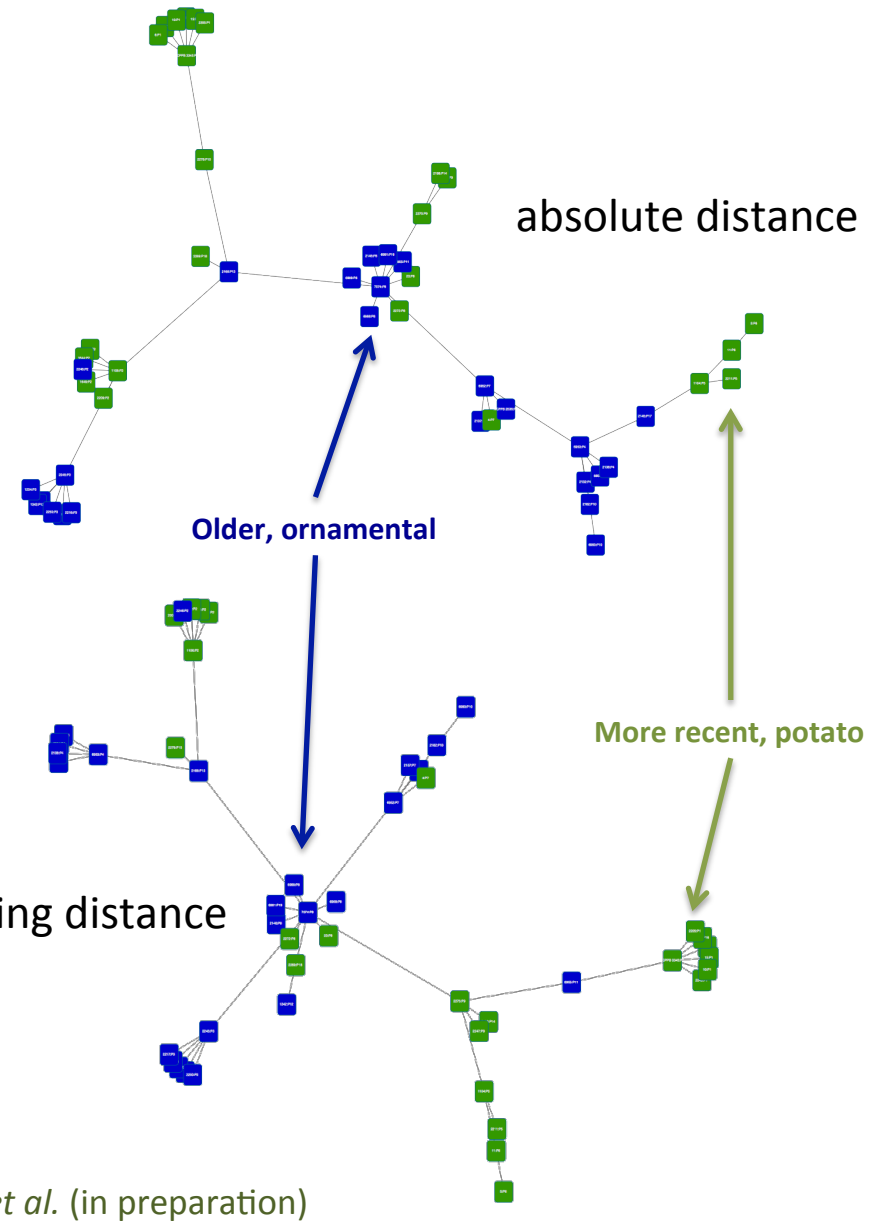
Blue: ornamental

Green: potato

Culture collection ref	Fera ref	Host	Country	Accession Date	19	19b	29	10	11b	28	VNTR Profile
PD 502	6	Potato	Unkown	Pre 1987	4	5	6	5	5	6	P1
PD 769	10	Potato	Unkown	Pre 1987	4	5	6	5	5	6	
PD 767	15	Potato	Unknown	Pre 1987	4	5	6	5	5	6	
NCPPB 3534	2205	Potato	Netherlands	1987	4	5	6	5	5	6	
PD 1022	2248	Potato	Netherlands	1987	4	5	6	5	5	6	
NCPPB 3345		Potato	France	1984	4	5	6	5	5	6	
NCPPB 3530	2209	Potato	Jersey	1987	3	4	7	5	3	5	P2
	1644	Potato	UK	1996	3	4	7	5	3	5	
	1649	Potato	UK	1996	3	4	7	5	3	5	
PD 484	2243	Potato	Netherlands	1984	3	4	7	5	3	5	
PD 788	2246	*Cichorium intybus*	Netherlands	1987	3	4	7	5	3	5	
	1106	Potato	UK	1992	3	4	7	5	3	5	
PD 554	1234	*Kalanchoe blossfeldiana*	Netherlands	1985	3	4	9	9	2	3	P3
PD 593	2245	*Kalanchoe blossfeldiana*	Netherlands	1985	3	4	9	9	2	3	
PD 1325	2250	*Kalanchoe sp.*	Netherlands	1989	3	4	9	9	2	3	
PD 1343	2251	*Kalanchoe sp.*	Netherlands	1989	3	4	9	9	2	3	
NCPPB 3730	2216	*Kalanchoe blossfeldiana*	Denmark	1990	3	4	9	9	2	3	
NCPPB 3729	2217	*Kalanchoe blossfeldiana*	France	1990	3	4	9	9	2	3	
NCPPB 518	2132	*Dianthus caryophyllus*	Denmark	1957	3	4	3	2	3	5	P4
NCPPB 393	2138	*Dianthus caryophyllus*	UK	1957	3	4	3	2	3	5	
NCPPB 429	6953	*Dianthus caryophyllus*	UK	1957	3	4	3	2	3	5	
NCPPB 430	6954	*Dianthus caryophyllus*	UK	1958	3	4	3	2	3	5	
	1104	Potato	UK	1992	3	2	3	2	3	3	P5
