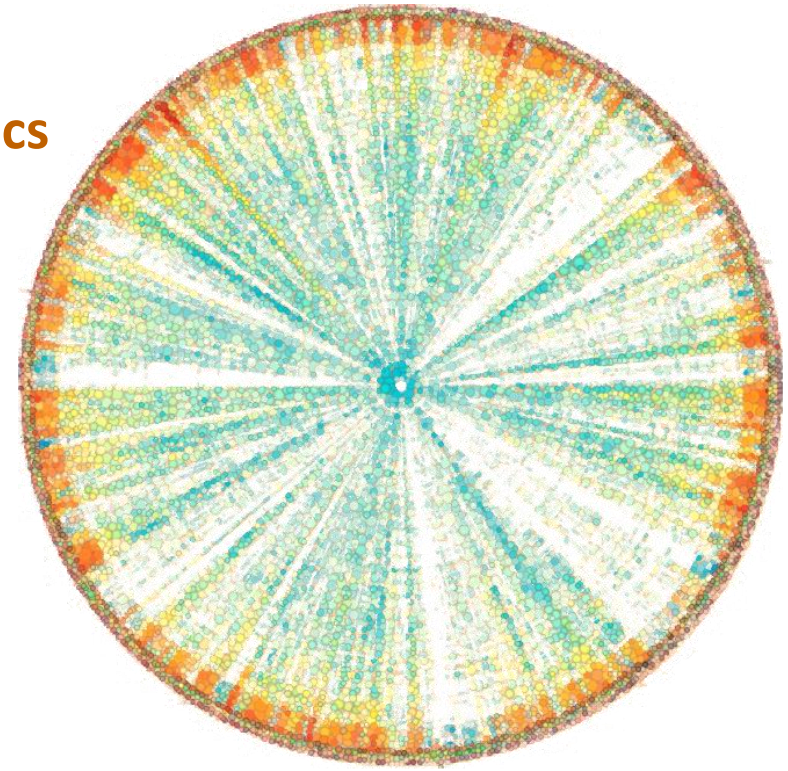
Comparative Genomics and Visualisation

- **What is comparative genomics?**
- **Levels of genome comparison**
  - bulk, whole sequence, features
- **Computational Comparative Genomics**
  - Bulk properties
  - Whole genome comparisons

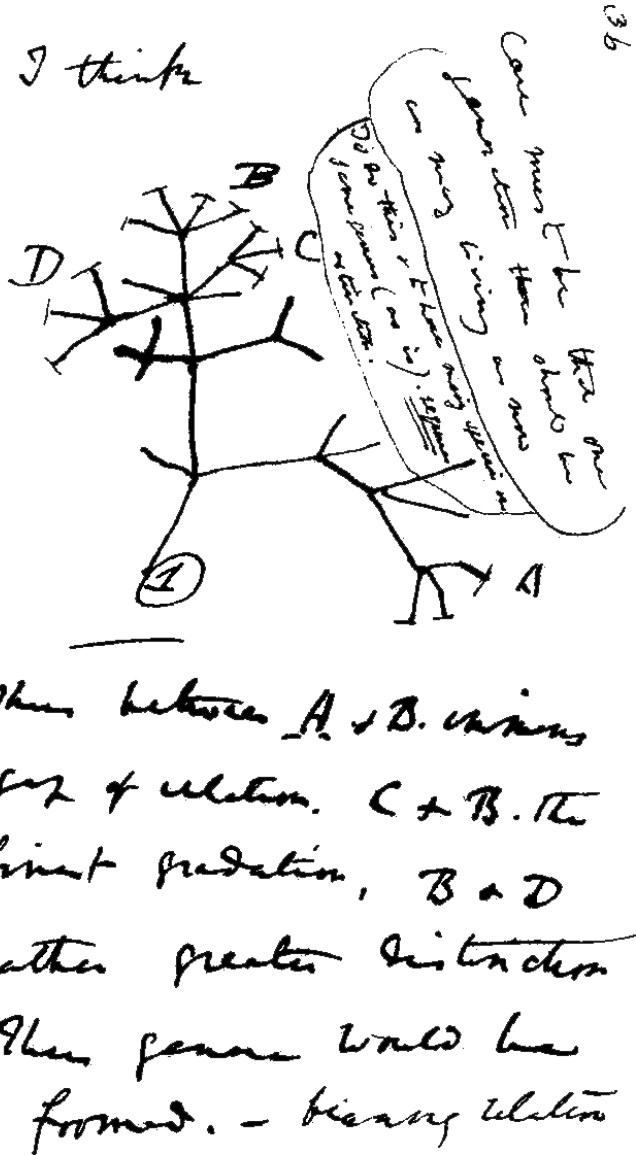


What is Comparative Genomics?

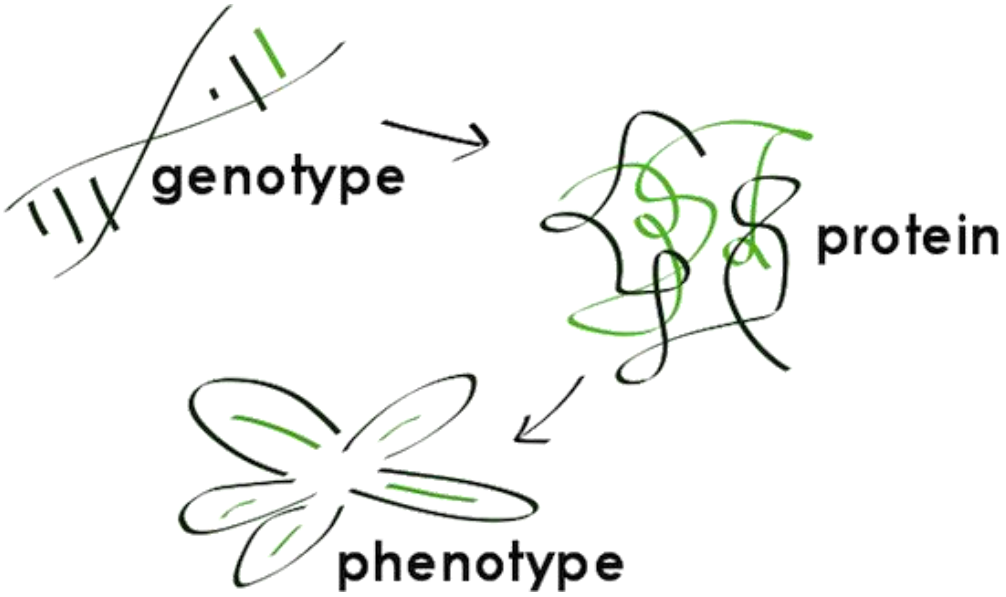
The combination of genomic data and comparative and evolutionary biology to address questions of genome structure, evolution and function.

