# Pathogen Genome Data

## EMBL-EBI Bioinformatics of Plants and Plant Pathogens 23$^{rd}$ May 2016

The James Hutton Institute

Leighton Pritchard[1,2,3]
[1]Information and Computational Sciences,
[2]Centre for Human and Animal Pathogens in the Environment,
[3]Dundee Effector Consortium,
The James Hutton Institute, Invergowrie, Dundee, Scotland, DD2 5DA

# Acceptable Use Policy

The James Hutton Institute

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

The James Hutton Institute


1 **Introduction**
- Pathogen Genome Data

2 **Public Data Sources**
- Online Resources

3 **Comparative Genomics**
- Why Comparative Genomics?

4 **Genome Comparisons**
- Whole Genome Comparisons

# Introduction

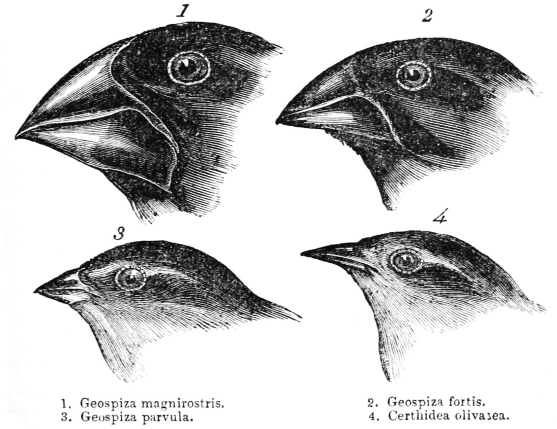
The James Hutton Institute

## 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

The James Hutton Institute


1 Introduction
- Pathogen Genome Data

2 Public Data Sources
- Online Resources

3 Comparative Genomics
- Why Comparative Genomics?

4 Genome Comparisons
- Whole Genome Comparisons

# NCBI

NCBI

The James Hutton Institute

## 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 | - |
| GCF_000197855.1_ASM1../ | 03-Sep-2015 06:48 | - |
| GCF_000212635.3_ASM2../ | 17-May-2016 19:06 | - |
| GCF_000215325.1_ASM2../ | 17-May-2016 21:15 | - |

**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 | - |

# GenBank vs RefSeq

The James Hutton Institute

## 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

# Ensembl

## 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 resource**
  - 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
  - Ensembl Bacteria: Broad Institute - now retiring their online resources
- **Specialist databases, e.g.**
  - 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

The James Hutton Institute

**`worksheets/01-downloading_data_biopython.ipynb`**

Downloading genome data from NCBI with Biopython

- MyBinder link

# Table of Contents

The James Hutton Institute


1 Introduction
- Pathogen Genome Data

2 Public Data Sources
- Online Resources

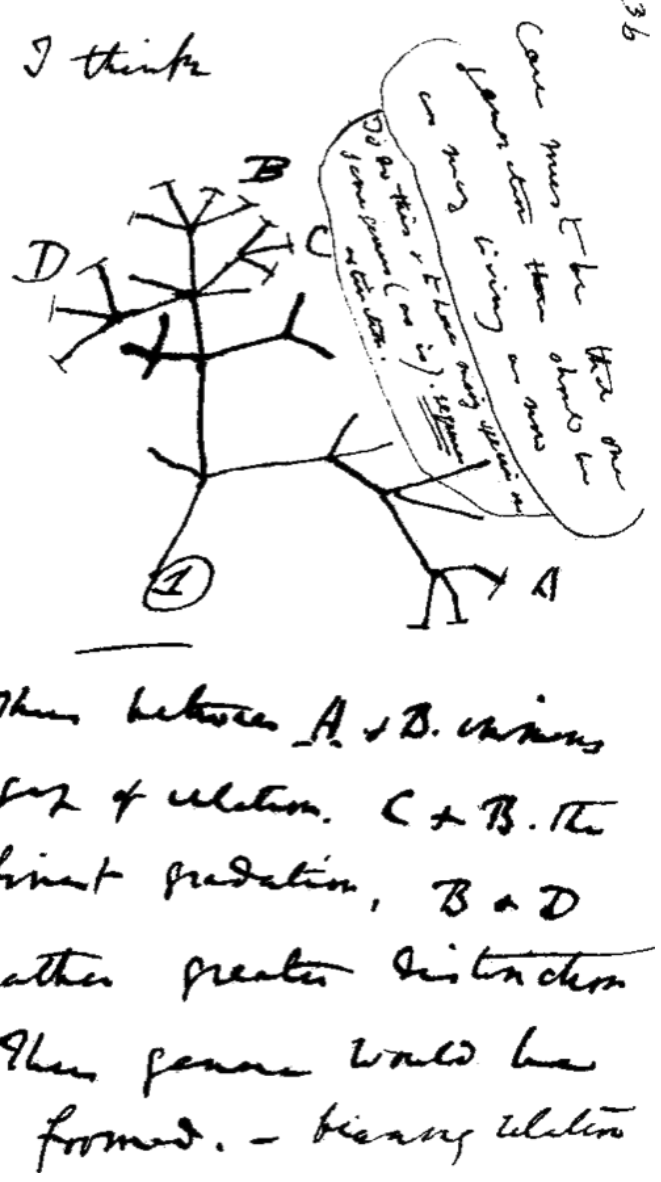
3 Comparative Genomics
- Why Comparative Genomics?

