Comparative Genomics and Visualisation BS32010

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-Dundee-BS32010

Table of Contents

The James Hutton Institute

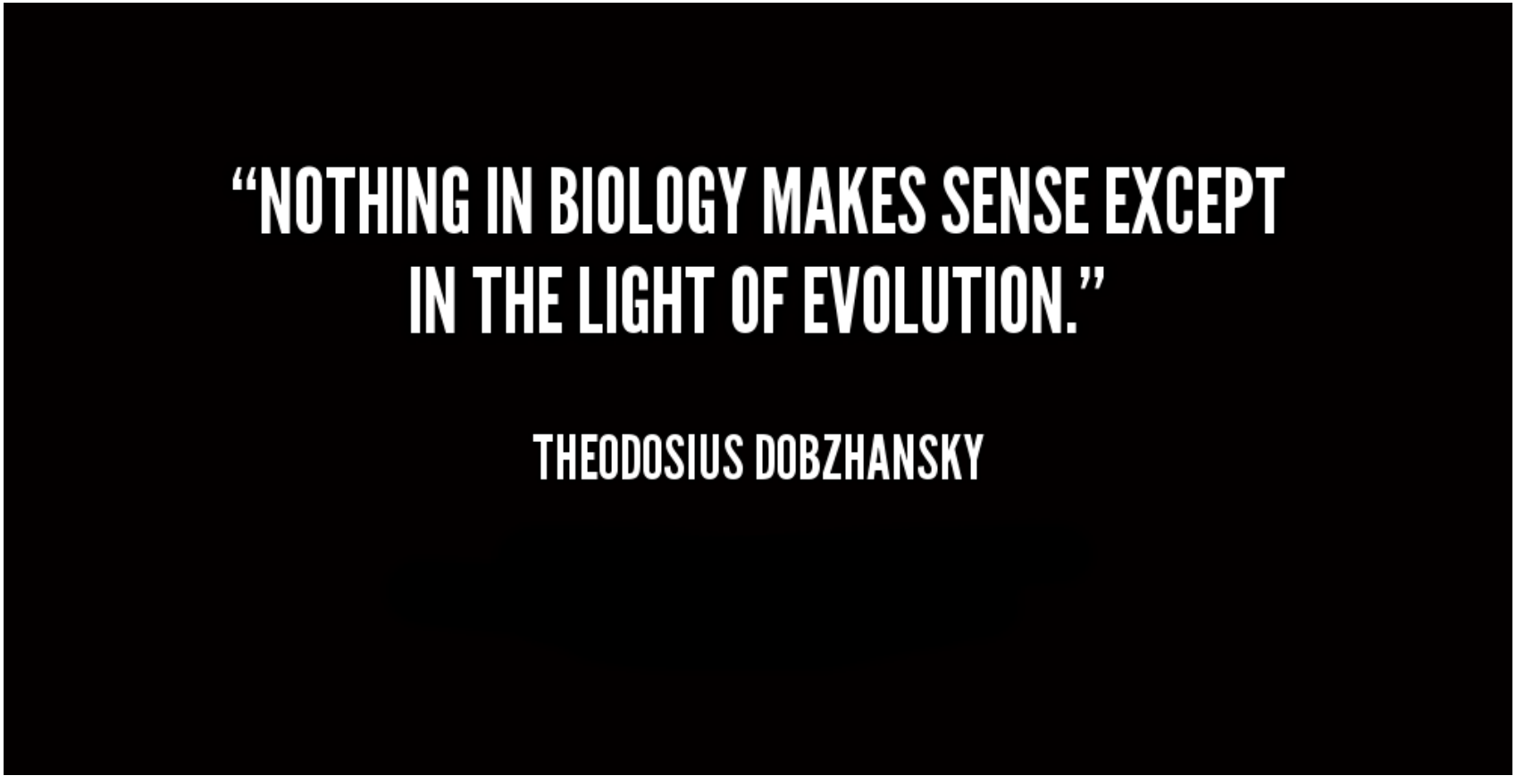
What Is Comparative Genomics?

The James Hutton Institute

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

Evolution is the central concept

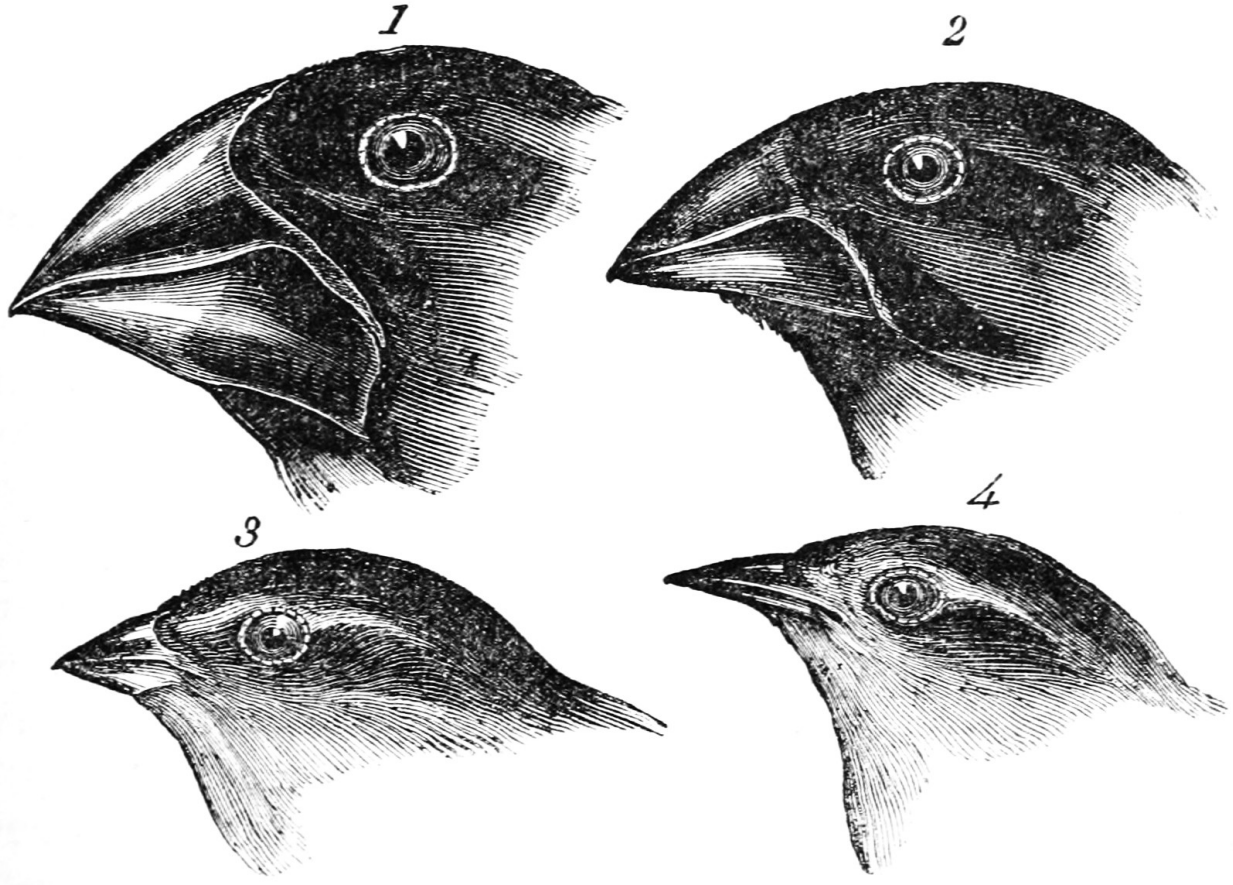
The James Hutton Institute



Comparison of physical features

The James Hutton Institute

How do we determine that features are related, and evolved?

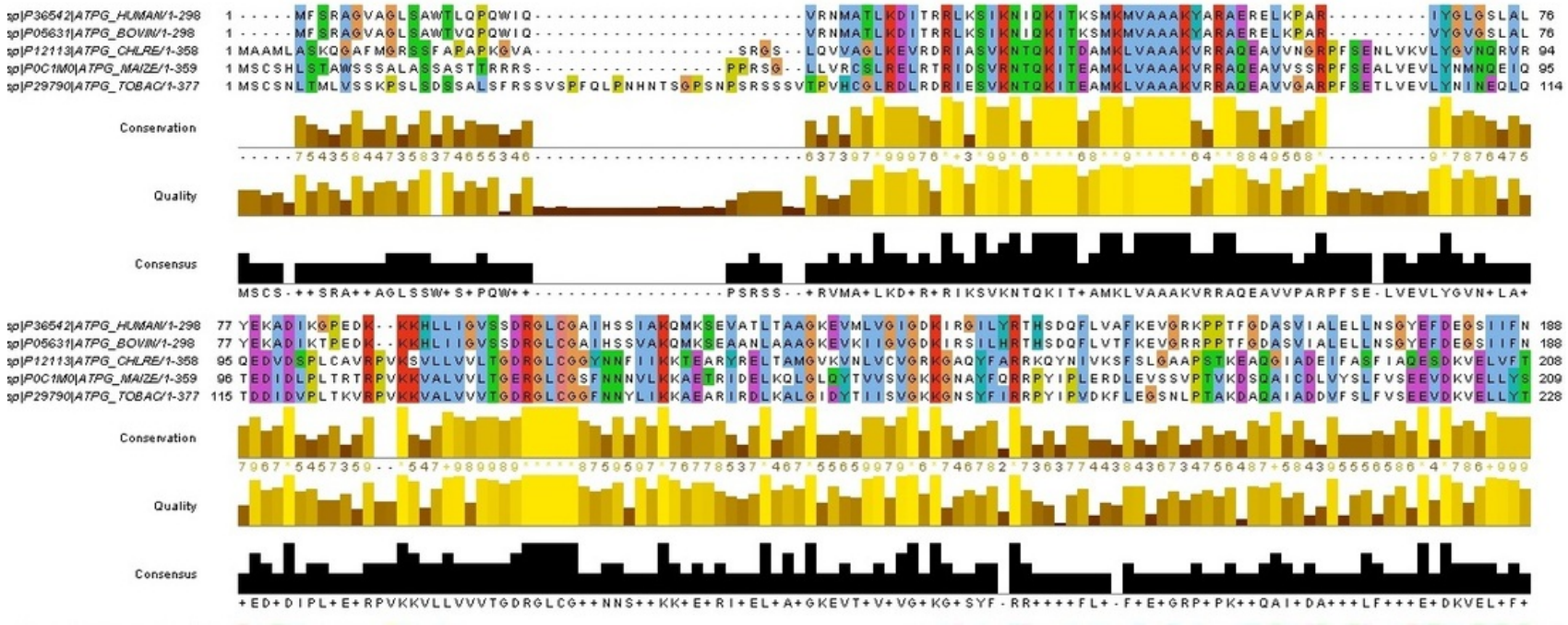


1. Geospiza magnirostris.
2. Geospiza fortis.
3. Geospiza parvula.
4. Certhidea olivasea.

Comparison of sequence features

The James Hutton Institute

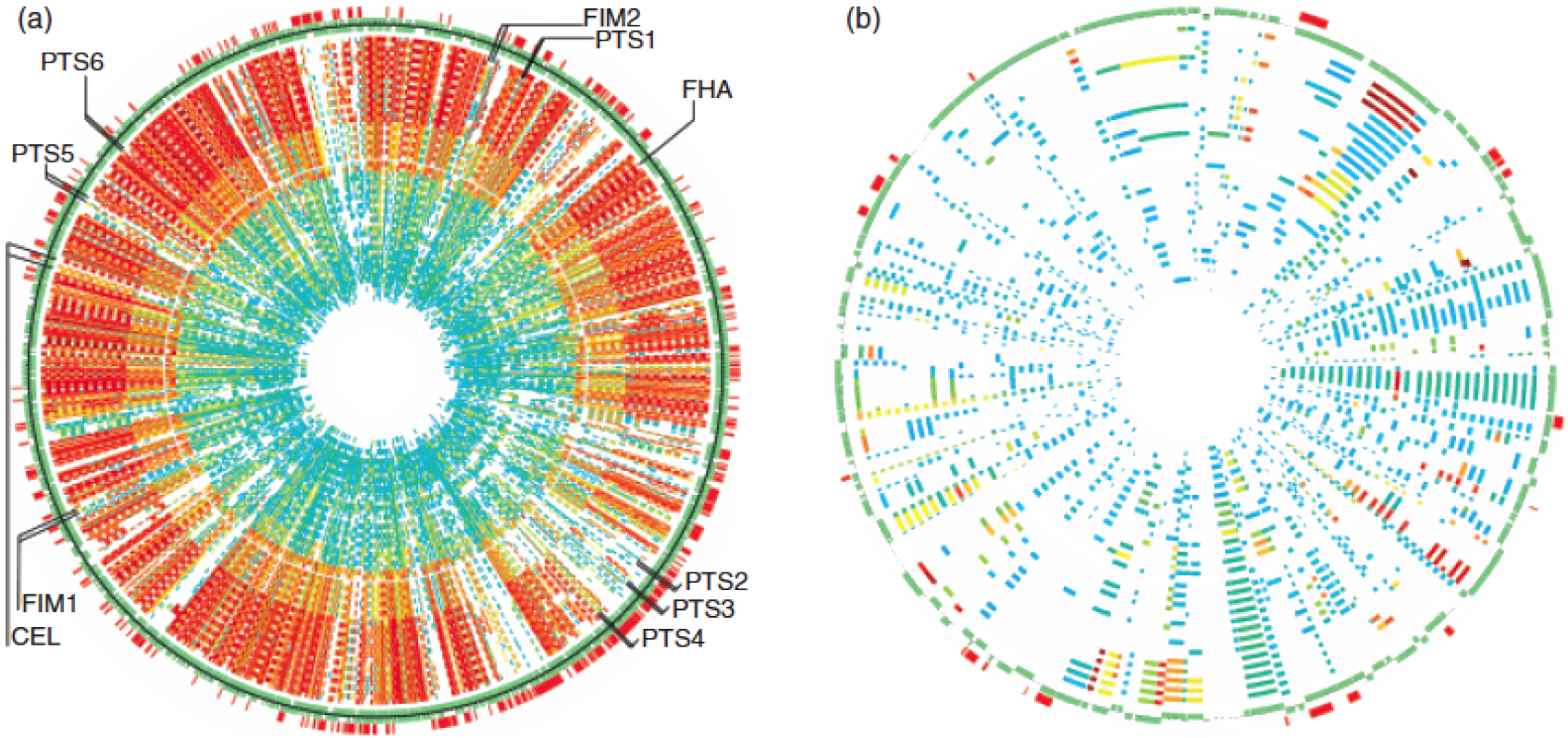
Multiple sequence alignment of ATP synthase



