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

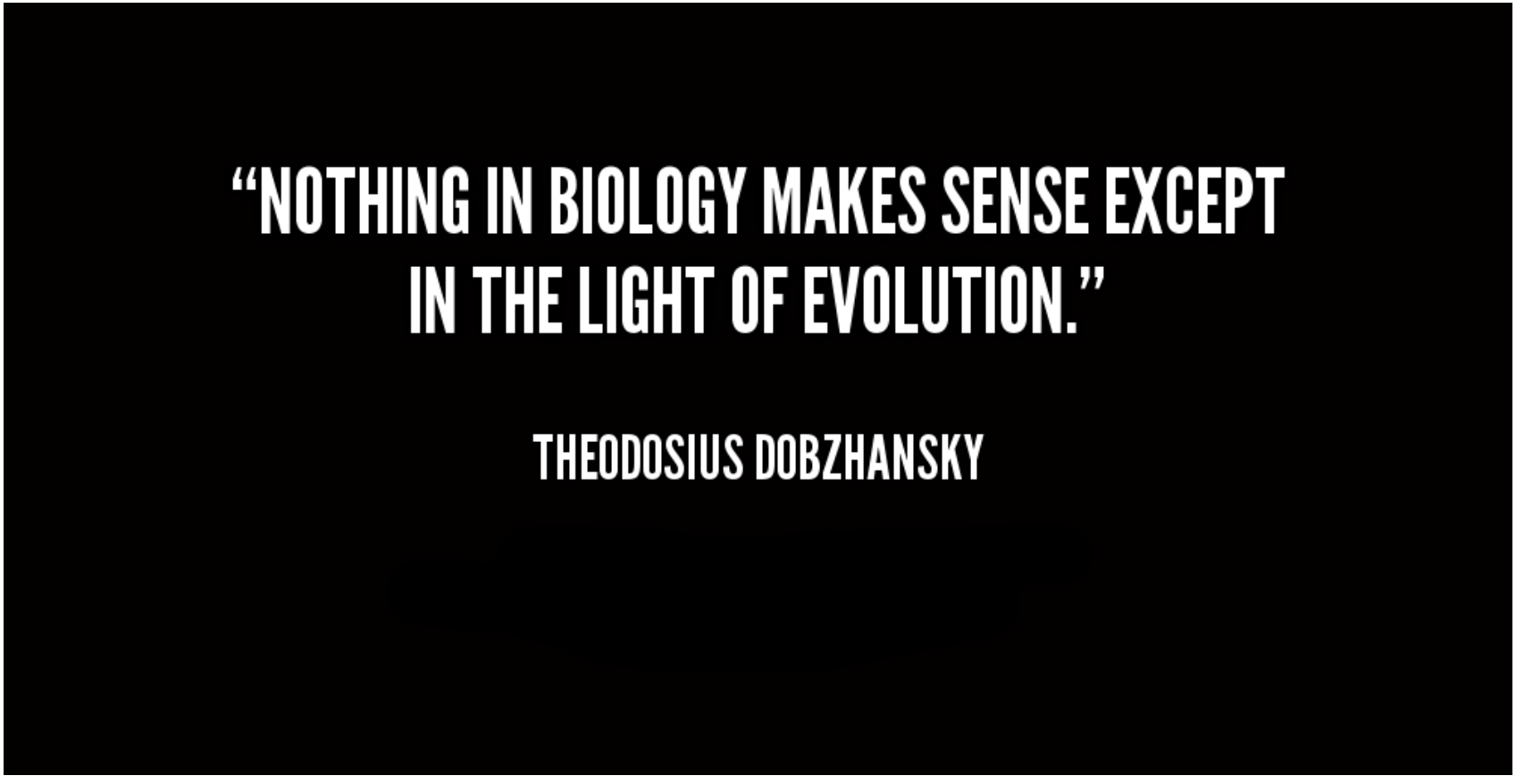
What Is Comparative Genomics?

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

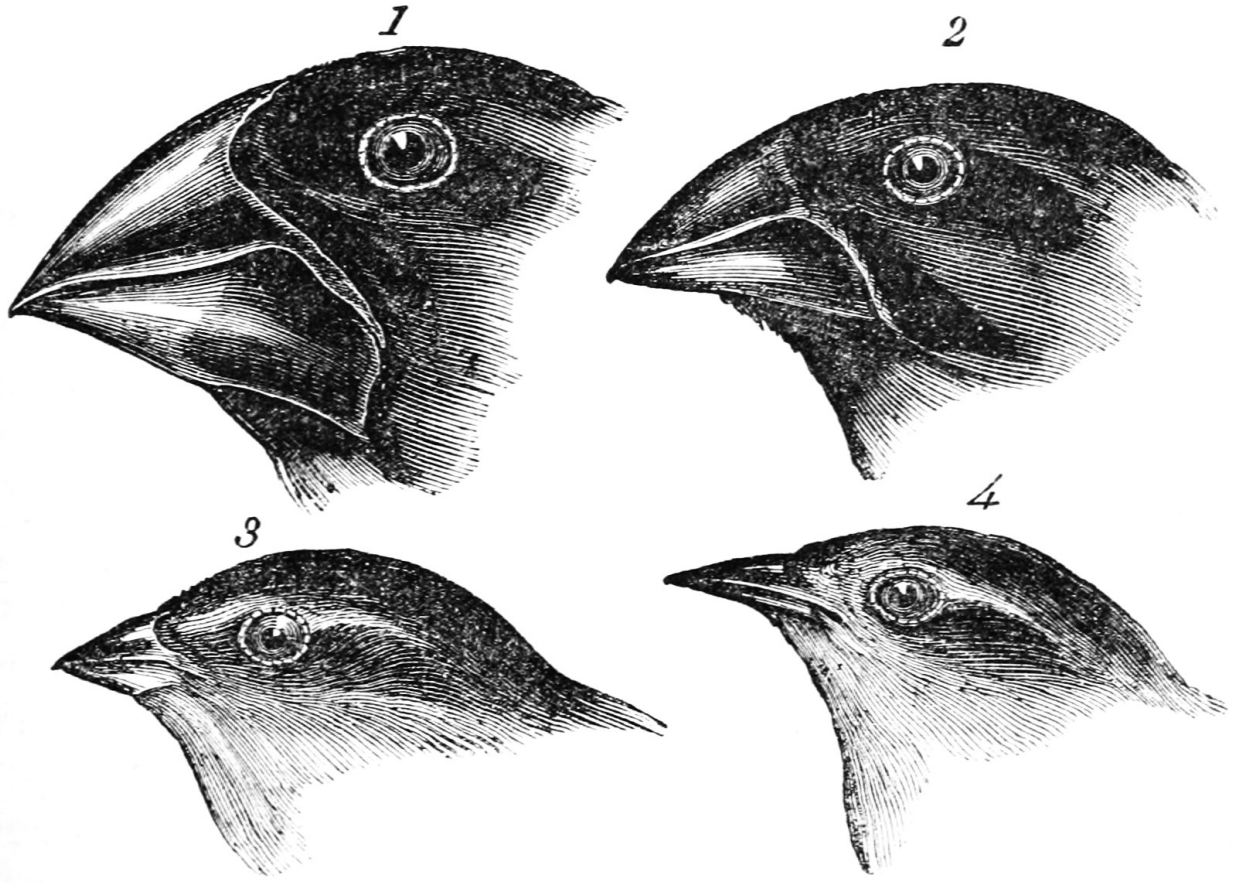
The James Hutton Institute



Comparison of physical features

The James Hutton Institute

How do we determine that features share a common ancestor?



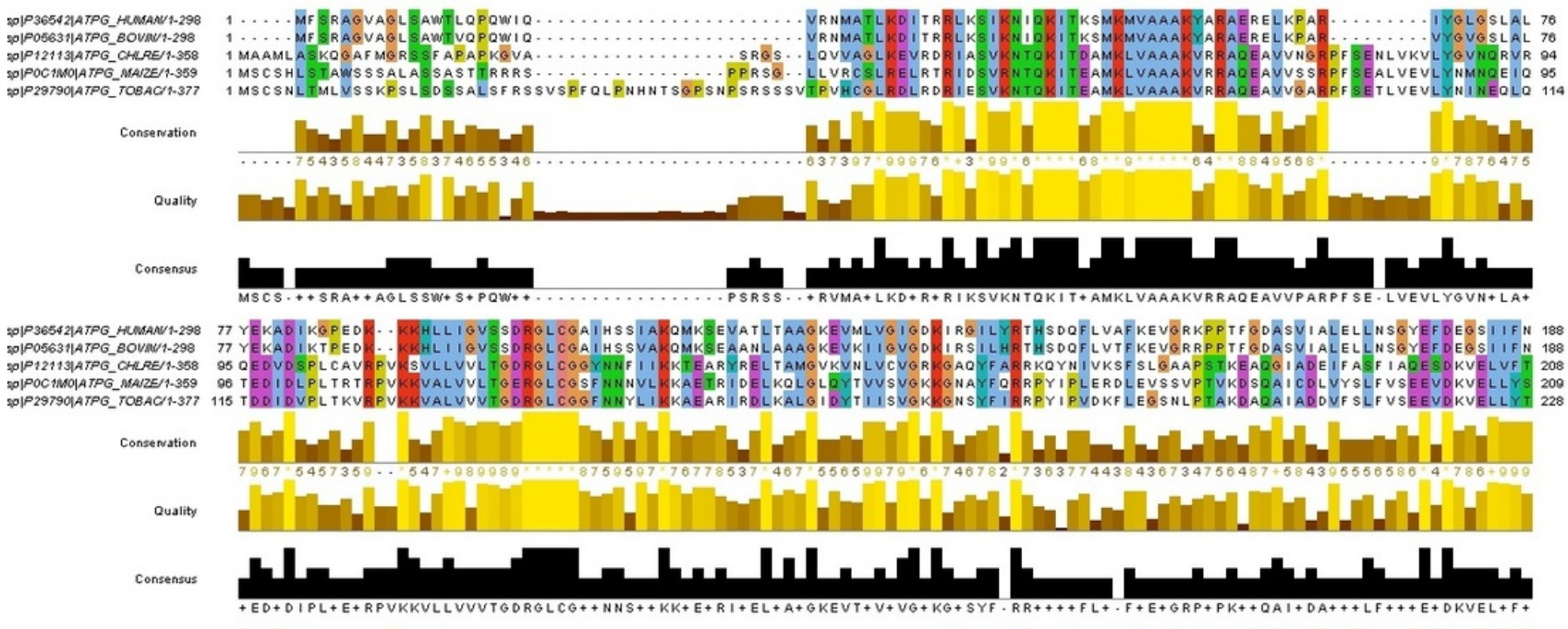
1. Geospiza magnirostris.
2. Geospiza fortis.
3. Geospiza parvula.
4. Certhidea olivacea.

Comparison of sequence features

The James Hutton Institute

How do we determine that features share a common ancestor?

Multiple sequence alignment of ATP synthase

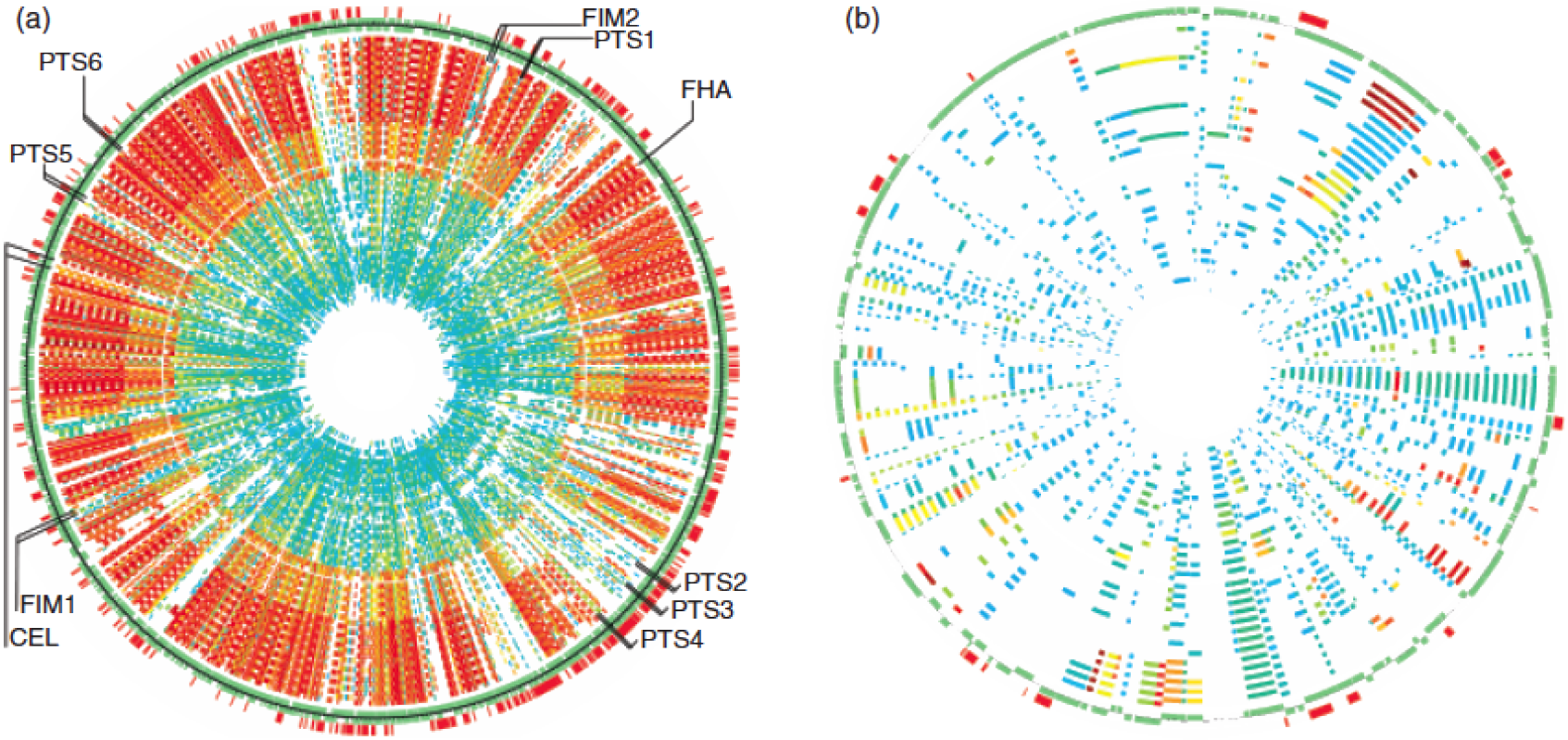


Comparison of genome features

The James Hutton Institute

How do we determine that features share a common ancestor?

Similarity of individual features (feature sequence)

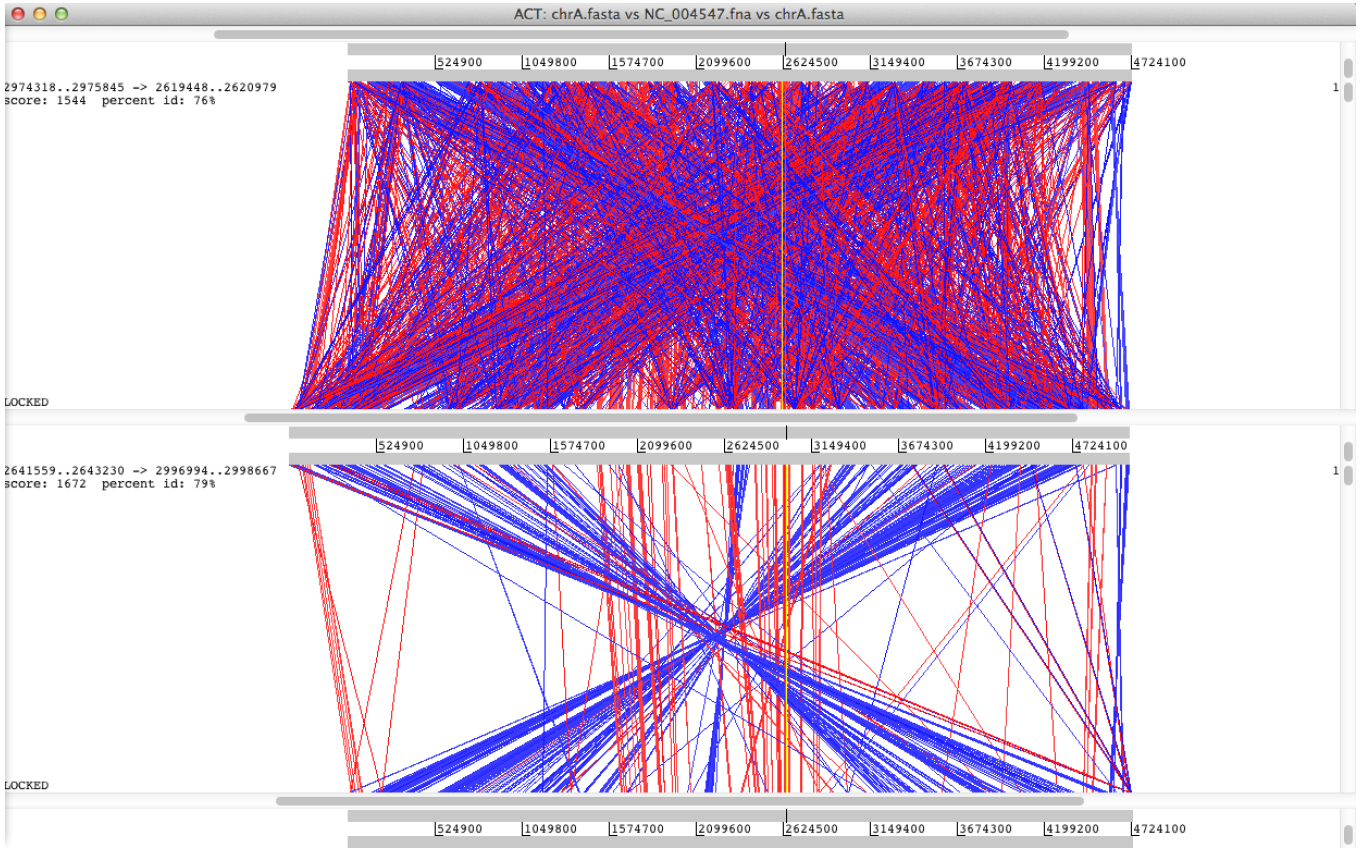


Comparison of genome features

The James Hutton Institute

How do we determine that features share a common ancestor?