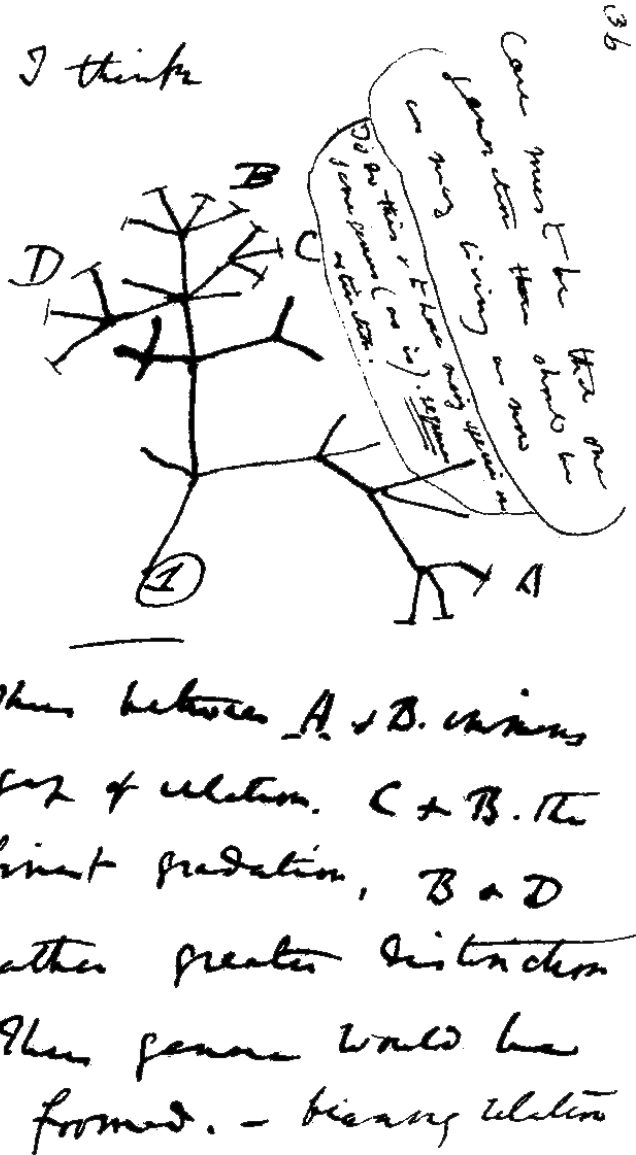
Why Comparative Genomics?

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



Why Comparative Genomics?

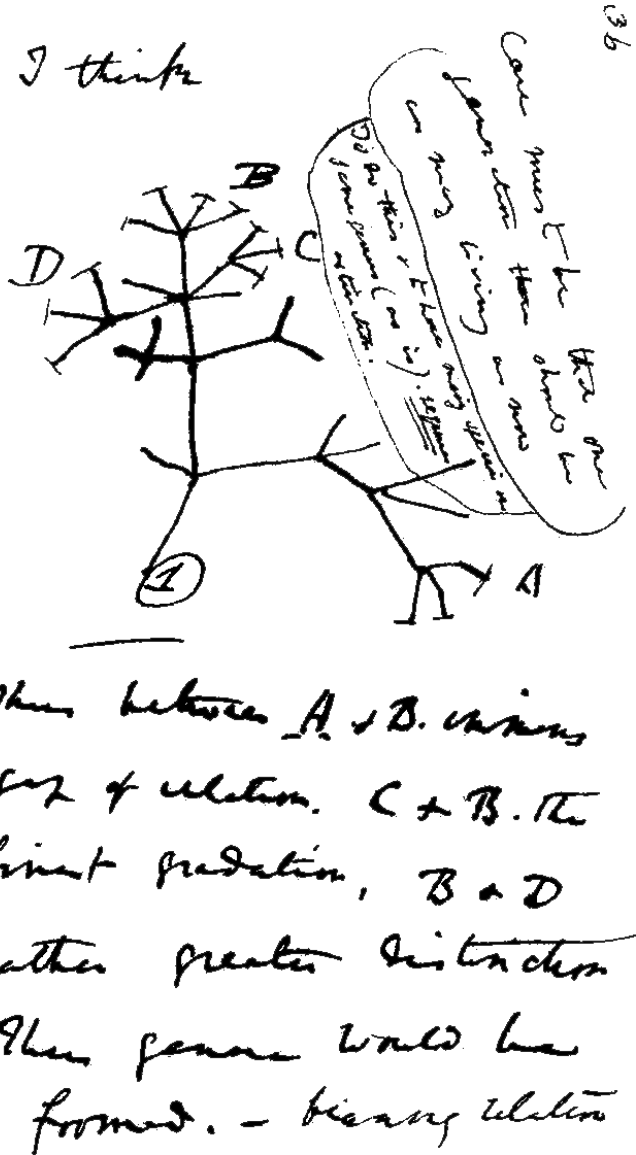




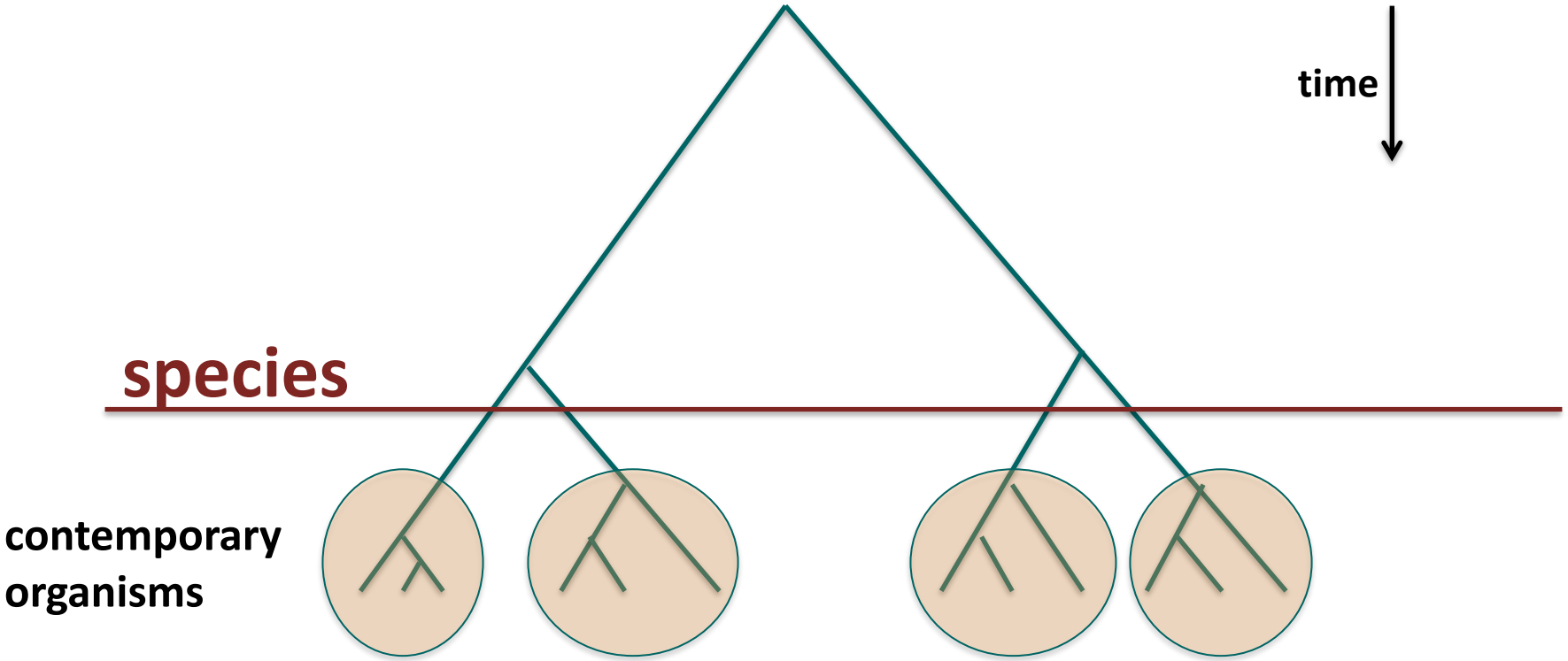
- **BUT:**
  - **Context:** epigenetics, tissue differentiation, mesoscale systems, etc.
  - **Phenotypic plasticity:** responses to temperature, stress, environment, etc.

Why Comparative Genomics?