4 Genome Comparisons
- Whole Genome Comparisons

# Why comparative genomics?

The James Hutton Institute

- 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



# Genomes aren't everything...

The James Hutton Institute

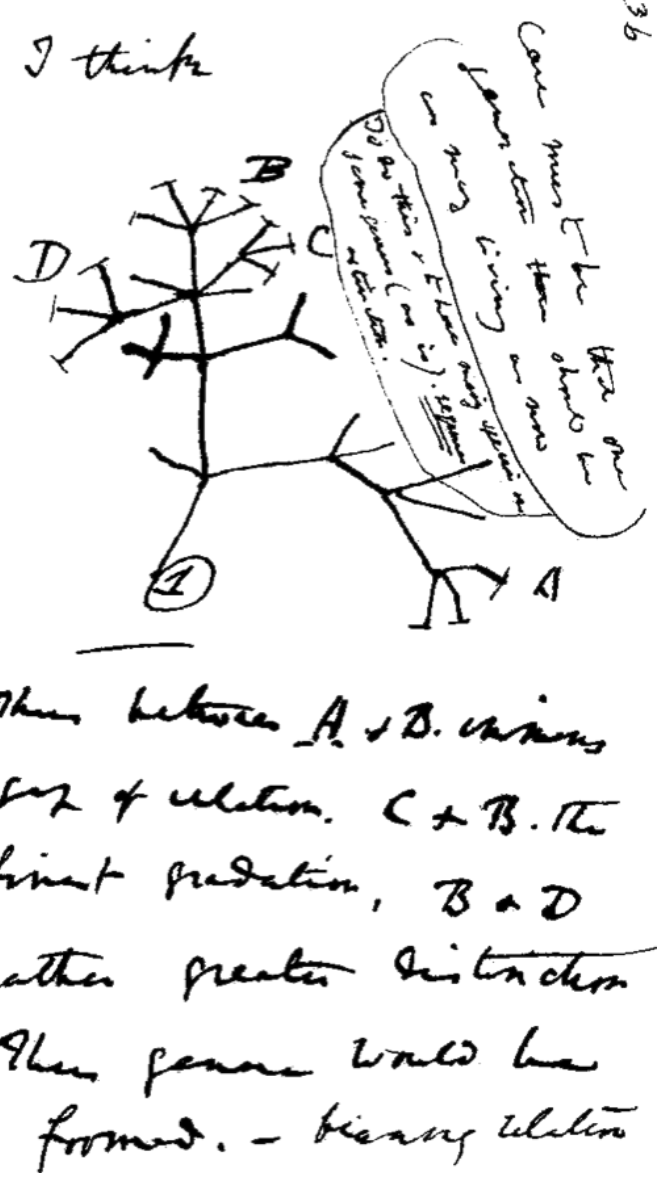
## Context

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

## Phenotypic plasticity, responses to

- temperature
- stress
- community, etc.