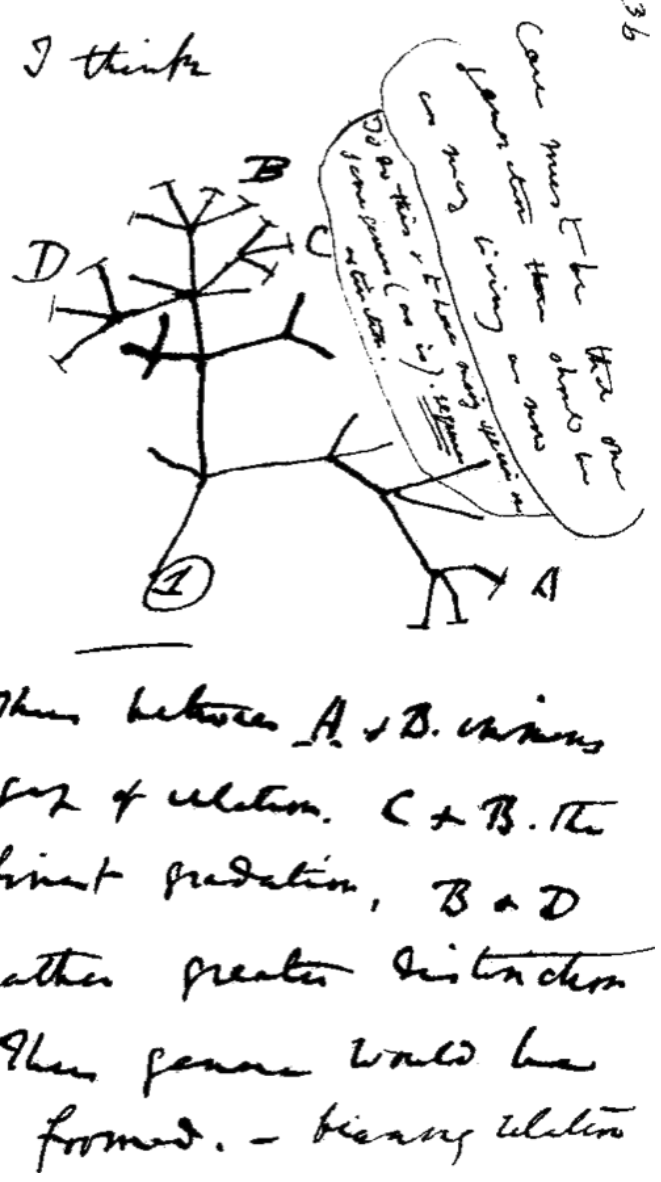
Similarity of individual features (ordering and arrangement)



Why comparative genomics?

The James Hutton Institute

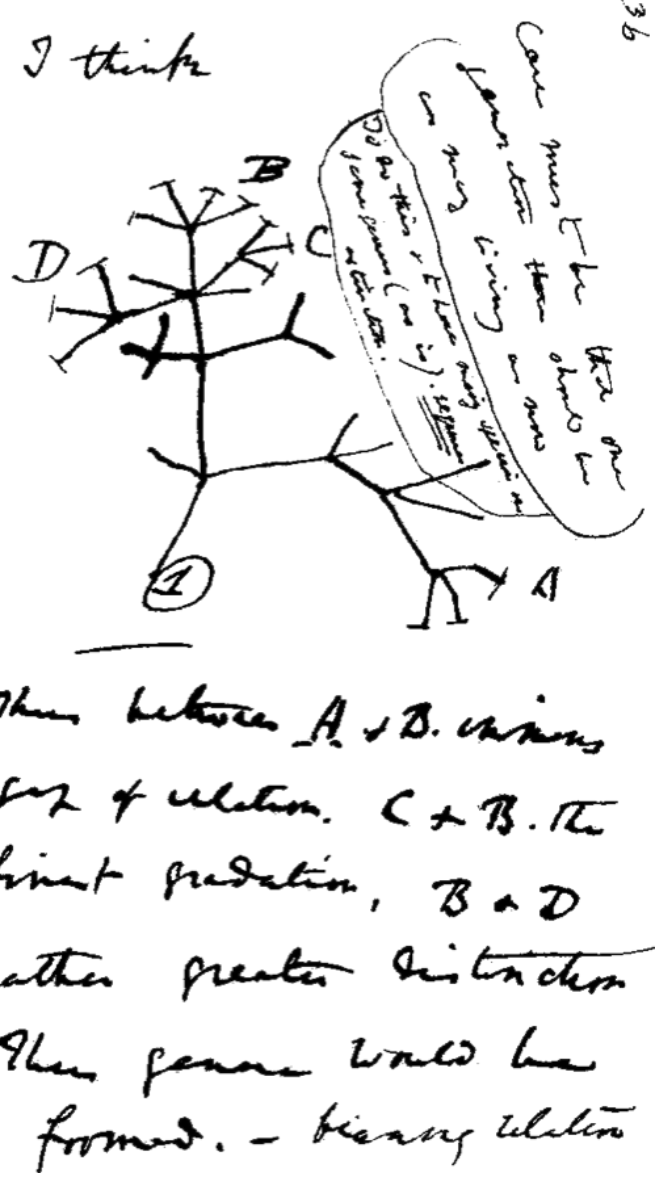
- Genome features are heritable characteristics
- Related organisms share ancestral genomes
- Related organisms inherit common genome features
- Genome similarity $\propto$ relatedness? (phylogenomics)



Why comparative genomics?

The James Hutton Institute

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



Why comparative genomics?

- Functionally-optimised elements are conserved
- (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

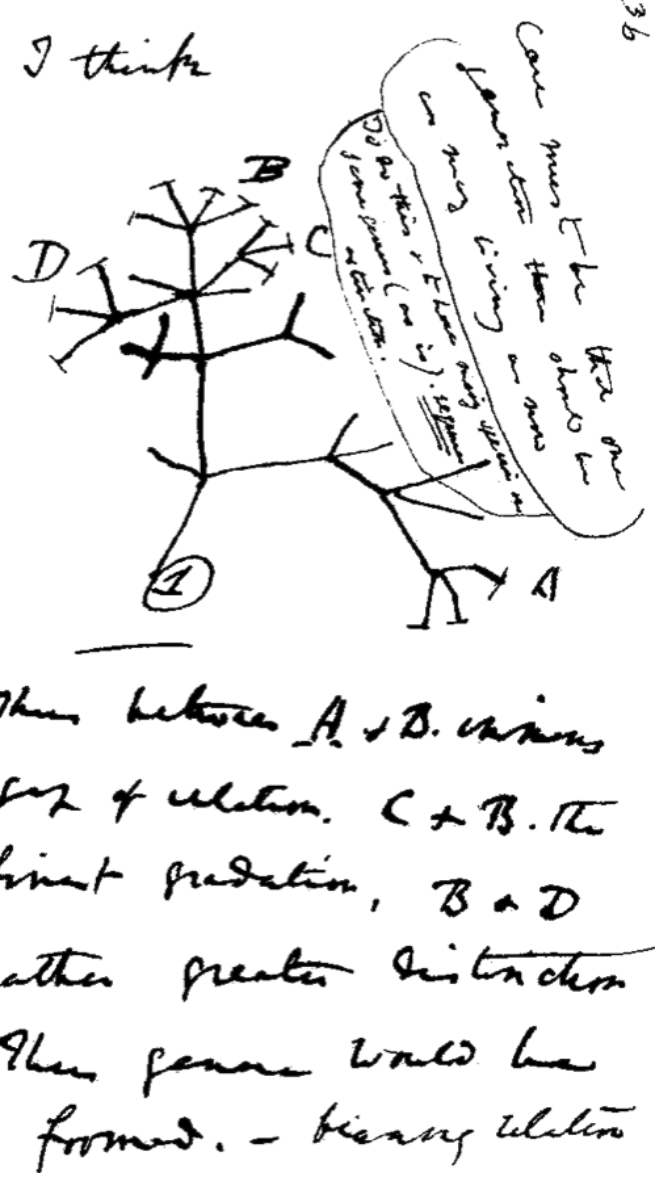


Table of Contents

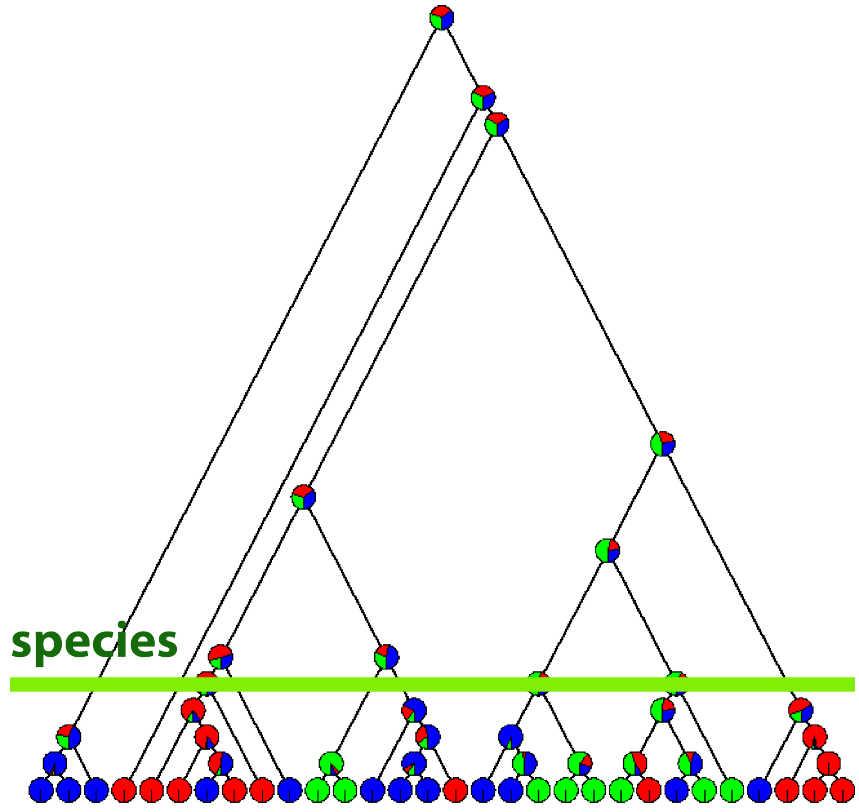
The James Hutton Institute

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?

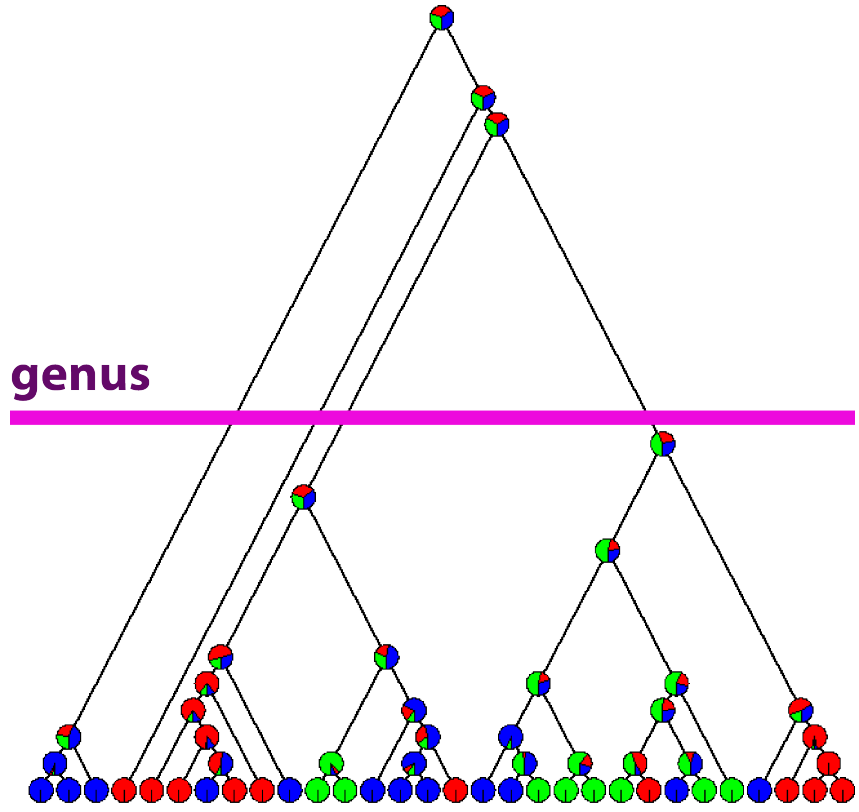


Types of comparison

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/adaptation?

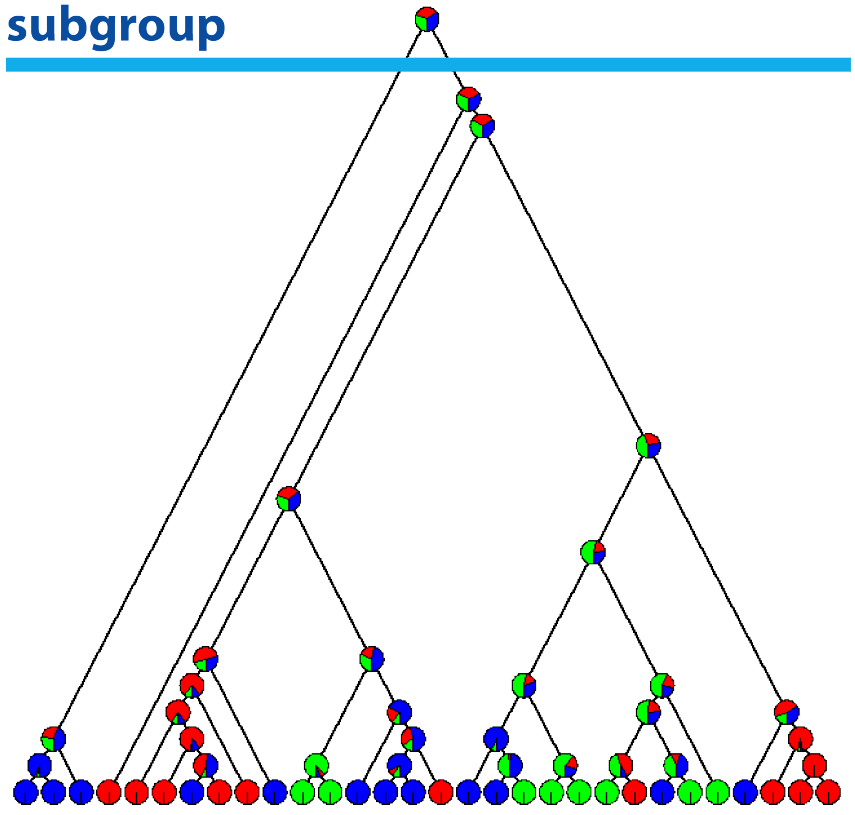


Types of comparison

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]

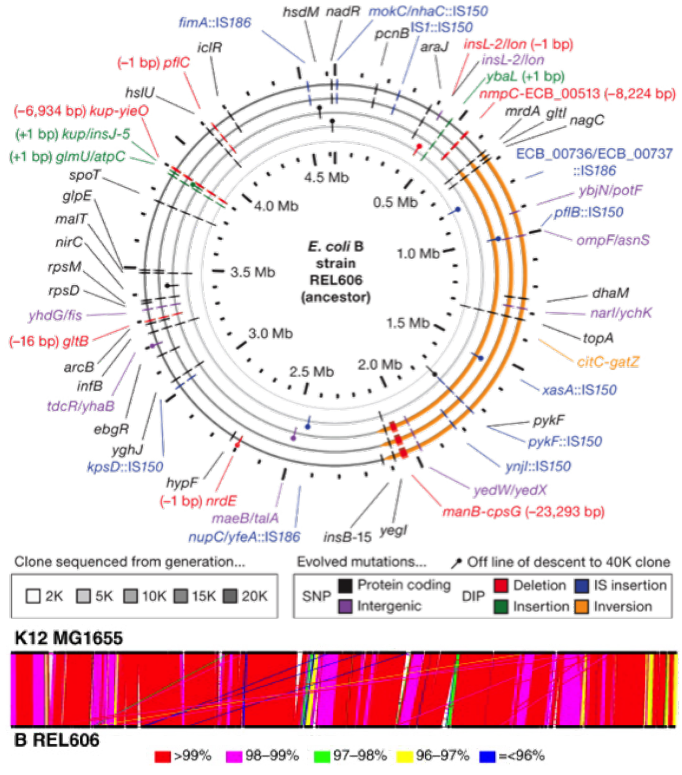
[a] Jeong *et al.* (2009) *J. Mol. Biol.* doi:10.1016/j.jmb.2009.09.052

[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

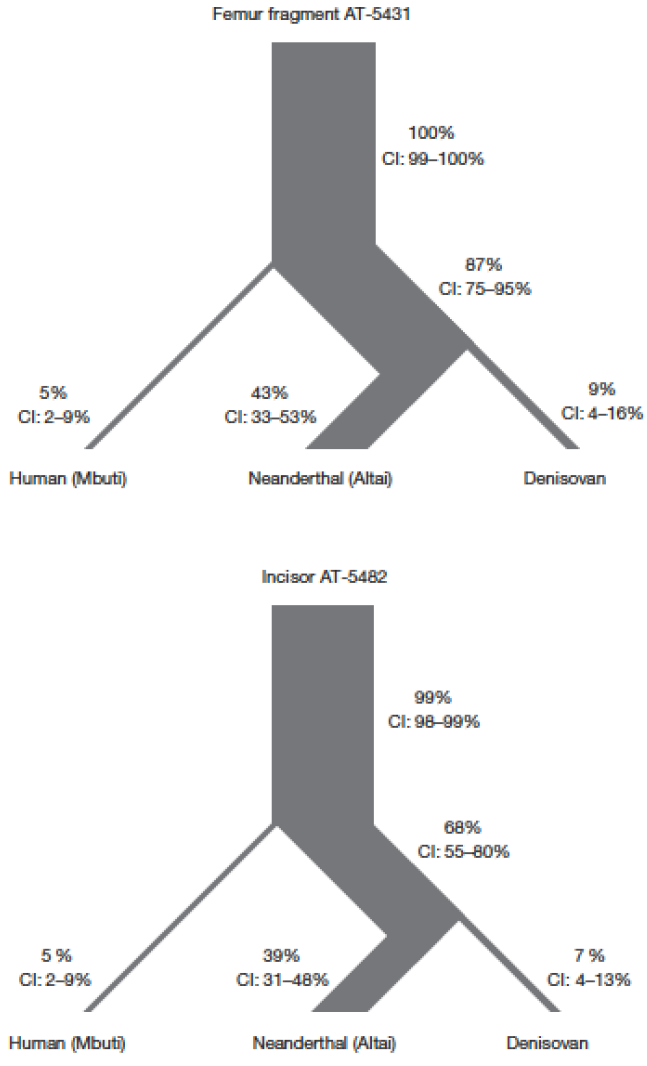


Comparative genomics in the news [a] [b]

[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?