Comparison of genome features

The James Hutton Institute

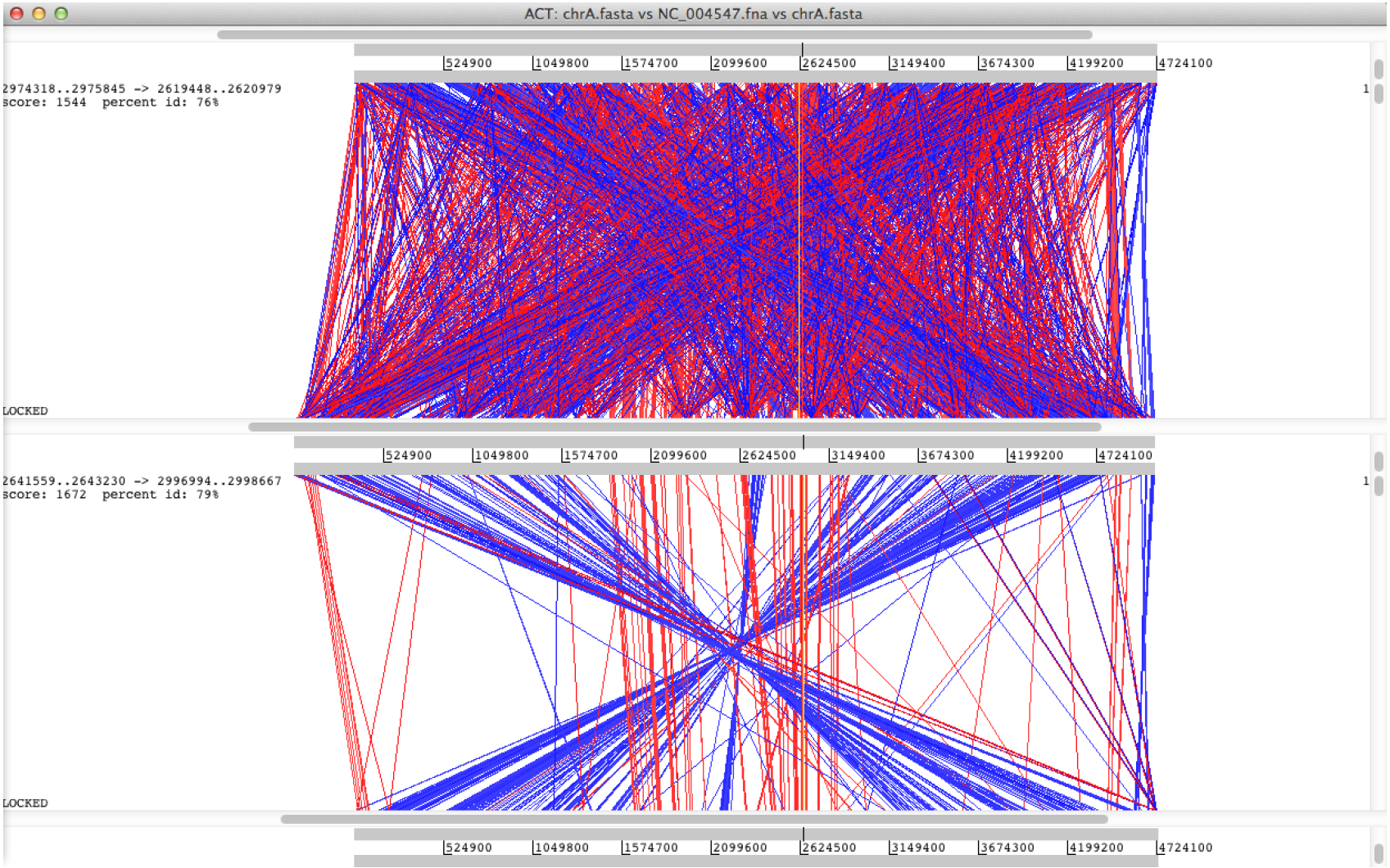
Similarity of individual features (feature sequence)



Comparison of genome features

The James Hutton Institute

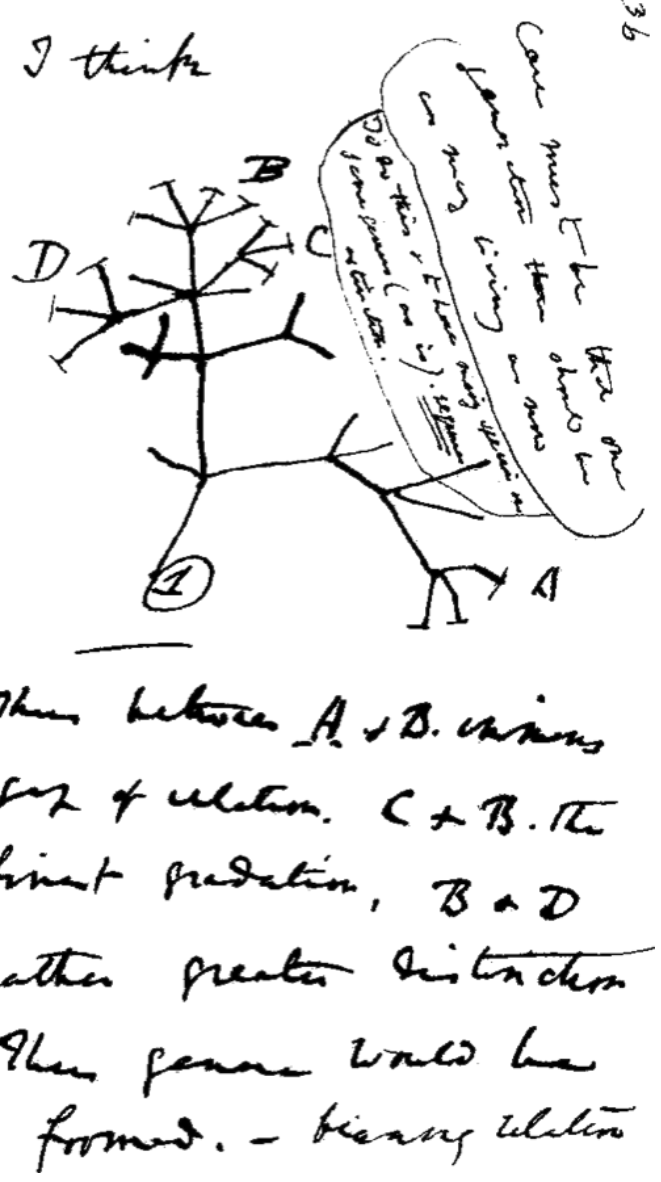
Similarity of individual features (ordering and arrangement)



Why comparative genomics?

The James Hutton Institute

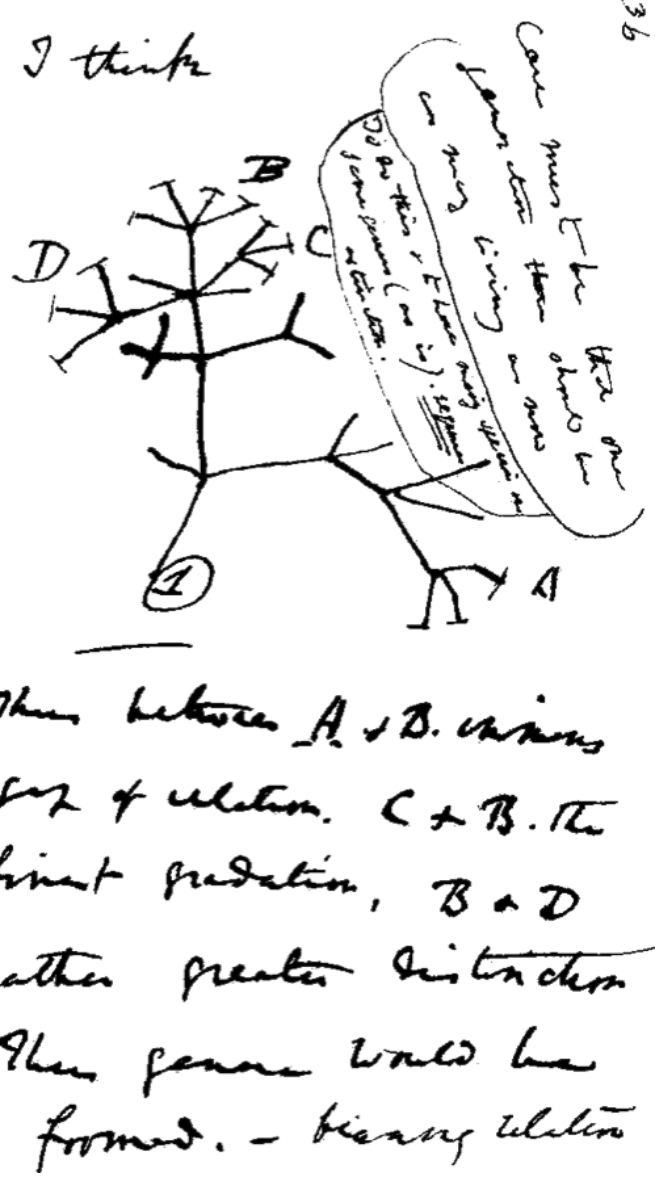
- Genome features are heritable characteristics
- Related organisms share ancestral genomes
- Related organisms inherit common genome features
- Genome similarity $\propto$ relatedness? (phylogenomics)



Why comparative genomics?

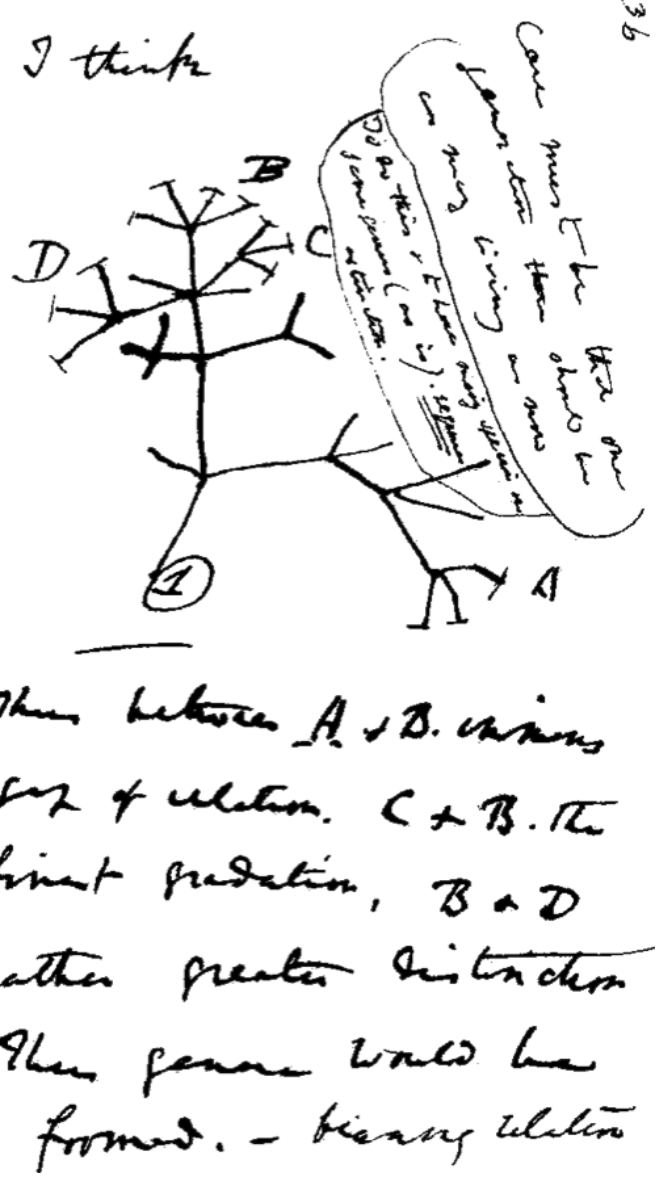
The James Hutton Institute

- Genomes carry functional elements under selection pressure
- Related organisms carry similar functional elements
- Deleterious functional elements are lost through selection
- Genome similarity $\propto$ phenotype? (functional genomics)



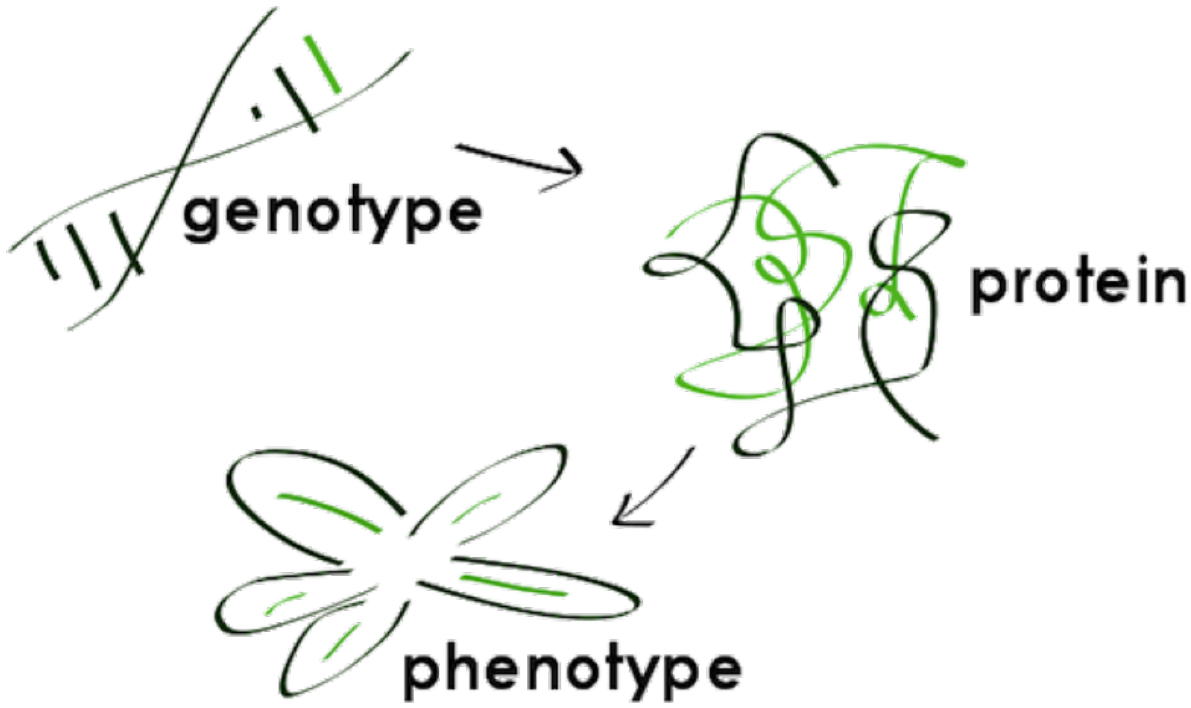
Why comparative genomics?