...and therefore systems biology...



# Levels of comparison

The James Hutton Institute

## Bulk Properties

e.g. *k*-mer profiles (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 Data Sources**
- Online Resources

3 **Comparative Genomics**
- Why Comparative Genomics?

4 **Genome Comparisons**
- Whole Genome Comparisons

# Whole genome comparisons

The James Hutton Institute

**Whole genome comparison**

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

Minimum requirement: **two genomes**

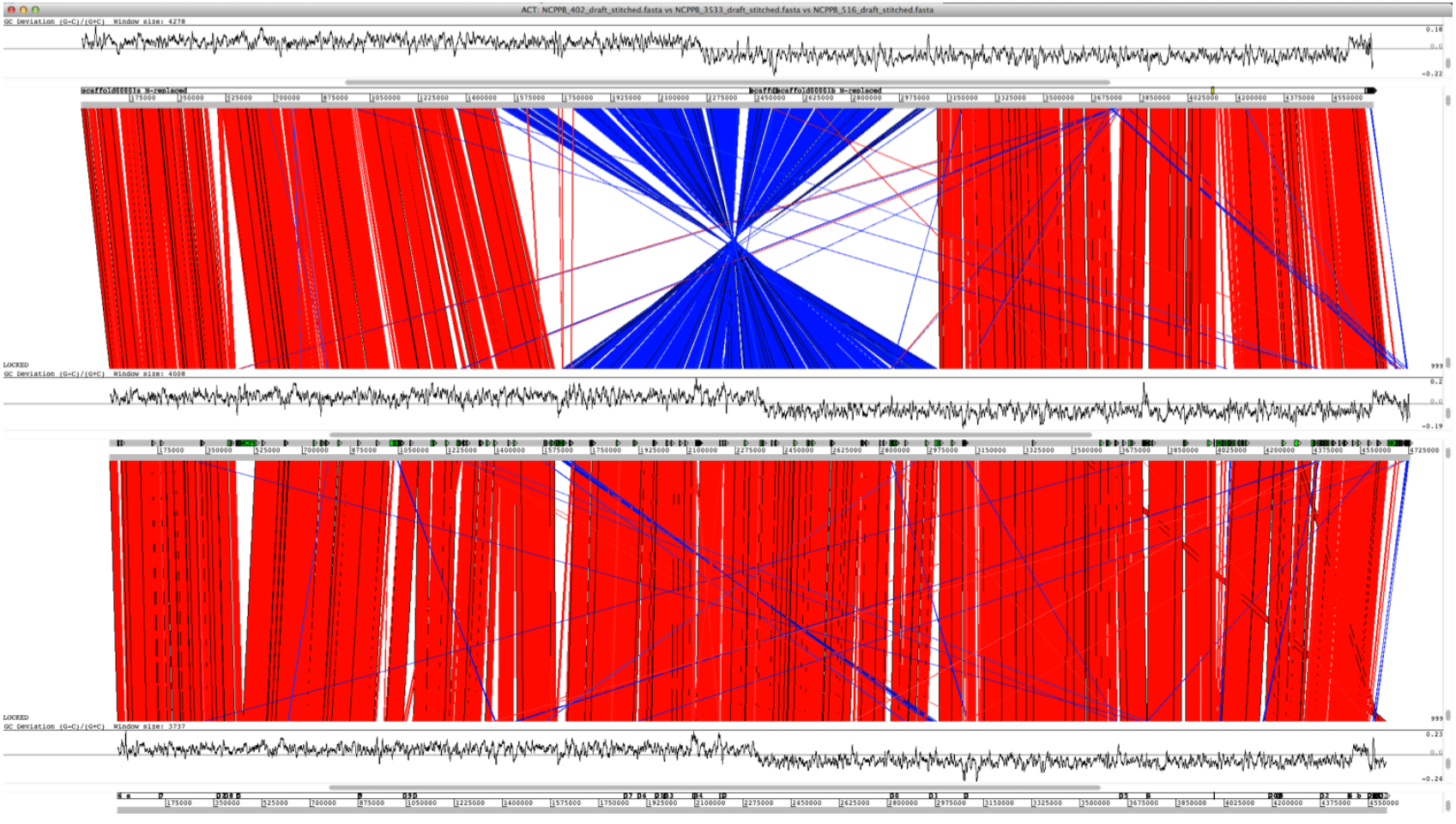
- Reference Genome
- Comparator Genome

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

# Pairwise genome alignments

The James Hutton Institute

Pairwise comparisons produce alignments of similar regions.

