Comparative Genomics and Visualisation
BS32010
3.Whole Genome Comparisons

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-2015-03-17-UoD_compgenvis

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

Table of Contents

The James Hutton Institute

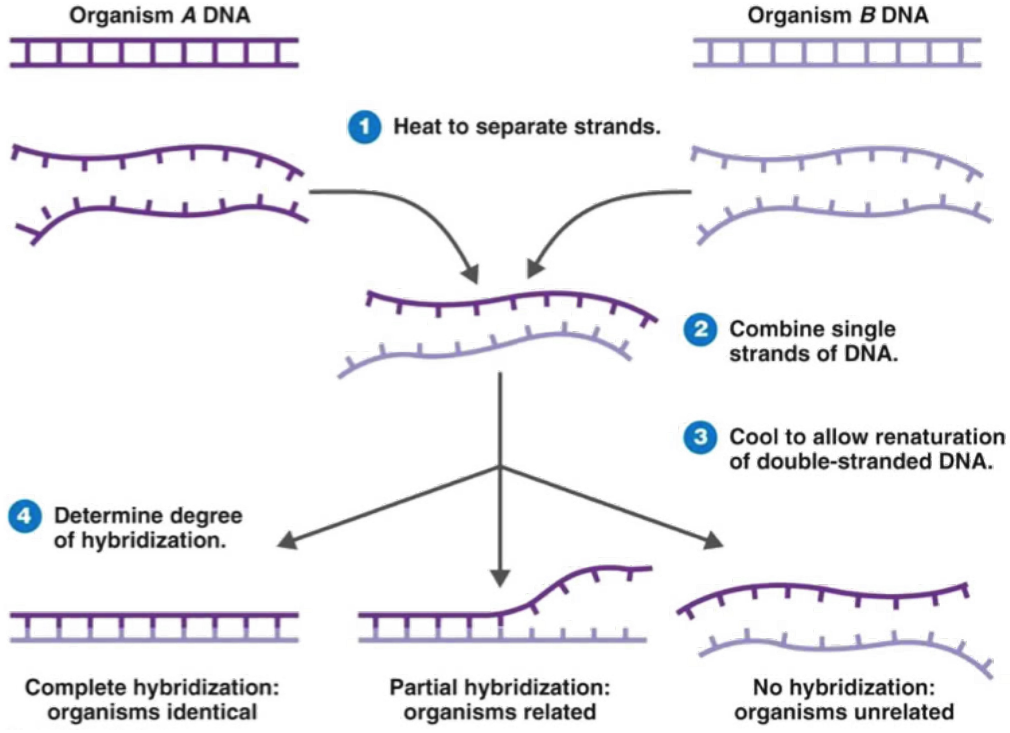
DNA-DNA hybridisation (DDH) [a]

[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

DNA-DNA hybridisation (DDH) [a]

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

- Find homoduplex T_{m1} for reference A
- Denature gDNA mixture for reference A, comparator B, and mix
- Allow gDNA to anneal; hybrids result
- Find heteroduplex T_{m2} for mixture
- $\Delta T_m = T_{m1} - T_{m2}$
- High $\Delta T \implies$ genomic difference (fewer H-bonds)

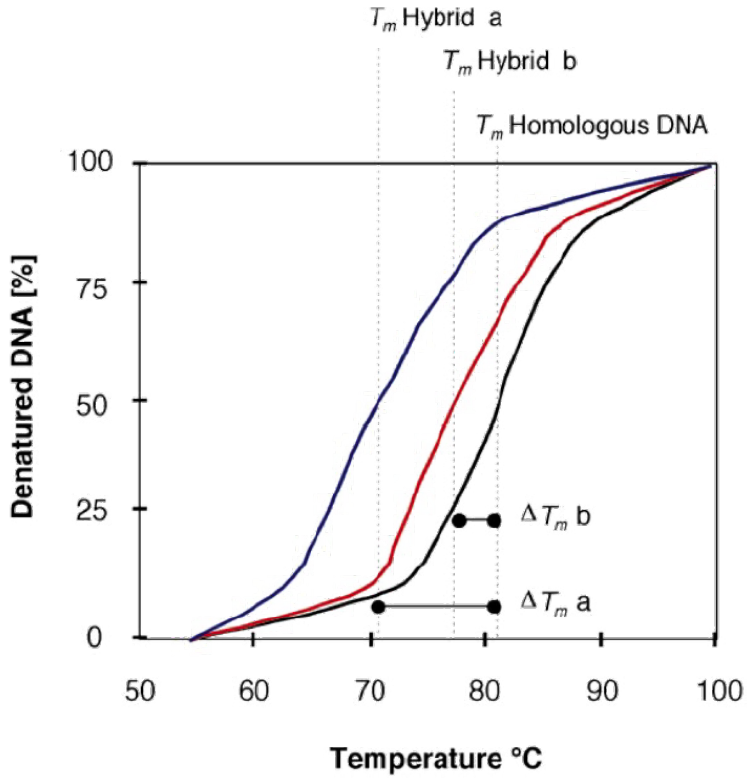


Fig. 2. Thermal denaturation curves of a homoduplex DNA and two heteroduplex DNAs.

Proxy for sequence similarity

DNA-DNA hybridisation (DDH) [a]

[a] Sibley & Ahlquist (1984) *J. Mol. Evol.* doi:10.1007/BF02101980

Used for taxonomic classification in prokaryotes since '60s

Redefined accepted relationships for birds, primates in 1980s

Controversial (at the time) proof that *Homo* shares more recent common ancestor with *Pan* than with *Gorilla*

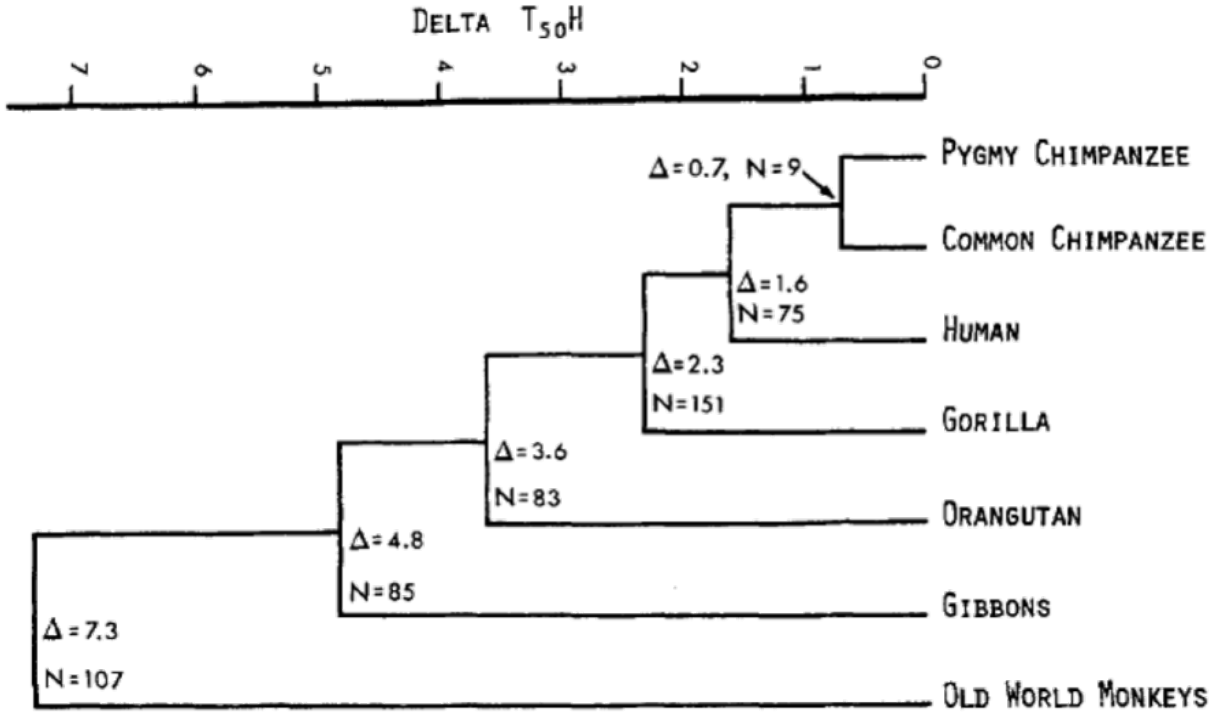


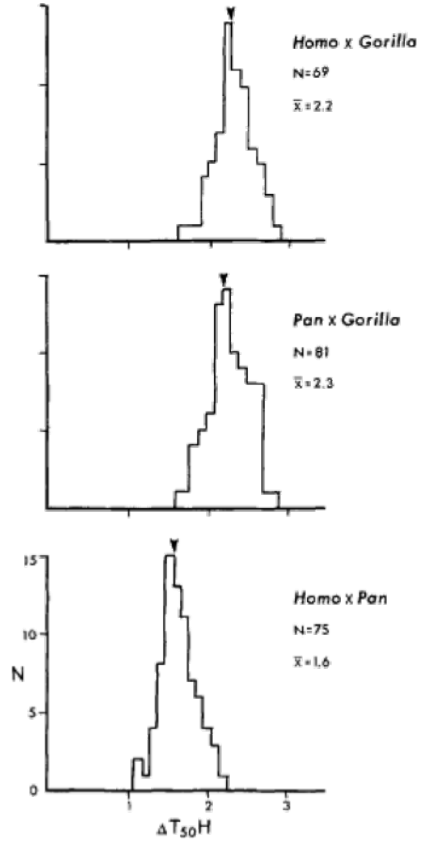
Fig. 3. Phylogeny of the hominoid primates as determined by average linkage clustering of delta $T_{50}H$ values derived from DNA–DNA hybridization