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



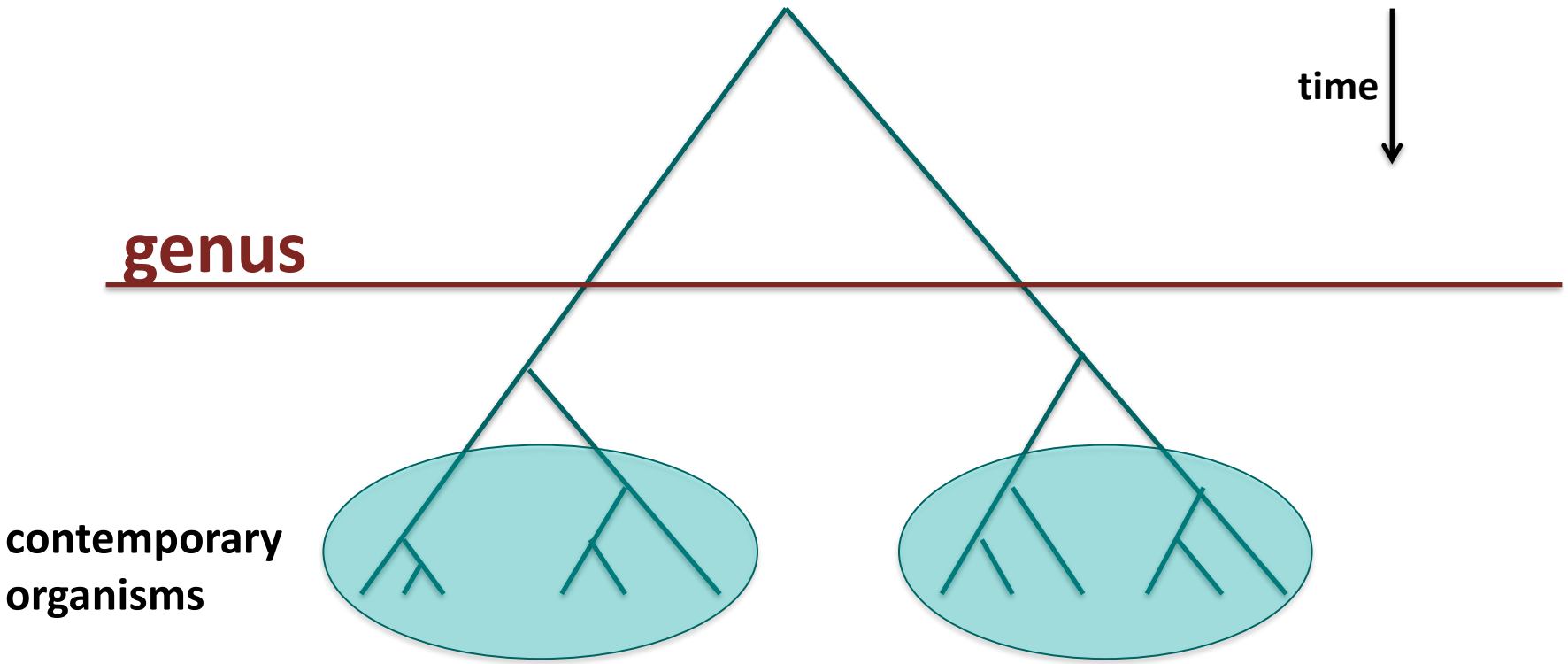
Why Comparative Genomics?



- Comparison within species (e.g. isolate-level – or even within individuals): which genome features may account for unique characteristics of organisms/tumours? Epigenetics in an individual.

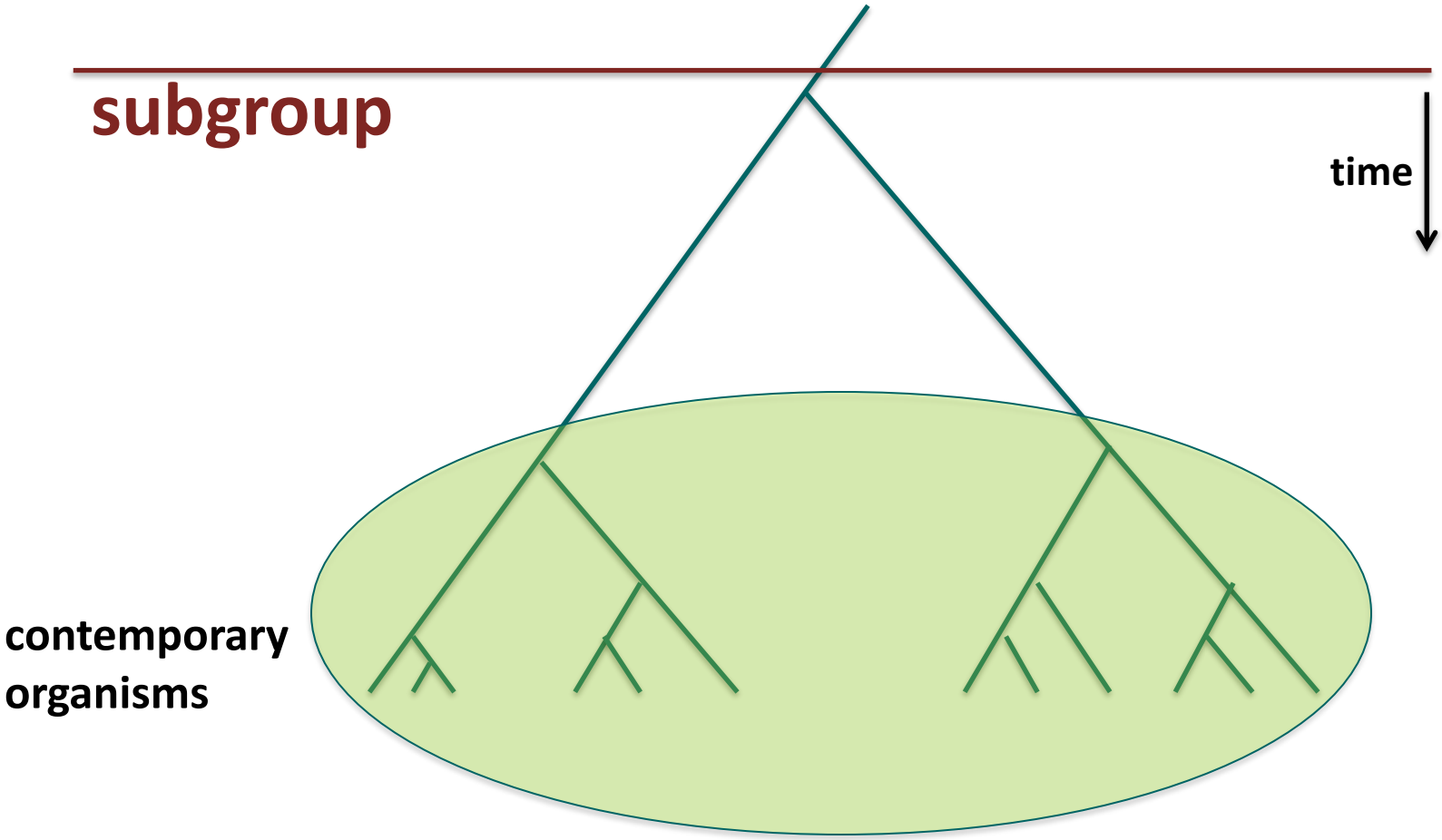
Adapted from Hardison (2003) *PLoS Biol.* doi:10.1371/journal.pbio.0000058

Why Comparative Genomics?



- Comparison within genus (e.g. species-level): what genome features show evidence of selective pressure, and in which species?