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

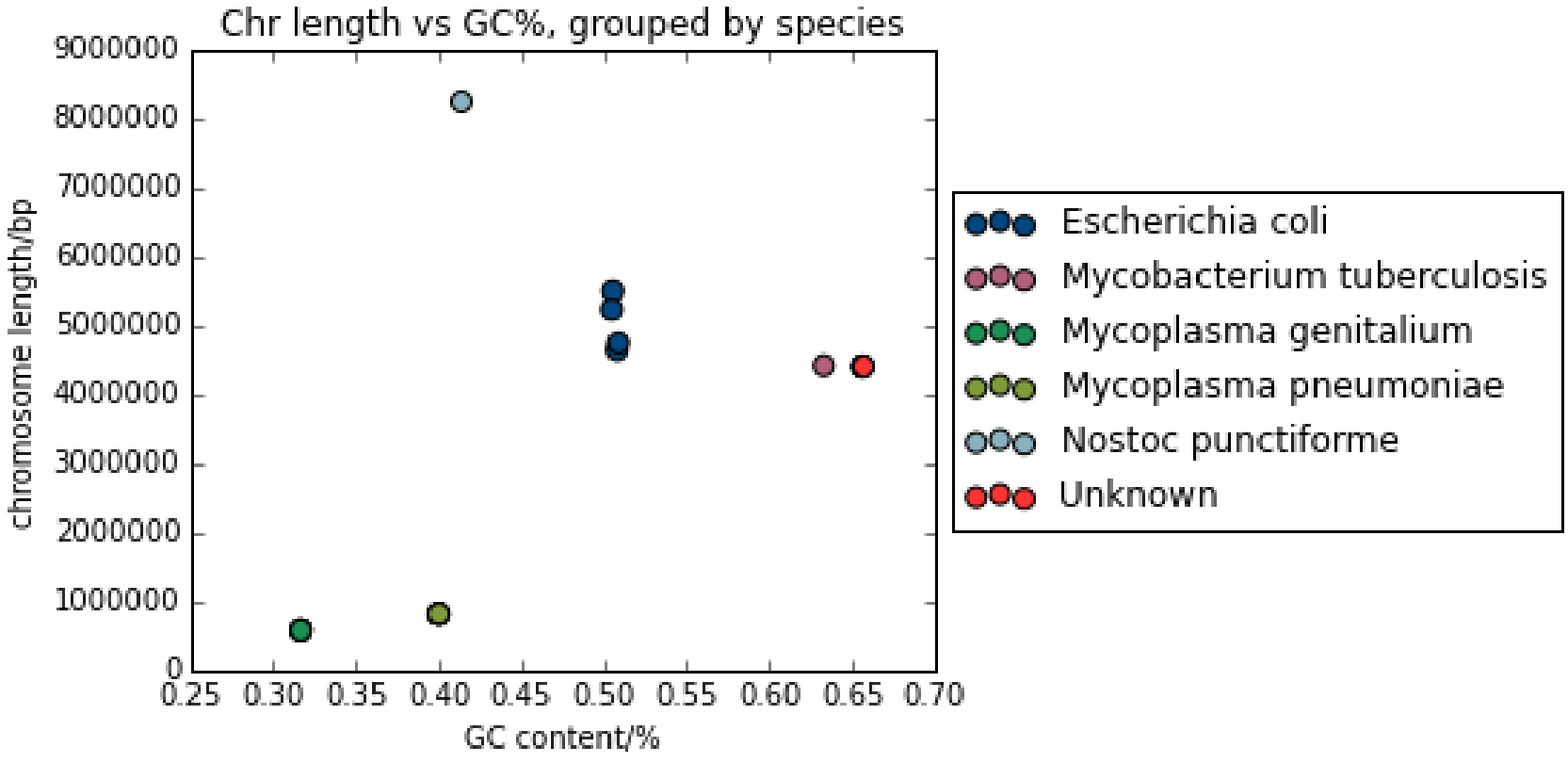
The James Hutton Institute

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 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

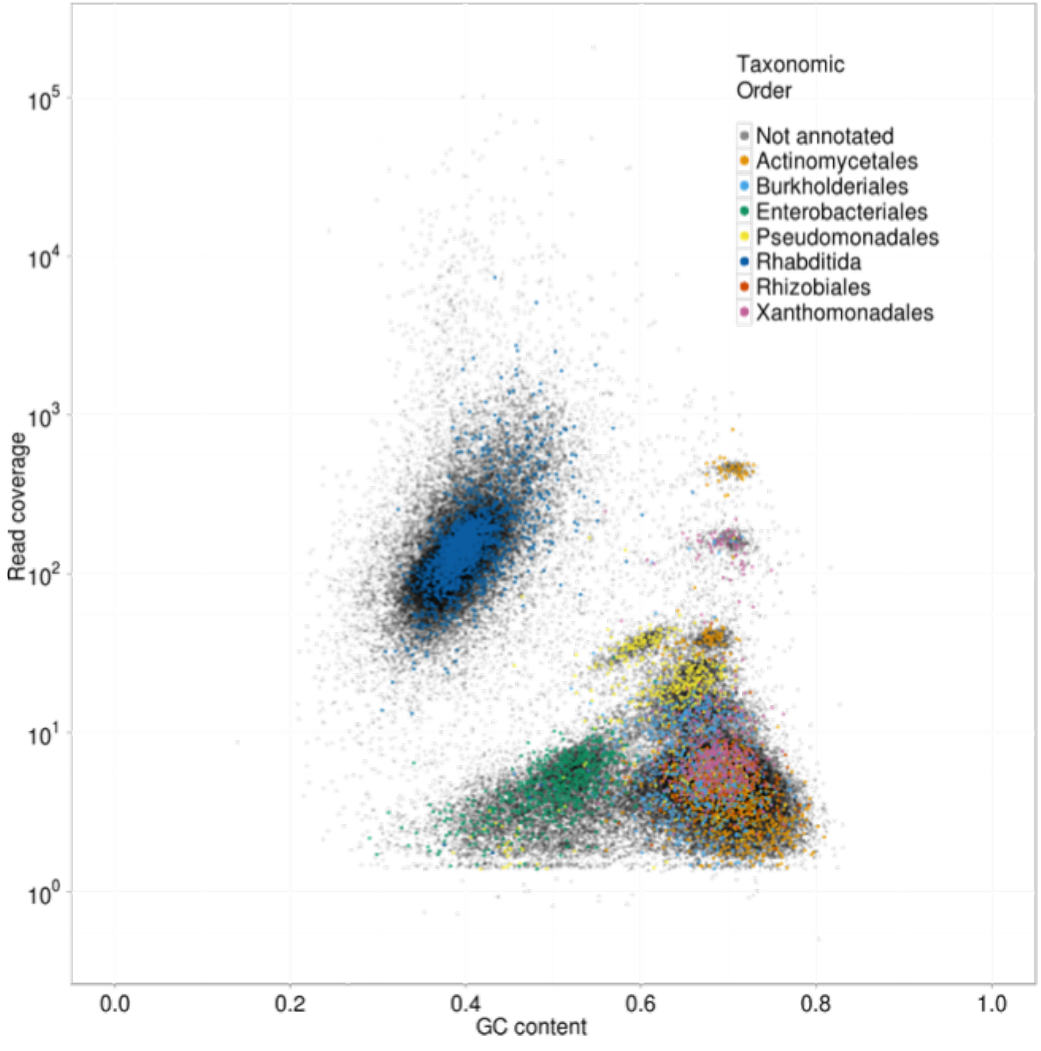
- Host and symbiont DNA have different %GC
- Host and symbiont DNA differ in coverage

- 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

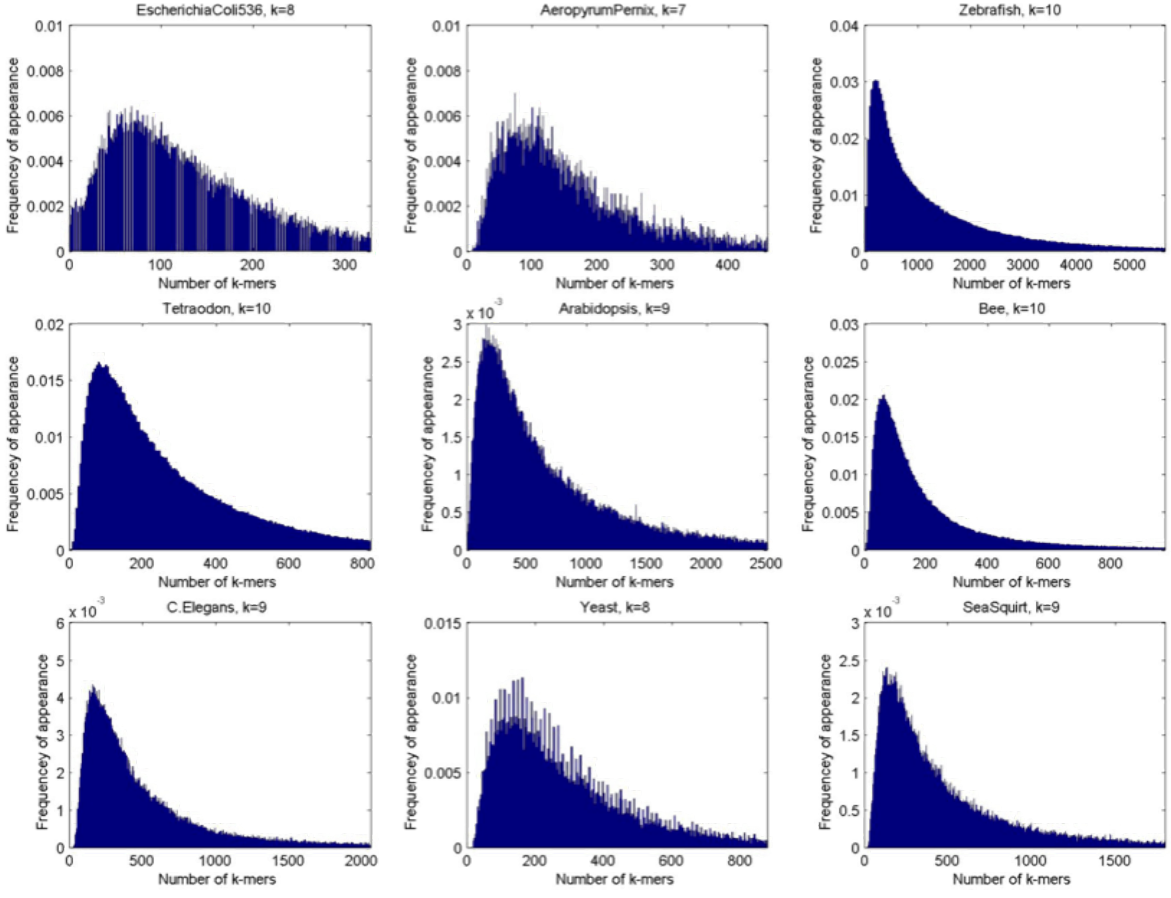
k-mer spectra [a]

[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 ($k \approx 9$)



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

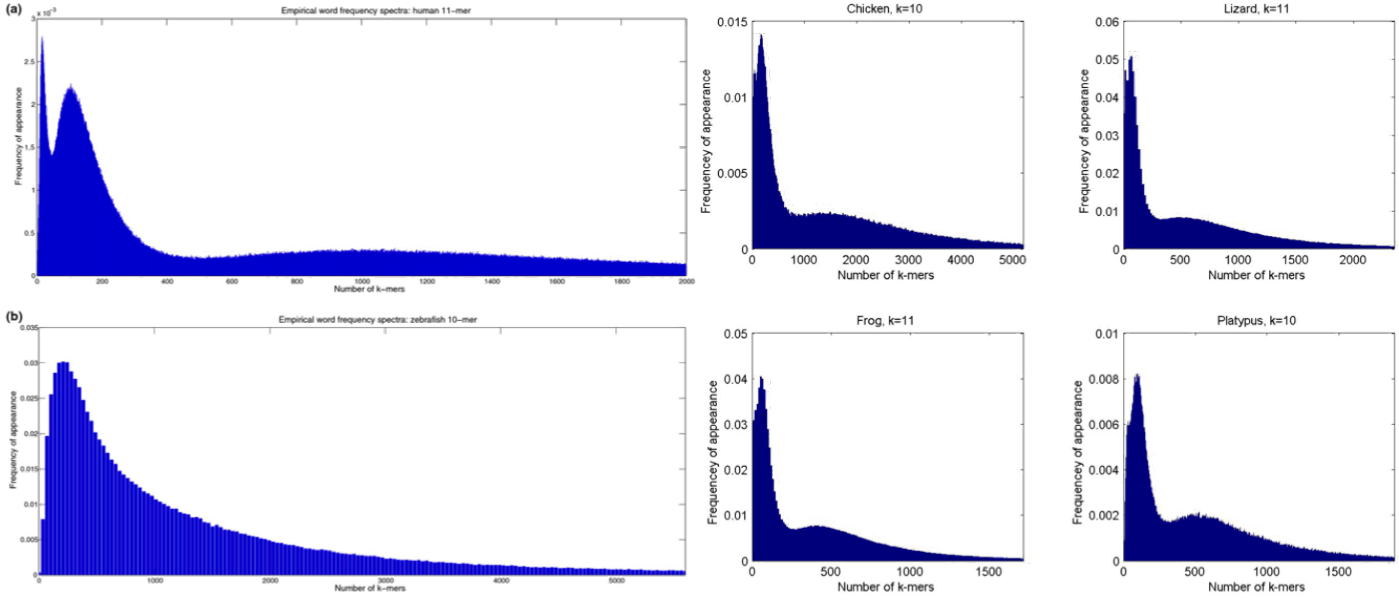


Table of Contents

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

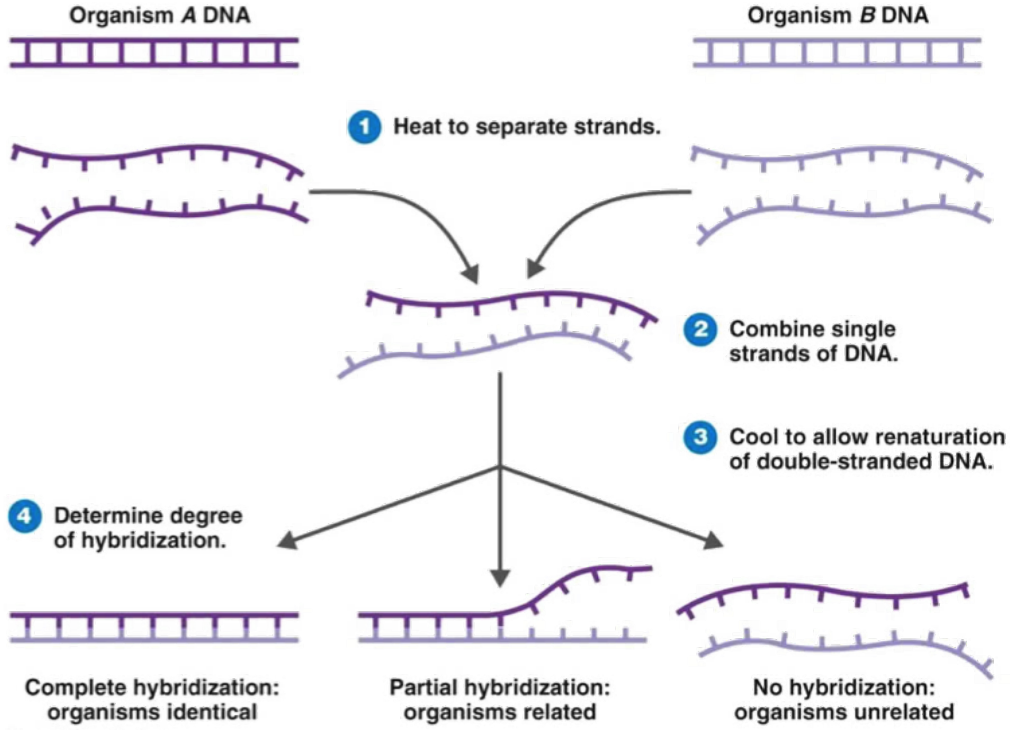
DNA-DNA hybridisation (DDH) [a]