DNA-DNA hybridisation (DDH) [a] [b]

[a] Sibley & Ahlquist (1984) *J. Mol. Evol.* doi:10.1007/BF02101980
[b] http://personal.uncc.edu/jmarks/DNAHYB/dnahyb2.html

Controversial (at the time) proof that *Homo* shares more recent common ancestor with *Pan* than with *Gorilla*

- Allegations of data manipulation (see link *b*)
- Close evolutionary relationships difficult to resolve due to *paralogy*
- Still the *de facto* gold standard for microbiological taxonomic classification



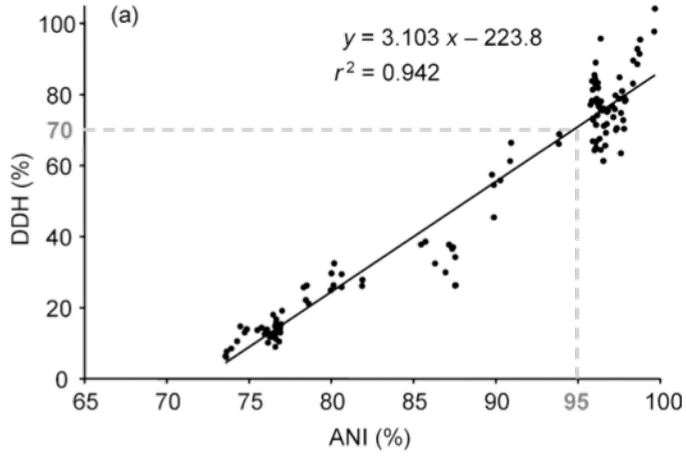
Average Nucleotide Identity (ANI) [a]

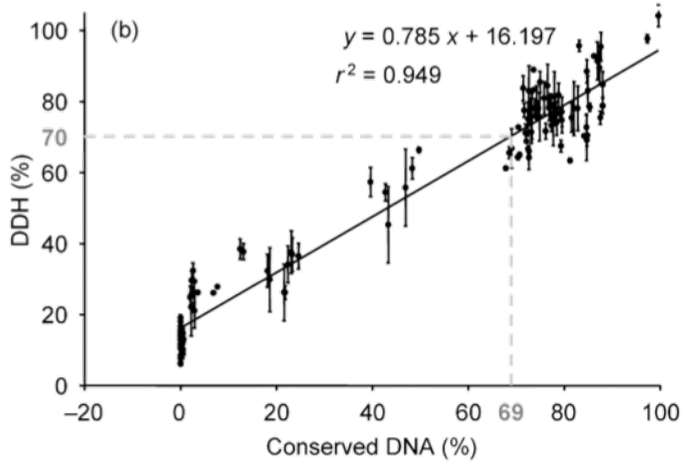
[a] Goris *et al.* (2007) *Int. J. System. Evol. Biol.* doi:10.1099/ijs.0.64483-0

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)





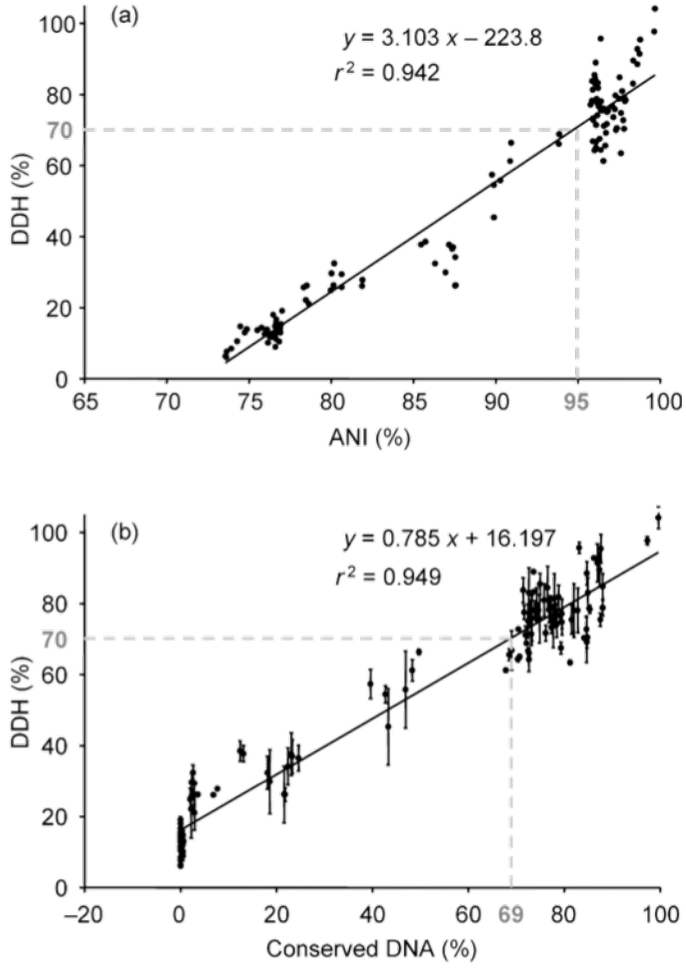
Average Nucleotide Identity (ANI) [a]

[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]

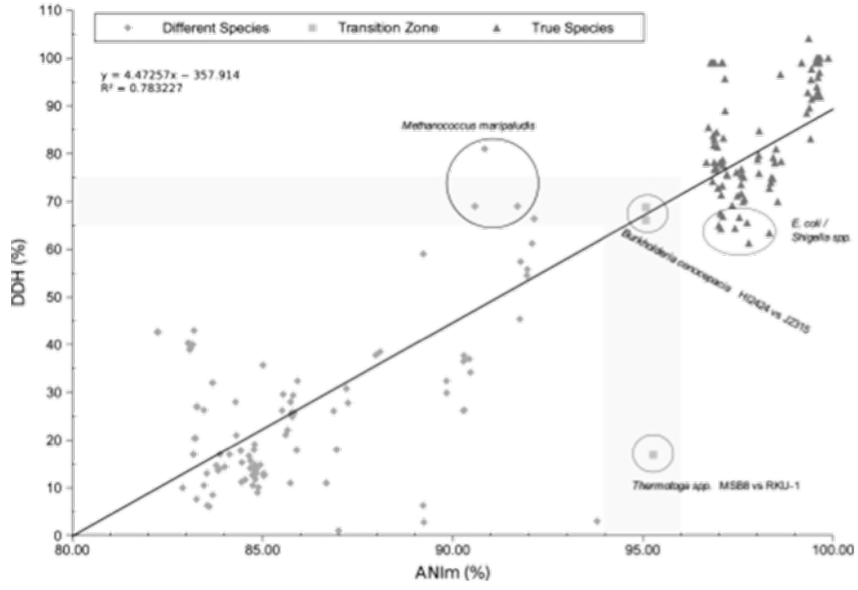
[a] Richter & Rosselló-Móra *et al.* (2009) *Proc. Natl. Acad. Sci. USA* doi:10.1073/pnas.0906412106

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

Bulk genome comparisons

The James Hutton Institute

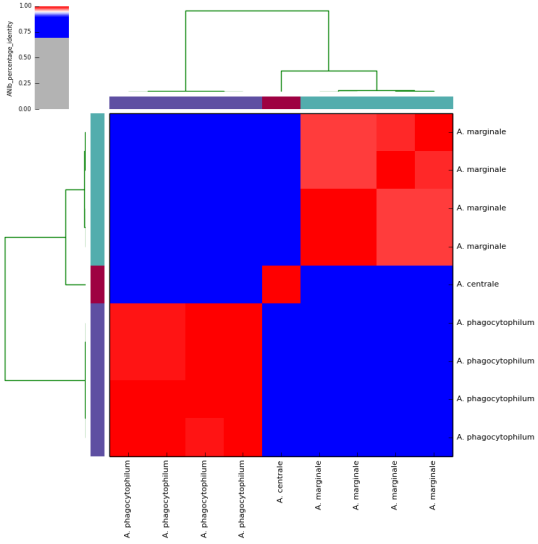
EXERCISE 3:

`ex03_ani.ipynb`

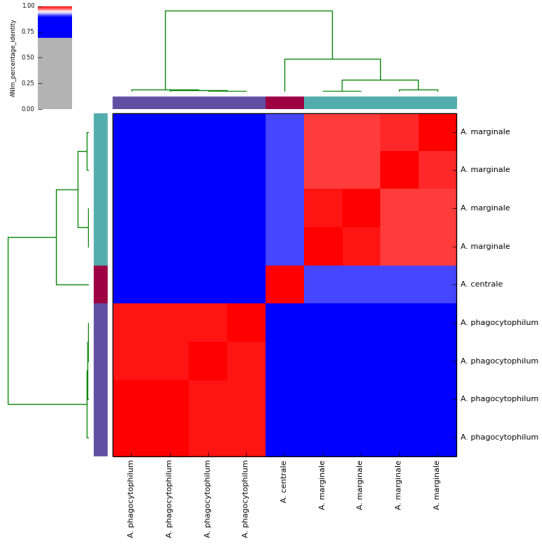
Bulk genome comparisons

Anaplasma spp. comparisons

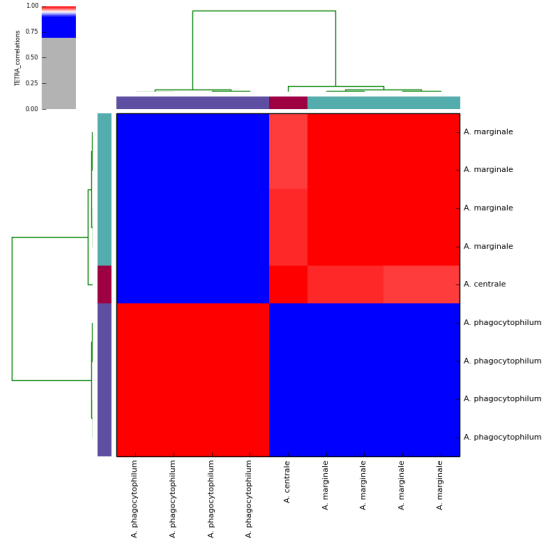
ANIb



ANIm



TETRA

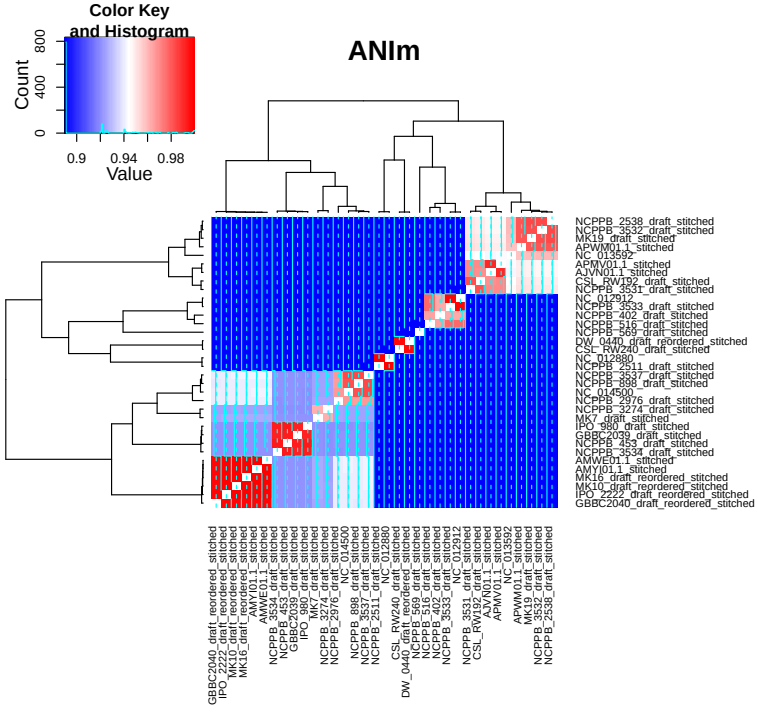


ANI in practice [a]

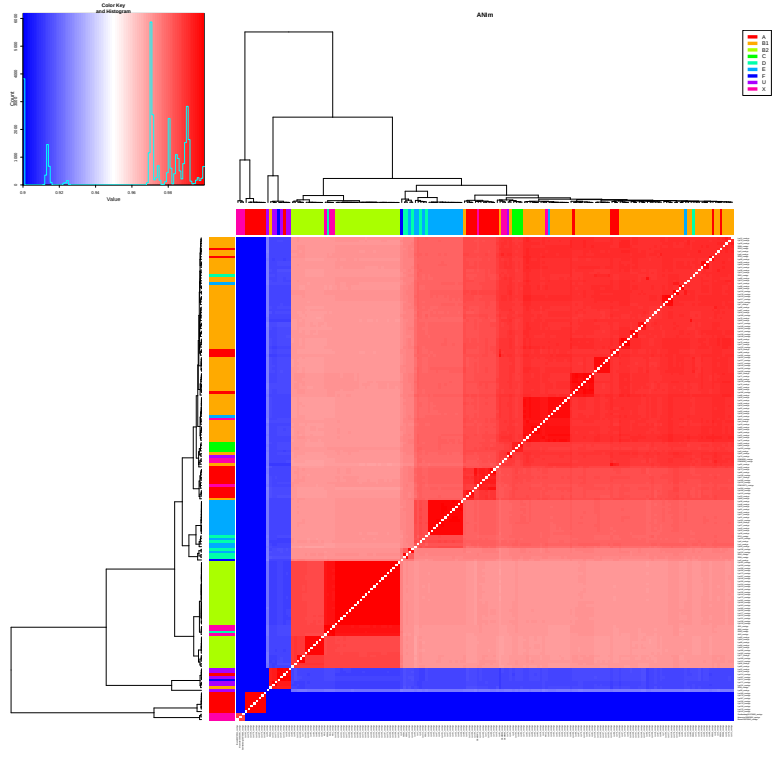
[a] van der Wolf *et al.* (2014) *Int. J. Syst. Evol. Micr.* **64**:768-774 doi:10.1099/ijs.0.052944-0

The James Hutton Institute

29 *Dickeya* isolates: species structure



180 *E.coli* isolates: subtyping



Average Nucleotide Identity (ANI)

The James Hutton Institute

ANIb vs ANIm

- ANIb can split a single matching region into two fragments that do not match criteria
- $\Longrightarrow$ ANIb may discard useful information that ANIm retains

ANIb/ANIm vs TETRA

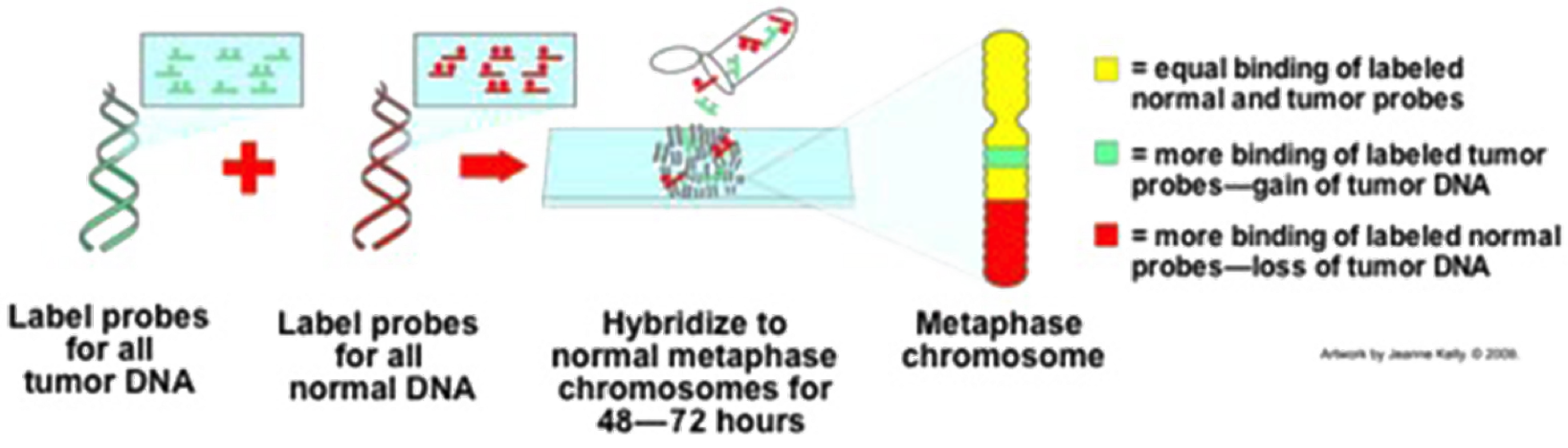
- ANIb/ANIm reflect sequence matching (analogous to hybridisation)
- TETRA reflects statistical deviation of a bulk genome measure (4-mer frequency)
- TETRA may be prone to false positives (saying two sequences are the same species when they are not)

Table of Contents

The James Hutton Institute

Comparative Genomic Hybridisation

- Two genomes: reference and test fragmented & labelled
- Hybridise reference and test against a third "normal" genome.



- Differences in red/green intensity mapped by microscopy
- Colour intensity differences correspond to hybridisation, sequence similarity, and **copy number variations (CNV)**.
- Can compare *within* species/individuals (e.g. tumours), but labour-intensive, low-resolution

CGH: epigenetics [a] [b]

[a] Kallionemi *et al.* (1992) *Science* doi:10.1099/10.1126/science.1359641

[b] Fraga *et al.* (2005) *Proc. Natl. Acad. Sci. USA* doi:10.1073/pnas.0500398102

The James Hutton Institute

Measurements taken using image analysis - intensity on medial axis

Used for epigenetics: hybridisation of methylated DNA

Comparative Genomic Hybridization for Molecular Cytogenetic Analysis of Solid Tumors

Anne Kallioniemi,* Olli-P. Kallioniemi, Damir Sudar, Denis Rutovitz, Joe W. Gray, Fred Waldman, Dan Pinkel

Comparative genomic hybridization produces a map of DNA sequence copy number as a function of chromosomal location throughout the entire genome. Differentially labeled test DNA and normal reference DNA are hybridized simultaneously to normal chromosome spreads. The hybridization is detected with two different fluorochromes. Regions of gain or loss of DNA sequences, such as deletions, duplications, or amplifications, are seen as changes in the ratio of the intensities of the two fluorochromes along the target chromosomes. Analysis of tumor cell lines and primary bladder tumors identified 16 different regions of amplification, many in loci not previously known to be amplified.