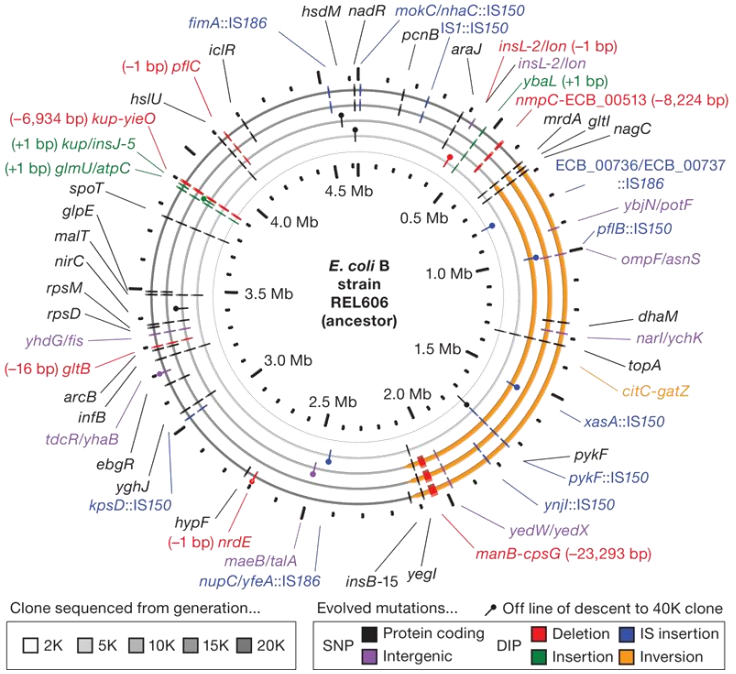
Why Comparative Genomics?

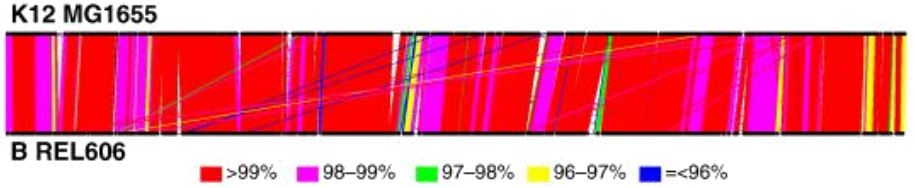


- Comparison within subgroup (e.g. genus-level): what are the core set of genome features that define a subgroup or genus?

The *E.coli* long-term evolution experiment

- 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 *Escherichia coli*!
  - Cultures propagated every day
  - Every 500 generations (75 days), mixed-population samples stored
  - Mean fitness estimated at 500 generation intervals





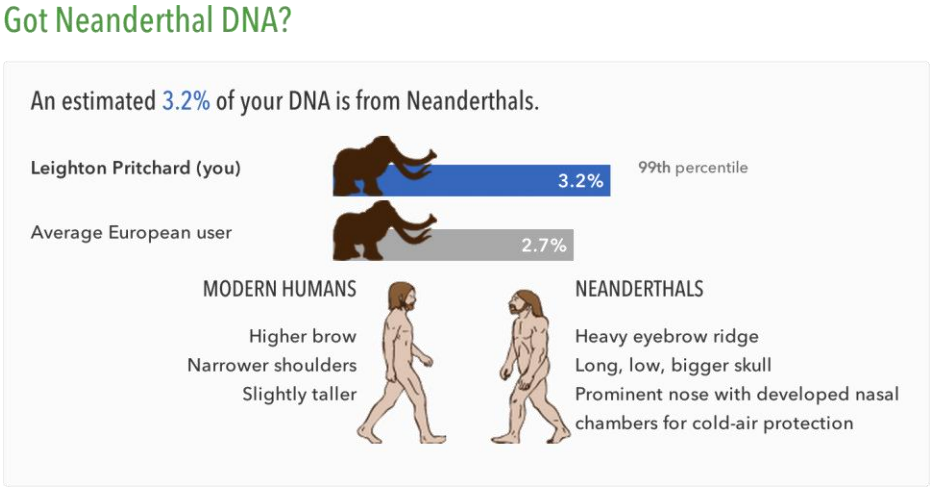
Jeong *et al.* (2009) *J. Mol. Biol.* doi:10.1016/j.jmb.2009.09.052

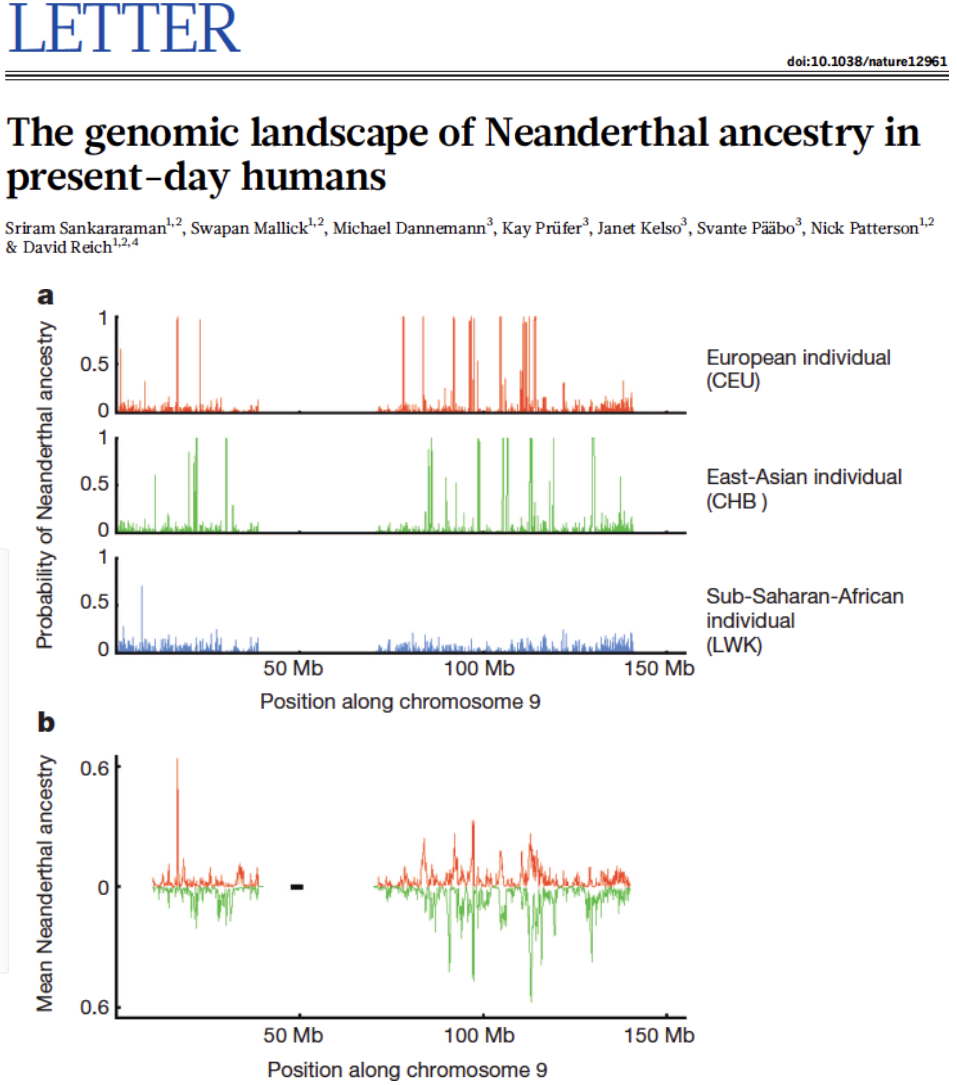
Barrick *et al.* (2009) Nature doi:10.1038/nature08480

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

Comparative Genomics in the News

- Neanderthal alleles:
  - Aid adaptation outwith Africa
  - Associated with disease risk
  - Reduce male fertility





Sankaraman *et al.* (2014) *Nature.* doi:10.1038/nature12961

Levels of Genome Comparison

Genomes are complex, and can be compared on a range of conceptual levels - both practically and *in silico*.

Three broad levels of comparison

- 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.
- Bulk Properties
  - chromosome/plasmid counts and sizes,
  - nucleotide content, etc.
- Whole Genome Sequence
  - sequence similarity
  - organisation of genomic regions (synteny), etc.