- Functional elements are selected over many generations
- Optimised functional elements may not change greatly
- (Functional elements can be transferred non-heritably)
- Genome feature similarity $\Longrightarrow$ common function? (genome annotation)
- Transfer functional information from model systems (*E. coli*, *A. thaliana*, *D. melanogaster*) to non-model systems



Genomes are informative, but...

BIOLOGICAL CONTEXT: epigenetics, tissue differentiation, mesoscale systems, etc.



PHENOTYPIC PLASTICITY: responses to temperature, stress, environment, etc.

Genomes to systems

The James Hutton Institute

Functional Genomics

- Genome differences underpin phenotype (morphological, physiological) differences.
- If phenotypes known, genome comparisons can indicate genomic origin of differences (e.g. GWAS).
- Genome comparisons can reveal evolutionary processes and constraints, and suggest phenotypic outcomes.

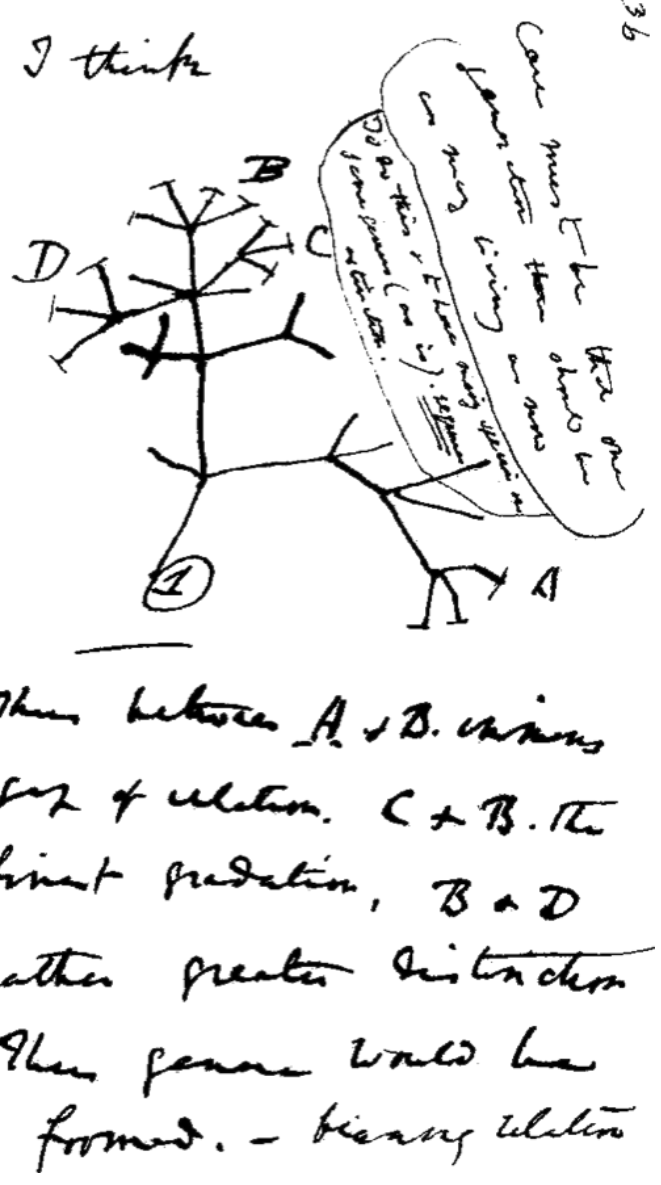


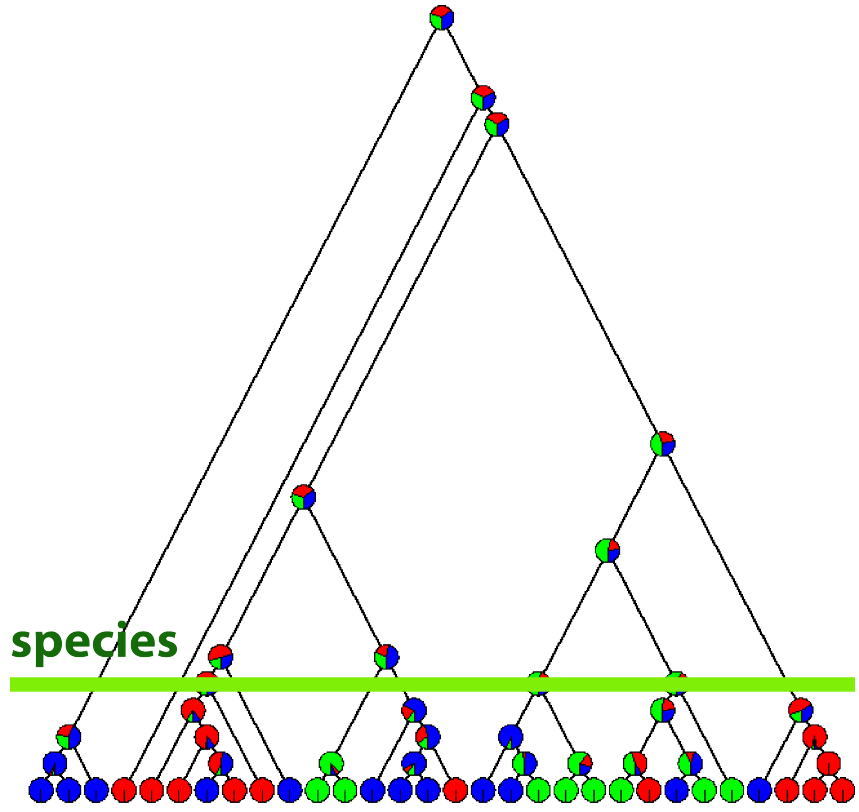
Table of Contents

Types of comparison

The James Hutton Institute

Within species

- e.g. between isolates/individuals (or between tissues...)
- Which genome features may account for unique characteristics of organisms or cell-types (e.g. tumours)?
- what epigenetic changes occur in an individual?

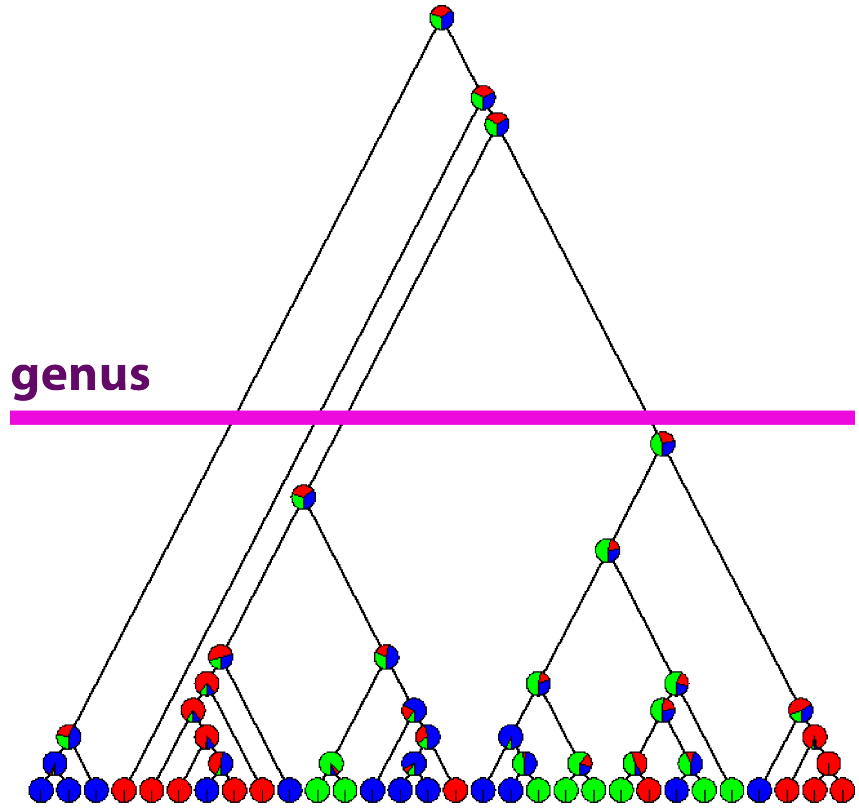


Types of comparison

The James Hutton Institute

Within genera/between species

- comparison between groups of individuals
- what genome features show evidence of selective pressure?
- which features/changes are associated with species phenotype?

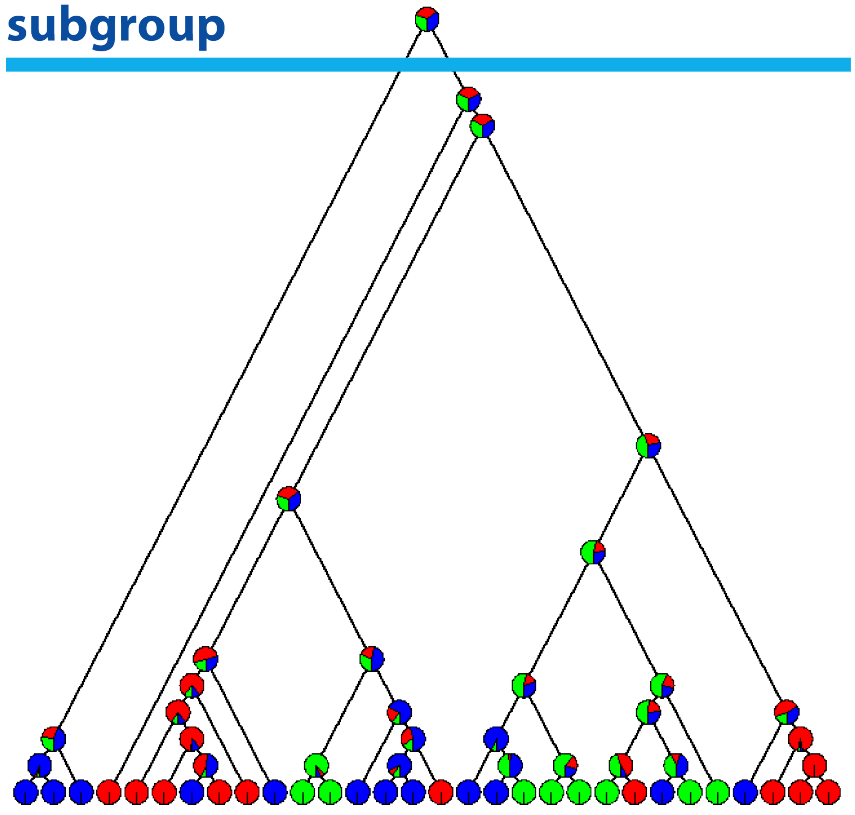


Types of comparison

The James Hutton Institute

Between subgroups

- e.g. comparisons across many diverse individuals
- what are the *core set* of genome features that define a subgroup or genus?
- what functions are present/absent between groups?



E. coli LTEE [a] [b] [c]

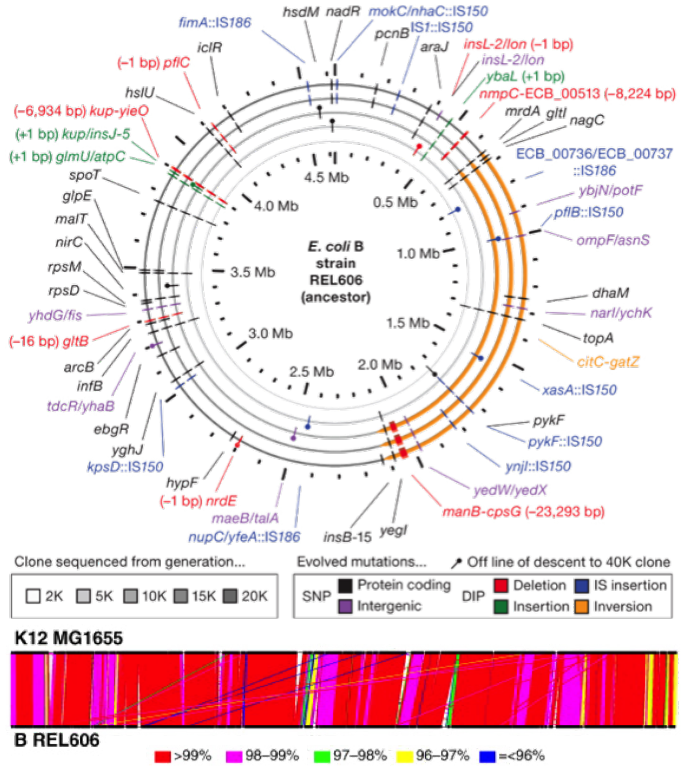
[a] Jeong *et al.* (2009) *J. Mol. Biol.* doi:10.1016/j.jmb.2009.09.052

[b] Barrick *et al.* (2009) *Nature* doi:10.1038/nature08480

[c] Wiser *et al.* (2013) *Science* doi:10.1126/science.1243357

The James Hutton Institute

- Run by the Lenski lab, Michigan State University since 1988 (http://myxo.css.msu.edu/ecoli/)
- 12 flasks, citrate usage selection
- >50,000 generations of *E coli*!
  - Cultures propagated every day
  - Every 500 generations (75 days), mixed-population samples stored
  - Mean fitness estimated at 500 generation intervals

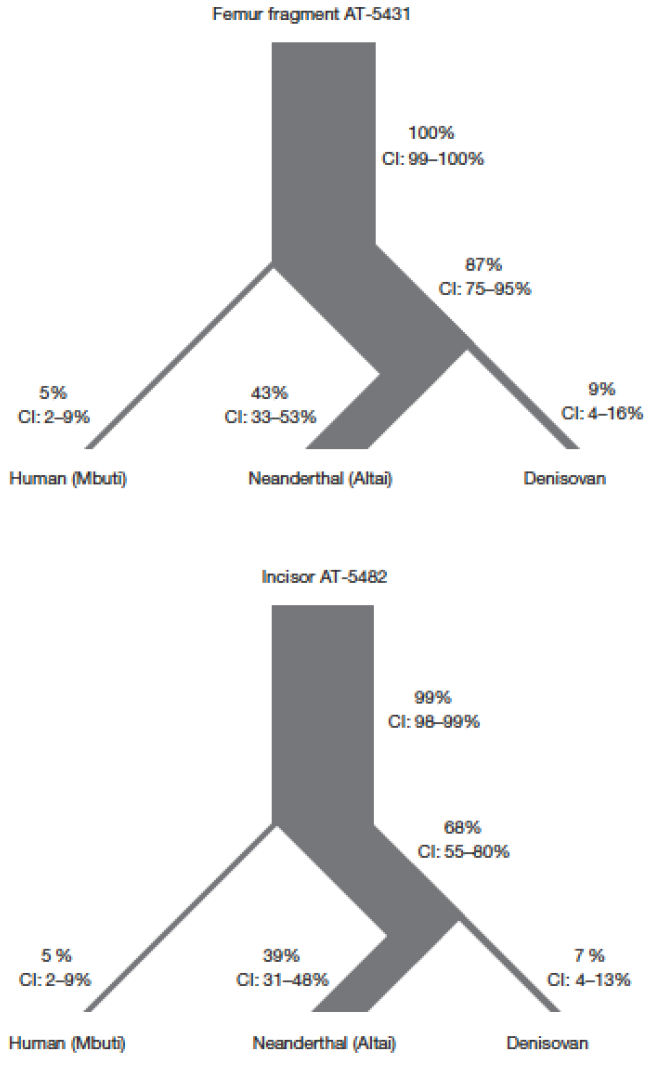


Comparative genomics in the news [a] [b]

[a] BBC News 15/3/2016

[b] Meyer *et al.* (2016) *Nature* doi:10.1038/nature17405

- Oldest DNA ever recovered from a human (430kya) - 0.1% of genome
- 28 individuals, Sima de los Huesos, N. Spain
- mitoDNA more similar to Siberian Denisovans than to modern humans
- Modern humans derived from wave out of Africa 250kya, with mitochondrial turnover?