[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

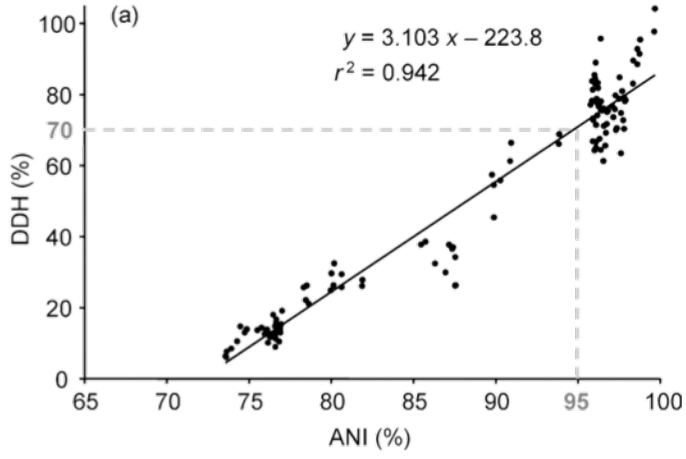
Average Nucleotide Identity (ANI) [a]

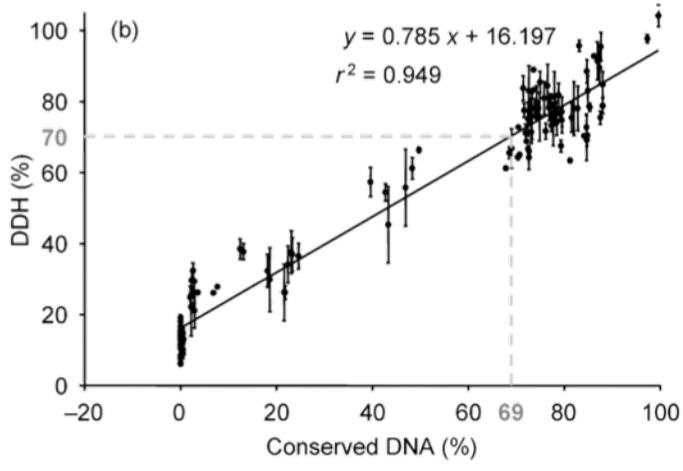
[a] Goris *et al.* (2007) *Int. J. System. Evol. Biol.* doi:10.1099/ijs.0.64483-0

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]

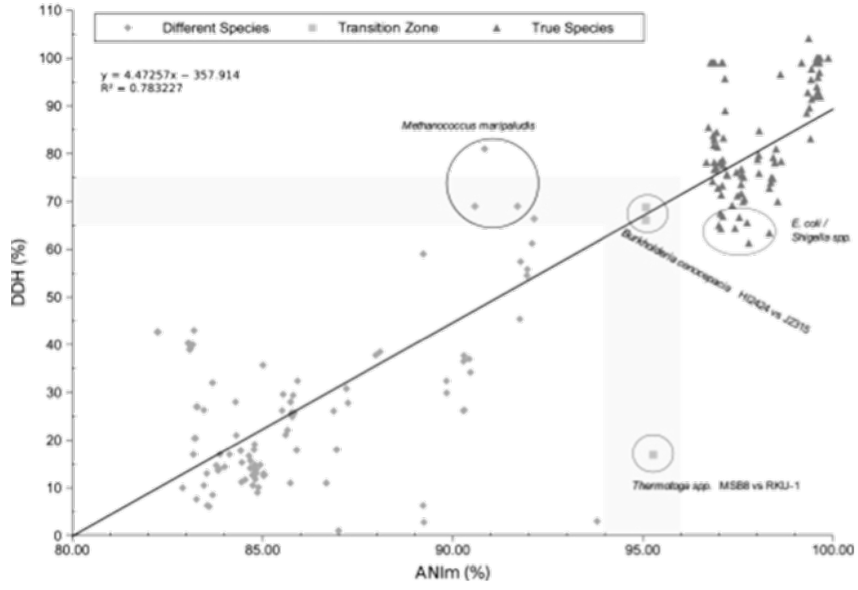
[a] Richter & Rosselló-Móra *et al.* (2009) *Proc. Natl. Acad. Sci. USA* doi:10.1073/pnas.0906412106

The James Hutton Institute

ANIm and TETRA variants introduced in 2009:

ANIm

1. Align sequences with NUCmer
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

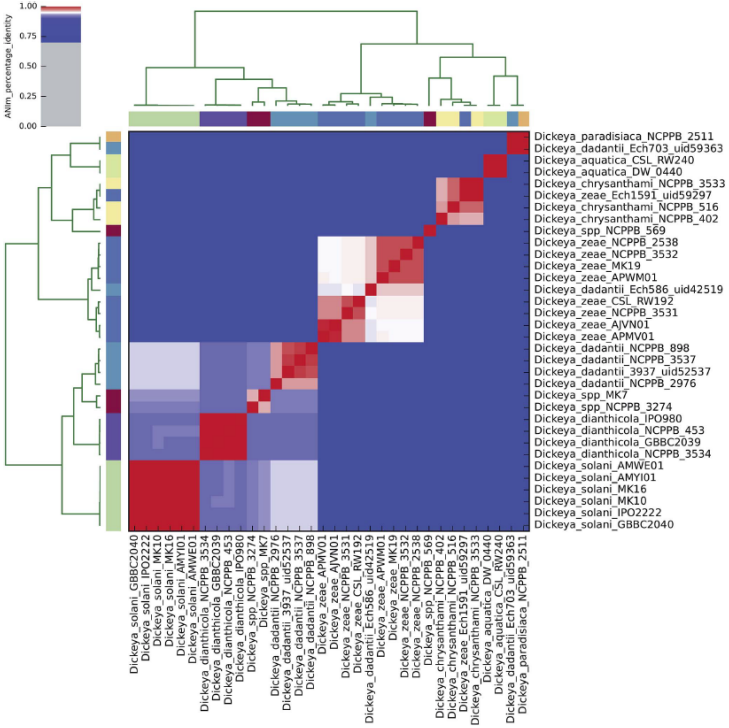
ANI in practice [a] [b]

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

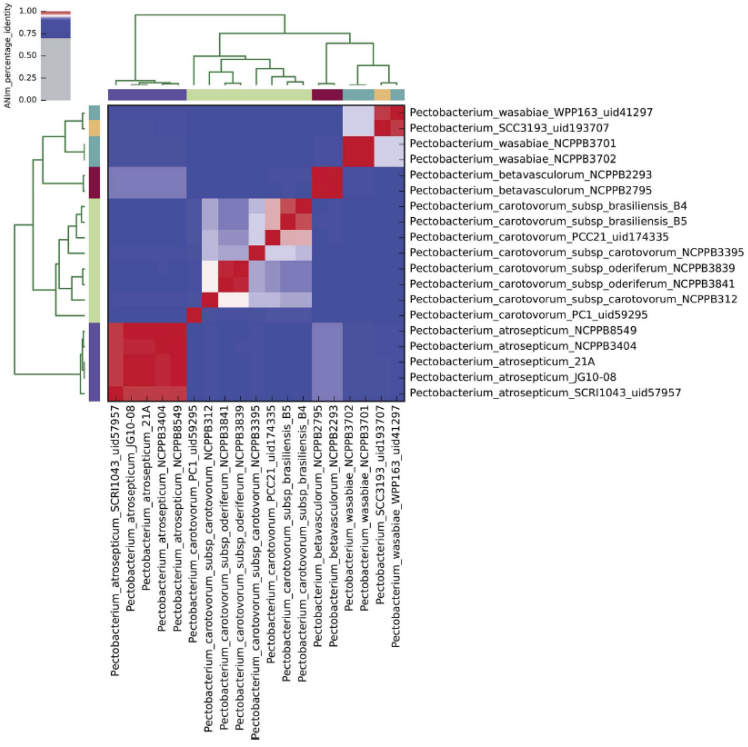
[b] Pritchard *et al.* (2016) *Anal. Methods* **8**:12-24 doi:10.1039/C5AY02550H

The James Hutton Institute

Dickeya species structure



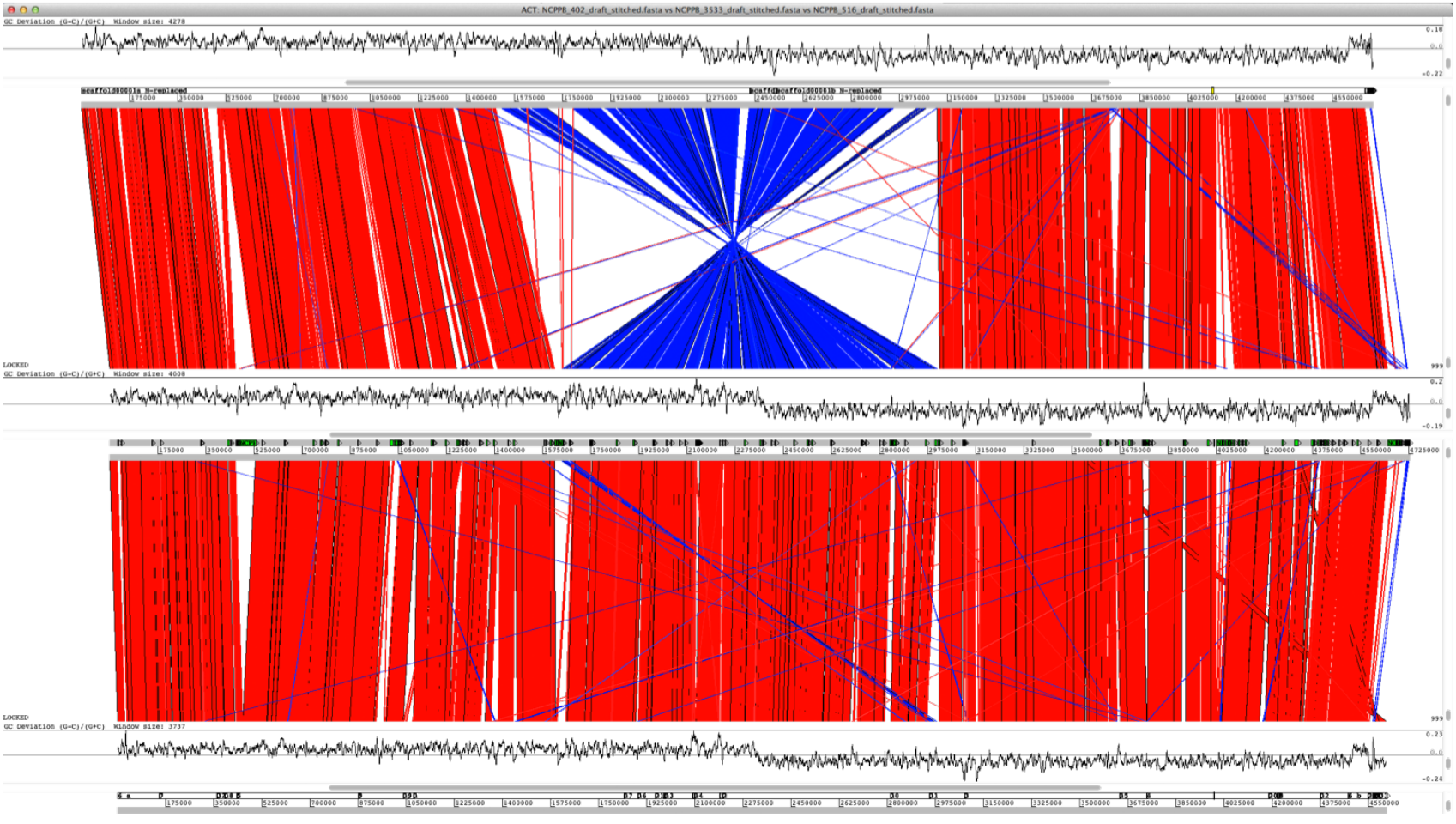
Pectobacterium species structure:



Pairwise genome alignments

The James Hutton Institute

Pairwise comparisons require alignment of similar regions.

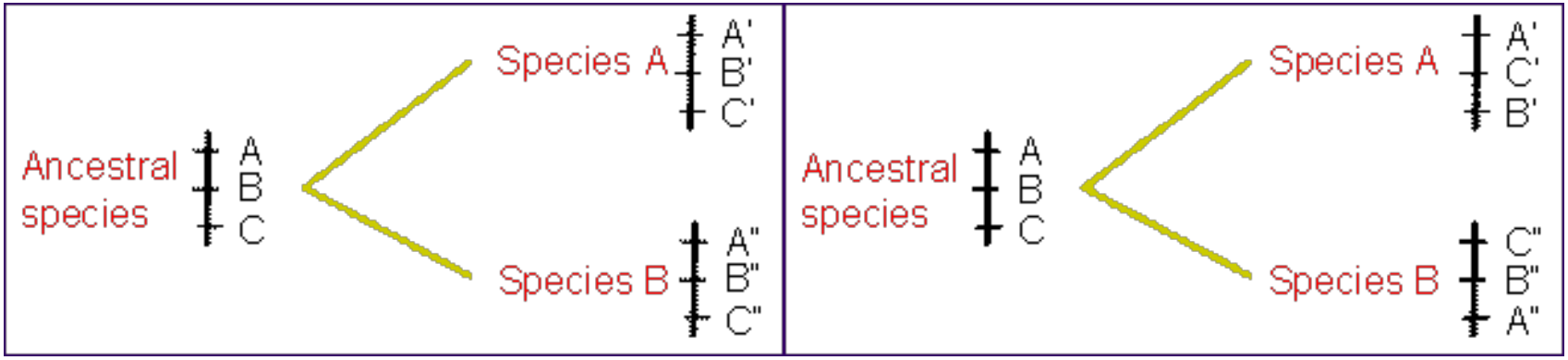


Synteny and Collinearity

The James Hutton Institute

Genome rearrangements may occur post-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 a functional constraint (and help determine *orthology*)

Alignment algorithms/programs

The James Hutton Institute

I assume you're familiar with BLAST...

BLASTN and naïve alignment algorithms are not appropriate for whole-genome alignment:

- 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

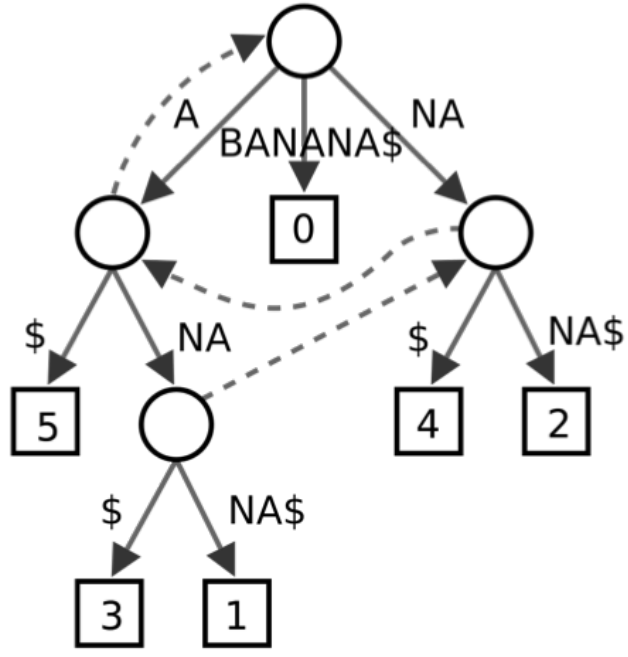
Conceptually completely different to BLAST/BLAT/megaBLAST

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$`



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

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.

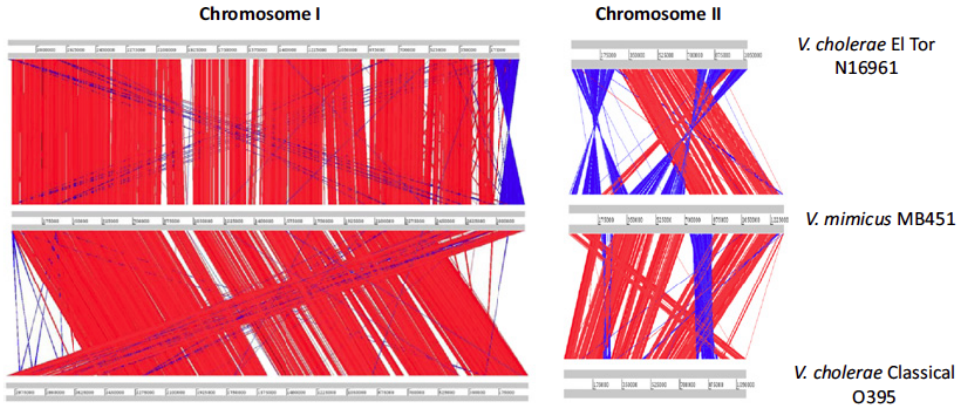


Fig. 2. Linear pairwise comparison of the *Vibrio mimicus* genome by Artemis Comparison Toll. Regions with similarity are highlighted by connecting red or blue lines between the genomes; red lines indicate homologous blocks of sequence, and blue lines indicate inversions. Gaps indicate unique DNA. The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

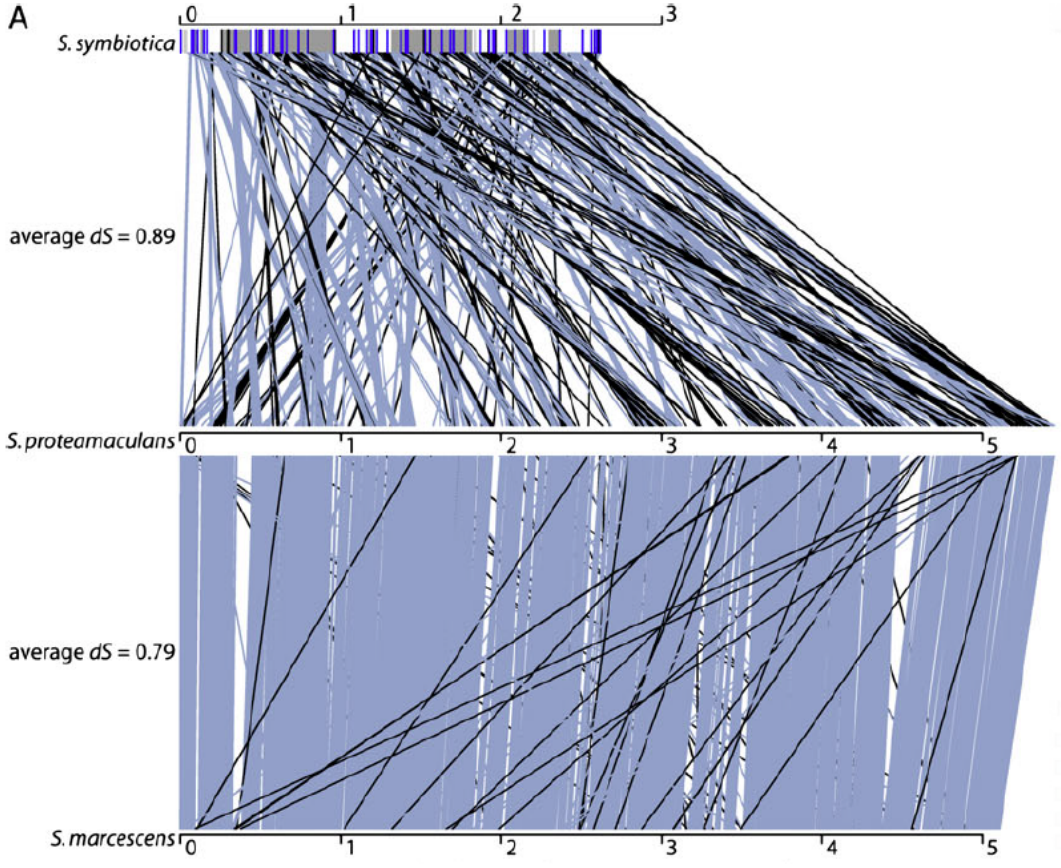
Suggests modularity of genome organisation, as a mechanism for adaptation (HGT, two-speed genome).