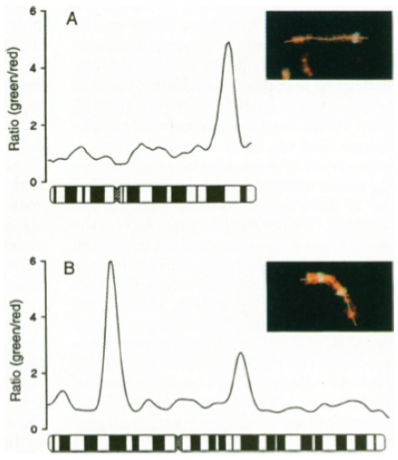
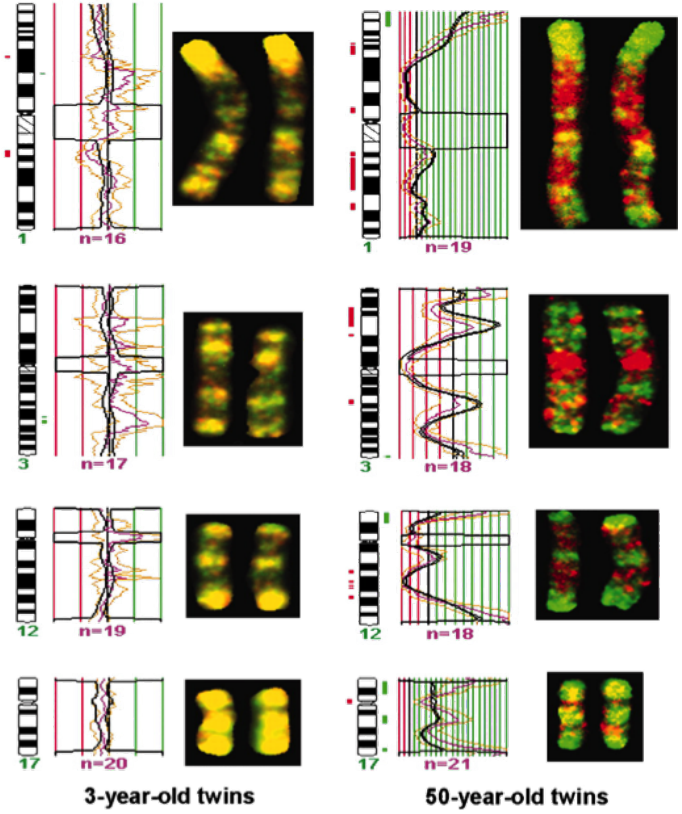


Fig. 3. Green-to-red fluorescence ratio profiles of chromosome 8 (**A**) and chromosome 2 (**B**) after hybridization with COLO 320HSR and NCI-H69 cell line DNAs, respectively (green). Normal reference DNA included in the hybridization is shown in red. The inserts show the overlaid green and red fluorescence images of the chromosomes and the chromosomal medial axis drawn by the image analysis program. In (A), the *myc* locus at 8q24 shows a highly elevated green-to-red ratio, which is compatible with the known high-level amplification of *myc* in the COLO 320HSR cell line. In (B), three regions of amplification are seen on chromosome 2. The signal at 2p24 corresponds to the location of N-*myc* known to be amplified in the NCI-H69 cell line. The two other regions with a highly increased fluorescence ratio, at 2p21 and 2q21, were not known to be amplified.



Array CGH (aCGH) [a]

[a] Pollack *et al.* (1999) *Nat. Genet.* doi:10.1099/10.1038/12640

The James Hutton Institute

Uses DNA microarrays: 1000s short, immobilised DNA probes gDNA, cDNA etc.fluorescently-labelled and hybridised to the array

- Smaller sample sizes than CGH
- Automatable, high-throughput, high-resolution
- Can identify copy number variation, segmental duplication, presence/absence

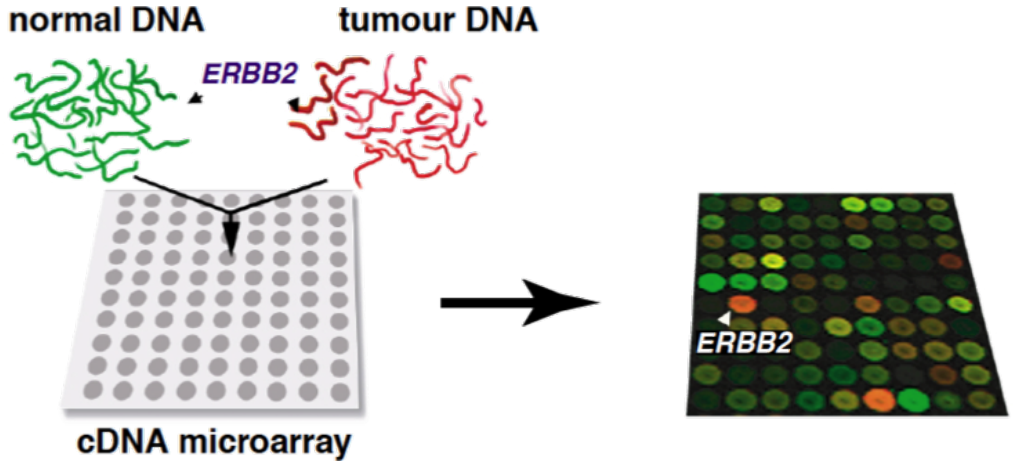


Table of Contents

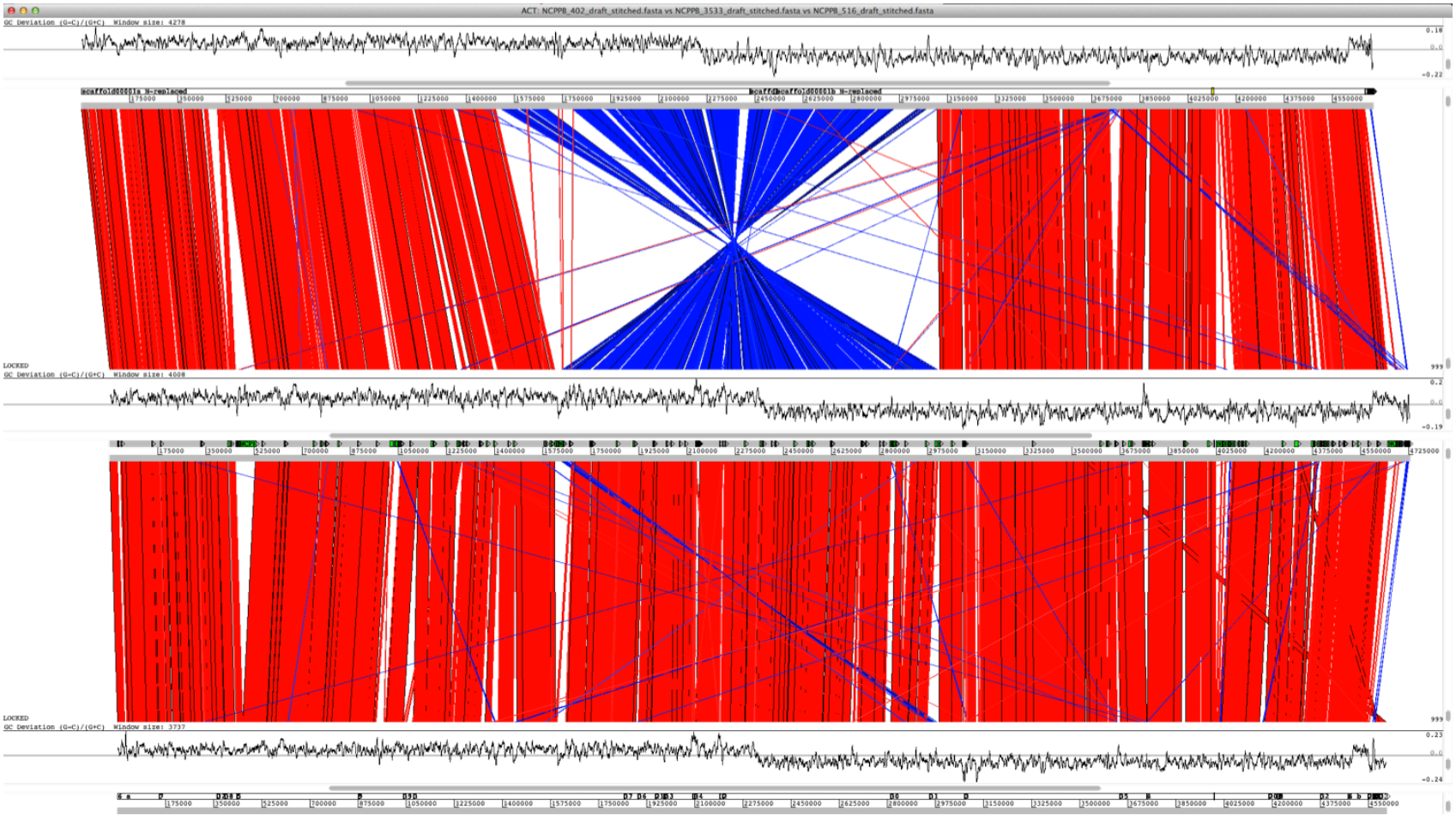
The James Hutton Institute

Pairwise genome alignments

The James Hutton Institute

Genome sequence data gives much more detail and power

Pairwise comparisons require alignment of similar regions.

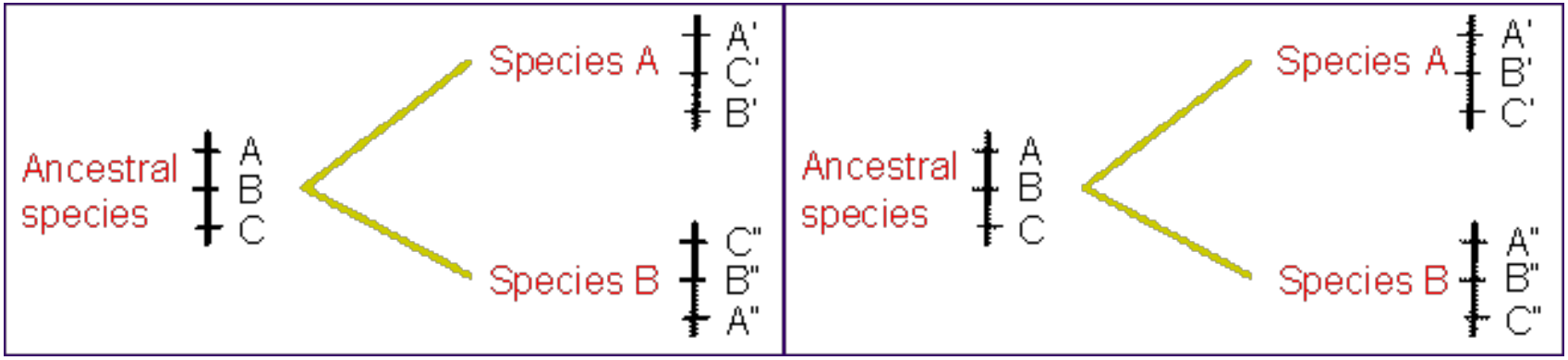


Synteny and Collinearity

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

Signs of evolutionary constraint (e.g. synteny) may indicate a functional region

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

(but, if not, see `supporting_information` subdirectory)

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?