NCPPB 3881	2211	Potato	UK	1993	3	2	3	2	3	3	
PD 771	5	Potato	Netherlands	1984	1	2	3	2	3	3	P6
PD 256	11	Potato	Netherlands	1975	1	2	3	2	3	3	
CPPB 394	2137	Chrysanthemum morifolium	USA	1957	3	4	3	3	3	6	P7
NCPPB 426	6952	*Dianthus caryophyllus*	UK	1957	3	4	3	3	3	6	
NCPPB 2536		*Dianthus caryophyllus*	UK	1973	3	4	3	3	3	6	
PD 765	4	Potato	Unknown	Pre 1987	3	4	3	3	3	6	
NCPPB 1111	2146	*Dianthus caryophyllus*	UK	1961	3	4	3	4	3	6	P8
NCPPB 453	7074	*Dianthus caryophyllus*	UK	1957	3	4	3	4	3	6	
NCPPB 1609	6968	*Dahlia sp.*	Netherlands	1964	3	4	3	4	3	6	
NCPPB 1956	6969	*Dahlia sp.*	Netherlands	1964	3	4	3	4	3	6	
	2272	Potato	Bangladesh	1988	3	4	3	4	3	6	
NCPPB 3529	23	Potato	UK	1987	3	4	3	4	3	6	
PD 1406	2247	Potato	Netherlands	1989	3	2	3	5	3	6	P9
PD 482	2275	Potato	Netherlands	1984	3	2	3	5	3	6	
NCPPB 3138	2182	*Dianthus caryophyllus*	Greece	1980	4	4	3	3	4	4	P10
NCPPB 3139	6983	*Dianthus caryophyllus*	Greece	1980	4	4	3	3	4	4	
NCPPB 452	6955	*Dianthus caryophyllus*	UK	1957	3	5	3	5	3	6	P11
	1242	*Dianthus sp.*	UK	1993	3	3	9	8	3	2	P12
NCPPB 2421	2166	*Begonia bertinii*	Netherlands	1971	3	4	3	6	3	5	P13
NCPPB 3344	2198	Potato	France	1984	3	2	2	6	3	6	P14
	2279	Potato	Netherlands	1989	4	4	3	5	5	5	P15
NCPPB 3710	2220	Potato	Netherlands	1990	4	5	5	4	5	6	P16
NCPPB 1385	2145	*Dahlia sp.*	Romania	1962	3	3	3	2	3	5	P17
PD 1405	2269	Potato	Netherlands	1989	3	4	3	8	3	6	P18
NCPPB 4305	6991	*Sedum*	UK	2002	3	4	3	4	6	6	P19