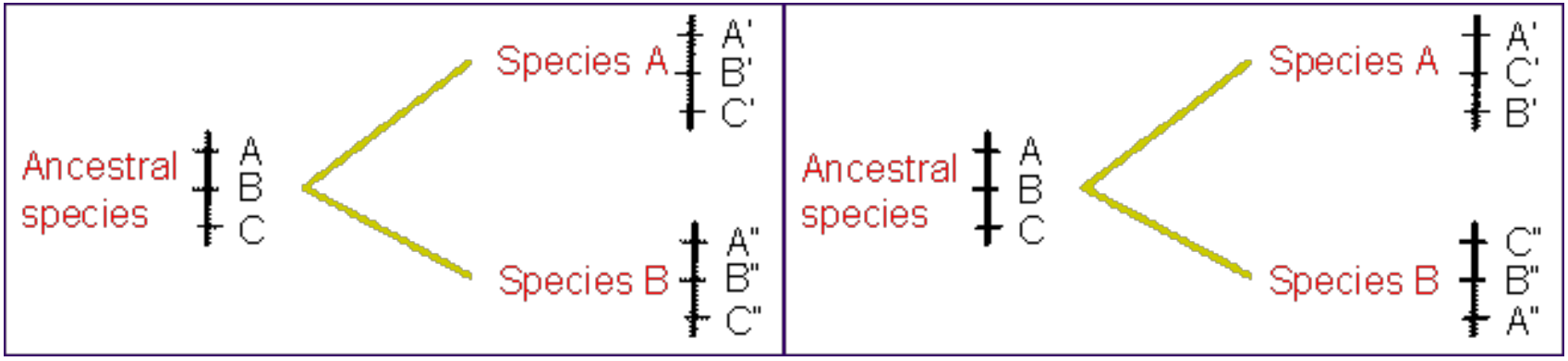


# Synteny and Collinearity

The James Hutton Institute

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

The James Hutton Institute

Chromosome C-II: environmental adaptation; C-I: virulence genes.
C-II has undergone extensive rearrangement; C-I has not.

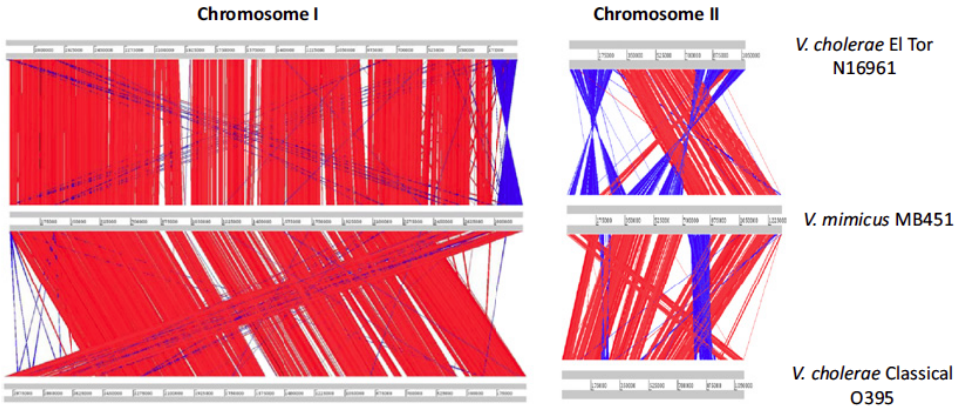



**Fig. 2.** Linear pairwise comparison of the *Vibrio mimicus* genome by Artemis Comparison Toll. Regions with similarity are highlighted by connecting red or blue lines between the genomes; red lines indicate homologous blocks of sequence, and blue lines indicate inversions. Gaps indicate unique DNA.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