Question:

The James Hutton Institute

Which of these questions can be answered by comparative genomics?

- A: Are genes in a metabolic pathway always collocated, in species X?
- B: What genes does a transcription factor regulate?
- C: Which genes in a genome are under positive selective pressure?
- D: What genome features can be used to define a species?

Table of Contents

The James Hutton Institute

Levels of comparison

The James Hutton Institute

Bulk Properties

- chromosome/plasmid counts and sizes, nucleotide content, etc.

Whole Genome Sequence

- sequence similarity
- organisation of genomic regions (synteny), 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.

Table of Contents

The James Hutton Institute

Bulk genome comparisons

The James Hutton Institute

You don't have to sequence genomes to compare them
(but it helps)

Genome comparisons predate NGS

The James Hutton Institute

- Sequence data wasn't always cheap and abundant
- Practical, experimental genome comparisons were needed

Bulk genome comparisons

The James Hutton Institute

Calculate values for individual genomes, then compare them.

- Number of chromosomes
- Ploidy
- Chromosome size
- Nucleotide (A,C,G,T) frequency

Nucleotide frequency/genome size

The James Hutton Institute

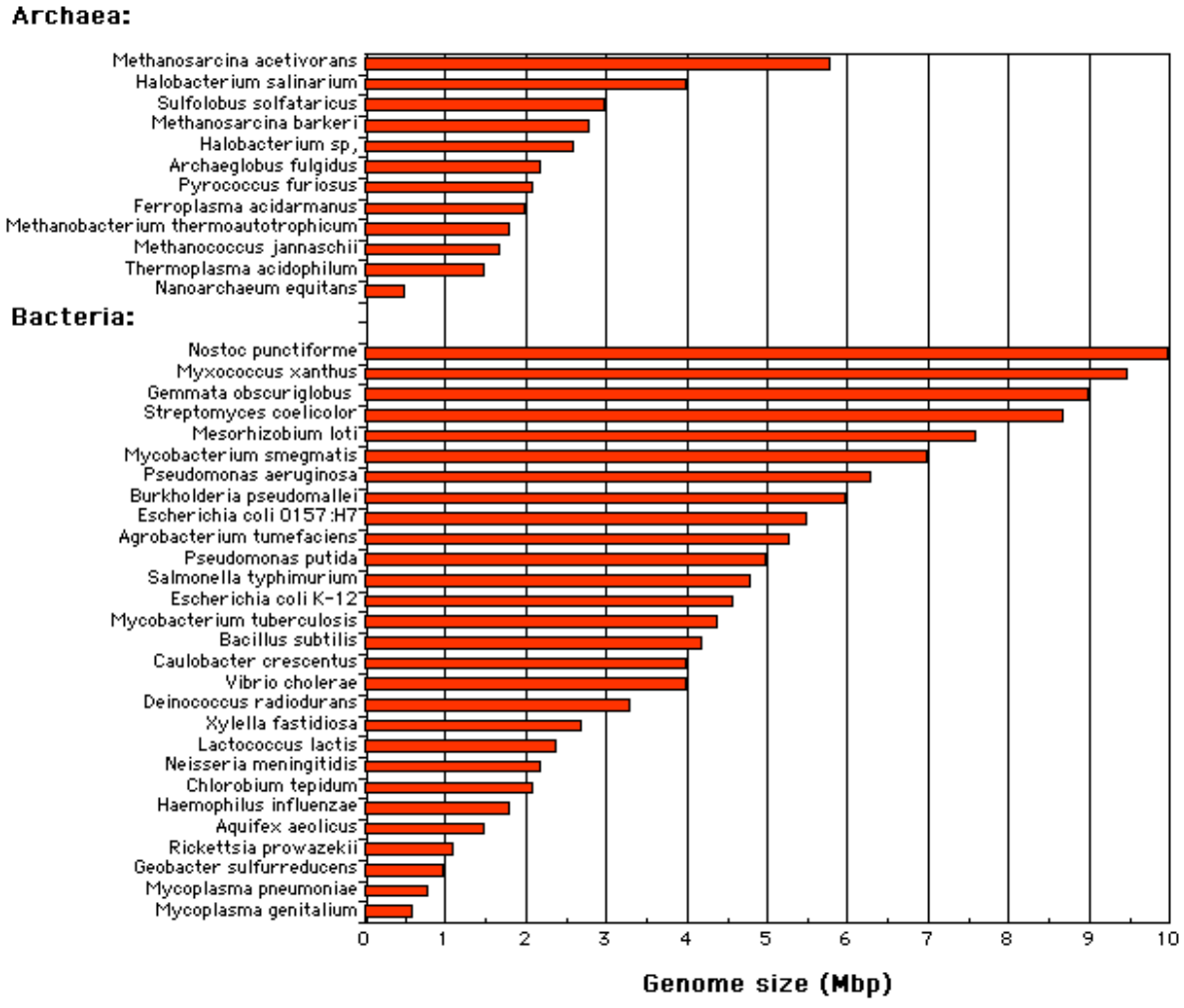
Very easy to calculate from complete or draft genome sequence

```
In [1]: from Bio import SeqIO
In [2]: s = SeqIO.read("data/NC_000912.fna", "fasta")
In [3]: a, c, g, t = s.seq.count("A"), s.seq.count("C"), s.seq.count("G"), s.seq.count("T")
In [4]: float(g + c)/len(s)
Out[4]: 0.40008010837904245
In [5]: float(g - c)/(g+c)
Out[5]: 0.002397259225467894
```

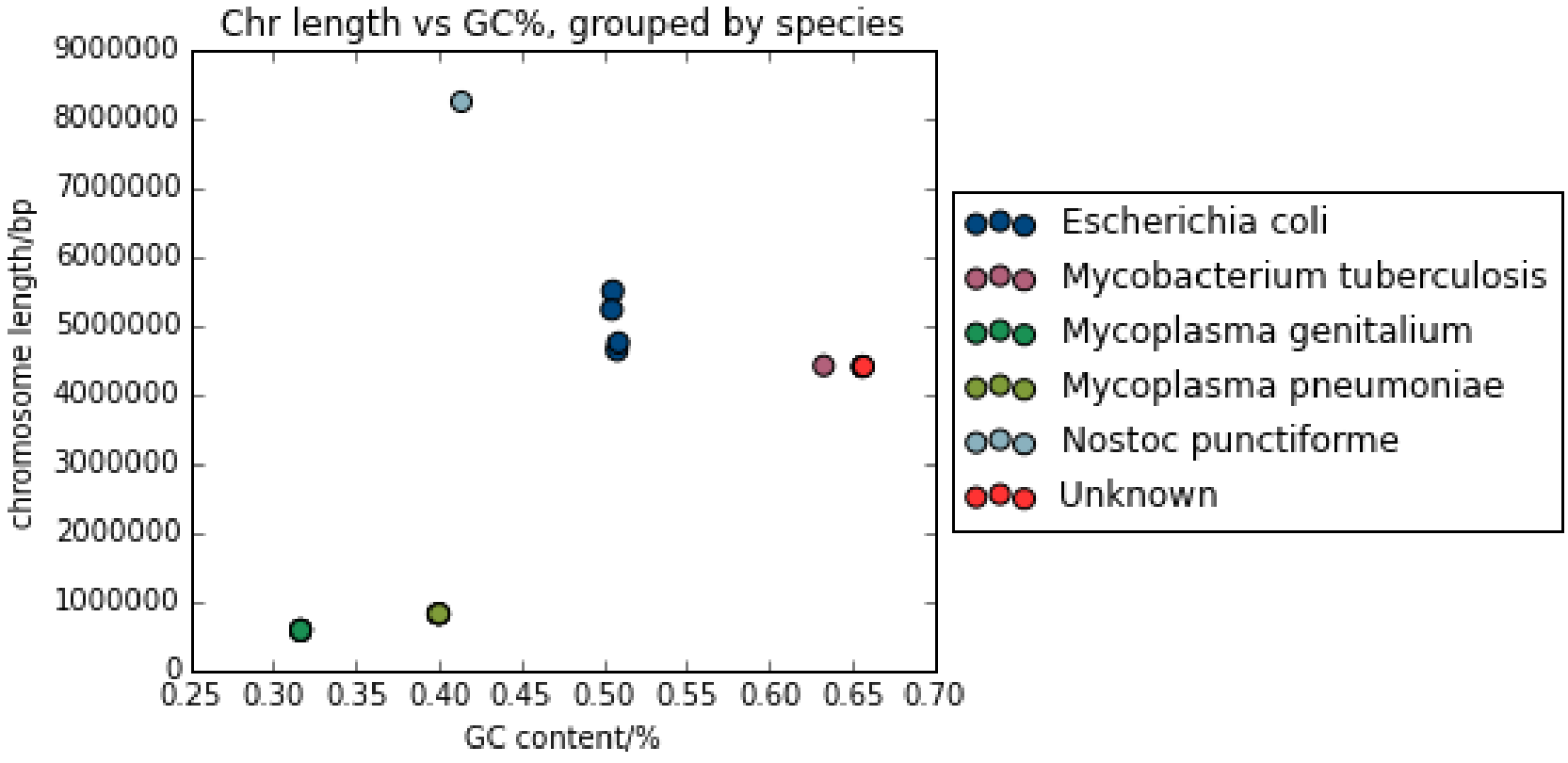
GC content, chromosome size can be characteristic of an organism.

Genome Size

The James Hutton Institute



Genome Size and GC%



Blobology [a] [b]

[a] Kumar & Blaxter (2011) *Symbiosis* doi:10.1007/s13199-012-0154-6
[b] http://nematodes.org/bioinformatics/blobology/

The James Hutton Institute

Sequence data can be contaminated by other organisms

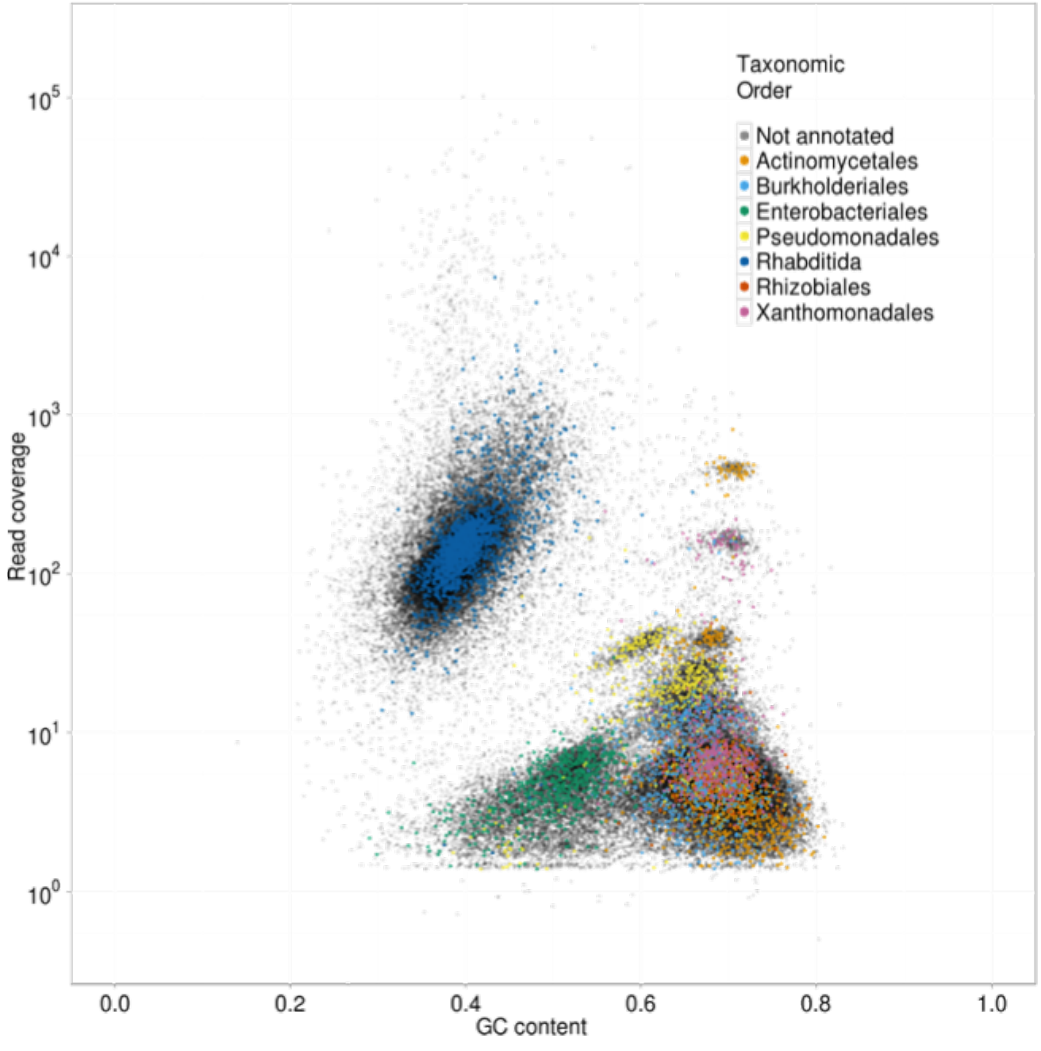
- Host and symbiont DNA have different %GC
- Host and symbiont DNA differ in coverage

- Assemble genome
- Map reads
- Plot coverage against %GC

Blobology [a] [b]

[a] Kumar & Blaxter (2011) *Symbiosis* doi:10.1007/s13199-012-0154-6

[b] http://nematodes.org/bioinformatics/blobology/



The James Hutton Institute

Nucleotide k-mers

The James Hutton Institute

Sequence data is necessary to determine k-mers/frequencies
Not possible by experiment