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

Many tools use either positional homology or glocal alignment

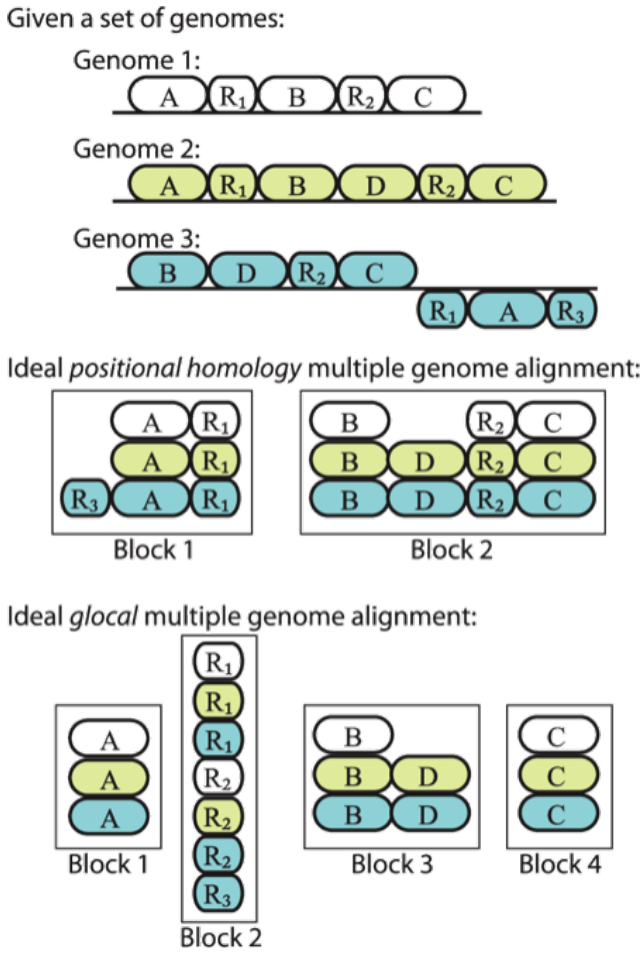
Several tools available:

- 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/)



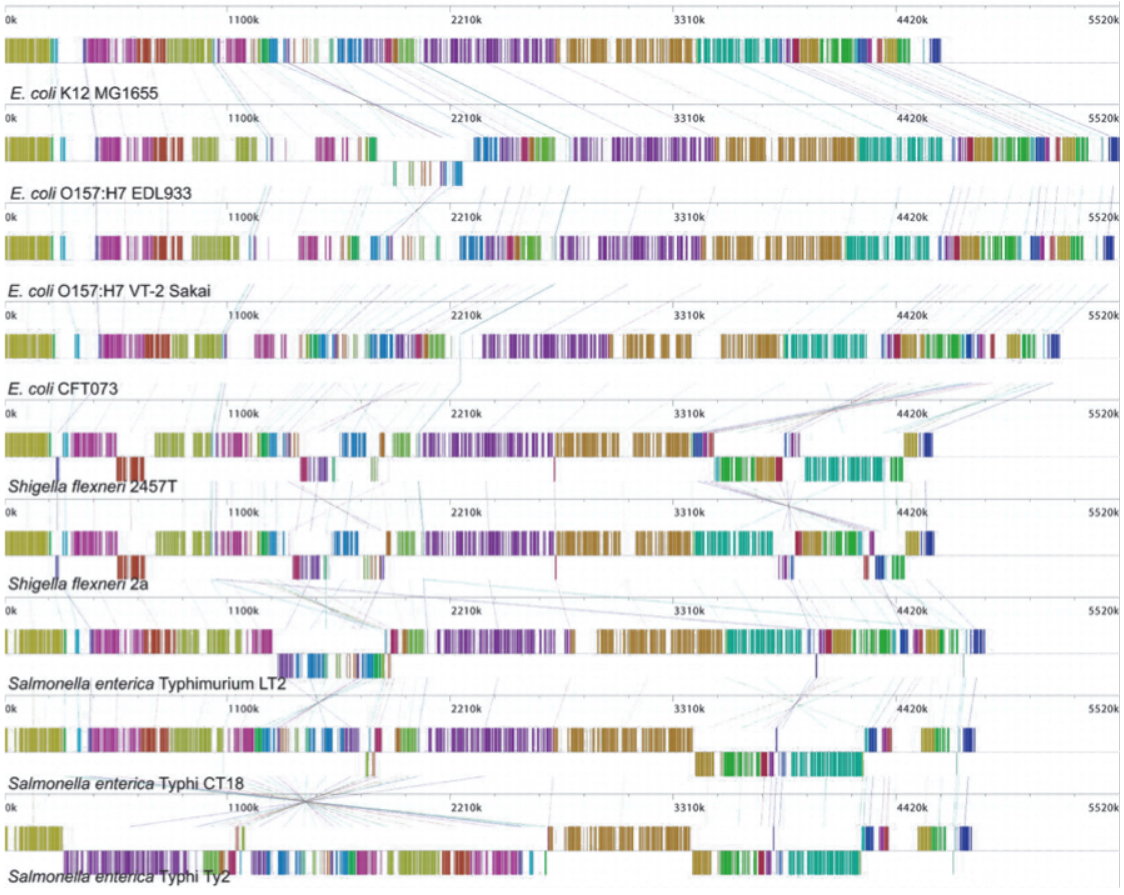
MAUVE [a]

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

The James Hutton Institute

MAUVE alignment of nine enterobacterial genomes

Evidence for rearrangement of homologous backbone sequence



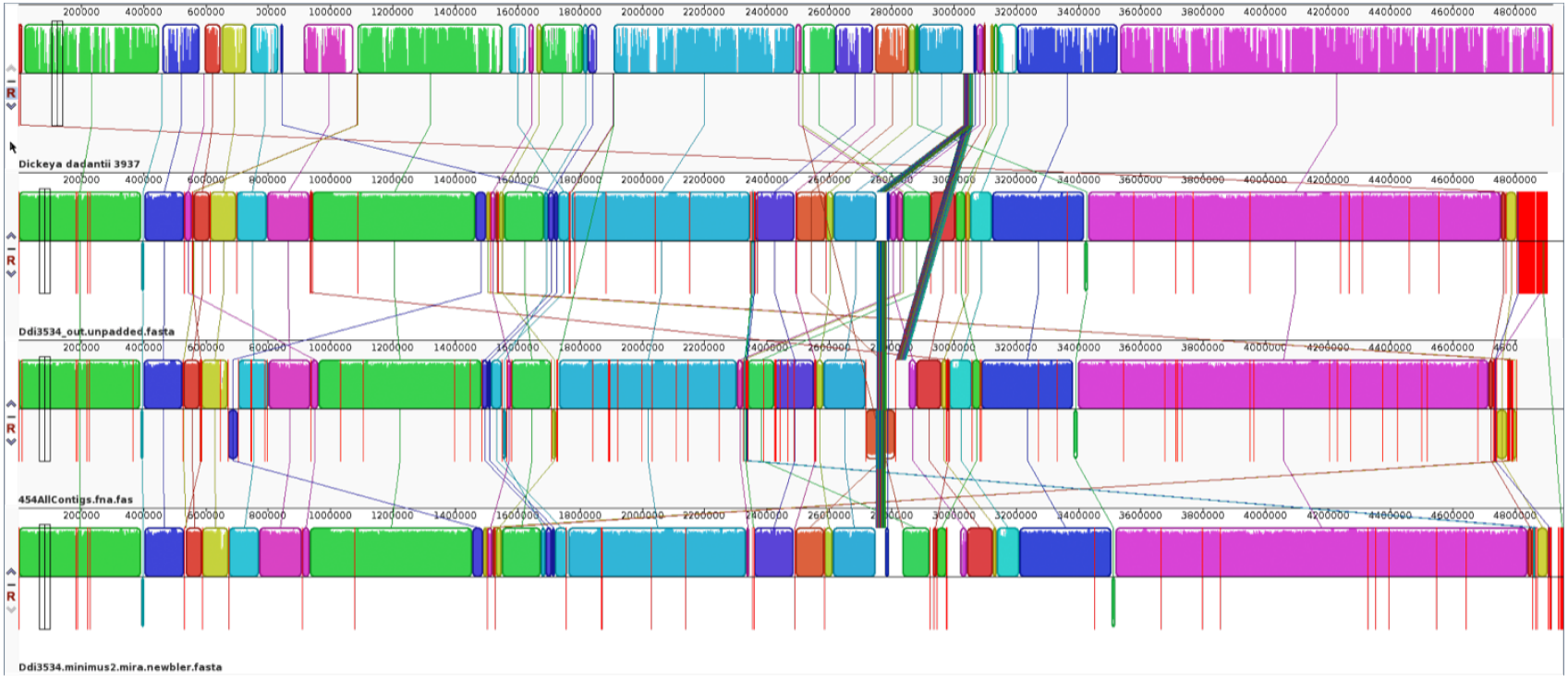
Draft genome alignment

The James Hutton Institute

High-throughput genome assemblies are often fragmented (contigs)

Contigs can be ordered (*scaffolded*):

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

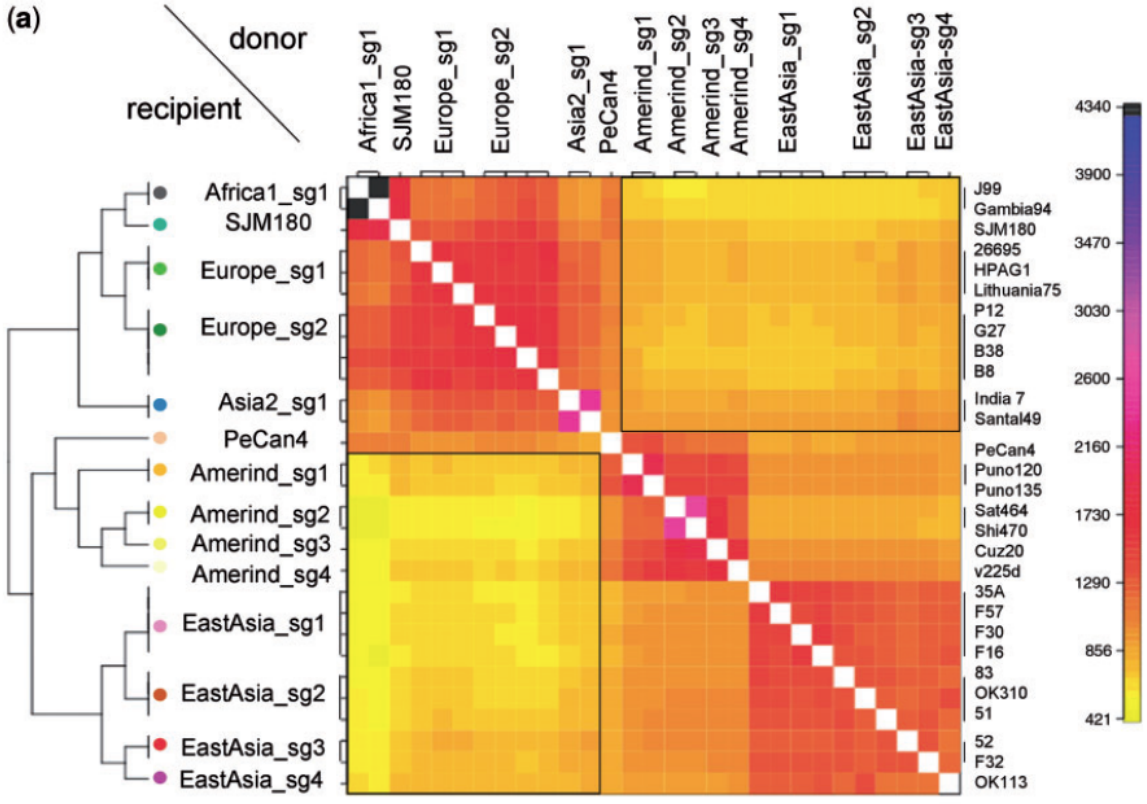
"Chromosome painting" infers recombination-derived 'chunks'
Genome's haplotype constructed in terms of recombination events from a 'donor' to a 'recipient' genome

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.



Whole Genome Comparisons

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

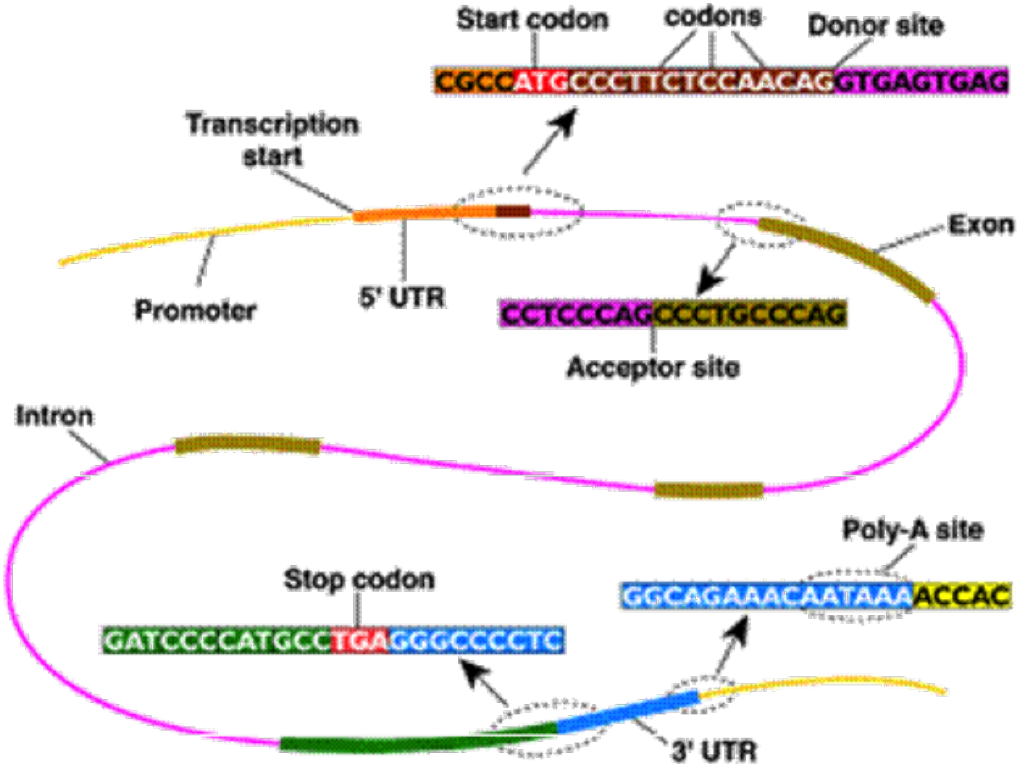
Table of Contents

The James Hutton Institute

Gene features

Significant substructure, especially in eukaryotes

- translation start
- introns
- exons
- translation stop
- translation terminator