Parkinson *et al.* (in preparation)

VNTR Minimum Spanning Trees



- Central nodes are associated with:
  - older accessions
  - ornamental accessions
- Recent accessions are associated with distal nodes
- Potato accessions are associated with distal nodes

Parkinson *et al.* (in preparation)

VNTR suggests ornamental introduction

The James Hutton Institute

- Earliest strains identified (1957) show VNTR diversity
- No *Dickeya* infection of potato observed until 1975
- Ornamental isolate diversity is greater than potato isolate diversity
- Cross-infection of ornamentals and potato supported by three VNTR profiles, found in both host types
- There is strict prohibition and regulation of potato import into the EU; there is not equivalent regulation of ornamental imports
- Ornamental plants may have provided a route for introduction of the broad host range pathogen *Dickeya* to EU potato crops.

Parkinson *et al.* (in preparation)

Systems Biology

The James Hutton Institute

- *Dickeya* spp. have a wide host range, and a broad set of phenotypes. Can we infer organism-level properties from sequenced genomes, and gain insight into pathogenicity and host range?

Mapping Genome Features to Metabolism

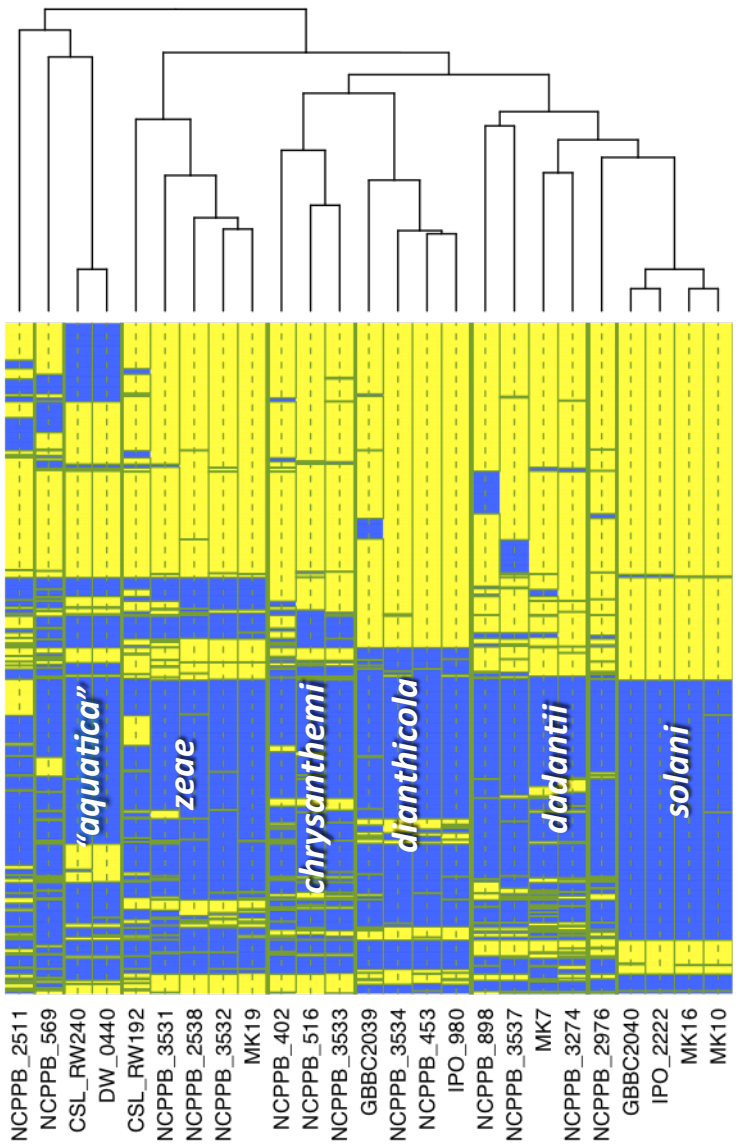
The James Hutton Institute

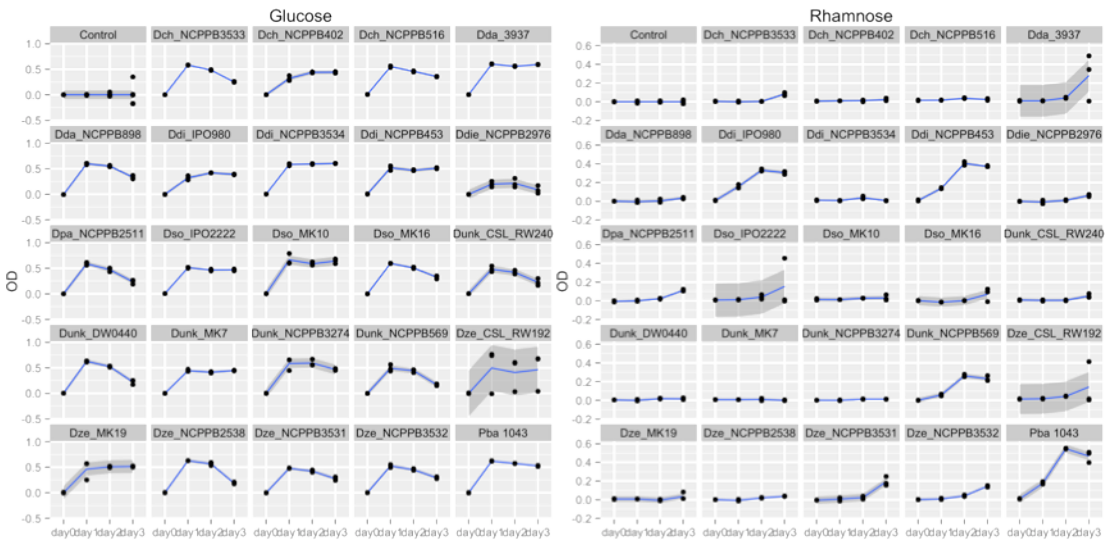
- Many annotation tools associate genome features with metabolic roles (but not all metabolic roles have known genes…):
    - BioCyc/EcoCyc/PathwayTools
      (http://biocyc.org/)
    - KEGG/KAAS
      (http://www.genome.jp/kegg/kaas/)
    - CanOE
      (http://www.genoscope.cns.fr/agc/microscope/metabolism/canoe.php)
    - UniPathway
      (http://www.unipathway.org/)

(this, again, is sequence classification)

- Having genome sequence also makes finding regulatory regions easier - potential for reconstructing regulatory networks