- Nucleotides, $k = 1$, 4x1-mers
  `A,C,G,T`
- Dinucleotides, $k = 2$, 16x2-mers
  `AA,AC,AG,AT,CA,CC,CG,CT,GA,GC,GG,GT,TA,TC,TG,TT`
- Triucleotides, $k = 1$, 64x3-mers
- k-nucleotides, 4^kxk-mers

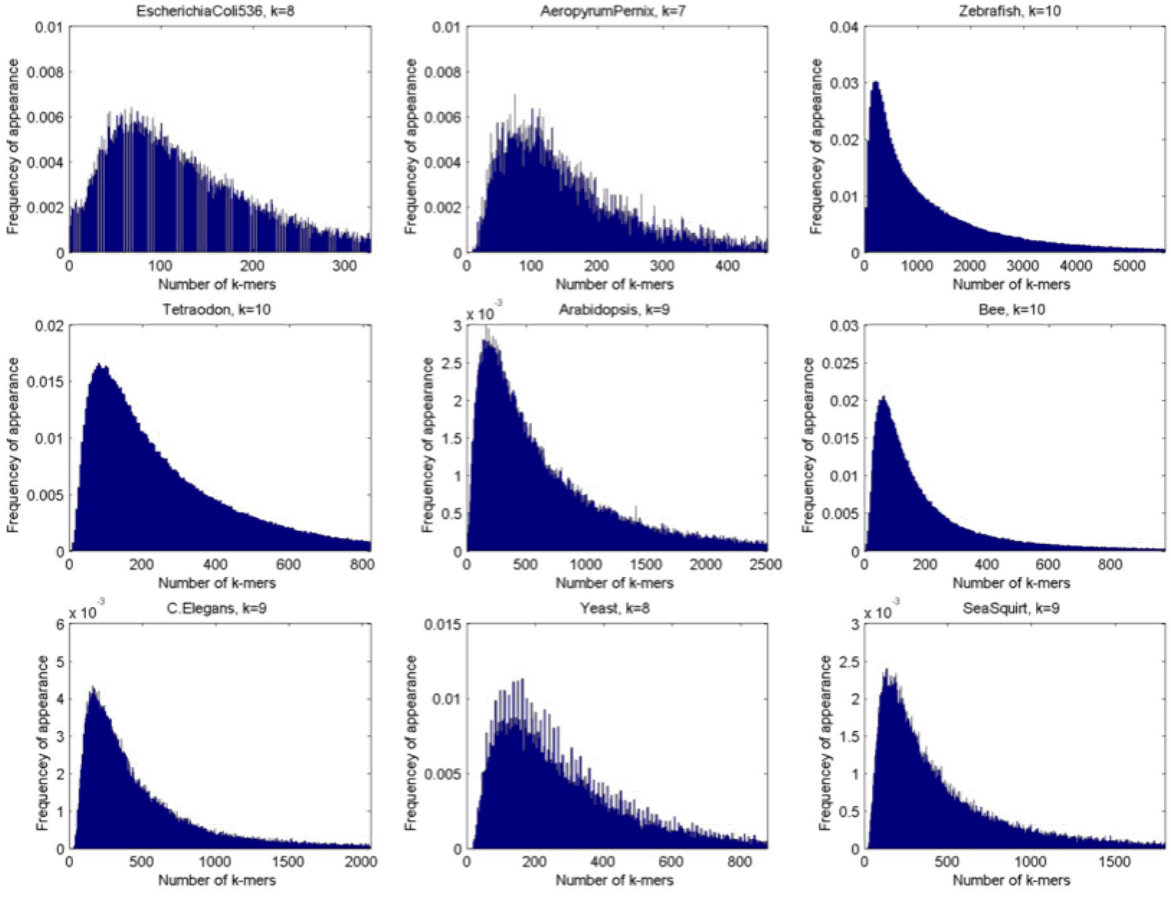
k-mer spectra [a]

[a] Chor *et al.* (2009) *Genome Biol.* doi:10.1186/gb-2009-10-10-r108

The James Hutton Institute

k-mer spectrum: frequency distribution of observed *k*-mer counts.

Most species have a unimodal *k*-mer spectrum



k-mer spectra [a]

[a] Chor *et al.* (2009) *Genome Biol.* doi:10.1186/gb-2009-10-10-r108

The James Hutton Institute

All mammals tested (and some other) species have *multimodal k-mer spectra*

Genomic regions also differ in this property

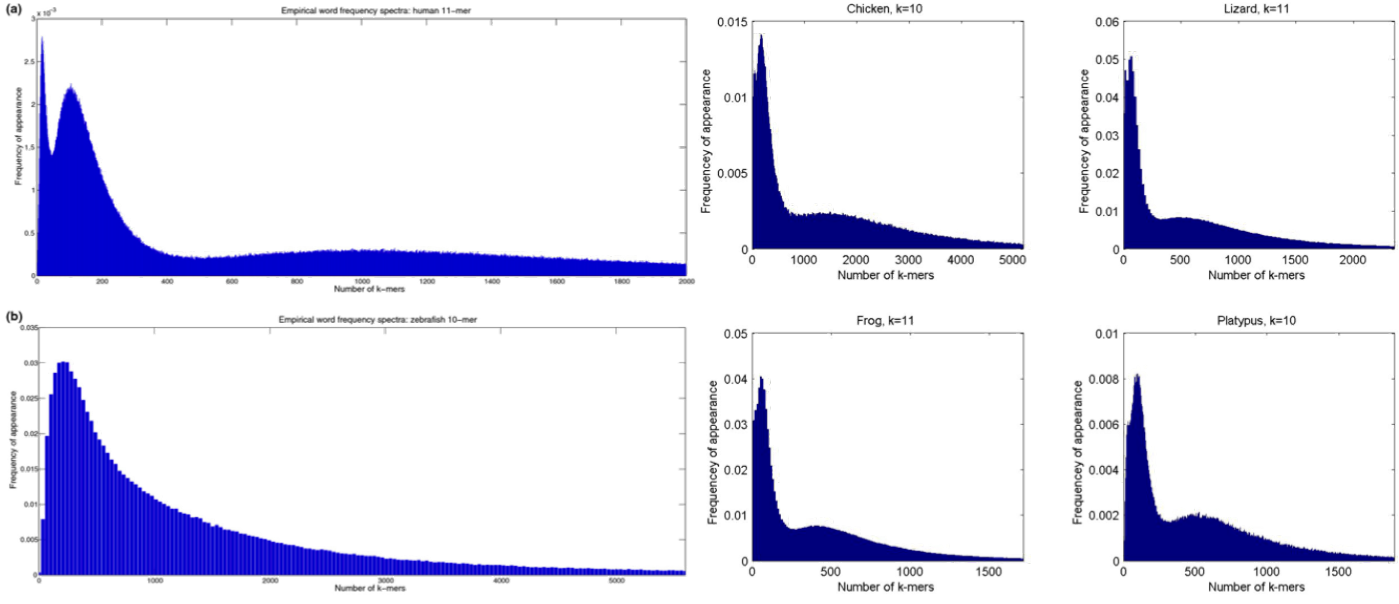


Table of Contents

The James Hutton Institute

Whole genome comparisons

The James Hutton Institute

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

Whole genome comparisons

The James Hutton Institute

Minimum requirement: **two genomes**

- Reference Genome
- Comparator Genome

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

Whole genome comparisons

The James Hutton Institute

Experimental methods mostly involve direct or indirect DNA hybridisation

- DNA-DNA hybridisation (DDH)
- Comparative Genomic Hybridisation (CGH)
- Array Comparative Genomic Hybridisation (aCGH)

Whole genome comparisons

The James Hutton Institute

Analogously, *in silico* methods mostly involve sequence alignment

- Average Nucleotide Identity (ANI)
- Pairwise genome alignment
- Multiple genome alignment

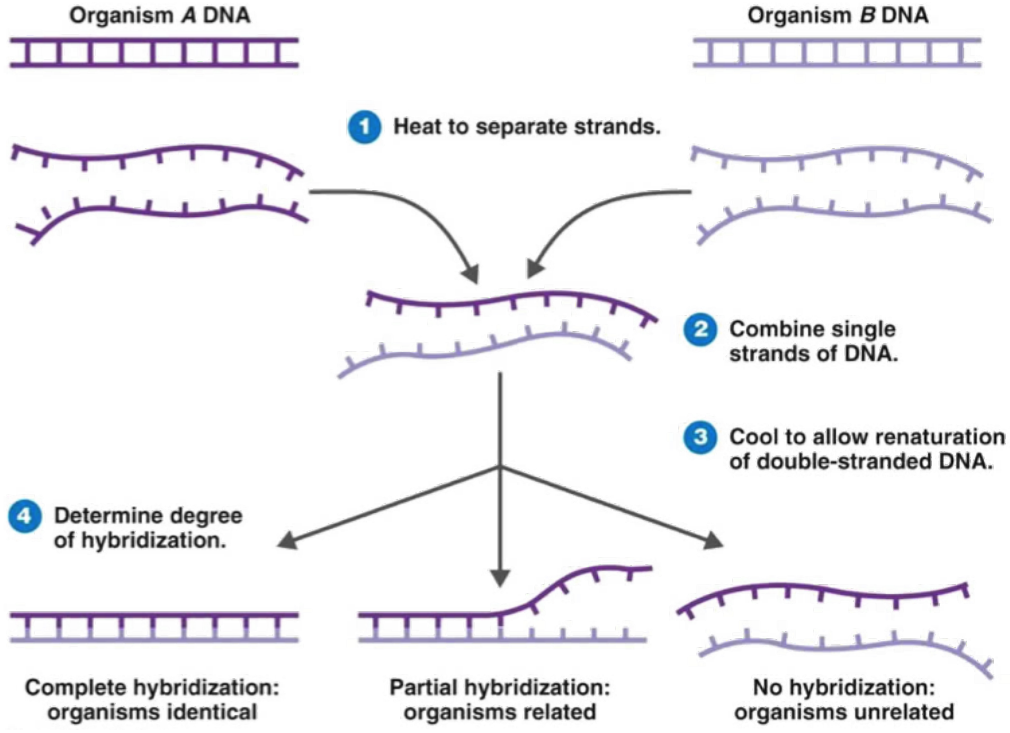
DNA-DNA hybridisation (DDH) [a]

[a] Morelló-Mora & Amann (2011) *FEMS Microbiol. Rev.* doi:10.1016/S0168-6445(00)00040-1

The James Hutton Institute

Several similar methods based on the same principle

- Denature gDNA mixture for organisms *A*, *B*
- Allow gDNA to anneal; hybrids result



Reassociation of gDNA $\approx$ sequence similarity

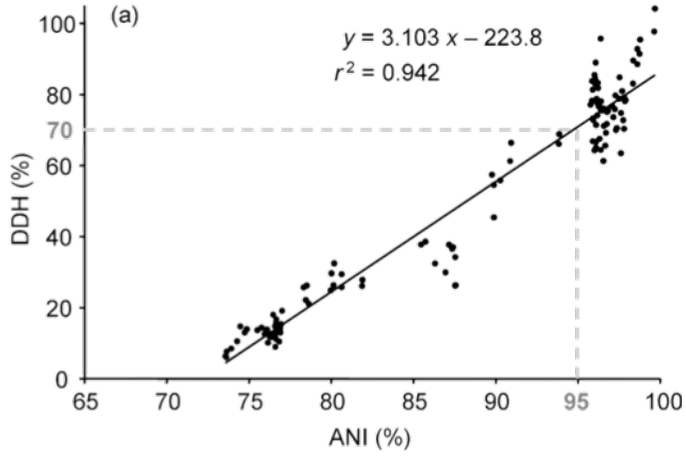
Average Nucleotide Identity (ANI) [a]

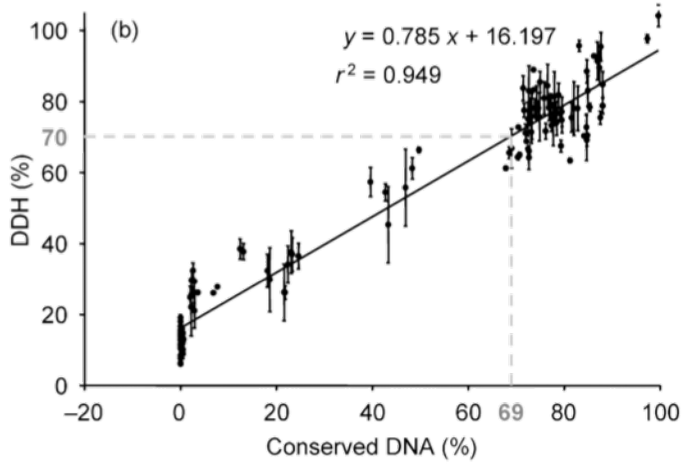
[a] Goris *et al.* (2007) *Int. J. System. Evol. Biol.* doi:10.1099/ijs.0.64483-0

The James Hutton Institute

Introduced as an *in silico* substitute for DDH in 2007:

- 70% identity (DDH) = "gold standard" prokaryotic species boundary
- 70% identity (DDH) ≈ 95% identity (ANI)





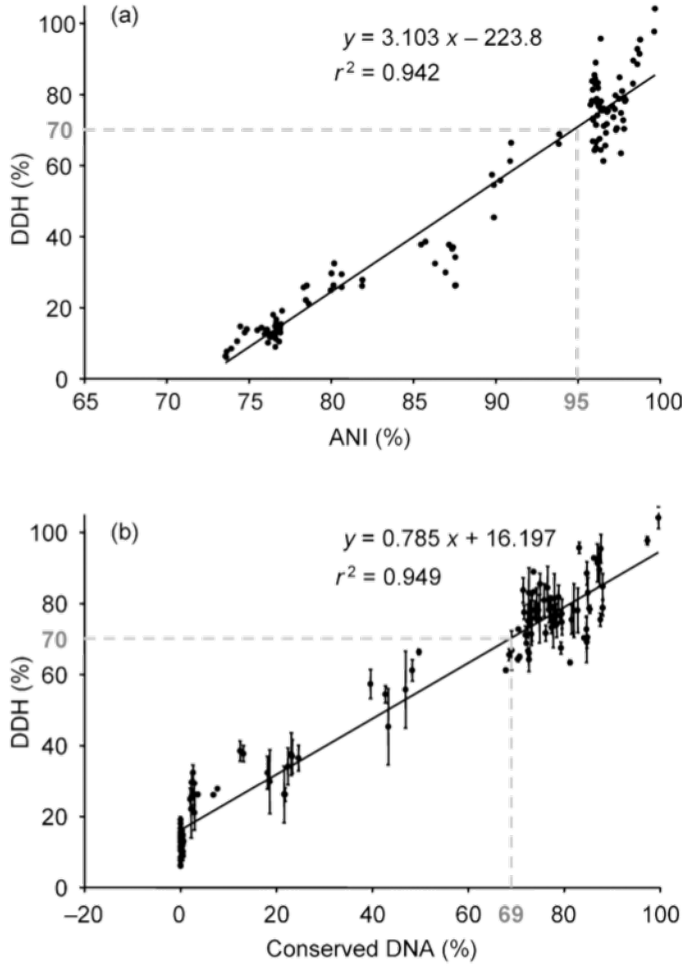
Average Nucleotide Identity (ANI) [a]