BLAT [a]

[a] Kent (2002) *Genome Res.* doi:10.1101/gr.229202

The James Hutton Institute

Broadly similar to BLAST

Main differences:

- optimised to find only exact or near-exact matches (speed)
- indexes the subject genome, and *scans the query*
- connects homologous match regions into a single alignment (BLAST reports these separately)
- reports mRNA match intron-exon bounds exactly (BLAST tends to extend beyond bounds)

ADVANTAGES: fast, exact exon bounds, UCSC integration
DISADVANTAGES: less sensitive on remote/divergent sequences

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

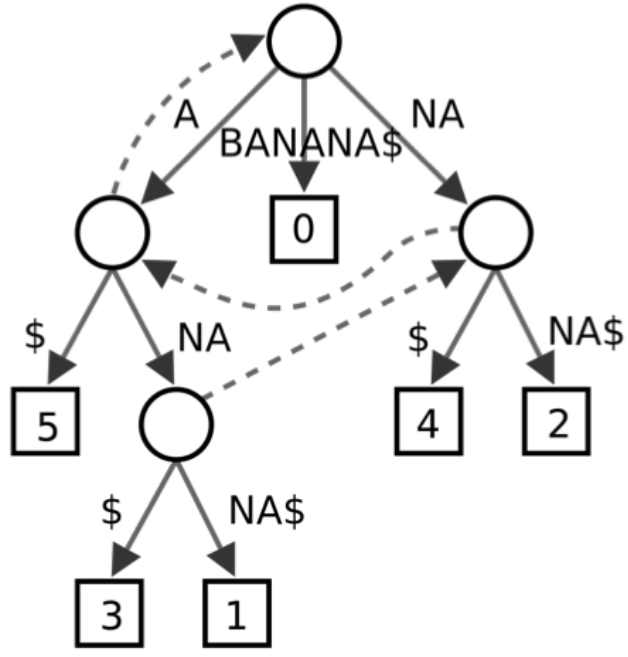
Conceptually completely different to BLAST/BLAT/megaBLAST

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`, ...)

Whole genome comparisons

The James Hutton Institute

EXERCISE 4:

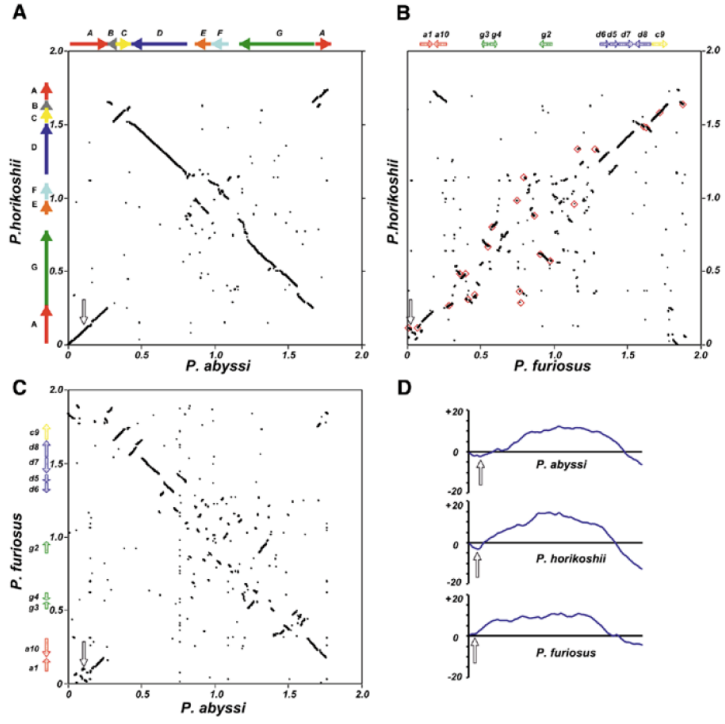
whole_genome_alignment/whole_genome_alignments_A.md

Pyrococcus spp. [a]

[a] Zivanovic *et al.* (2002) *Nuc. Acids Res.* **30**:1902-1910 doi:10.1093/nar/30.9.1902

The James Hutton Institute

Comparison of *Pyrococcus* genomes (*P. horikoshii*, *P. abyssi*, *P. furiosus*) shows chromosome-shuffling.



Transposition a major cause of genomic disruption

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.

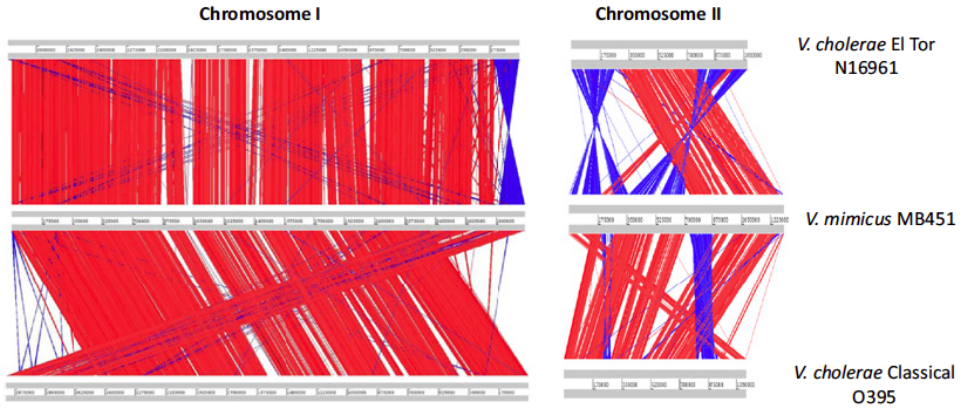


Fig. 2. Linear pairwise comparison of the *Vibrio mimicus* genome by Artemis Comparison Toll. Regions with similarity are highlighted by connecting red or blue lines between the genomes; red lines indicate homologous blocks of sequence, and blue lines indicate inversions. Gaps indicate unique DNA. The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