Reactions differ by species



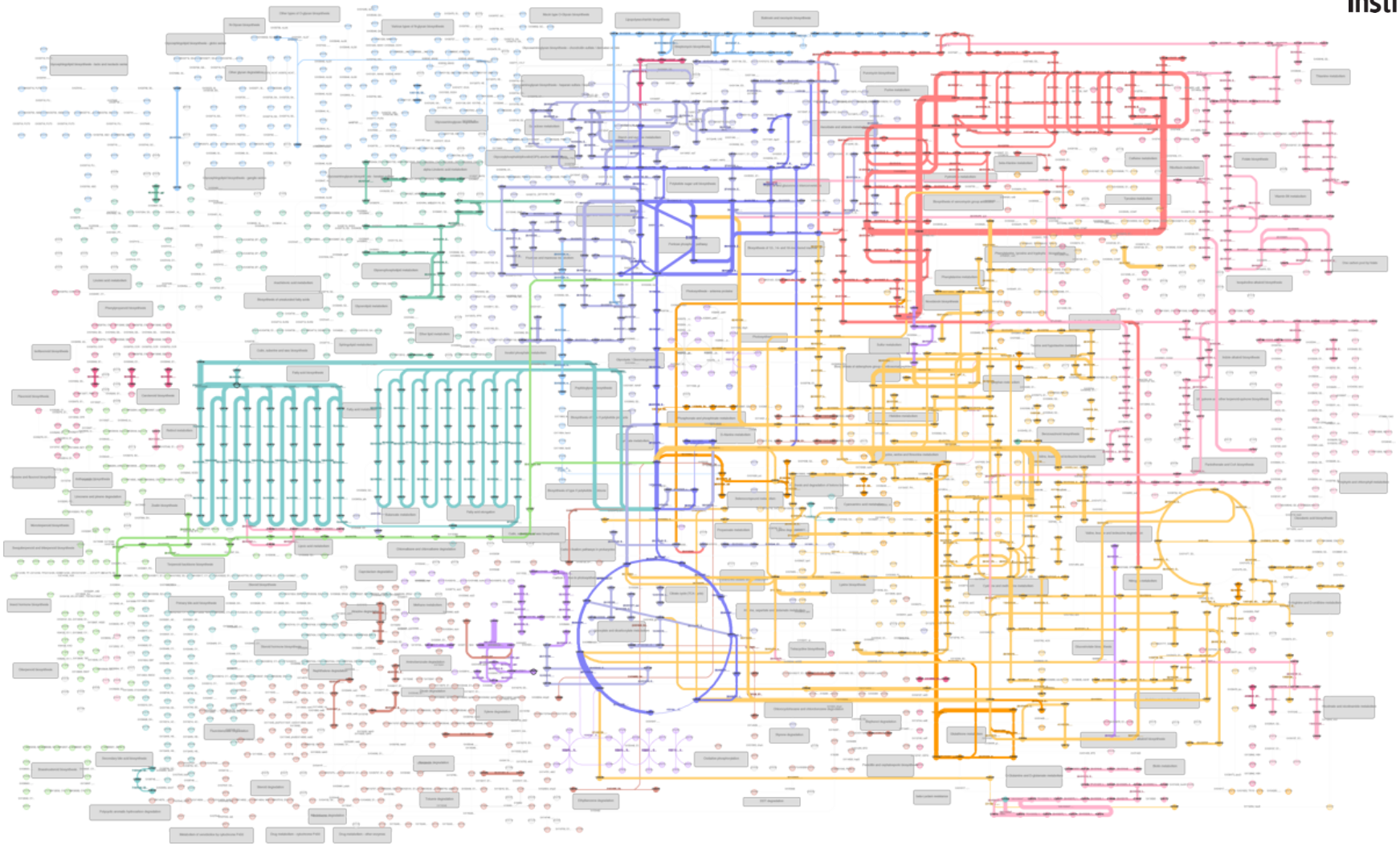


- Presence/absence of reactions associates with species
- Growth curves/substrate preference associates with species