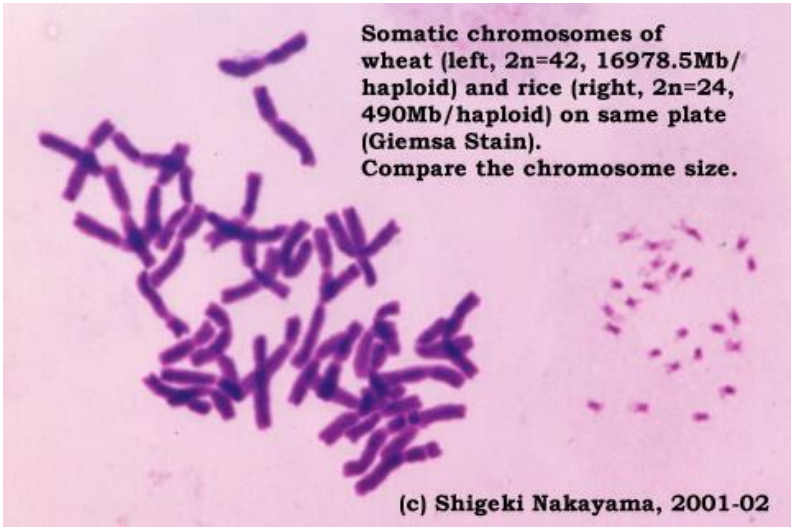
Bulk Genome Properties

- Large-scale summary measurements
- Measure genomes independently – compare values later
  - Number of chromosomes
  - Ploidy
  - Chromosome size
  - Nucleotide (A, C, G, T) frequency/percentage

Chromosome Counts/Size

- The chromosome counts/ploidy of organisms can vary widely
  - *Escherichia coli*: 1 (but plasmids…)
  - Rice (*Oryza sativa*): 24 (but mitochondria, plastids etc…)
  - Human (*Homo sapiens*): 46, diploid
  - Adders-tongue (*Ophioglossum reticulatum*): up to 1260
  - Domestic (but not wild) wheat somatic cells hexaploid, gametes haploid

- Physical genome size (related to sequence length) can also vary greatly
- Genome size and chromosome count do **not** indicate organism 'complexity'
- Still surprises to be found in physical study of chromosomes! (e.g. Hi-C)



Kamisugi *et al.* (1993) *Chromosome Res.* 1(3): 189-96
Wang *et al.* (2013) *Nature Rev Genet.* doi:10.1038/nrg3375

Nucleotide Content

- Experimental approaches for accurate measurement
  - e.g. use radiolabelled monophosphates, calculate proportions using chromatography

Nucleic Acids Research, Vol. 19, No. 19 5181–5185

Rapid determination of nucleotide content and its application to the study of genome structure

Dan E.Krane, Daniel L.Hartl and Howard Ochman*
Department of Genetics, Box 8232, Washington University School of Medicine, St Louis, MO 63110, USA

Received July 26, 1991; Revised and Accepted September 4, 1991

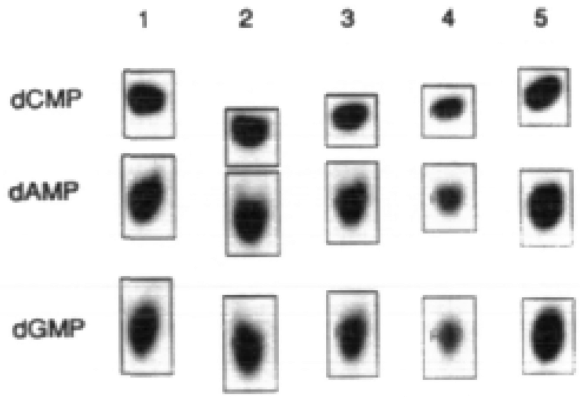


Table 4. Base composititon of yeast artificial chromosomes containing of human DNA inserts.

YAC	Replicate	Corrected G:A	Corrected C:A	%G+C from G:A	%G+C from C:A	Average GC content	Std dev.
yKM19-3	1	0 737	0.799	42.4	44.4	43.4	
(180 kb)	2	0.703	0.771	41.3	43.5	42.4	0 58
	3	0.707	0.768	41.4	43.4	42.4	
yCF-4	1	0.724	0.768	42.0	43.4	42 7	
(330 kb)	2	0.708	0.731	41.4	42.2	41 8	0.45
	3	0.701	0 759	41.2	43.2	42.2	
yW30-5	1	0.722	0 747	41.9	42.7	42.3	
(300 kb)	2	0 744	0.765	42.7	43.3	43.0	0.47
	3	0 713	0.739	41.6	42 5	42.1	
yJ311-2	1	0.725	0.747	42.0	42.8	42 4	
(230 kb)	2	0.728	0.757	42.1	42.4	42.6	0 10
	3	0.728	0.736	42.1	42.4	42.5	
yJ311-5	1	0.695	0 726	41.0	42.1	41.5	
(200 kb)	2	0 746	0.756	42.7	43.0	42.9	0.70
	3	0 731	0 733	42.2	42.3	42.3	
yHPRT	1	0 718	0.744	41.8	42 7	42.2	
(680 kb)	2	0 718	0.725	41.8	42.0	41 9	0.46
	3	0.694	0.713	41.0	41.6	41.3	

Karl (1980) *Microbiol. Rev.* 44(4) 739-796
Krane *et al.* (1991) *Nucl. Acids Res.* doi:10.1093/nar/19.19.5181

Chromosomal Rearrangements

- Genomes are dynamic, and undergo large-scale changes
- Hybridisation used to map genome rearrangement/duplication
  - Separate chromosomes electrophoretically
  - Apply single gene hybridising probes
  - Reciprocal hybridisations indicate translocations

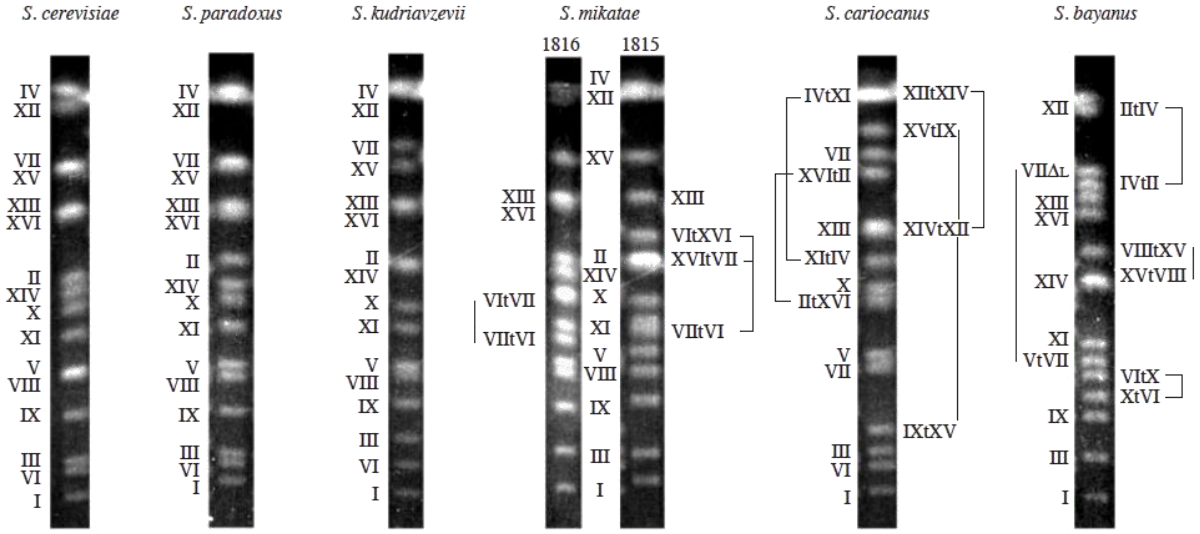


Figure 1 Electrophoretic karyotypes of the *Saccharomyces* 'sensu stricto' species. Strains presented here are *S. cerevisiae* Y55, *S. paradoxus* N17, *S. kudriavzevii* IFO 1802, *S. mikatae* IFO 1816 and IFO 1815, *S. cariocanus* IMUFRJ 50816 and S. *bayanus* CBS 7001. Chromosomes are labelled from I to XVI according to the *S. cerevisiae* nomenclature. Bands showing double intensities correspond to doublets where two non-homologous chromosomes run at the same position. A triplet involving chromosomes II, XIV and XVI t VII is present in the *S. mikatae* IFO 1815 karyotype. Pairs of chromosomes involved in a reciprocal translocation are connected. In *S. bayanus*, the non-reciprocal translocation event is depicted as V t VII connected to VII ΔL (deletion of the left arm of chromosome VII).