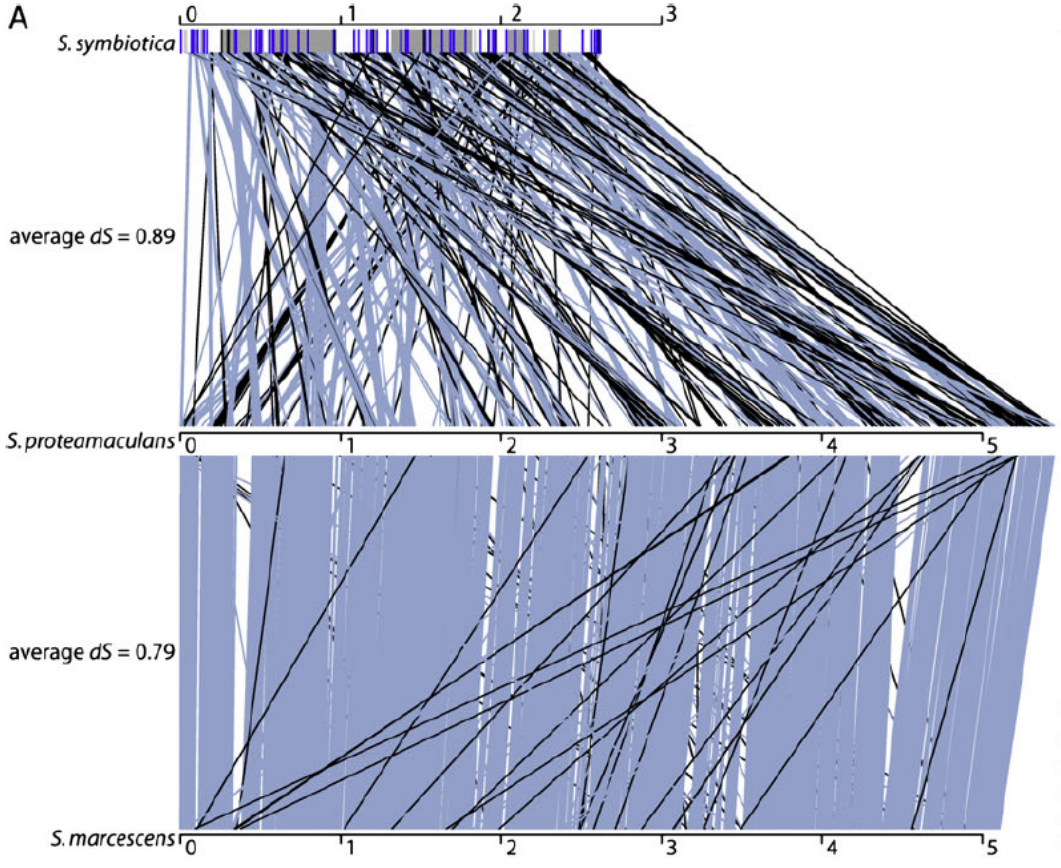


Table of Contents

The James Hutton Institute

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

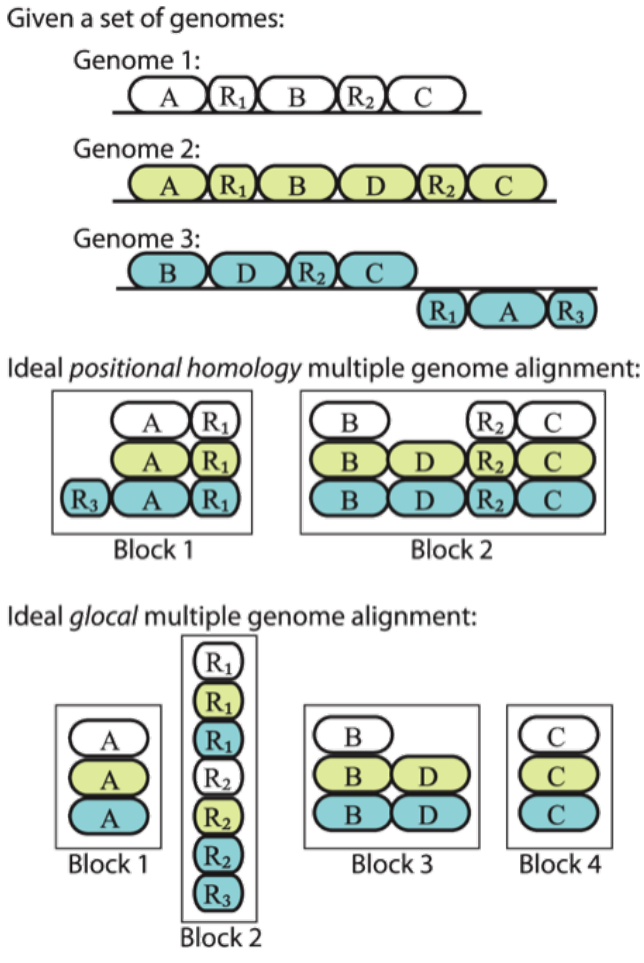
Several tools:

- 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/)



LAGAN [a]

[a] Brudno *et al.* (2003) *Genome Res.* doi:10.1101/gr.926603

The James Hutton Institute

Rapid pairwise alignment of homologous genomes

Algorithm:

1. Generate local alignments (*anchors*, B)
2. Construct rough global map (*max-scoring ordered subset*, C)
3. Join anchors that lie within a threshold distance
4. Compute global alignment within each anchor (*Needleman-Wunsch*, D)

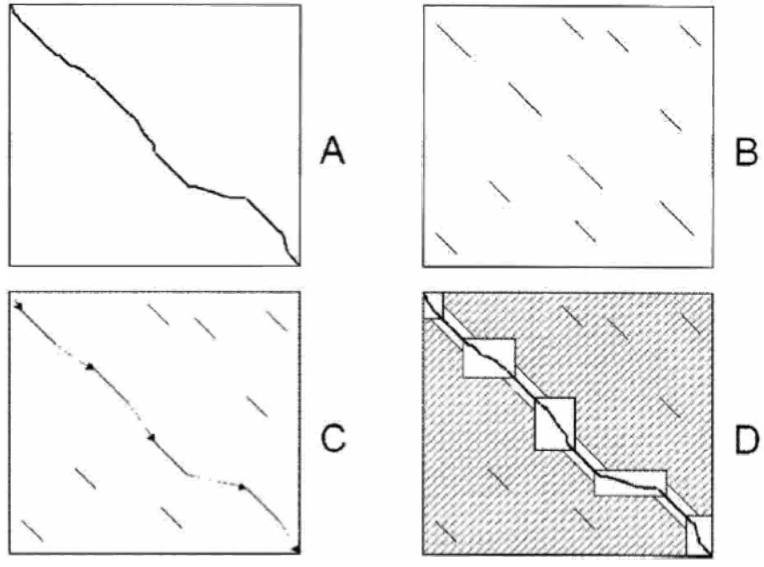


Figure 1 The LAGAN algorithm. (*A*) A global alignment between two sequences is a path between the top-left and the bottom-right corner of their alignment matrix. (*B*) LAGAN first finds all local alignments between the two sequences. (*C*) LAGAN computes a maximal-scoring ordered subset of the alignments, the *anchors*, and puts together a rough global map. (*D*) LAGAN limits the search for an optimal alignment to the area included in the boxes and around the anchors, and computes the optimal Needleman-Wunsch alignment limited to that area. LAGAN uses memory proportional to the area of the largest box plus the memory to hold the optimal alignment.

MLAGAN [a]

[a] Brudno *et al.* (2003) *Genome Res.* doi:10.1101/gr.926603

The James Hutton Institute

Multiple genome alignment of k genomes in $k-1$ alignment steps
Progressive/iterative, using phylogenetic tree (like `CLUSTAL`)

Algorithm:

1. Construct rough global maps between each pair of sequences (step C in LAGAN)
2. Progressive multiple alignment with anchors:
   - LAGAN alignment between closest pair of sequences: combined as *multi-sequences*
   - Find rough global maps of this multi-sequence to all other multi-sequences

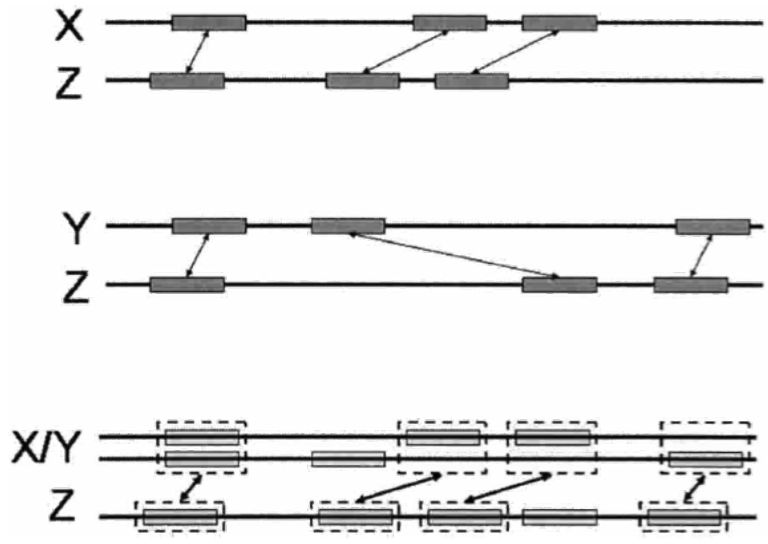
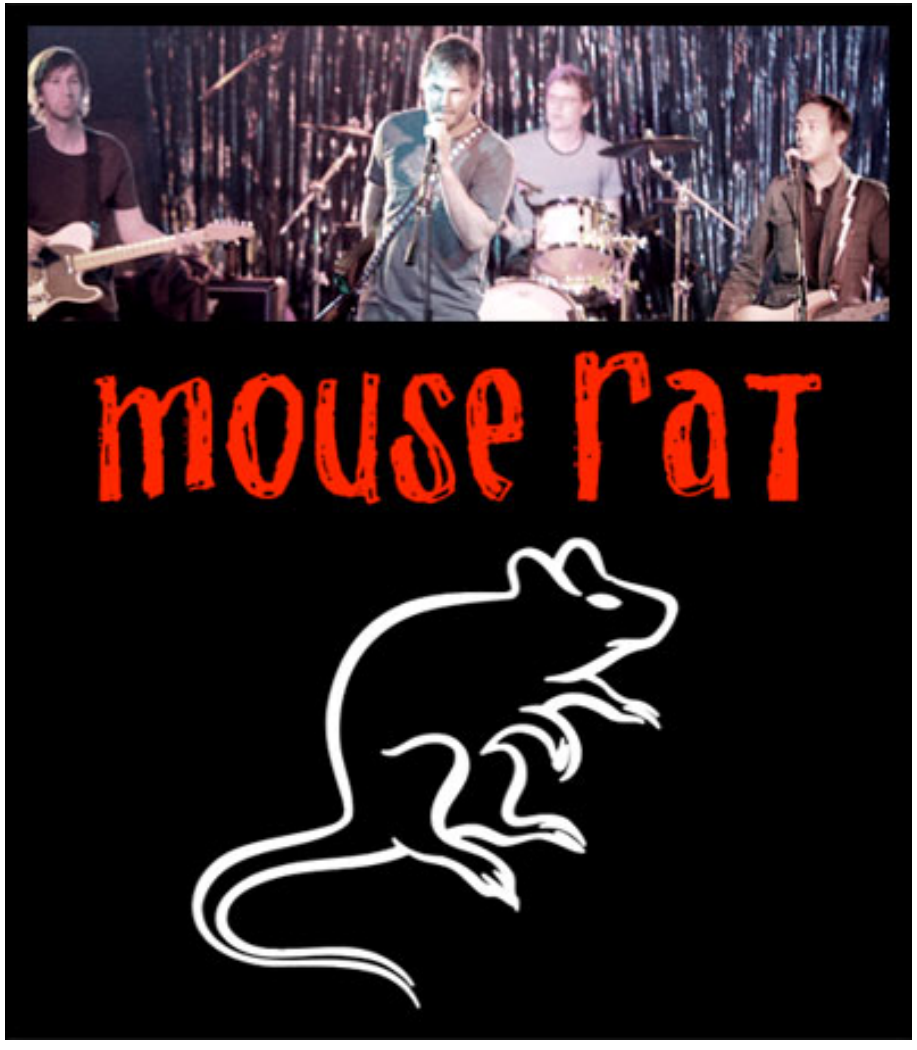


Figure 5 Generation of anchors during progressive alignment. Multi-sequence X/Y is aligned to sequence Z. Anchors between X and Z (*top*) and anchors between Y and Z (*middle*) are remapped to coordinates in the X/Y multi-sequence, and given a new score. Then, the Longest Increasing Subsequence algorithm is applied to select a subset of the remapped anchors, as the anchors between X/Y and Z.