[a] Goris *et al.* (2007) *Int. J. System. Evol. Biol.* doi:10.1099/ijs.0.64483-0

The James Hutton Institute

Original method emulated physical experiment:

1. break genome into 1020nt fragments
2. align all fragments with BLASTN
3. ANI = mean identity of all matches with $>$ 30% identity, $>$ 70% coverage



Average Nucleotide Identity (ANI) [a]

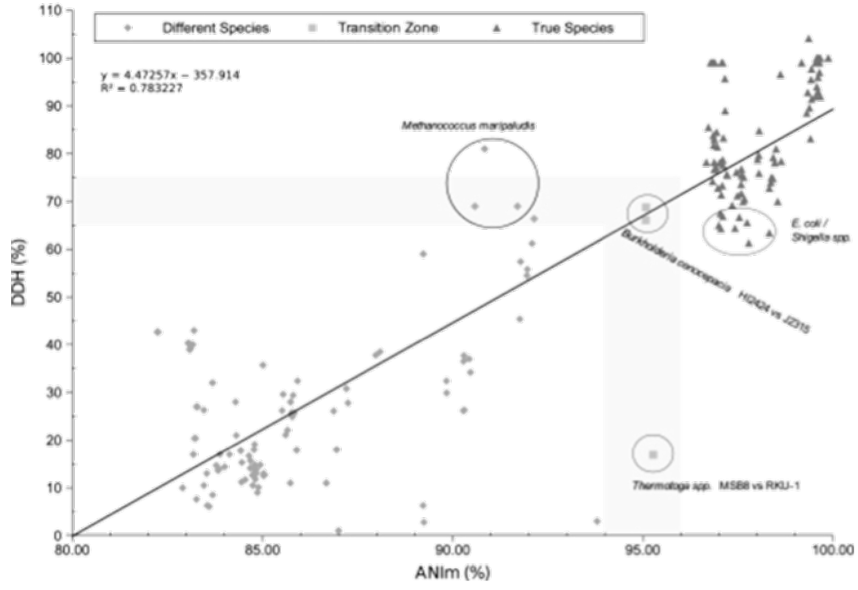
[a] Richter & Rosselló-Móra *et al.* (2009) *Proc. Natl. Acad. Sci. USA* doi:10.1073/pnas.0906412106

The James Hutton Institute

ANIm and TETRA variants introduced in 2009:

ANIm

1. Align sequences with NUCmer (no fragmentation)
2. ANI = mean identity of matches



TETRA (a bulk measure!)

1. Calculate 4-mer frequencies
2. Determine Z-score for 4-mer deviation from expected value, given %GC content
3. TETRA = Pearson correlation coefficient of Z-scores

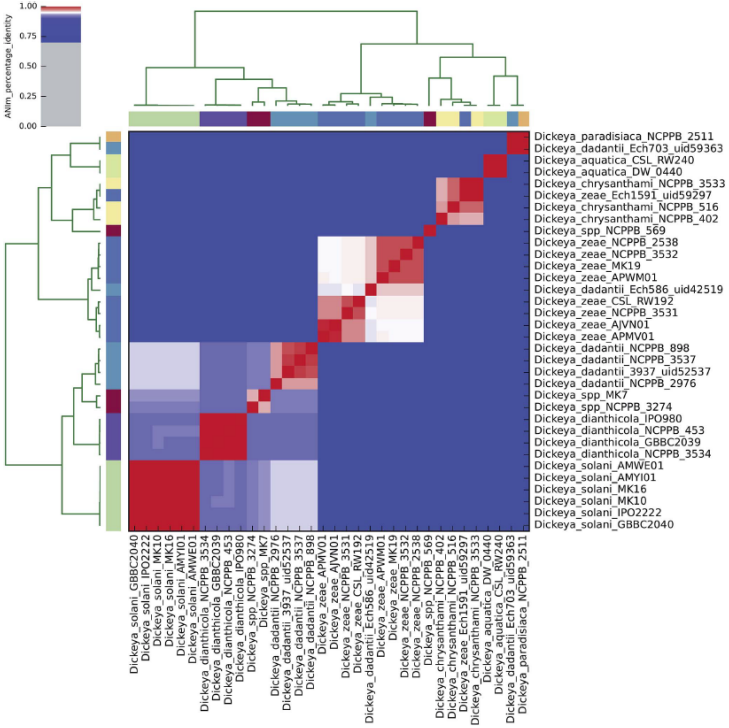
ANI in practice [a] [b]

[a] van der Wolf *et al.* (2014) *Int. J. Syst. Evol. Micr.* **64**:768-774 doi:10.1099/ijs.0.052944-0

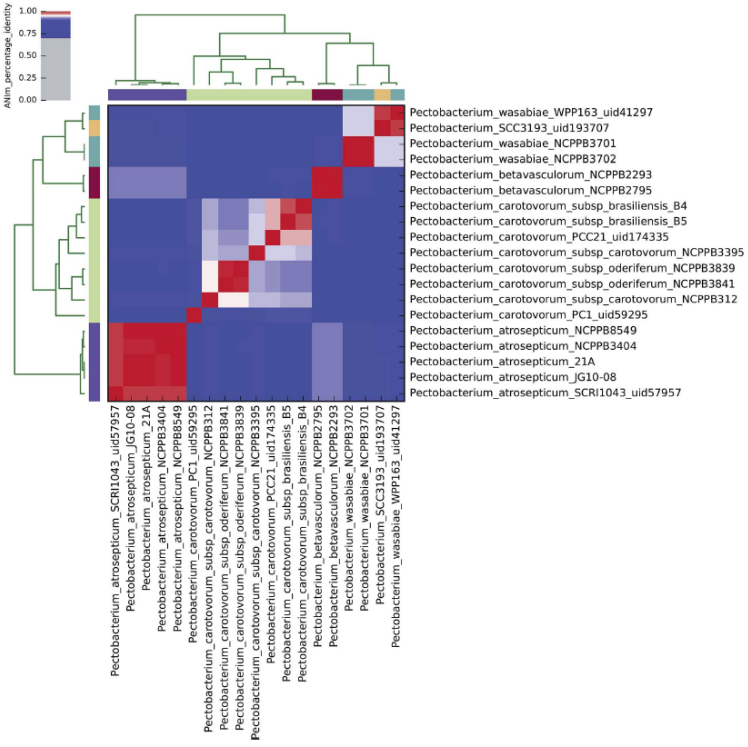
[b] Pritchard *et al.* (2016) *Anal. Methods* **8**:12-24 doi:10.1039/C5AY02550H

The James Hutton Institute

Dickeya species structure



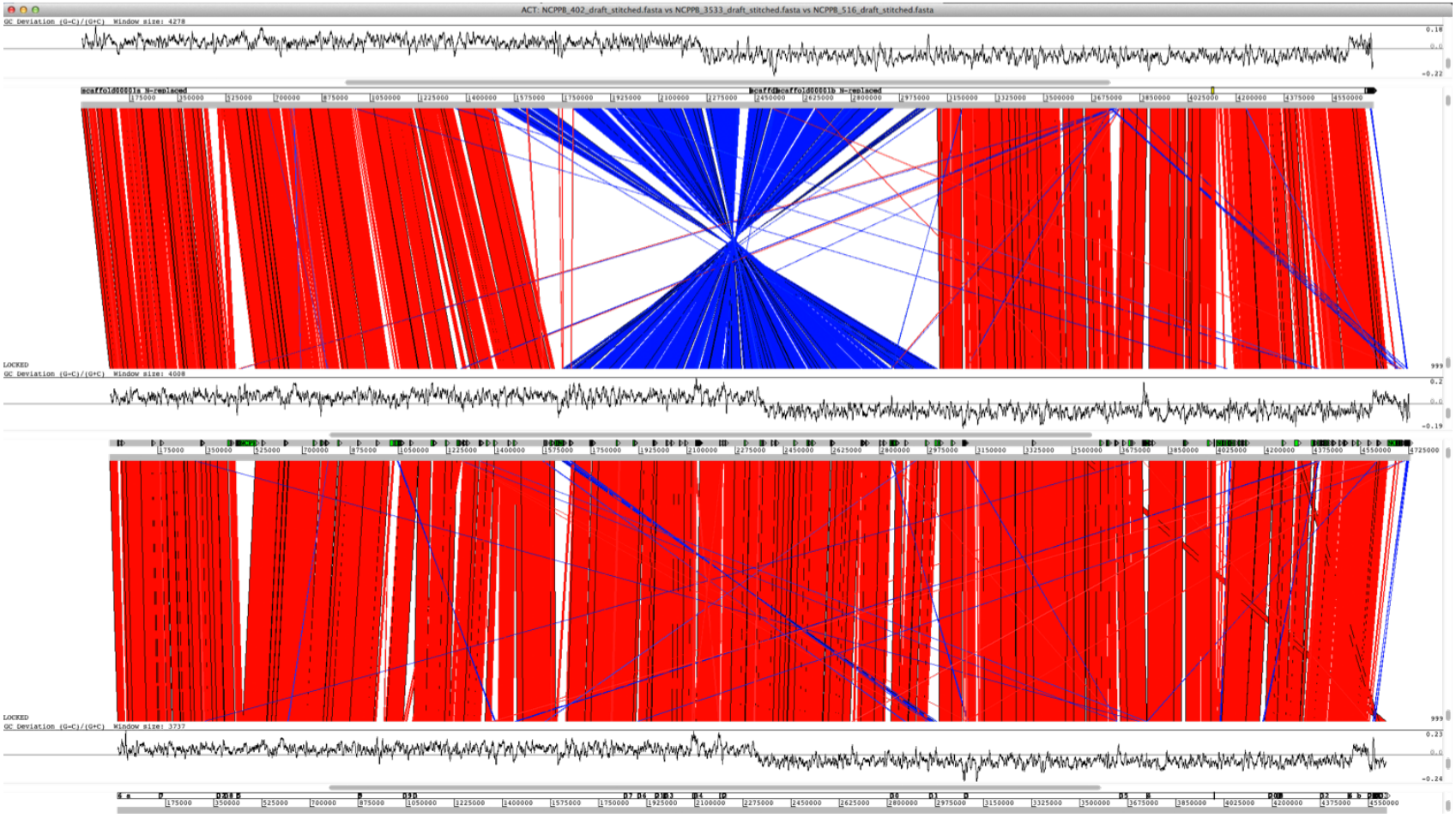
Pectobacterium species structure:



Pairwise genome alignments

The James Hutton Institute

Pairwise comparisons require alignment of similar regions.

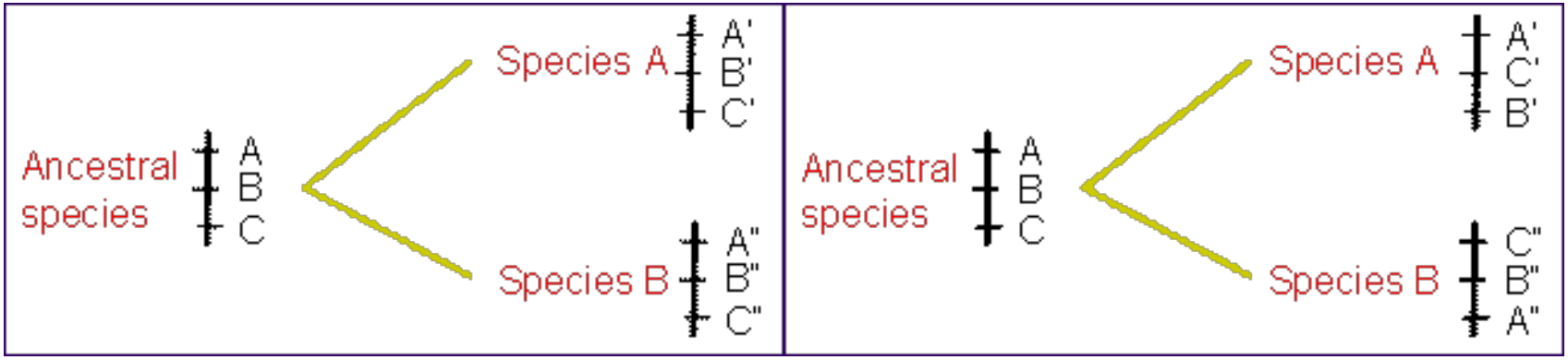


Synteny and Collinearity

The James Hutton Institute

Genome rearrangements may occur post-speciation

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 a functional constraint

Pairwise genome alignments

The James Hutton Institute

Which genomes should you align (or not bother with)?
For reasonable analysis, genomes should:

- derive from a sufficiently recent common ancestor, so that homologous regions can be identified
- derive from a sufficiently distant common ancestor, so that biologically meaningful changes are likely to be found

Alignment algorithms/programs

The James Hutton Institute

I assume you're familiar with BLAST

Naïve alignment algorithms are not appropriate:

- Needleman-Wunsch: optimal global alignment
- Smith-Waterman: optimal local alignment

Cannot handle rearrangement

Computationally expensive

Alignment algorithms/programs

The James Hutton Institute

Many whole-genome alignment algorithms proposed

Handle genome-scale evolutionary processes, scalable

- LASTZ (http://www.bx.psu.edu/~rsharris/lastz/)
- **BLAT** (http://genome.ucsc.edu/goldenPath)
- Mugsy (http://mugsy.sourceforge.net/)
- **megaBLAST** (http://www.ncbi.nlm.nih.gov/blast/)
- **MUMmer** (http://mummer.sourceforge.net/)
- LAGAN (http://lagan.stanford.edu/lagan_web/index.shtml)
- WABA, etc?

megaBLAST [a] [b]

[a] Zhang *et al.* (2000) *J. Comp. Biol.* **7**(1-2): 203-214
[b] Korf *et al.* (2003) *BLAST* O'Reilly & Associates, Sebastopol, CA

The James Hutton Institute

Optimised for:

- speed and genome-level searching
- queries on large sequence sets: "query-packing"
- long alignments of very similar sequences (`dc-megablast` for divergent sequences)

Uses Zhang et al. greedy algorithm, **not BLAST algorithm**

BLASTN+ defaults to megaBLAST algorithm
(see http://www.ncbi.nlm.nih.gov/blast/Why.shtml)

MUMmer [a]

[a] Kurtz *et al.* (2004) *Genome Biol.* doi:10.1186/gb-2004-5-2-r12

The James Hutton Institute

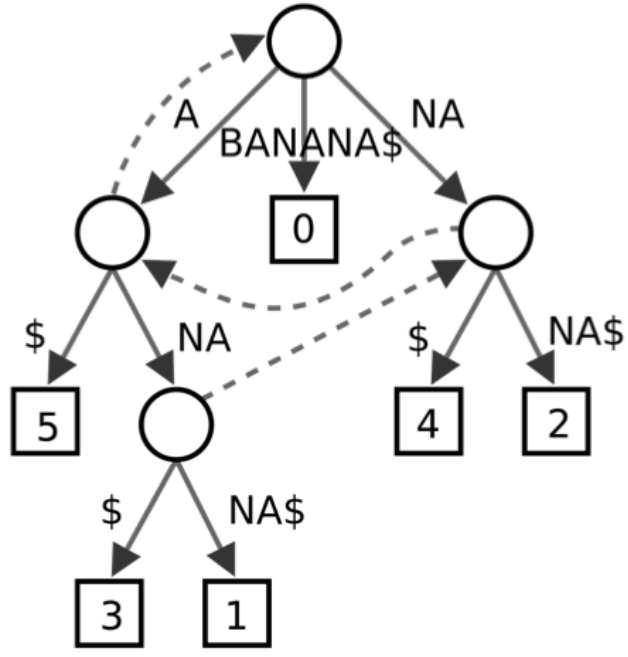
Conceptually completely different to BLAST/BLAT/megaBLAST

Uses *suffix trees* for pattern matching

- Finds maximal exact matches
- Memory use depends only on reference sequence size

Suffix Tree:

- Constructed and searched in $O(n)$ time
- Useful algorithms are nontrivial
- `BANANA$`



MUMmer [a]

[a] Kurtz *et al.* (2004) *Genome Biol.* doi:10.1186/gb-2004-5-2-r12

The James Hutton Institute

Process:

1. Identify non-overlapping set of maximal exact matches (MUMs: *maximal unique matches*)
2. Cluster MUMs into *alignment anchors*
3. Extend between anchors to produce final (gapped) alignment

The approach is very flexible

1. Used in a suite of programs (`mummer`, `nucmer`, `promer`, ...)
2. Nucleotide and "conceptual protein" (sensitive!) alignments
3. Used for comparisons, scaffolding, repeat detection
4. Basis of other aligners and assemblers, (`Mugsy`, `AMOS`, ...)