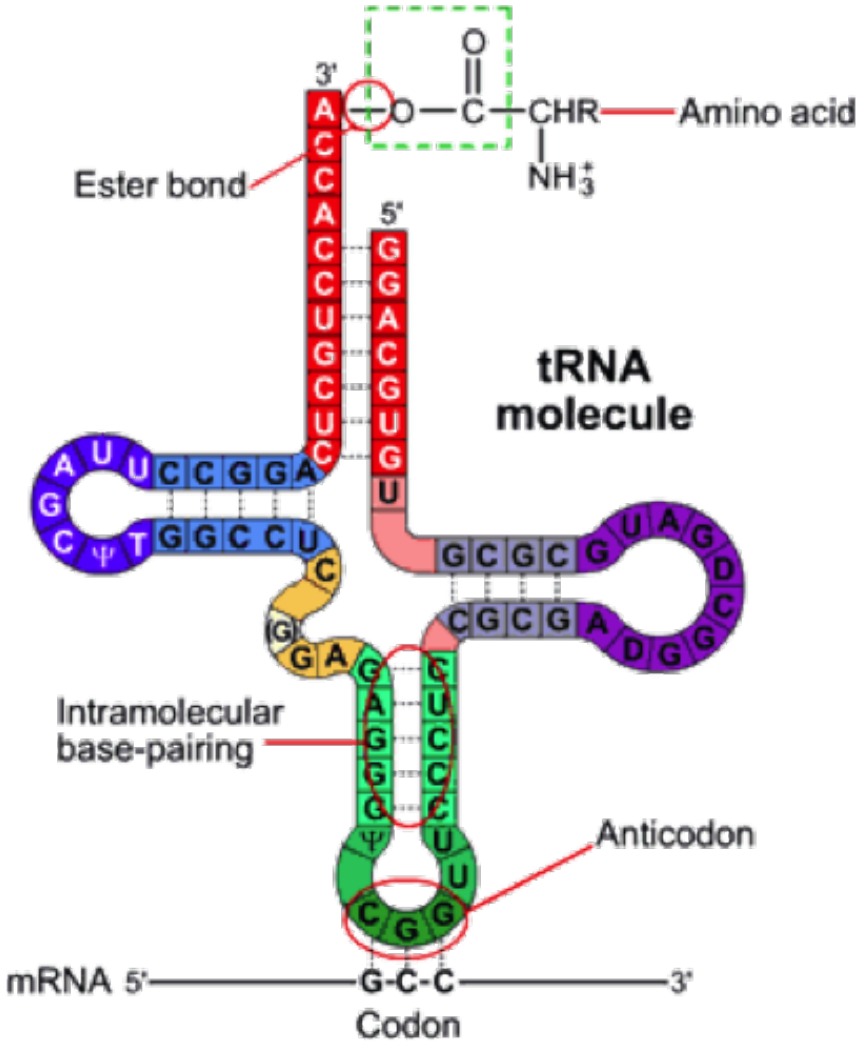


RNA features

The James Hutton Institute

RNA/ncRNA: characterised by complex secondary structure

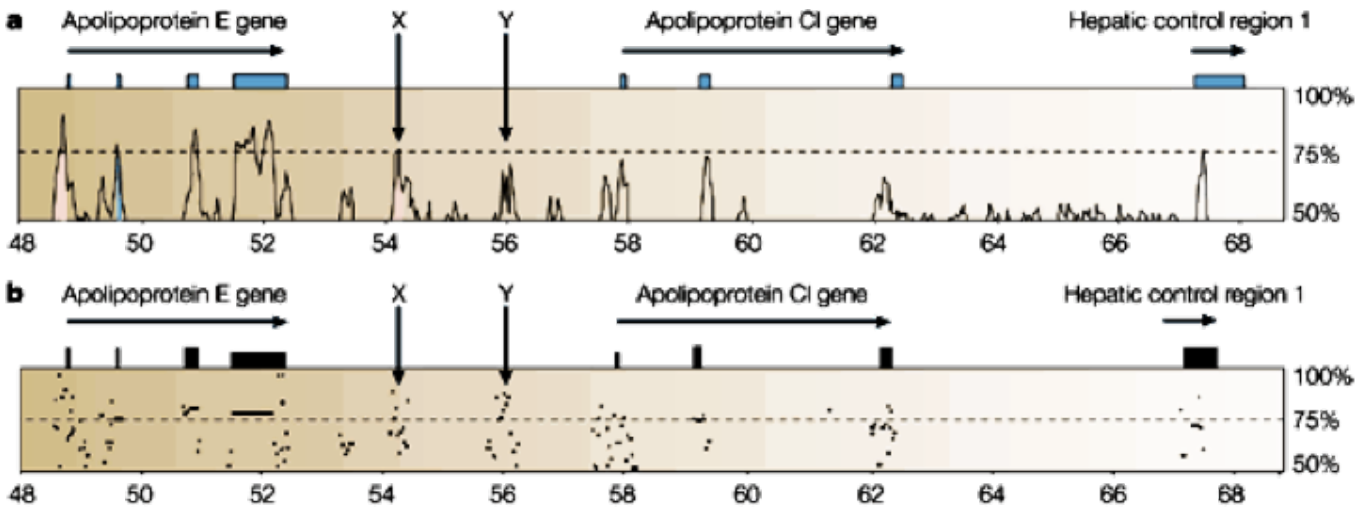
- tRNA - transfer RNA
- rRNA - ribosomal RNA
- CRISPRs - prokaryotic defence, and genome editing
- many other functional classes, including enhancers



Regulatory features [a]

[a] Pennacchio & Rubin (2001) *Nature Rev. Genet.* doi:10.1038/35052548

- transcription start sites (TSS)
- RNA polymerase (RNAp) binding sites
- transcription factor binding sites (TFBS)
- core, proximal and distal promoter regions



Nature Reviews | Genetics

human vs mouse comparison

Gene finding [a] [b] [c]

[a] Liang *et al.* (2009) *Genome Res.* doi:10.1101/gr.088997.108
[b] Brent (2007) *Nat. Biotech.* doi:10.1038/nbt0807-883
[c] Korf (2004) *BMC Bioinf.* doi:10.1186/1471-2105-5-59

The James Hutton Institute

At genome scales, we need to automate functional prediction

Empirical (evidence-based) methods:

- Inference from known protein/cDNA/mRNA/EST sequence
- Interference from mapped RNA reads (e.g. RNAseq)

Ab initio methods:

- Prediction on the basis of gene features (TSS, CpG islands, Shine-Dalgarno sequence, stop codons, nucleotide composition, etc.)

Inference from genome comparisons/sequence conservation

Regulatory element finding [a] [b] [c]

[a] Zhang *et al.* (2011) *BMC Bioinf.* doi:10.1186/1471-2105-12-238
[b] Kilic *et al.* (2013) *Nucl. Acids Re.* doi:10.1093/nar/gkt1123
[c] Vavouri & Elgar (2005) *Curr. Op. Genet. Deve.* doi:10.1016/j.gde.2005.05.002

The James Hutton Institute

Empirical (evidence-based) methods:

- Inference from protein-DNA binding experiments
- Interference from co-expression

Ab initio methods:

- Identification of regulatory motifs (profile/other methods; TATA, σ-factor binding sites, etc.)
- Statistical overrepresentation of motifs
- Identification from sequence properties

Inference from genome comparisons/sequence conservation

Equivalent genome features

The James Hutton Institute

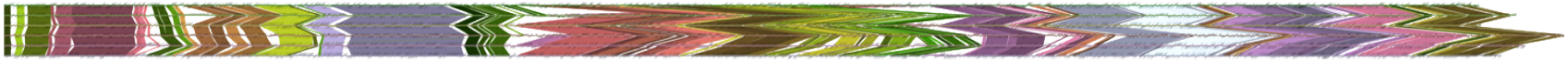
When comparing two features (e.g. genes) between two or more genomes, there must be some basis for making the comparison. They have to be *equivalent* in some way, such as:

- common evolutionary origin
- functional similarity
- a family-based relationship

It's common to define equivalence of genome features in terms of evolutionary relationship.

Why look at equivalent features?

The James Hutton Institute



The real power of genomics is comparative genomics!

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

Who let the -logues out?

The James Hutton Institute

Genome features can have complex evolutionary relationships

We have precise terms to describe these relationships

The -logues drop [a]

[a] Fitch *et al.* (1970) *Syst. Zool.* doi:10.2307/2412448

The James Hutton Institute

How do we understand the relationships between features in more than one genome?

- Functional similarity: **analogy**
- Evolutionary common origin: **homology, orthology, etc.**
- Evolutionary/functional/family relationship: **paralogy**

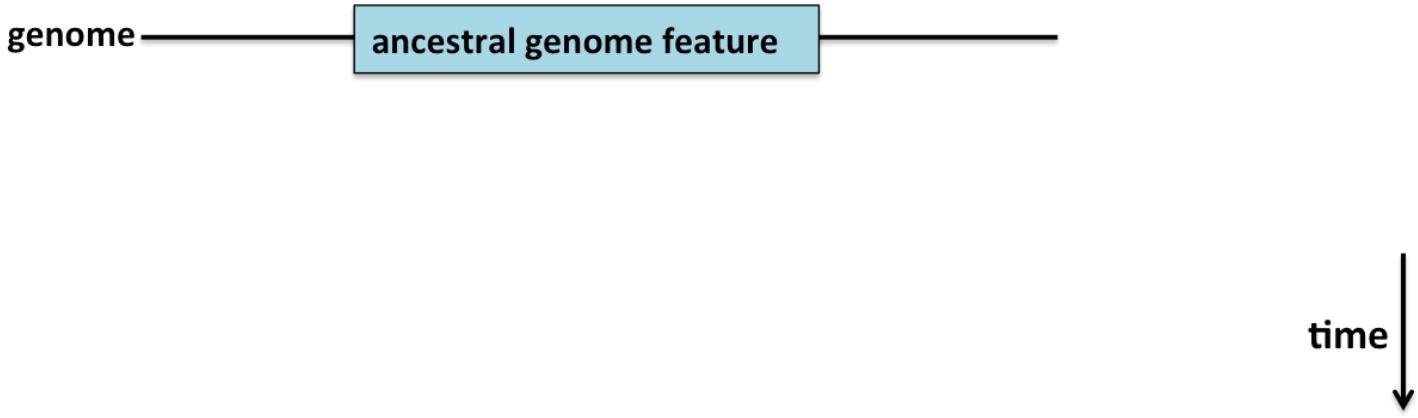
DISTINGUISHING HOMOLOGOUS FROM ANALOGOUS PROTEINS

WALTER M. FITCH

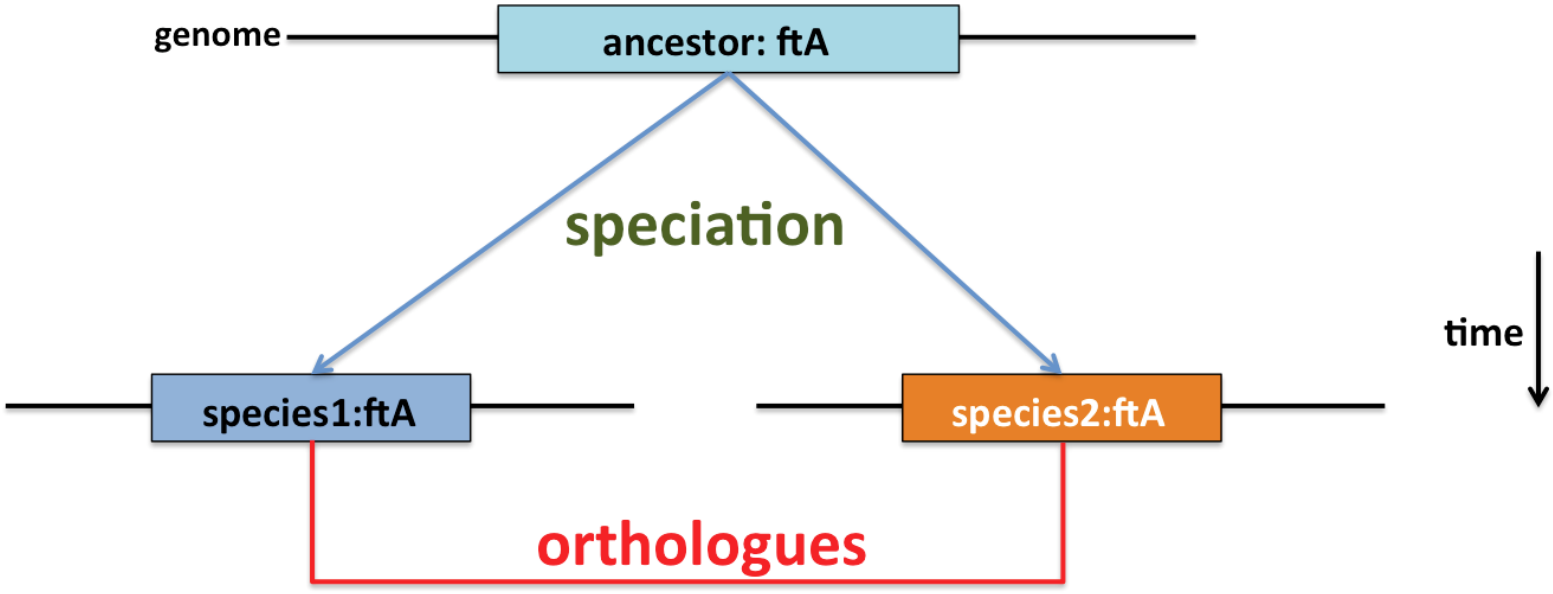
Abstract

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

Who let the -logues out?

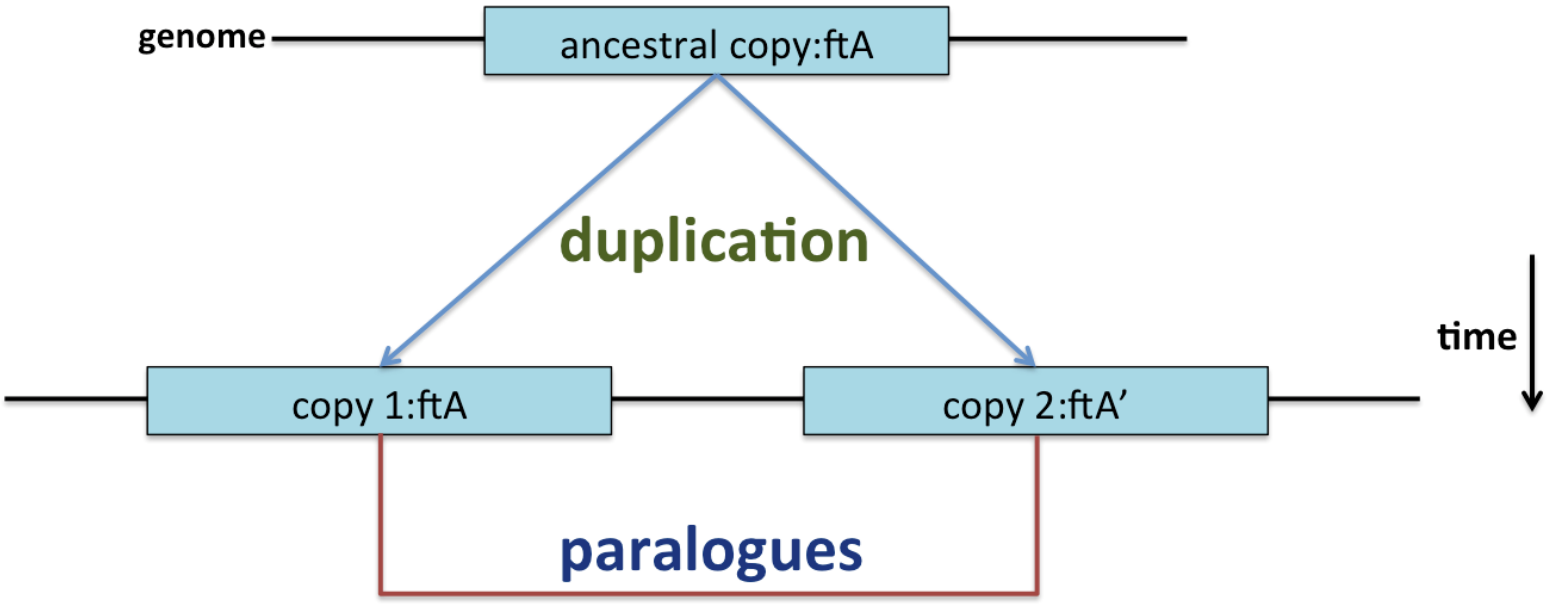


Who let the -logues out?



- **Orthologues:** homologues that diverged through speciation

Who let the -logues out?



Paralogues: homologues that diverged through duplication within the same genome

Orthology [a]

[a] Storm & Sonnhammer (2002) *Bioinformatics* doi:10.1093/bioinformatics/18.1.92

The James Hutton Institute

- Frequently abused/misused as a term
- "Orthology" is an evolutionary relationship, bent into service as a functional descriptor
- Orthology is strictly defined *only for two species or clades*! (cf. OrthoMCL)
- Orthology is not transitive:
  (A is an orthologue of C, and B is an orthologue of C, does **not** imply that A is an orthologue of B

All classifications of orthology/paralogy are inferences!

The Ortholog Conjecture [a] [b]

[a] Nehrt *et al.* (2011) *PLoS Comp. Biol.* doi:10.1371/journal.pcbi.1002073
[b] Chen *et al.* (2012) *PLoS Comp. Biol.* doi:10.1371/journal.pcbi.1002784

The James Hutton Institute

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

Why focus on orthologues? [a] [b] [c]

[a] Chen and Zhang (2012) *PLoS Comp. Biol.* doi:10.1371/journal.pcbi.1002784
[b] Dessimoz (2011) *Brief. Bioinf.* doi:10.1093/bib/bbr057
[c] Altenhoff and Dessimoz (2009) *PLoS Comp. Biol.* **5**:e1000262 doi:10.1371/journal.pcbi.1000262

Formalisation of the idea of *corresponding genes* in different organisms.

Orthologues serve two purposes:

- **Evolutionary equivalence**
- **Functional equivalence** ("The Ortholog Conjecture")

Applications in comparative genomics, functional genomics and phylogenetics.