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

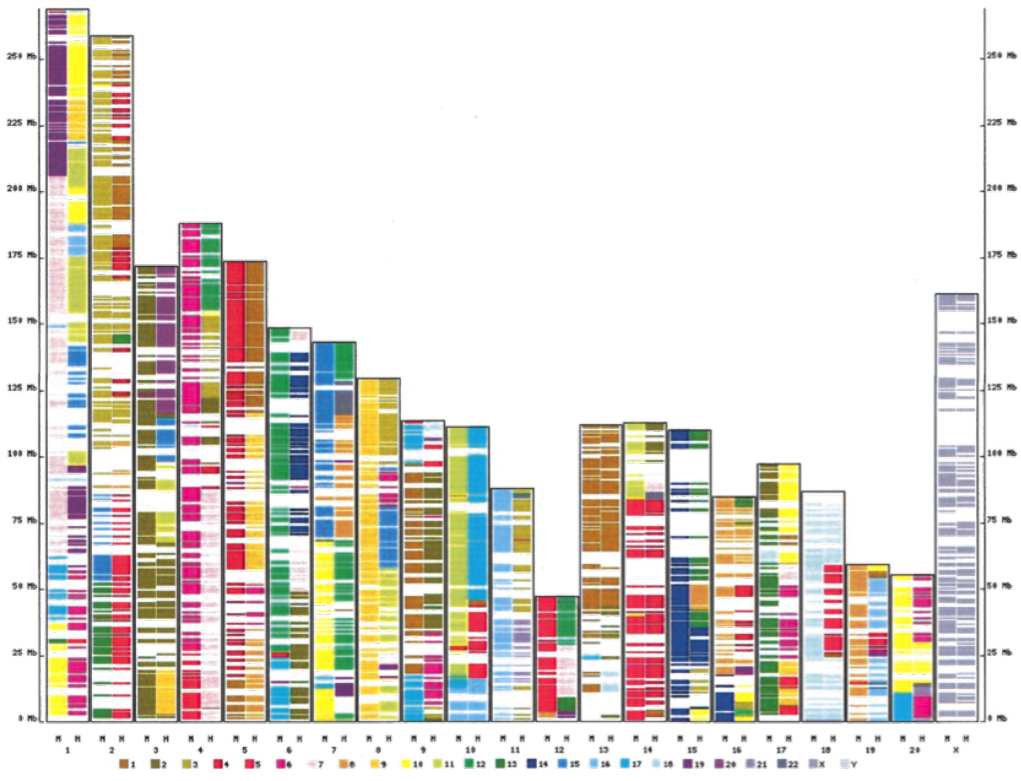
Three-way progressive alignment

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]

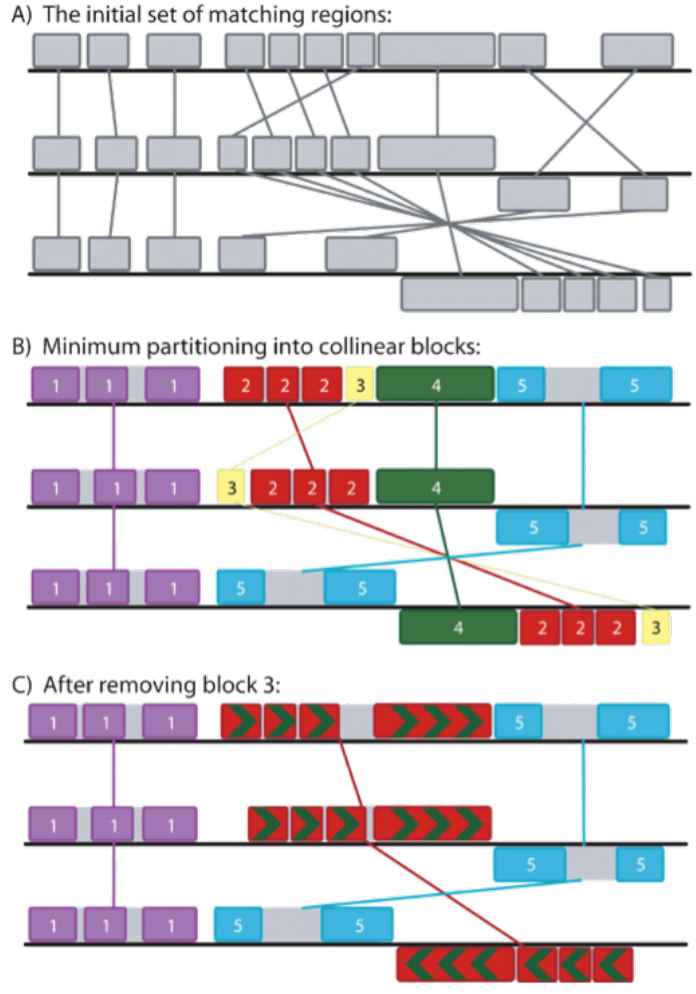
[a] Darling *et al.* (2003) *Genome Res.* doi:10.1101/gr.2289704

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



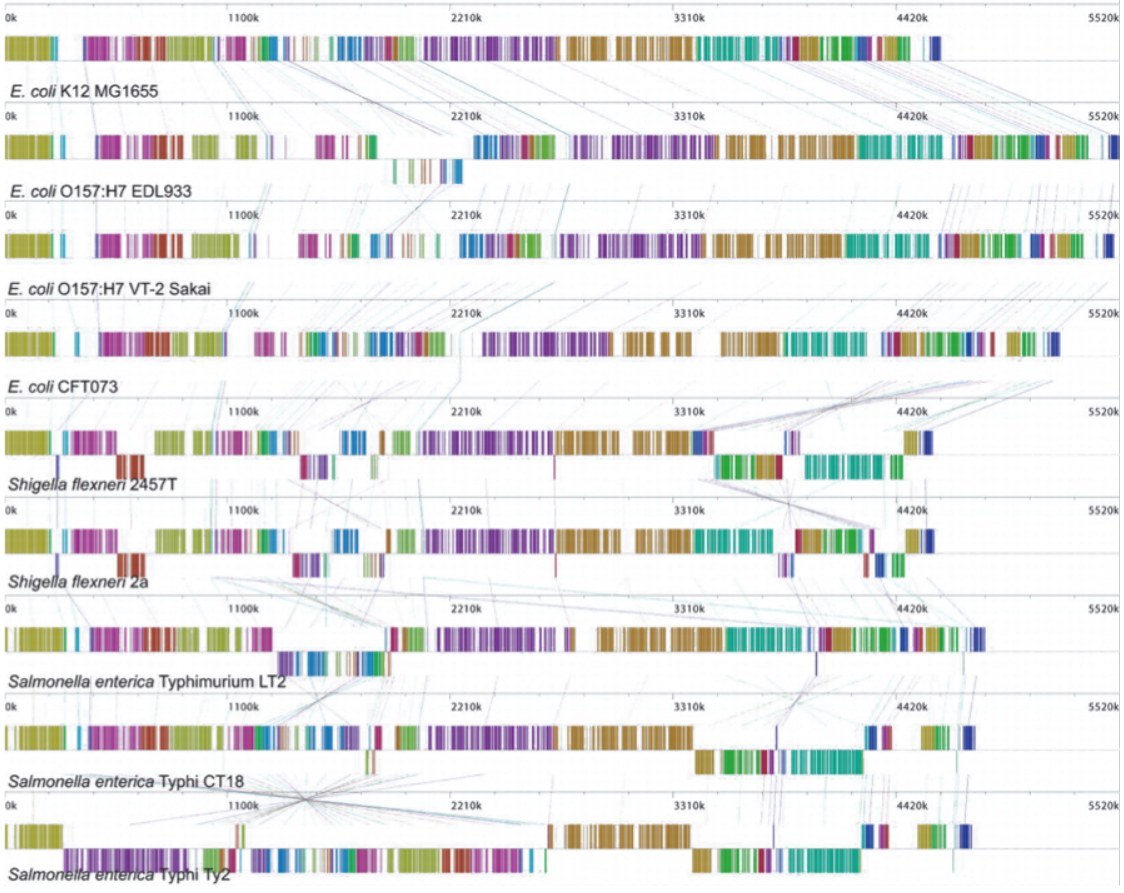
MAUVE [a]