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 carries genes associated with environmental adaptation; C-I carries virulence genes.
C-II has undergone extensive rearrangement; C-I has not.

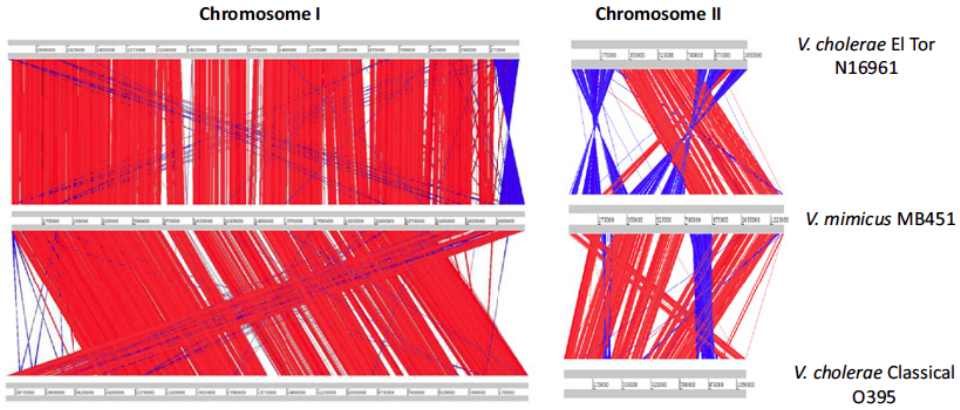


Fig. 2. Linear pairwise comparison of the *Vibrio mimicus* genome by Artemis Comparison Toll. Regions with similarity are highlighted by connecting red or blue lines between the genomes; red lines indicate homologous blocks of sequence, and blue lines indicate inversions. Gaps indicate unique DNA. The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

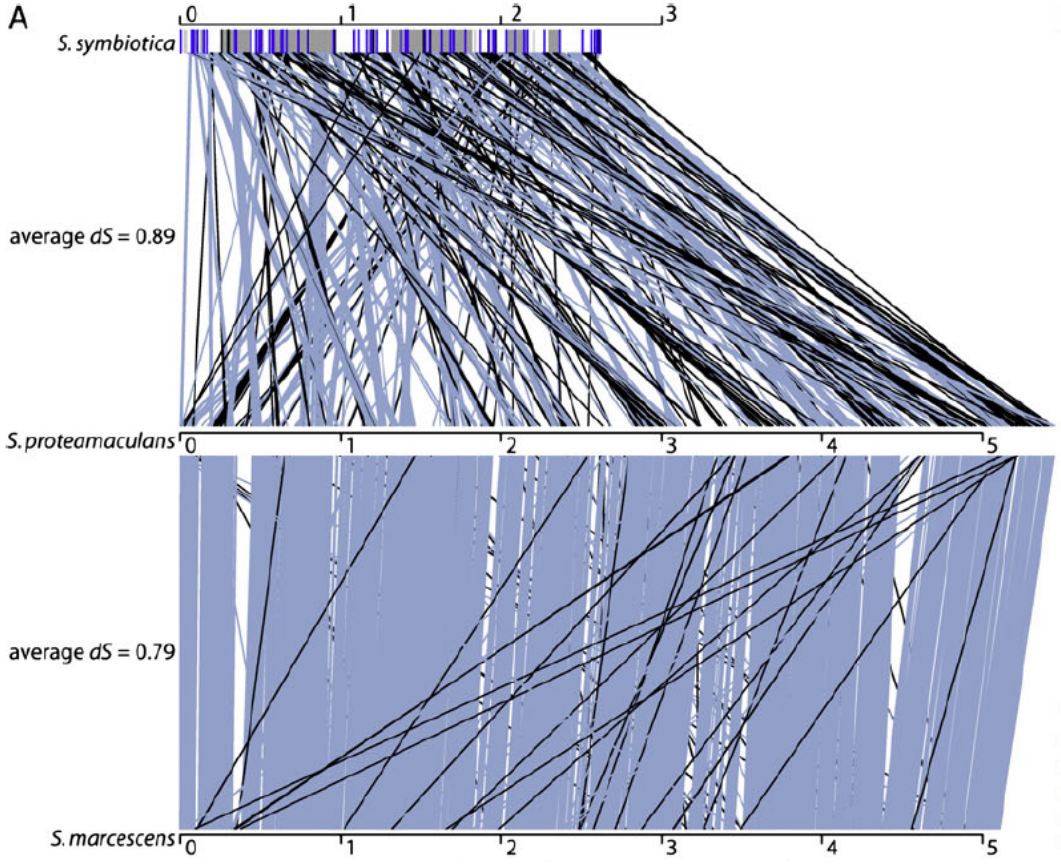
Suggests modularity of genome organisation, as a mechanism for adaptation (HGT, two-speed genome).

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 is an adaptation to the new environment.



Multiple genome alignments

The James Hutton Institute

Multiple genome alignments are "harder" than pairwise

- Computationally difficult to produce
- Lead to NP-complete optimisation problems!

Solutions: **heuristics**

- Progressive (build a tree, combine pairwise alignments)
- Iterative (realign initial sequences as new genomes added)
- Positional homology
- *Glocal* alignments

Multiple genome alignment

The James Hutton Institute

Many tools either positional homology or glocal alignment

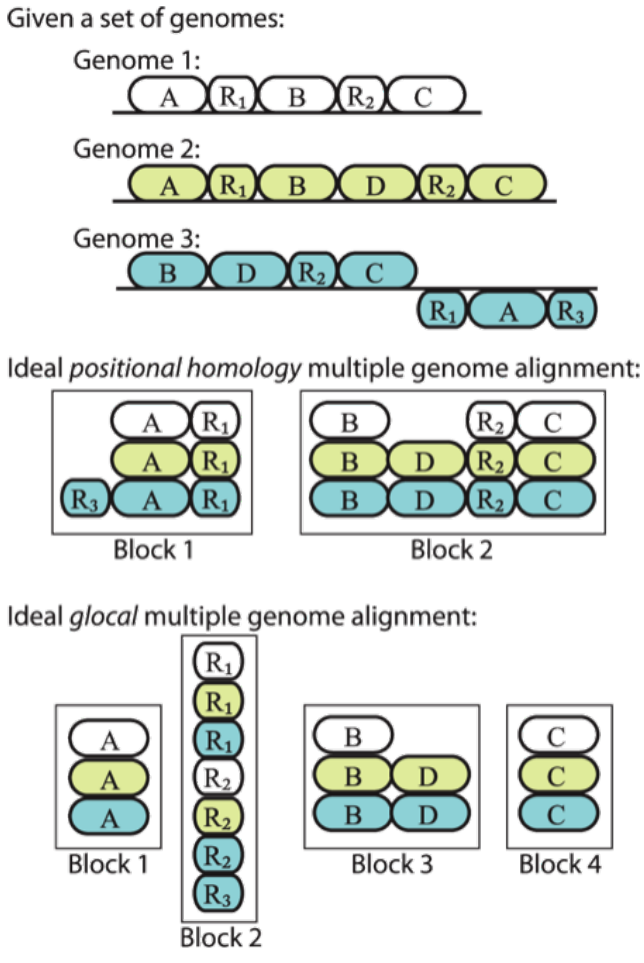
Several tools:

- Mugsy: (http://mugsy.sourceforge.net/)
- MLAGAN:

  (http://lagan.stanford.edu/lagan_web/index.shtml)
- TBA/MultiZ:

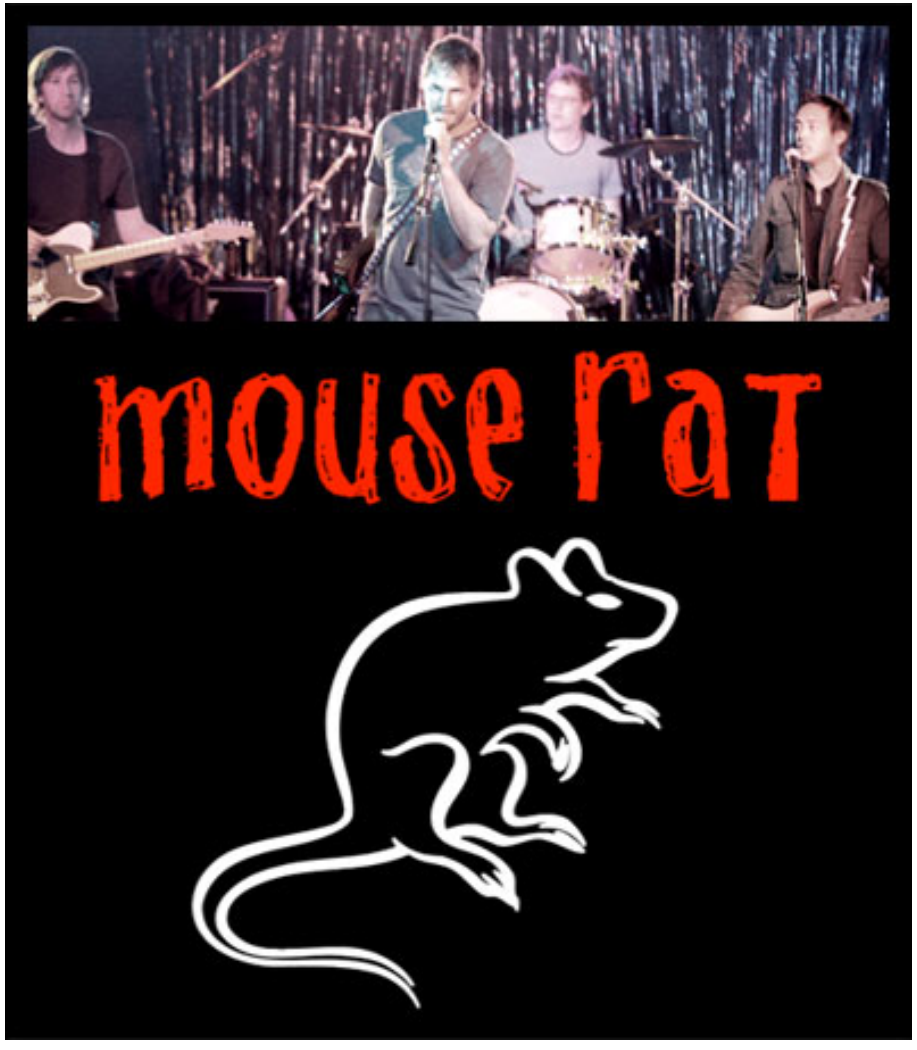
  (http://www.bx.psu.edu/miller_lab/)
- **Mauve**:

  (http://gel.ahabs.wisc.edu/mauve/)



Human:Mouse:Rat

The James Hutton Institute



Human:Mouse:Rat [a]

[a] Brudno *et al.* (2004) *Genome Res.* doi:10.1101/gr.2067704

The James Hutton Institute

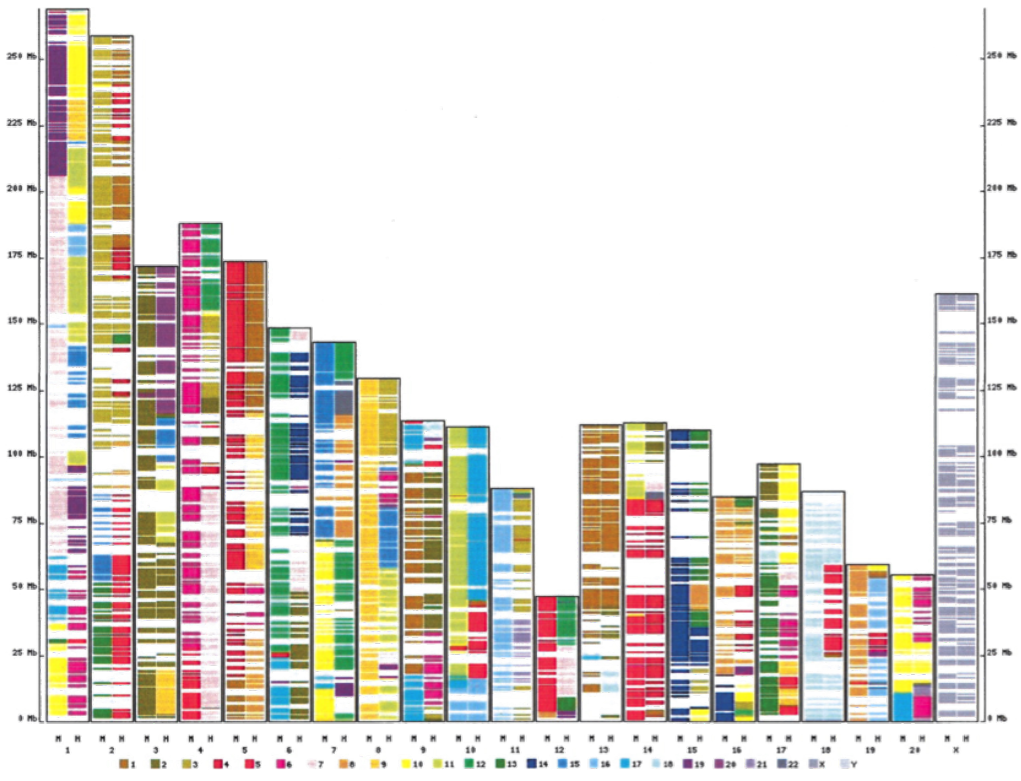
Three-way progressive alignment

Initial alignments: BLAT

Synteny: LAGAN/MLAGAN

Three-way synteny

- Homologous (HMR)
- Rodent-only (MR)
- Human-mouse (HM)
- Human-rat (HR)