http://dx.doi.org/10.6084/m9.figshare.767275

Presence/absence of KEGG reactions

The James Hutton Institute



https://github.com/widdowquinn/KGML
http://dx.doi.org/10.6084/m9.figshare.767275

Flux Balance Analysis models

- FBA models represent whole-organism metabolic flux

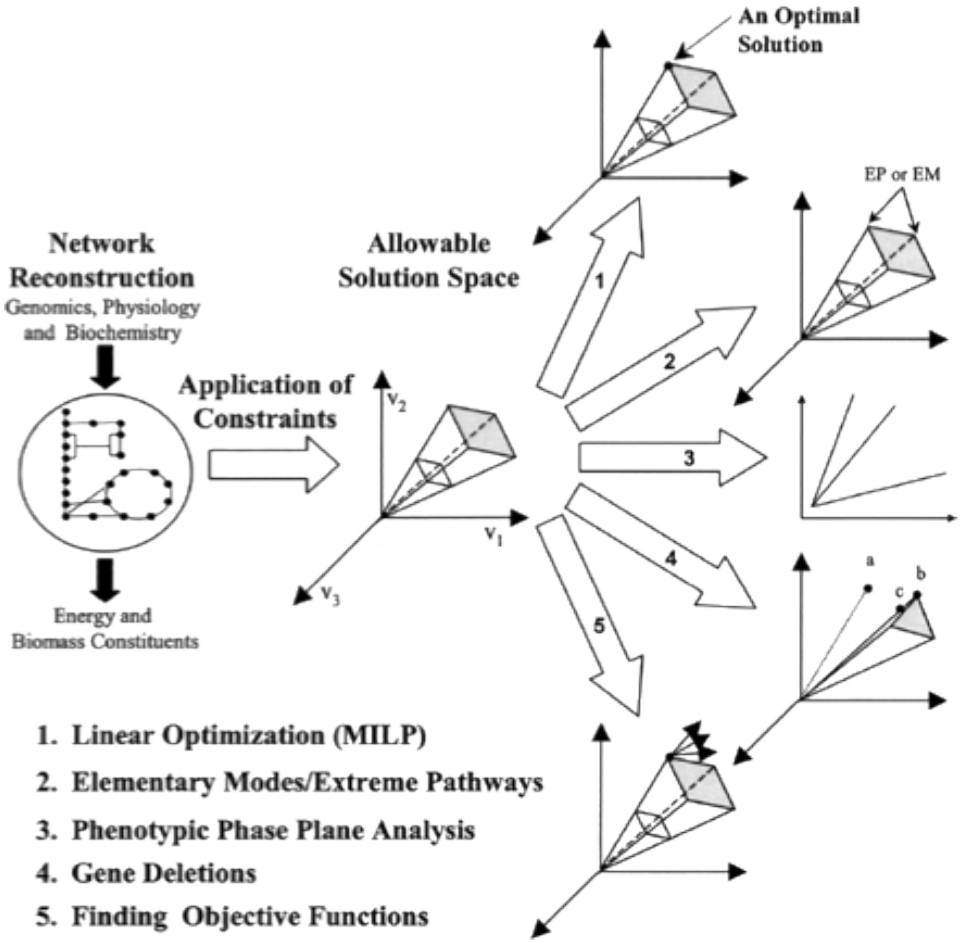
Elementary modes

Substrate usage

Predictive for knockouts

Basis for SynthBio & engineering

- FBA models constructed for each sequenced *Dickeya* isolate
- Transposon mutant libraries being constructed for representative isolates
- Regulatory network reconstruction in progress



Conclusions III

- Microbial pathogen genomes are much easier to obtain, but (at the moment) typically less complete than they used to be
- **Automated annotations (and databases…) have errors – be careful and critical**
- **Imperfect sequences and annotations are still amazingly useful**
- Whole genome information allows for more precise taxonomic classification and diagnostics
- Genome analyses provide a springboard for understanding (and engineering) the biology of pathogens

Acknowledgements

- **JHI *Dickeya***
  - Ian Toth
  - Sonia Humphris
  - Emma Campbell
  - Ashley Boath
  - Anne-Laure Lucquet
- **Edinburgh *Dickeya***
  - Ian Simpson
- **Oxford *Dickeya***
  - Gail Preston
  - Souvik Kusari
- **Fera *Dickeya***
  - Neil Parkinson
  - John Elphinstone
- **SASA Dickeya**
  - Gerry Saddler
  - Vince Mulholland
- **ILVO *Dickeya***
  - Steve Baeyen