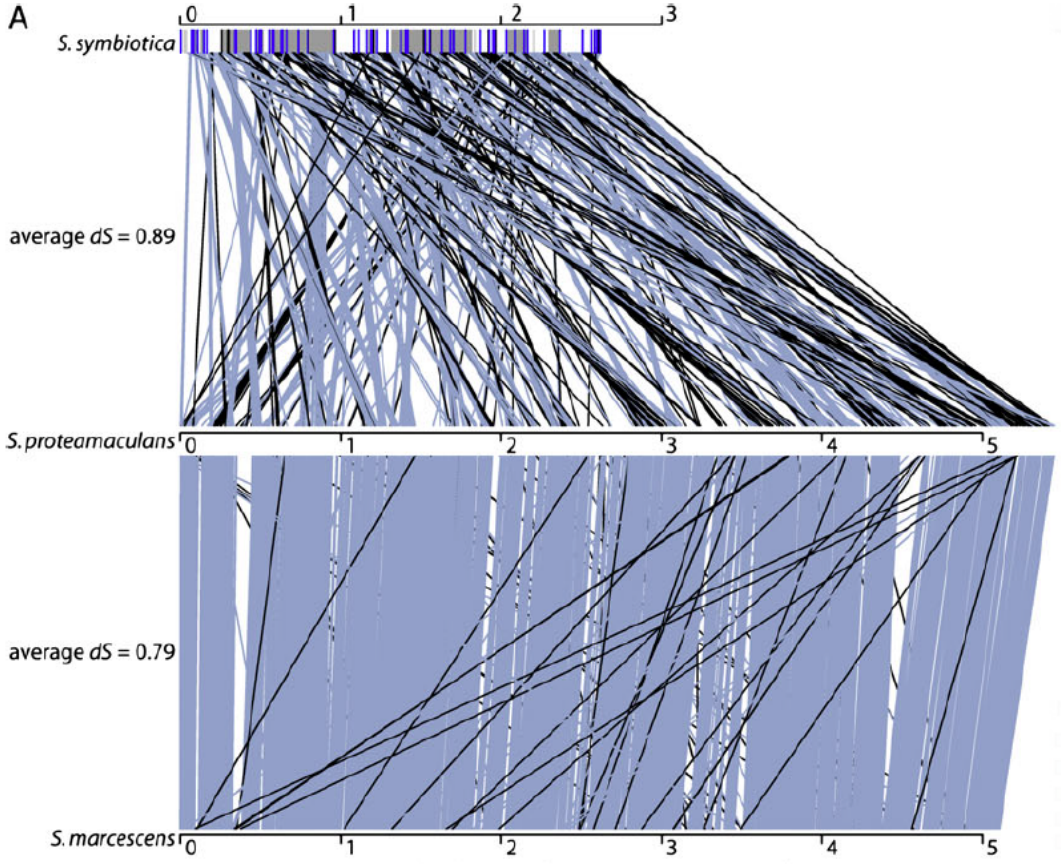
Suggests modularity of genome organisation, as a mechanism for adaptation (HGT, two-speed genome).

# Serratia symbiotica [a]

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

*S. symbiotica* is a recently evolved symbiont of aphids

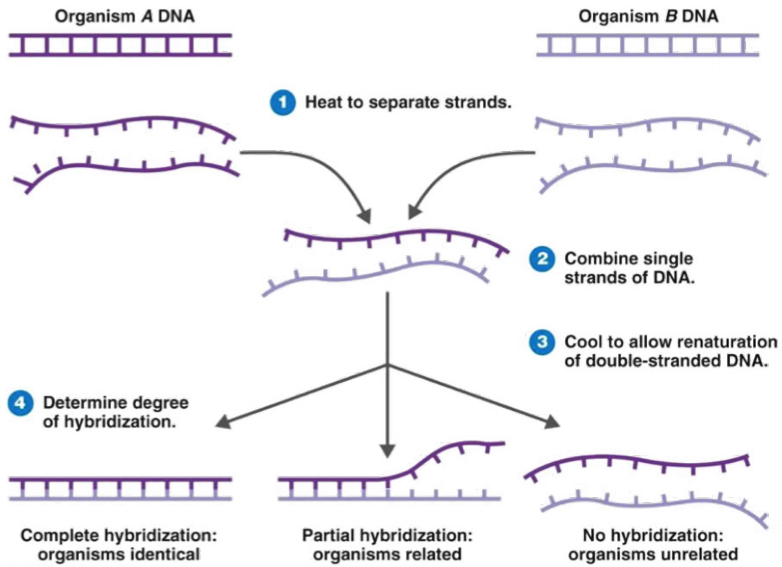
Massive genomic decay: consequence of adaptation