Over 30 databases attempt to describe orthologous relationships (http://questfororthologs.org/orthology_databases)

Finding orthologues [a] [b] [c]

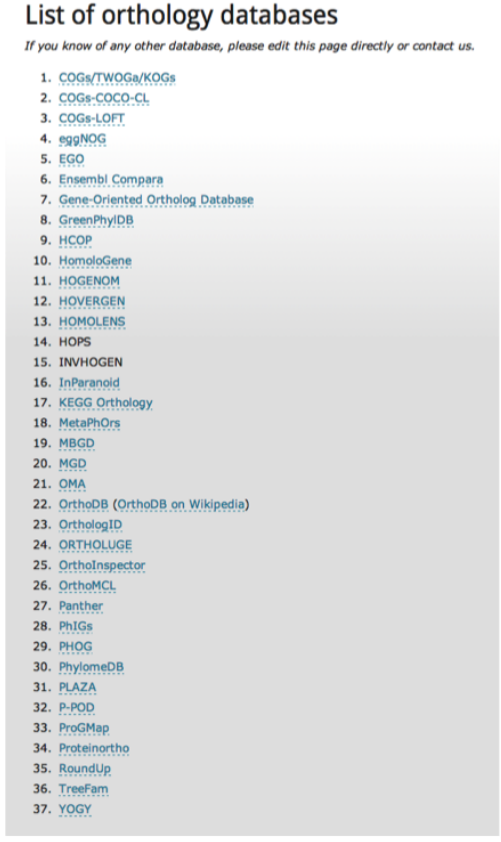
[a] Kristensen *et al.* (2011) *Brief. Bioinf.* **12**:379-391 doi:10.1093/bib/bbr030

[b] Trachana *et al.* (2011) *Bioessays* **33**:769-780 doi:10.1002/bies.201100062

[c] Salichos and Rokas (2011) *PLoS One* **6**:e18755 doi:10.1371/journal.pone.0018755.g006

Multiple methods and databases

- **Pairwise genome**
  - RBBH (aka BBH, RBH), RSD, InParanoid, RoundUp
- **Multi-genome**
  - *Graph-based*: COG, eggNOG, OrthoDB, OrthoMCL, OMA, MultiParanoid
  - *Tree-based*: TreeFam, Ensembl Compara, PhylomeDB, LOFT



List of orthology databases

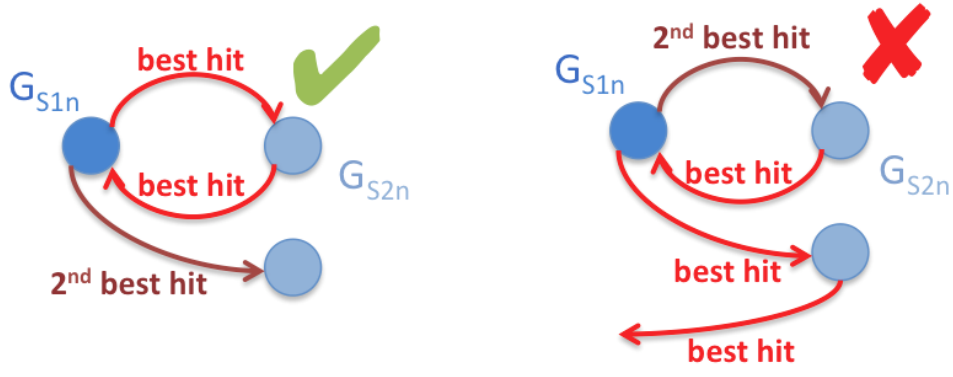
If you know of any other database, please edit this page directly or contact us.

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

Reciprocal Best BLAST Hits [a]

[a] On Reciprocal Best BLAST Hits 19/7/2012

- S1, S2 are the gene sequence sets from two organisms
- Use sequence search tool (BLAST/FASTA):
  - Query=S1, Subject=S2
  - Query=S2, Subject=S1



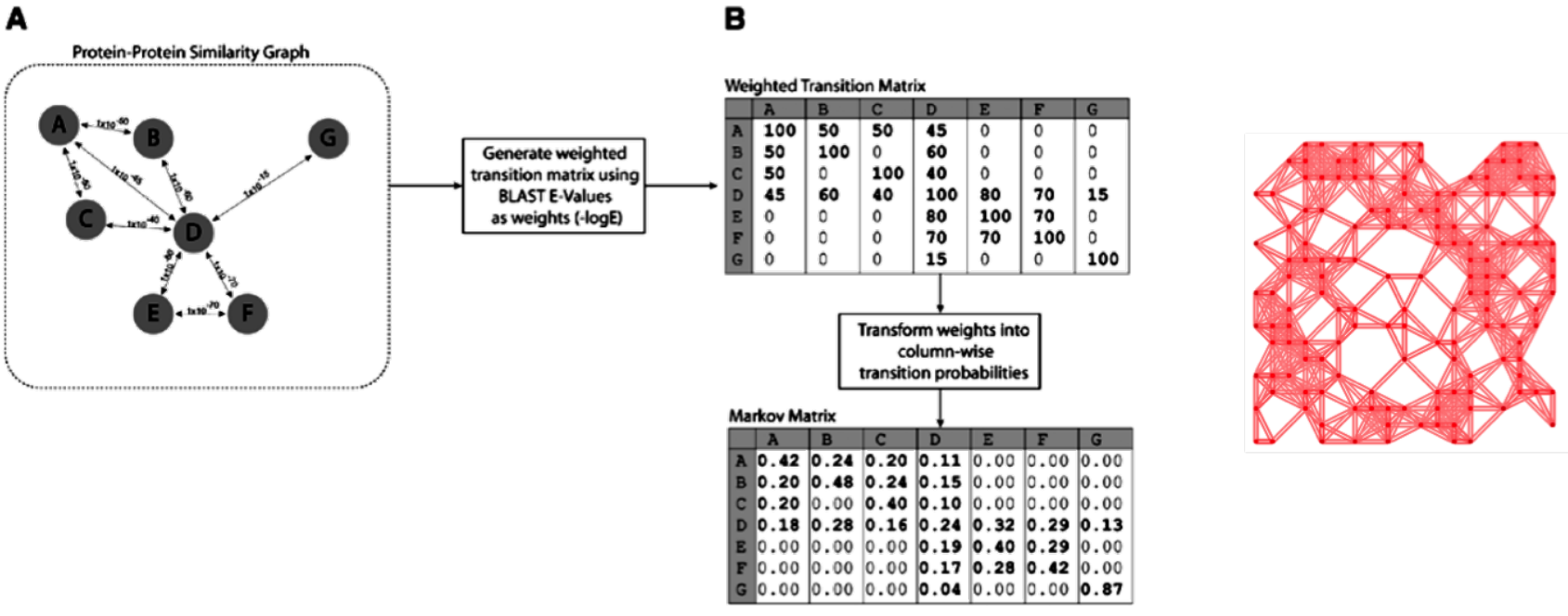
- Optionally filter hits (e.g. on %identity and %coverage)
- Find all pairs of sequences {G_{S1n}, G_{S2n}} in S1, S2 where G_{S1n} is the best BLAST match to G_{S2n} and G_{S2n} is the best BLAST match to G_{S1n}.

MCL [a]

[a] Enright *et al.* (2002) *Nucl. Acids Res.* doi:10.1093/nar/30.7.1575

The James Hutton Institute

- MCL constructs a network (*graph*) from all-against-all BLAST results
- Matrix operations (*expansion*, *inflation*) are applied
- Expansion, inflation iterated until the network converges



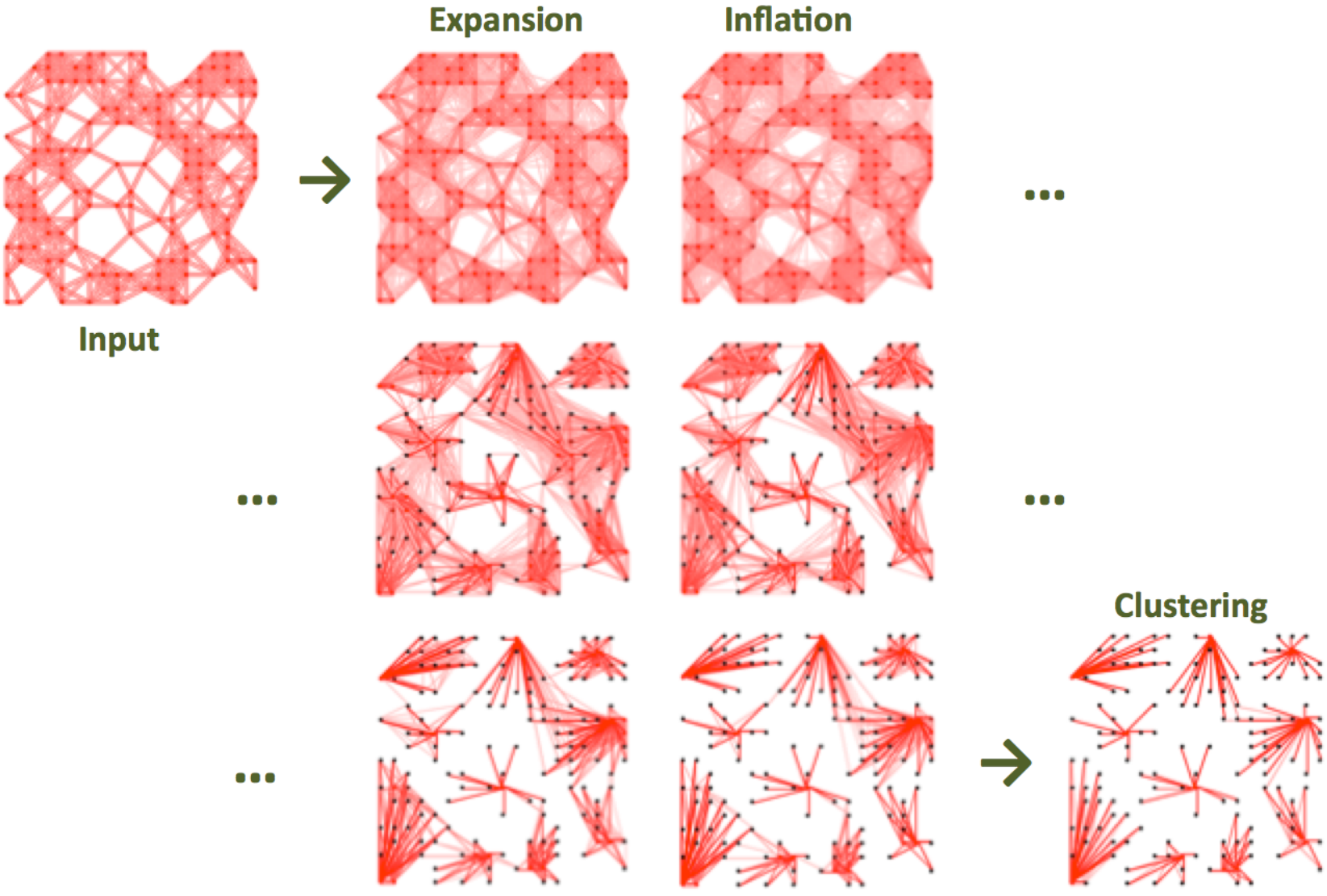
	A	B	C	D	E	F	G
A	100	50	50	45	0	0	0
B	50	100	0	60	0	0	0
C	50	0	100	40	0	0	0
D	45	60	40	100	80	70	15
E	0	0	0	80	100	70	0
F	0	0	0	70	70	100	0
G	0	0	0	15	0	0	100

	A	B	C	D	E	F	G
A	0.42	0.24	0.20	0.11	0.00	0.00	0.00
B	0.20	0.48	0.24	0.15	0.00	0.00	0.00
C	0.20	0.00	0.40	0.10	0.00	0.00	0.00
D	0.18	0.28	0.16	0.24	0.32	0.29	0.13
E	0.00	0.00	0.00	0.19	0.40	0.29	0.00
F	0.00	0.00	0.00	0.17	0.28	0.42	0.00
G	0.00	0.00	0.00	0.04	0.00	0.00	0.87

MCL [a]

[a] Enright *et al.* (2002) *Nucl. Acids Res.* doi:10.1093/nar/30.7.1575

The James Hutton Institute



Which prediction methods work best? [a]

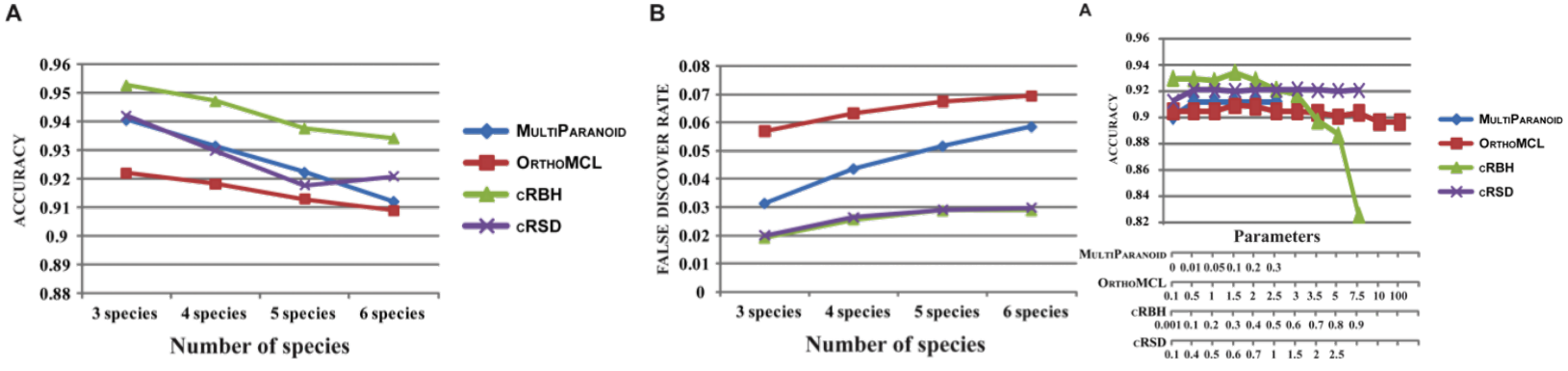
[a] Salichos and Rokas (2011) *PLoS One* **6**:e18755 doi:10.1371/journal.pone.0018755.g006

The James Hutton Institute

Four methods tested against 2,723 curated orthologues from six *Saccharomycetes*

- RBBH (and cRBH); RSD (and cRSD); MultiParanoid; OrthoMCL
- Rated by statistical performance metrics: sensitivity, specificity, accuracy, FDR

cRBH most accurate and specific, with lowest FDR.



Which prediction methods work best? [a] [b]

The James Hutton Institute

[a] Wolf and Koonin (2012) *Genome Biol. Evol.* **4**:1286-1294 doi:10.1093/gbe/evs100

[b] Altenhoff and Dessimoz (2009) *PLoS Comp. Biol.* **5**:e1000262 doi:10.1371/journal.pcbi.1000262

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

How orthologues help

Defining core groups of genes as "orthologues" allows analysis of groups of genes by:

- synteny/collocation
- gene neighbourhood changes (e.g. *genome expansion*)
- pan genome (core/accessory genomes)

and of individual genes within those groups, by:

- multiple alignment
- domain detection
- identification of functional sites
- inference of directional selection (stabilising/positive selection)

Genome expansion [a]

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

- Mobile/repeat elements reproduce and expand during evolution
- Generates a "sequence laboratory" for variation and experiment
- e.g. *Phytophthora infestans* effector protein expansion and arms race