  - Johan Van Vaerenbergh
  - Martine Maes
- **Wageningen *Dickeya***
  - Jan van der Wolf
- **JHI Bioinformatics**
  - Peter Cock
  - Sue Jones
  - Linda Milne
  - Iain Milne
- **JHI *E. coli***
  - Nicola Holden
  - Ian Toth
  - Ashleigh Holmes
- **Muenster *E. coli***
  - Helge Karch
  - Martina Bielaszewska
- **TSL Potato**
  - Dan MacLean
  - Graham Etherington
- **Sci/Art**
  - Ian Toth
  - Elaine Shemilt
  - Genevieve Murphy
  - Danny Hill
  - Michel Perombelon
- **Aberdeen *Campylobacter***
  - Ken Forbes
  - Norval Strachan
- **JHI *Phytophthora***
  - Paul Birch
  - Steve Whisson
  - David Cooke
  - Miles Armstrong
  - Eleanor Gilroy
  - Hazel Mclellan
  - Julie Squires
  - Susan Breen
- **Warwick *Phytophthora***
  - Jim Beynon
  - Laura Baxter
  - Rachel Clewes
  - Jens Steinbrenner
  - Daniel Tome
- **JHI Potato**
  - Ingo Hein
  - Florian Jupe
- **TSL Potato**
  - Dan MacLean
  - Graham Etherington
- **JHI *Pectobacterium***
  - Ian Toth
  - Paul Birch
  - Kenny Bell
  - Beth Hyman
  - Maria Holeva
- **Sanger *Pectobacterium***
  - Julian Parkhill
  - Mo Sebaihia
  - Matt Holden
  - Steve Bentley
  - Nick Thomson

AND MANY MORE…

How to get presentation materials

- All slides and example code/data/exercises at GitHub: https://github.com/widdowquinn/Teaching-EMBL-Plant-Path-Genomics
- Use Git application or command line:

```
git clone https://github.com/widdowquinn/Teaching-EMBL-Plant-Path-Genomics.git
```

- Also in the course materials directory