Fischer *et al.* (2000) *Nature.* doi:10.1038/35013058

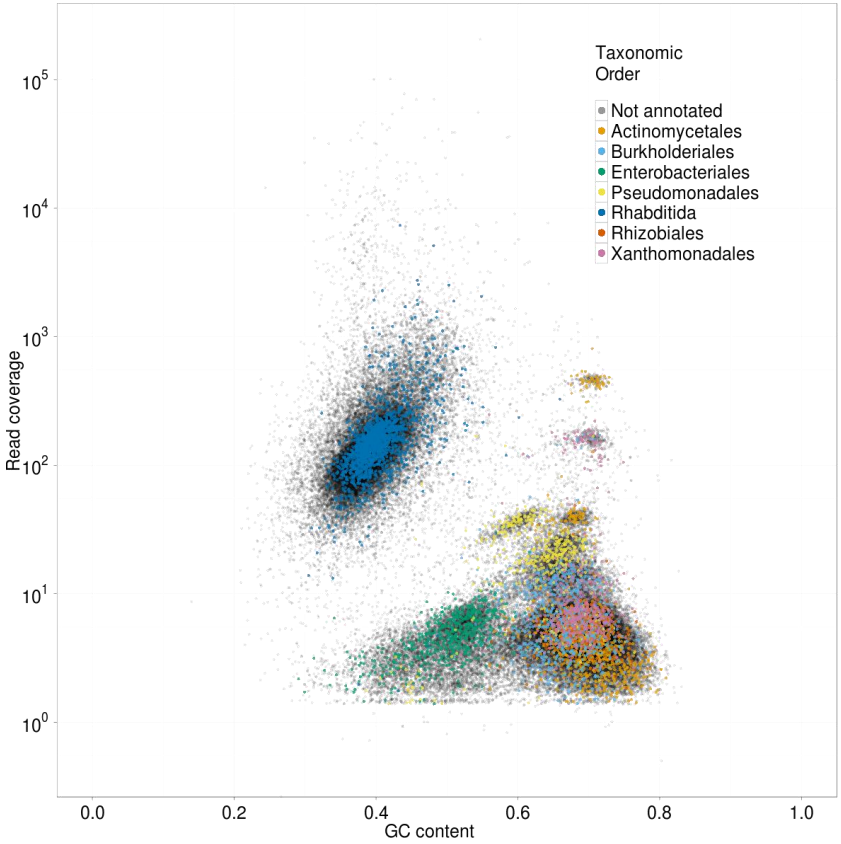
Nucleotide Frequencies/Genome Size

- Very easy to calculate from complete or draft genome sequence
  - (or in a region of genome sequence)

- GC content/chromosome size can be characteristic of an organism
- **[ACTIVITY]**
  - **`bacteria_size_gc`** iPython notebook
  - **`ipython notebook --pylab inline`** in **`bacteria_size`** directory

Blobology

- Metazoan sequence data can be contaminated by microbial symbionts.
  - Host and symbiont DNA have different %GC (and are present in different amounts/coverage)
  - Preliminary genome assembly, followed by read mapping
  - Plot contig coverage against %GC = **Blobology**



- http://nematodes.org/bioinformatics/blobology/

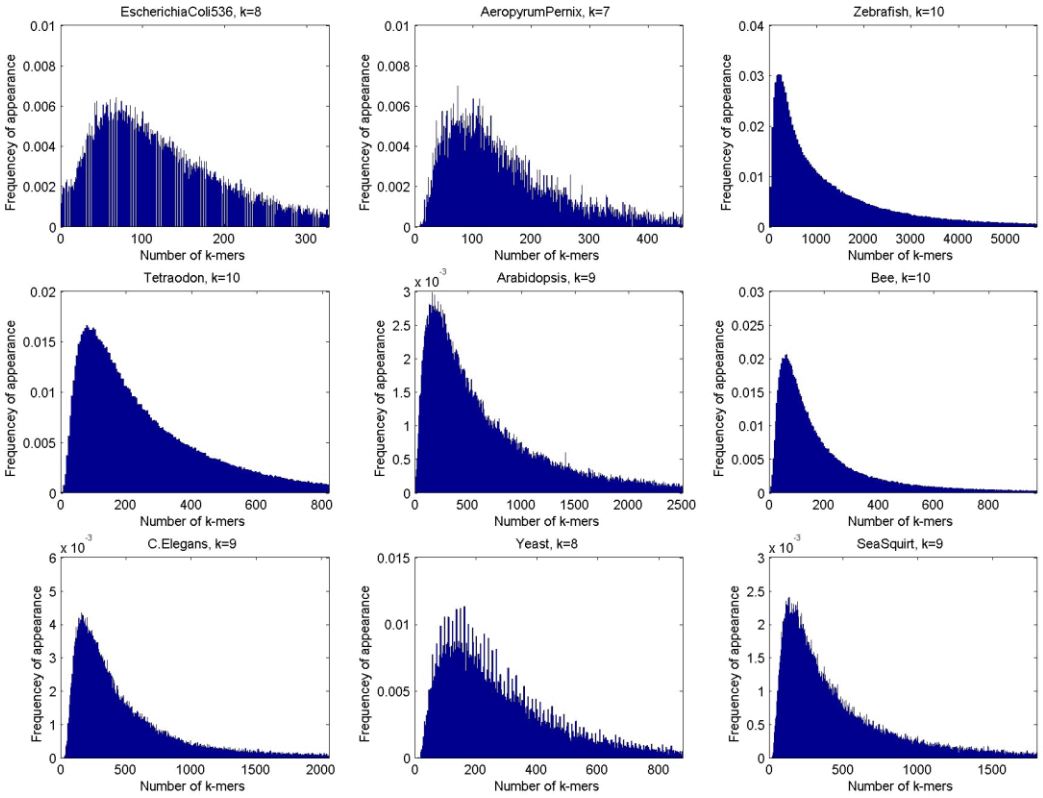
Kumar & Blaxter (2011) *Symbiosis* doi:10.1007/s13199-012-0154-6

Nucleotide *k*-mers

- **Sequence data is required to determine *k*-mers**
- Nucleotide frequencies:
  - A, C, G, T
- Dinucleotide frequencies:
  - AA, AC, AG, AT, CA, CC, CG, CT, GA, GC, GG, GT, TA, TC, TG, TT
- Trinucleotide frequencies:
  - 64 trinucleotides
- *k*-nucleotide frequencies:
  - 4^k *k*-mers
- **[ACTIVITY]**
  - `runApp("shiny/nucleotide_frequencies")` **in RStudio**

k-mer Spectra

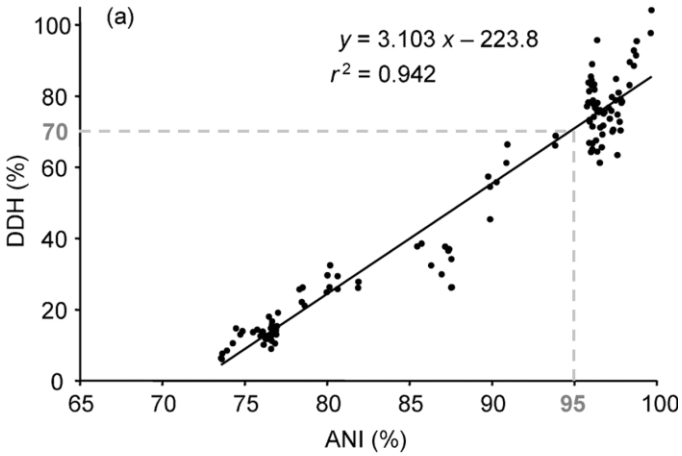
- ***k*-mer spectrum:**
  - Frequency distribution of observed *k*-mer counts
  - Most species have a unimodal *k*-mer spectrum