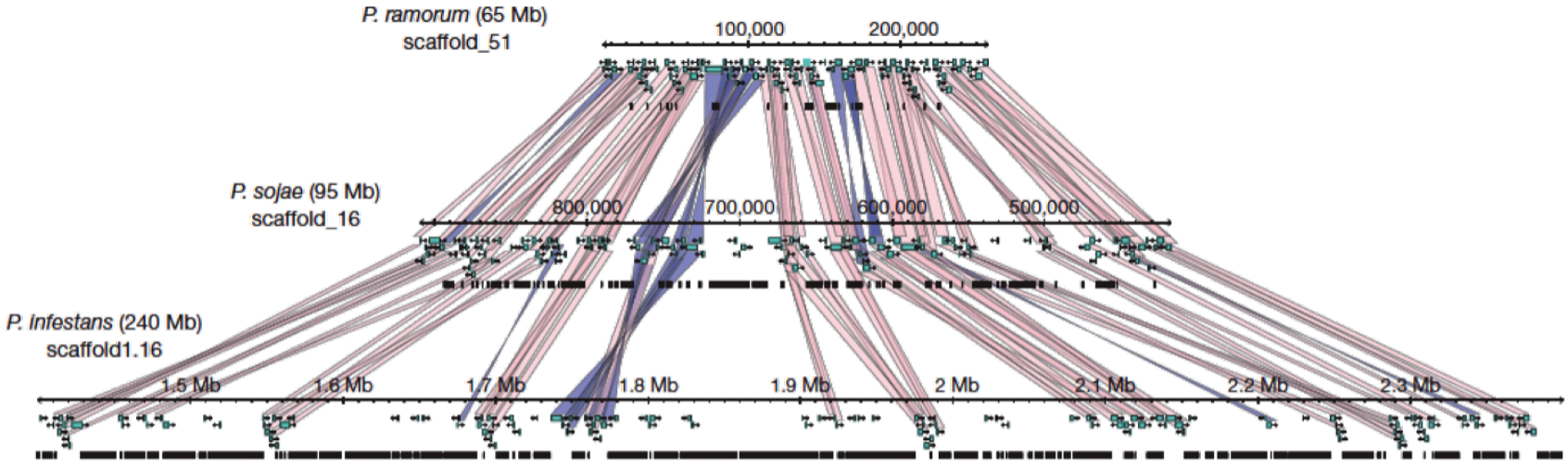


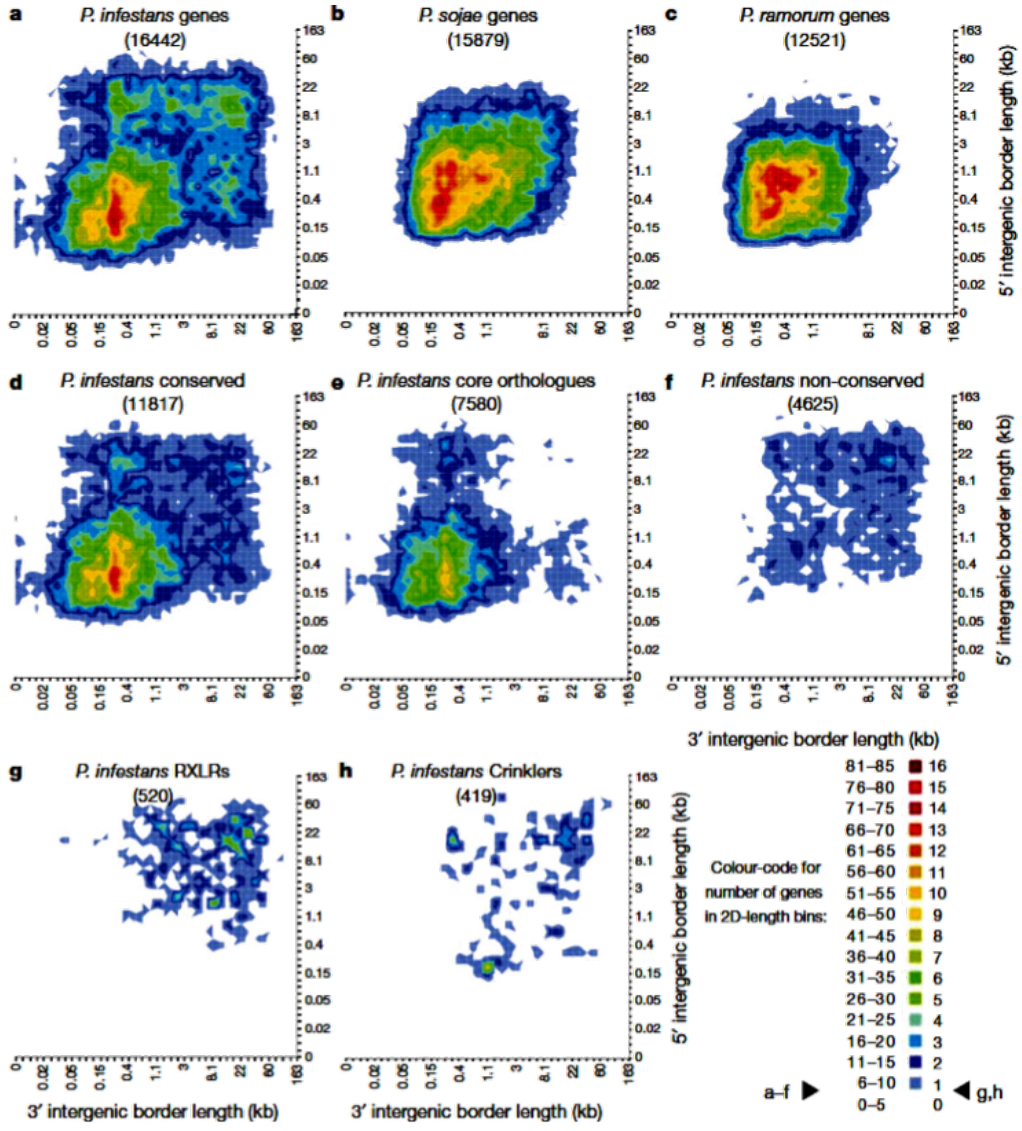
Figure 1 | Repeat-driven genome expansion in *Phytophthora infestans*. Conserved gene order across three homologous *Phytophthora* scaffolds. Genome expansion is evident in regions of conserved gene order, a consequence of repeat expansion in intergenic regions. Genes are shown as turquoise boxes, repeats as black boxes. Collinear orthologous gene pairs are connected by pink (direct) or blue (inverted) bands.

Genome expansion [a]

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

The James Hutton Institute

- Mobile elements (MEs) are large, and duplicate/carry genes with them
- Larger intergenic regions in MEs
- Effector proteins found preferentially in regions with large gaps
- Two-speed genome associated with adaptability



The Pangenome

The James Hutton Institute

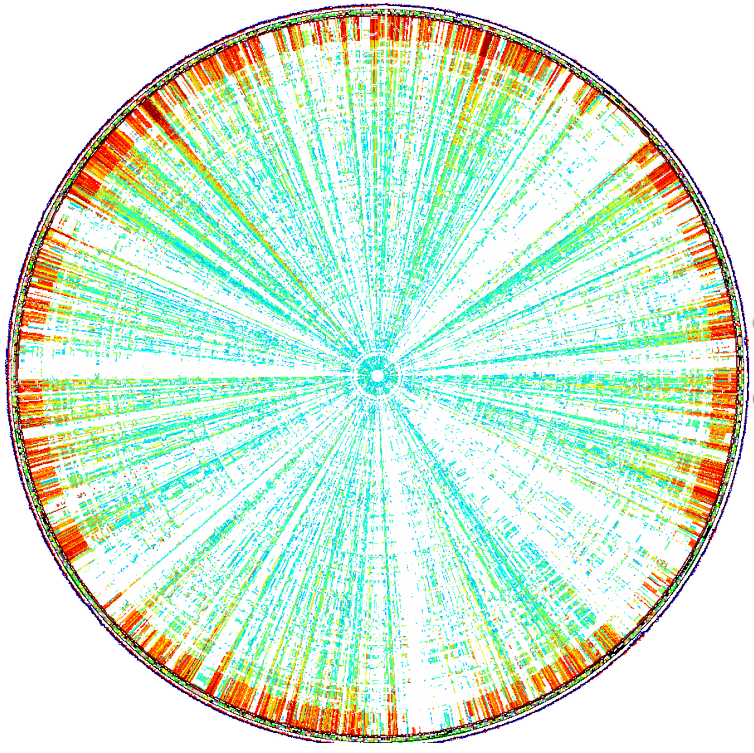
The Core Genome Hypothesis:
"The core genome is the primary cohesive unit defining a bacterial species"

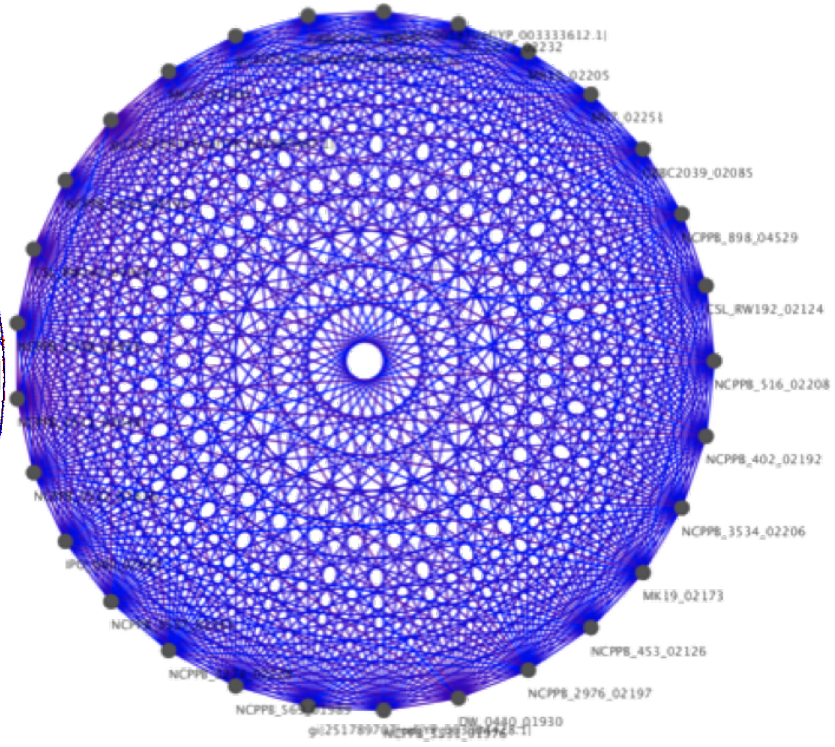
Core genome [a] [b]

[a] Laing (2010) *BMC Bioinf.* doi:10.1186/1471-2105-11-461

[b] Lefébure *et al.* (2010) *Genome Biol. Evol.* doi:10.1093/gbe/evq048

Once equivalent genes have been identified, those present in all related isolates can be identified: **the core genome**.





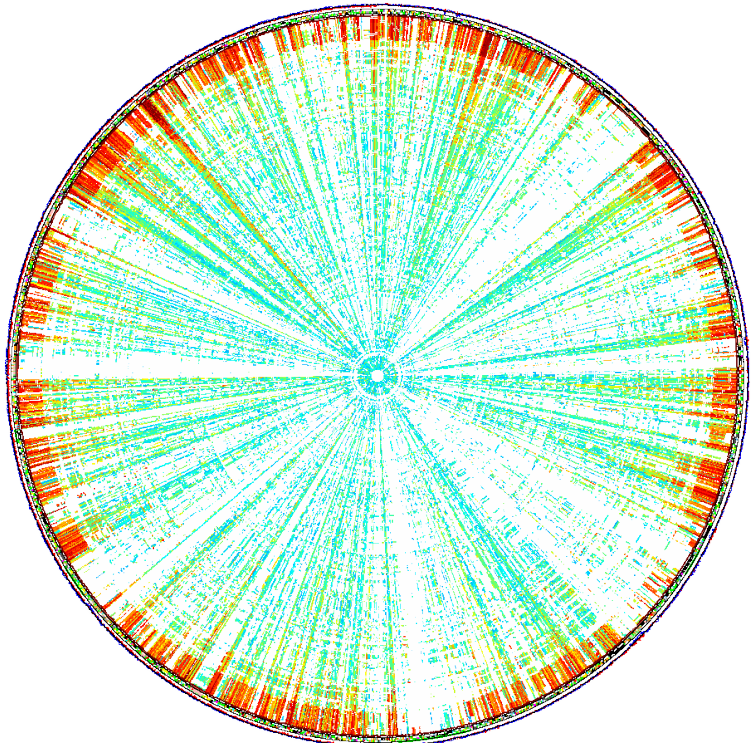
Accessory genome [a] [b]

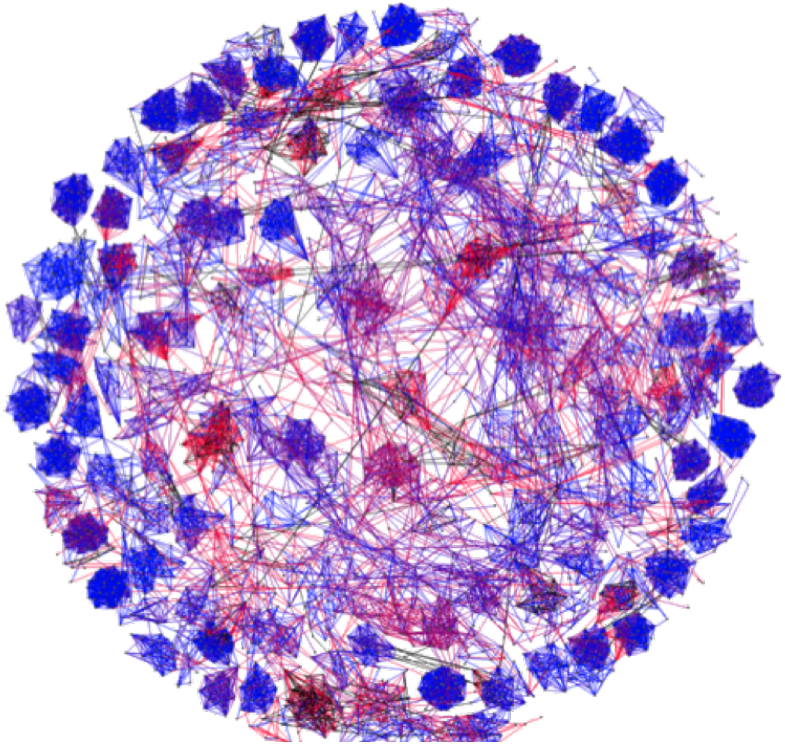
[a] Laing (2010) *BMC Bioinf.* doi:10.1186/1471-2105-11-461

[b] Lefébure *et al.* (2010) *Genome Biol. Evol.* doi:10.1093/gbe/evq048

The James Hutton Institute

The remaining genes are **the accessory genome**, and are expected to mediate function that distinguishes between isolates.





Accessory genome [a] [b]

[a] Croll and Mcdonald (2012) *PLoS Path.* **8**:e1002608 doi:10.1371/journal.ppat.1002608

[b] Baltrus *et al.* (2011) *PLoS Path.* **7**:e1002132 doi:10.1371/journal.ppat.1002132.t002

Accessory genomes are a cradle for adaptive evolution

This is particularly so for bacterial pathogens, such as *Pseudomonas* spp.

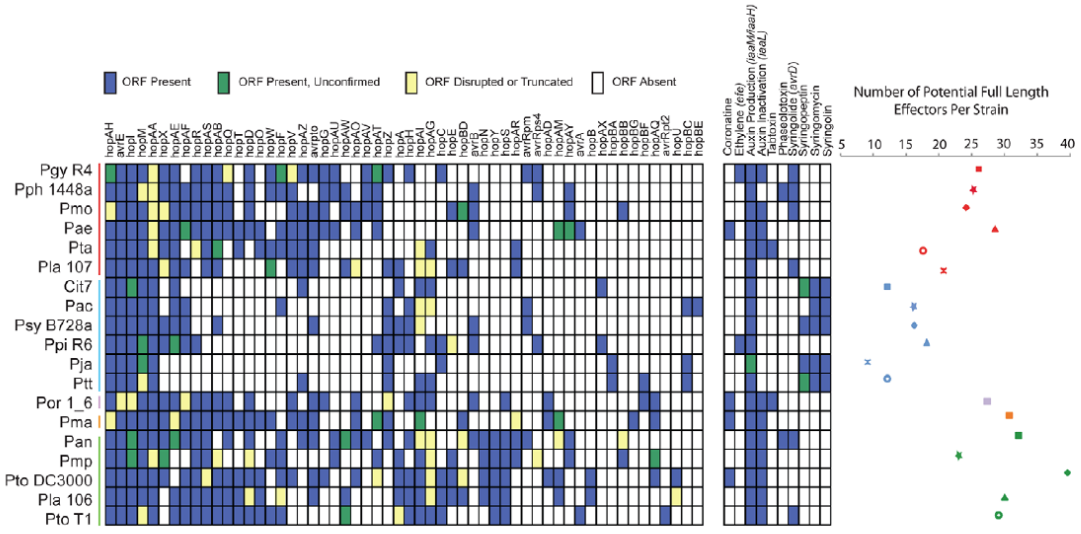


Figure 3. ***P. syringae*** **isolates harbor extensive diversity in virulence gene repertoires.** TTE, toxin, and plant hormone biosynthesis genes are listed across the top, *P. syringae* genomes, color-coded by phylogenetic group as in Figure 1. At the left, a blue box indicates presence of full-length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

Identifying the Pangenome [a]

[a] Page *et al.* (2015) *Bioinf.* **31**:3691-3693 doi:10.1093/bioinformatics/btv421

The James Hutton Institute

`Roary` can produce pangenomes for 1000s of prokaryotes on a desktop machine

- Pre-cluster with CD-HIT (reduce input size)
- All-against-all on reduced sequence set
- MCL clustering
- Merge clusters and use synteny to identify orthologues

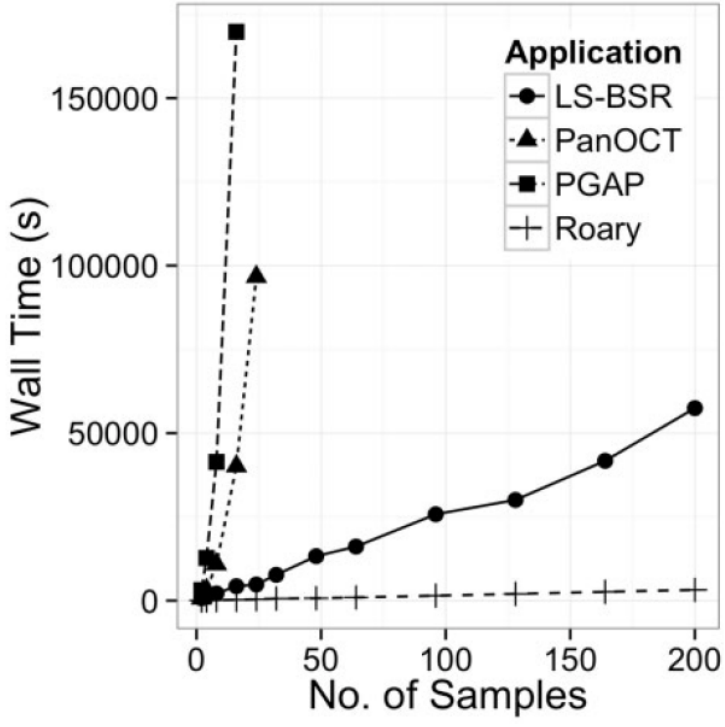


Fig. 1. Effect of dataset size on the wall time of multiple applications. Only analysis that completed within 2 days and 60 GB of RAM is shown

What didn't I get to?

The James Hutton Institute

- Genome-Wide Association Studies (GWAS)
  - Try http://genenetwork.org/ to play with some data
- Prediction of regulatory elements, e.g.
  - Kellis *et al.* (2003) *Nature* doi:10.1038/nature01644
  - King *et al.* (2007) *Genome Res.* doi:10.1101/gr.5592107
  - Chaivorapol *et al.* (2008) *BMC Bioinf.* doi:10.1186/1471-2105-9-455
  - CompMOBY http://genome.ucsf.edu/compmoby
- Detection of Horizontal/Lateral Gene Transfer (HGT/LGT), e.g.
  - Tsirigos & Rigoutsos (2005) *Nucl. Acids Res.* doi:10.1093/nar/gki187
- Phylogenomics, e.g.
  - Delsuc *et al.* (2005) *Nat. rev. Genet.* doi:10.1038/nrg1603
  - AMPHORA https://phylogenomics.wordpress.com/software/amphora/

Messages to take away

The James Hutton Institute

- Comparative genomics is a powerful set of techniques for:
  - Understanding and identifying evolutionary processes and mechanisms
  - Reconstructing detailed evolutionary history
  - Identifying and understanding common genomic features
  - Providing hypotheses about gene function for experimental investigation

Messages to take away

The James Hutton Institute

- Comparative genomics is comparisons
  - What is *similar* between two genomes?
  - What is *different* between two genomes?
- Comparative genomics *is* evolutionary genomics
  - Lots of scope for improvement in tools
- Tools that 'do the same thing' can give different output
  - BLAST vs MUMmer
  - RBBH vs MCL
  - The choice of application matters for correctness and interpretation

Licence: CC-BY-SA

The James Hutton Institute

By: Leighton Pritchard