# 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 $\Delta T$ 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

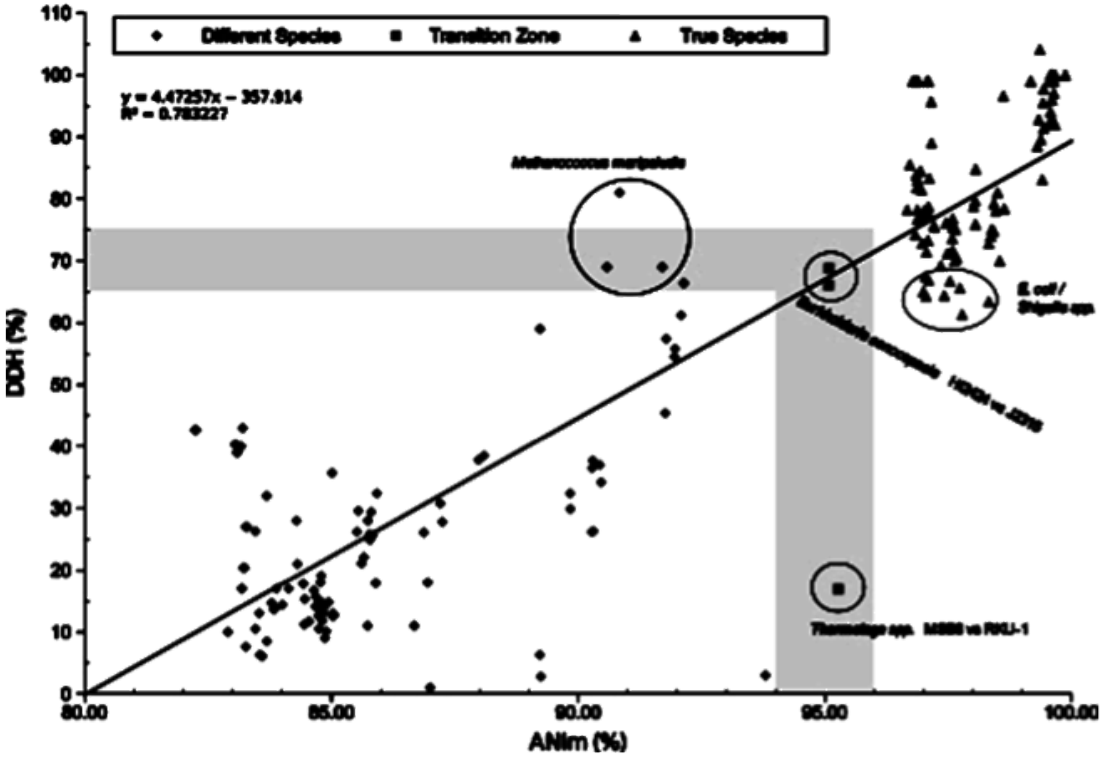
# Average Nucleotide Identity (ANIm)[a]

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

The James Hutton Institute

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

- DDH:ANIm linear
- 70%ID ≈ 95%ANIb



# ANI

The James Hutton Institute

## Advantages

- Average identity of all 'homologous' regions
- Approximates limiting case of MLST/MLSA/multigene comparisons
- Classification not dependent on dataset composition (unlike tree methods)

## Criticisms

- 95% threshold 'arbitrary'
- Taxonomic classification, not phylogenetic reconstruction
- No functional (or gene-based) interpretation; still need pangenome classification and analysis

# EXERCISE

The James Hutton Institute

`exercises/01-whole_genome_comparisons.ipynb`

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

- MyBinder link

# Licence: CC-BY-SA

The James Hutton Institute

By: Leighton Pritchard