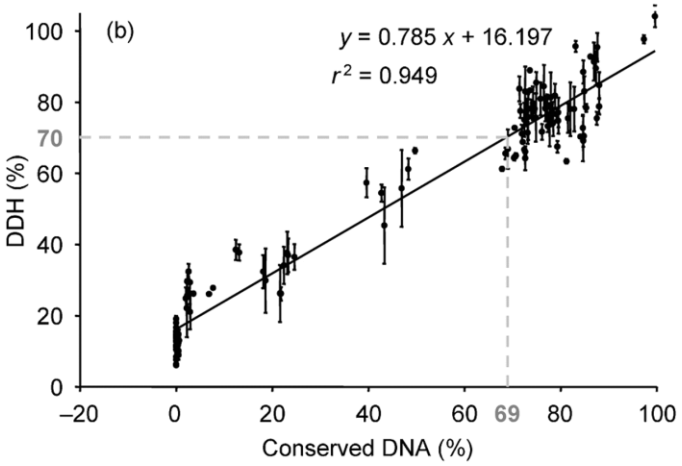


Chor *et al.* (2009) *Genome Biol.* doi:10.1186/gb-2009-10-10-r108

Average Nucleotide Identity (ANI)

- ANI introduced as a substitute for DDH in 2007:
  - 70% identity (DDH) = **"gold standard"** prokaryotic species boundary
  - **70%** identity (DDH) ≈ **95%** identity (ANI)
- Original method emulates physical experiment:
  1. break genome into 1020nt fragments
  2. align fragments using BLASTN
  3. ANI = mean identity of all BLASTN matches with >30% identity over 70% alignable length

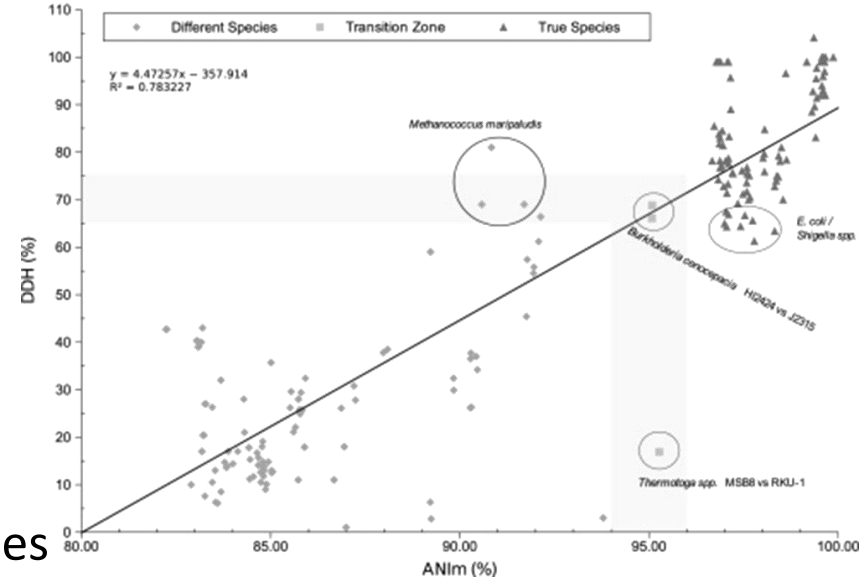




Goris *et al.* (2007) *Int. J. System. Evol. Biol.* doi:10.1099/ijs.0.64483-0

Average Nucleotide Identity (ANI)

- ANI introduced as a substitute for DDH in 2007:
  - 70% identity (DDH) = **"gold standard"** prokaryotic species boundary
  - **70%** identity (DDH) ≈ **95%** identity (ANI)

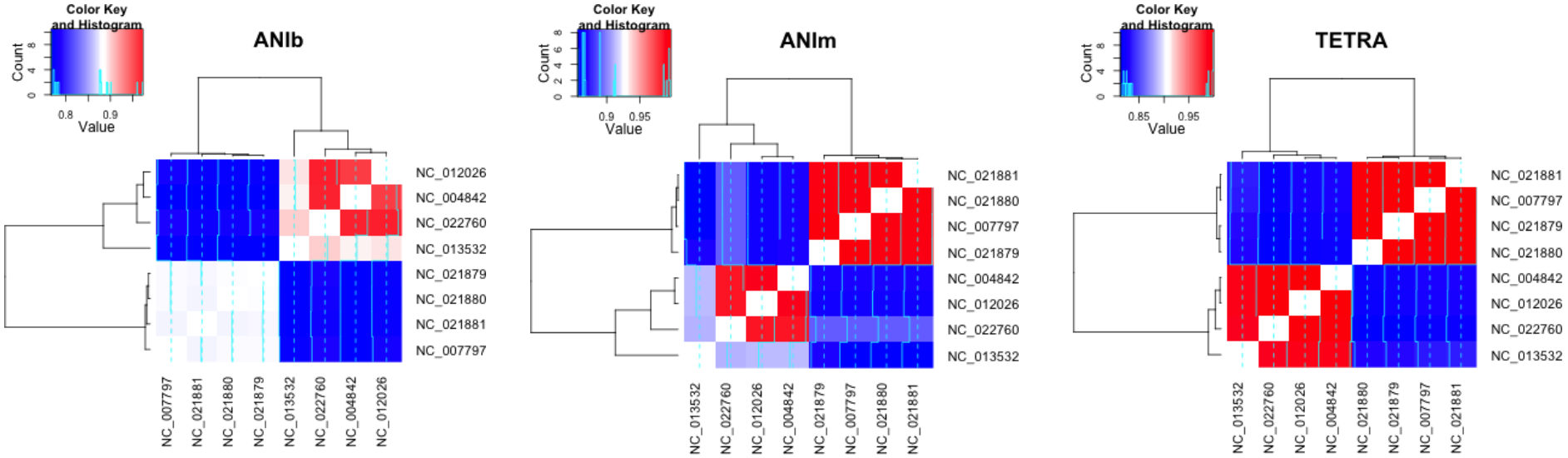


- ANIm and TETRA introduced (2009)
  1. Align sequences using NUCmer
  2. ANI = mean %identity of matches
- TETRA:
  1. Calculate tetranucleotide frequencies
  2. Determine each tetramer deviation from expectation (Z-score)
  3. TETRA = Pearson correlation coefficient of tetramer Z-scores

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

Average Nucleotide Identity (ANI)

- ANIb discards useful information that ANIm retains
- TETRA reflects bulk genome properties rather than selection on sequence



- Data for *Anaplasma marginale* (3), *A.phagocytophilum* (4), *A.centrale* (1)

- TETRA scores are prone to false positives; ANIb scores are prone to false negatives

Average Nucleotide Identity (ANI)