(synteny mapped to rat genome)

MAUVE [a]

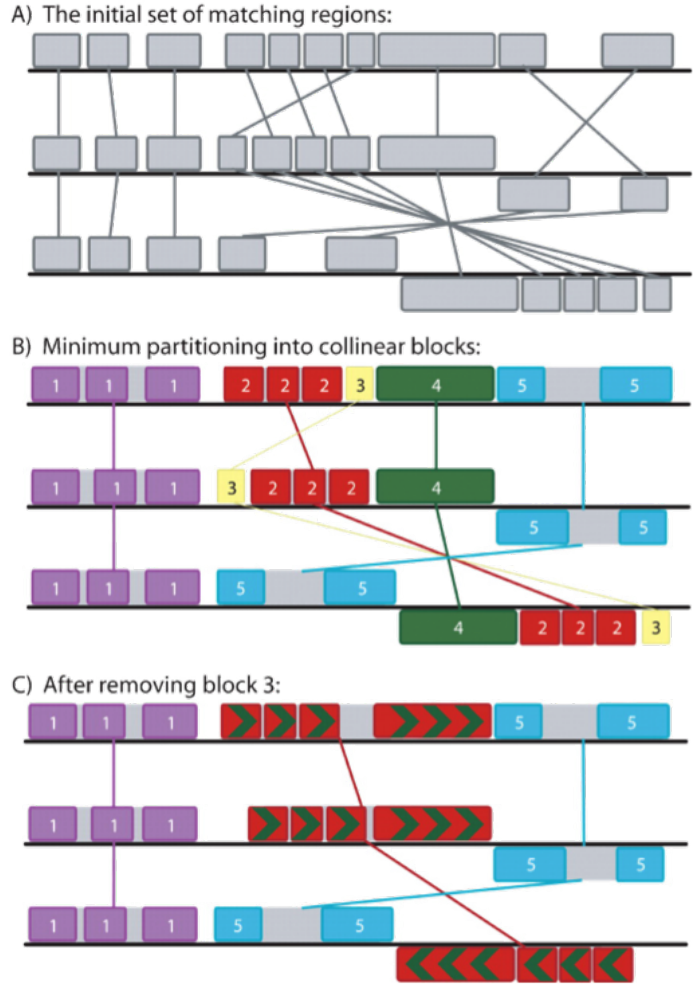
[a] Darling *et al.* (2003) *Genome Res.* doi:10.1101/gr.2289704

The James Hutton Institute

MAUVE/Progressive MAUVE: http://gel.ahabs.wisc.edu/mauve/

Algorithm:

1. Find local alignments (*multi-MUMs*, A)
2. Build guide tree from multi-MUMs
3. Select subset of multi-MUMs (*anchors*, B)
   - Partition into *local collinear blocks* (LCBs)
4. Recursive anchoring to refine anchors (C)
5. Progressive alignment against guide tree



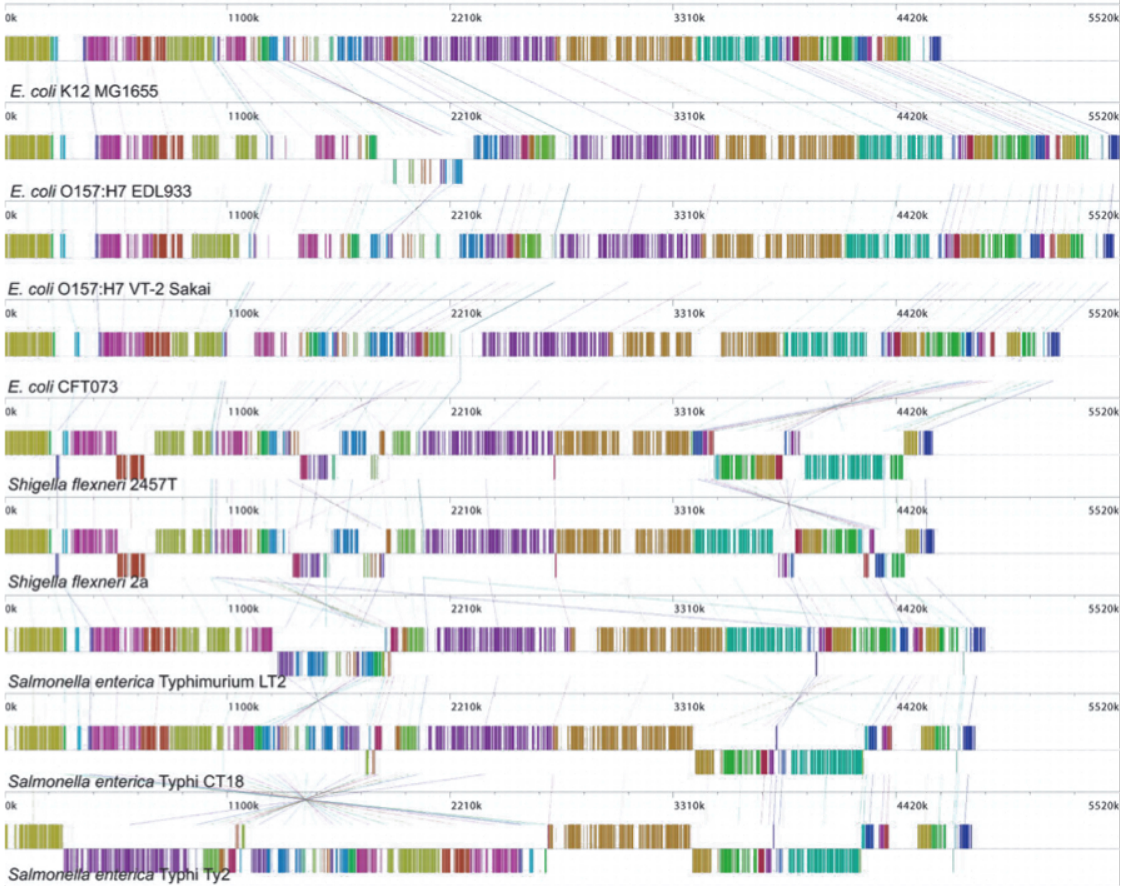
MAUVE [a]

[a] Darling *et al.* (2003) *Genome Res.* doi:10.1101/gr.2289704

The James Hutton Institute

MAUVE alignment of LCBs in nine enterobacterial genomes

Evidence for rearrangement of homologous backbone sequence



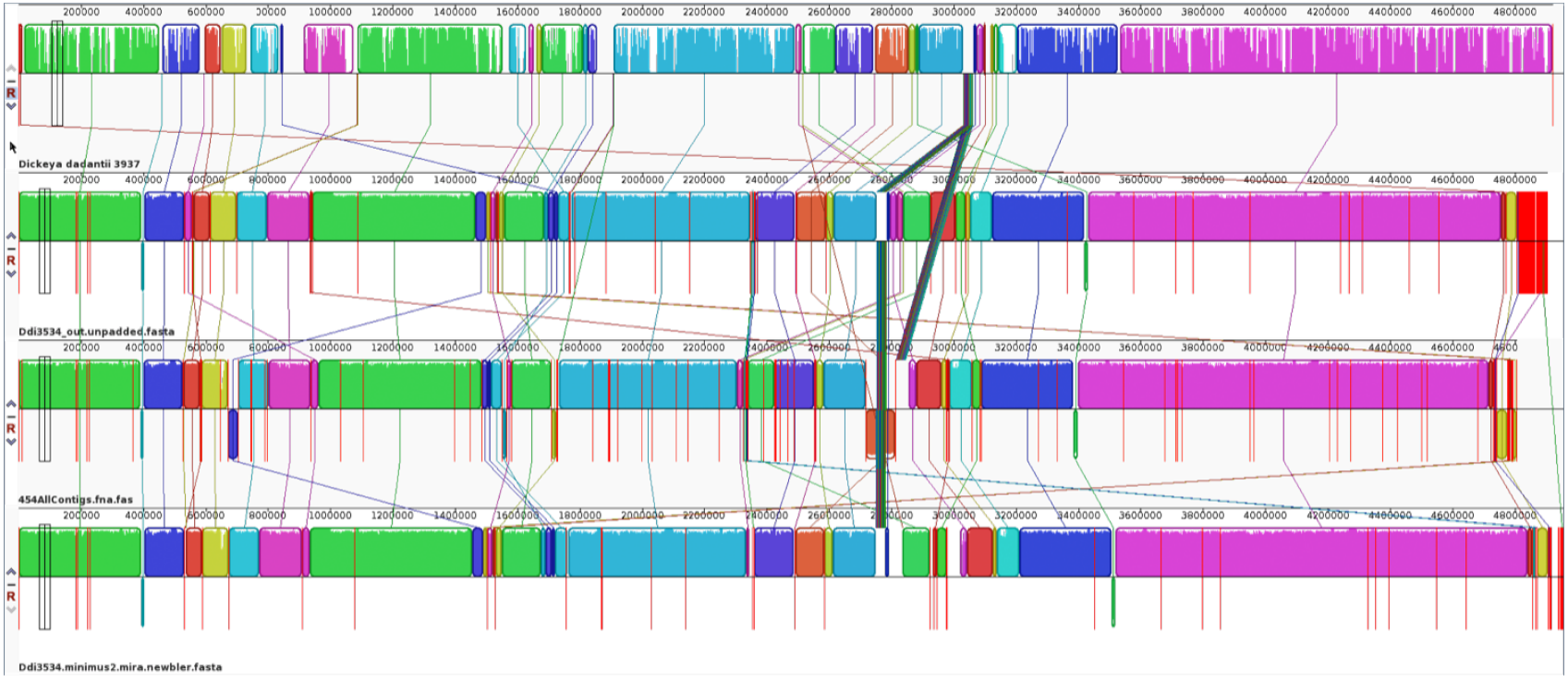
Draft genome alignment

The James Hutton Institute

High-throughput genome assemblies may be fragments (contigs)

Contigs can be ordered (*scaffolded*):

- without alignment, by long or paired-end reads
- by alignment, to complete *reference* genomes
- by alignment, to other draft incomplete genomes



Chromosome painting[a]

[a]Yahara *et al.* (2013) *Mol. Biol. Evol.* **30**:1454-1464 doi:10.1093/molbev/mst055

The James Hutton Institute

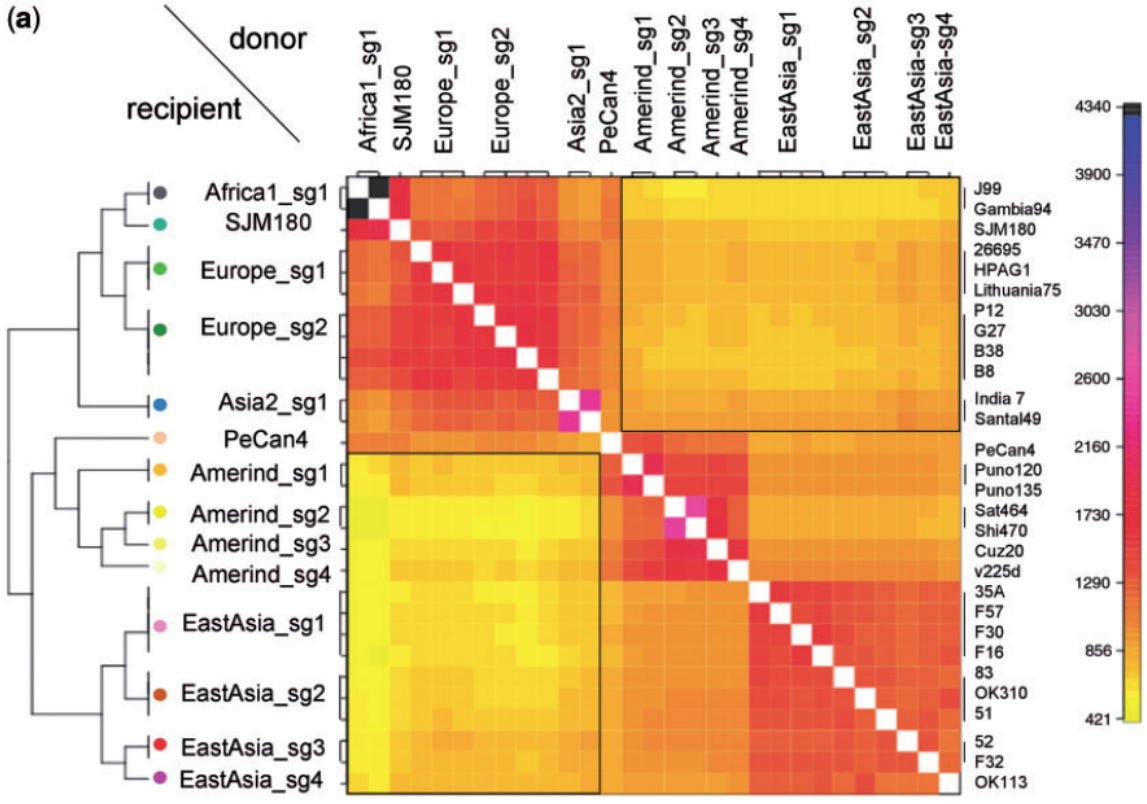
"Chromosome painting" infers recombination-derived 'chunks'
Genome's haplotype constructed in terms of recombination events from a 'donor' to a 'recipient' genome

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

Chromosome painting[a]

[a] Yahara *et al.* (2013) *Mol. Biol. Evol.* **30**:1454-1464 doi:10.1093/molbev/mst055

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



Conclusions

The James Hutton Institute

Physical and computational genome comparisons

- Similar biological questions
- $\therefore$ similar concepts

Modern biology: lots of sequence data

- Conservation $\approx$ evolutionary constraint
- Many choices of algorithms/software
- Many choices of visualisation tools/software

Table of Contents

The James Hutton Institute

Licence: CC-BY-SA

The James Hutton Institute

By: Leighton Pritchard