[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



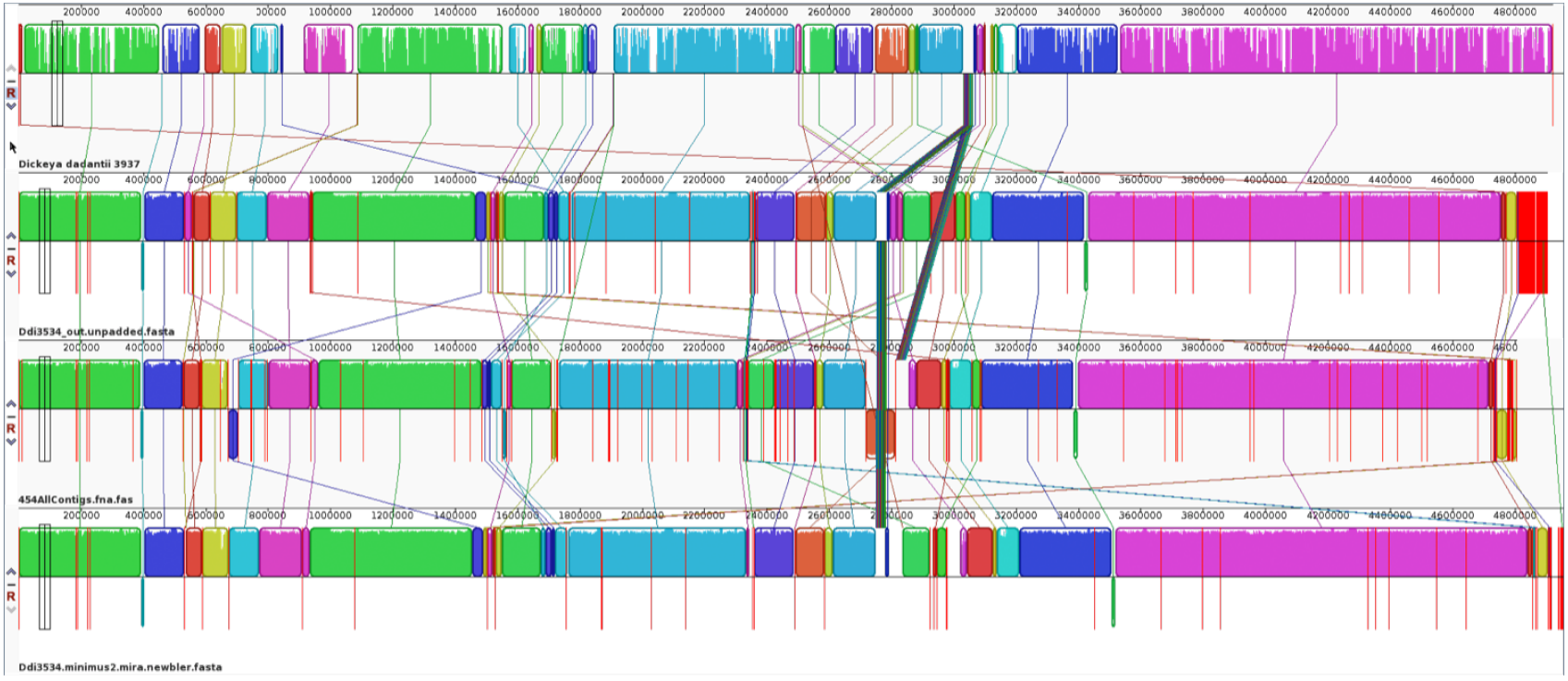
Draft genome alignment

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

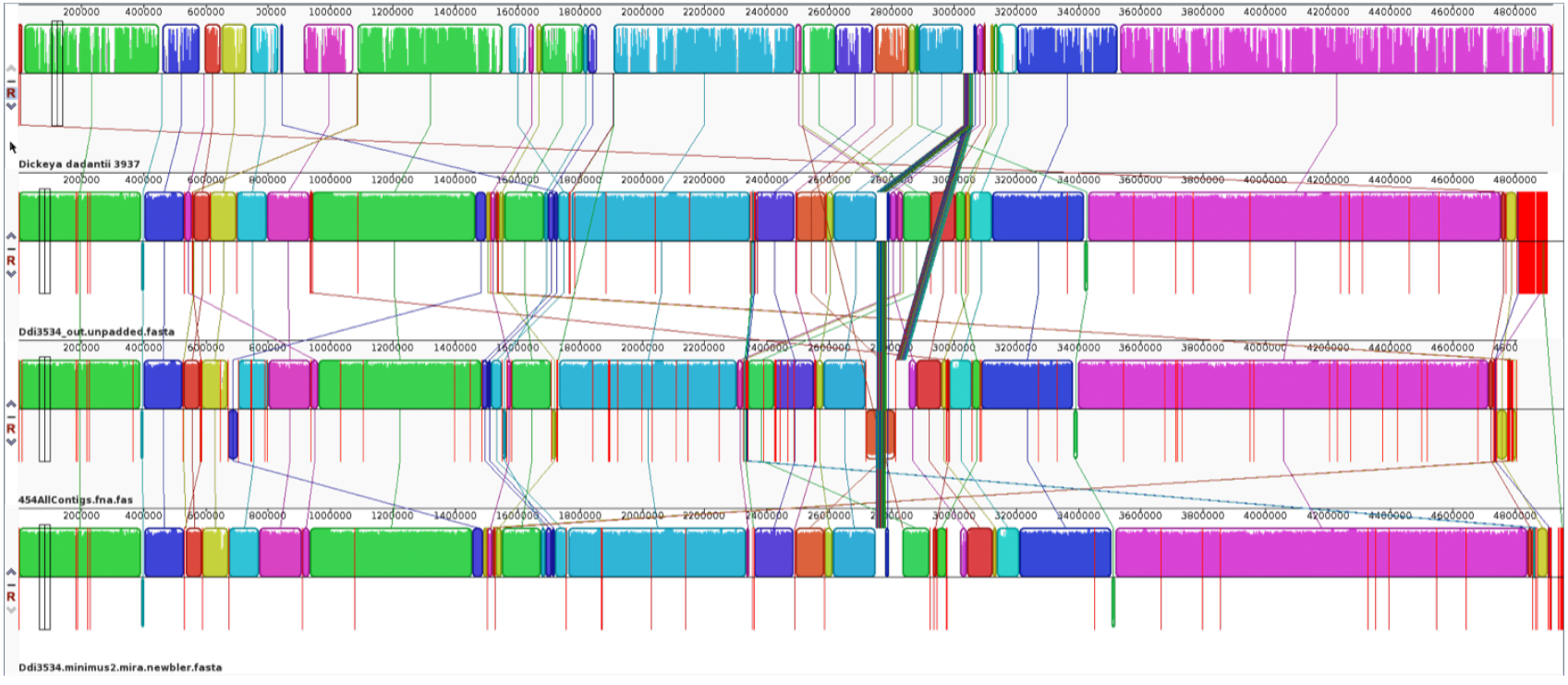


Draft genome alignment

The James Hutton Institute

Alignment-based scaffolding:

- MUMmer: `nucmer`, or `promer` (if divergent)
- MAUVE/Progressive MAUVE: Mauve Contig Mover (MCM)



Multiple genome alignment

The James Hutton Institute

EXERCISE 5:

`whole_genome_alignment/whole_genome_alignments_B.md`

Multiple genome alignment

The James Hutton Institute

EXERCISE 6:

`ex06_biopython_visualisation.ipynb`

Chromosome painting[a]

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

The James Hutton Institute

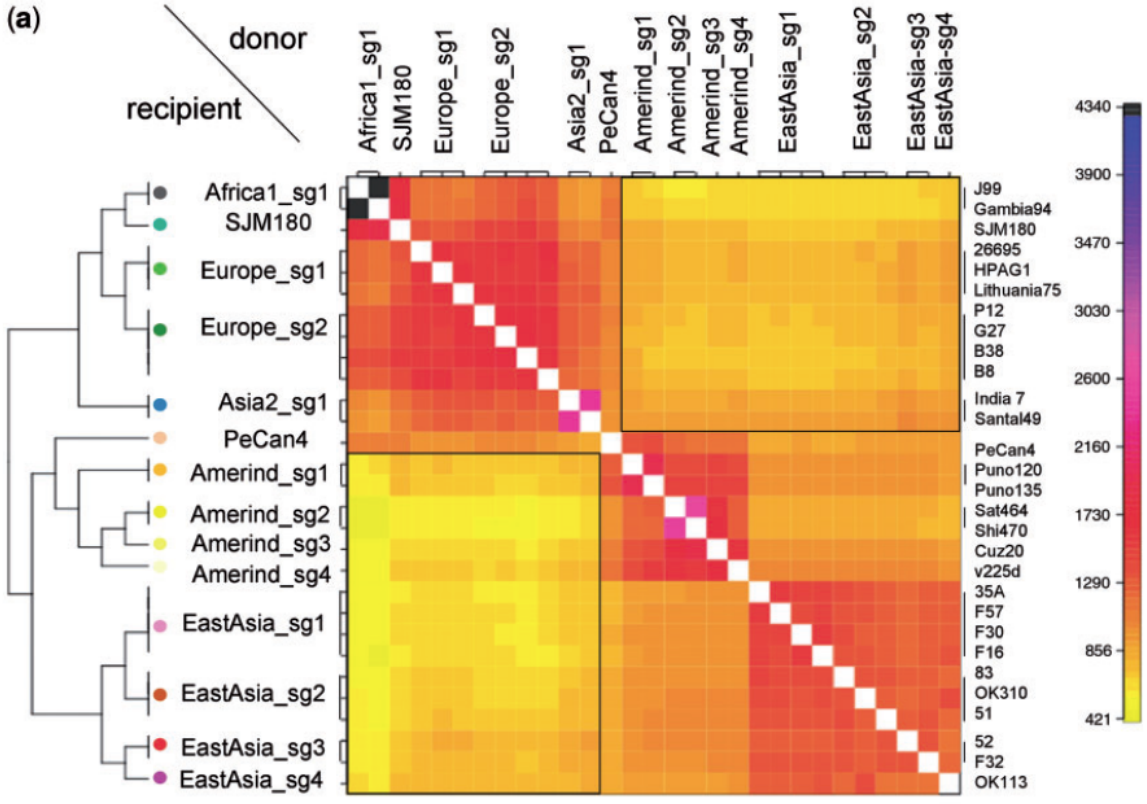
"Chromosome painting" infers recombination-derived 'chunks'
Genome's haplotype constructed in terms of recombination events from a 'donor' to a 'recipient' genome

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

Licence: CC-BY-SA

The James Hutton Institute

By: Leighton Pritchard