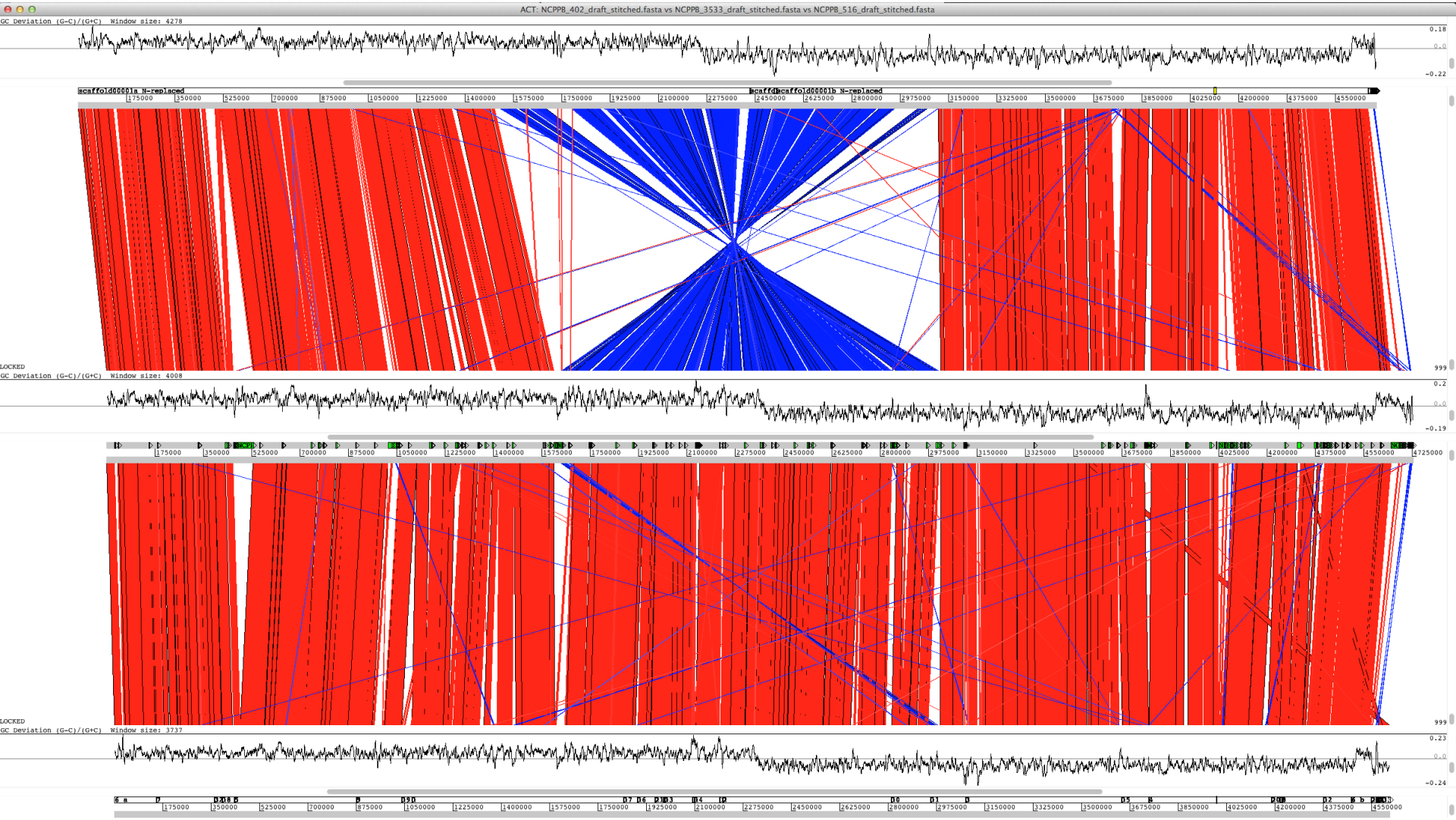
- Jspecies (http://www.imedea.uib.es/jspecies/)
  - WebStart
  - `java -jar -Xms1024m -Xmx1024m jspecies1.2.1.jar`
- Python script
  - `scripts/calculate_ani.py`
- **[ACTIVITY]**
  - `average_nucleotide_identity/README.md` Markdown

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

Whole Genome Sequence Comparisons

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

Whole Genome Alignment



Whole Genome Alignment

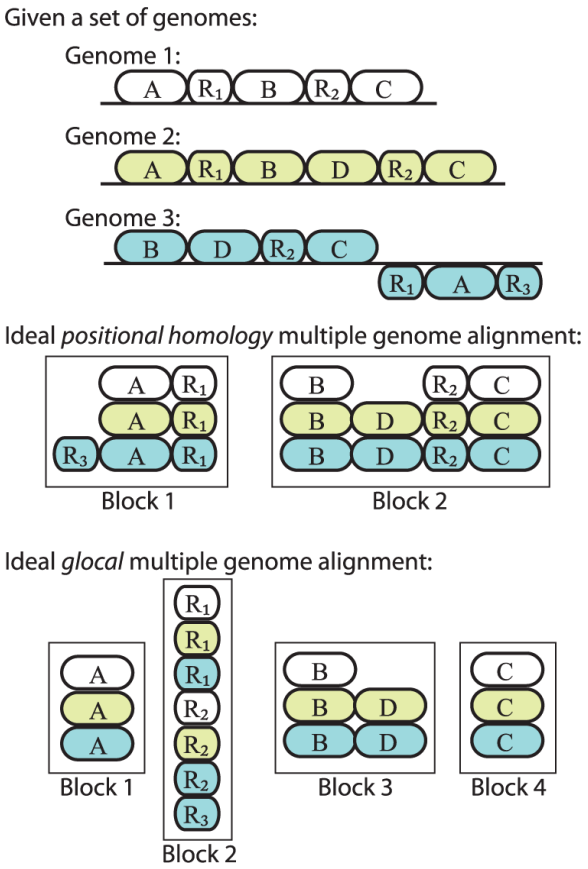
- Which genomes should you align? (or not bother aligning)
- For reasonable analysis, genomes should:
  - derive from a sufficiently **recent** common ancestor: so that homologous regions can be identified.
  - derive from a sufficiently **distant** common ancestor: so that sufficiently "interesting" changes are likely to have occurred
  - help answer your biological question:
    - is your question organism or phenotype specific?
    - are you investigating a process?
- This may be more involved for metazoans (vertebrates, arthropods, nematodes, etc.) than prokaryotes…

Whole Genome Alignment

- Naïve alignment algorithms (e.g. Needleman-Wunsch/Smith-Waterman) are not appropriate:
  - Do not handle rearrangements
  - Computationally expensive on large sequences
- Many whole-genome alignment algorithms proposed, including:
  - LASTZ (http://www.bx.psu.edu/~rsharris/lastz/)
  - **BLAT** (http://genome.ucsc.edu/goldenPath/help/blatSpec.html)
  - Mugsy (http://mugsy.sourceforge.net/)
  - **megaBLAST** (http://www.ncbi.nlm.nih.gov/blast/html/megablast.html)
  - **MUMmer** (http://mummer.sourceforge.net/)
  - LAGAN (http://lagan.stanford.edu/lagan_web/index.shtml)
  - WABA, etc…

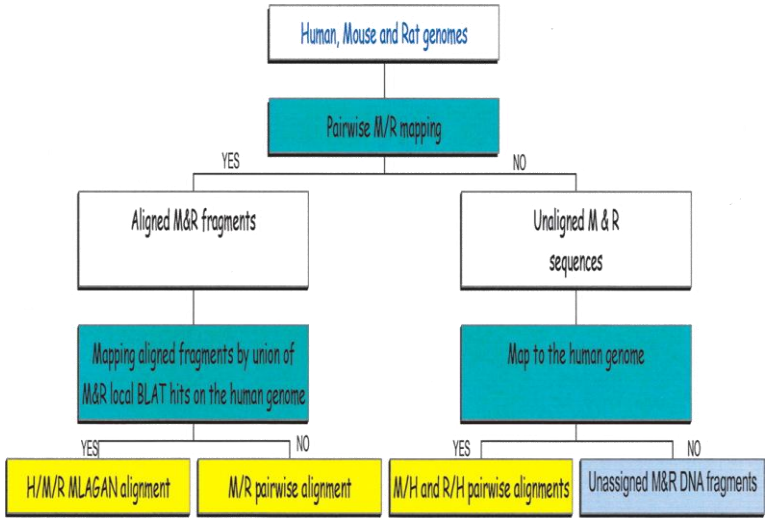
Multiple Genome Alignment

- 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/)
- Positional homology vs. *glocal*

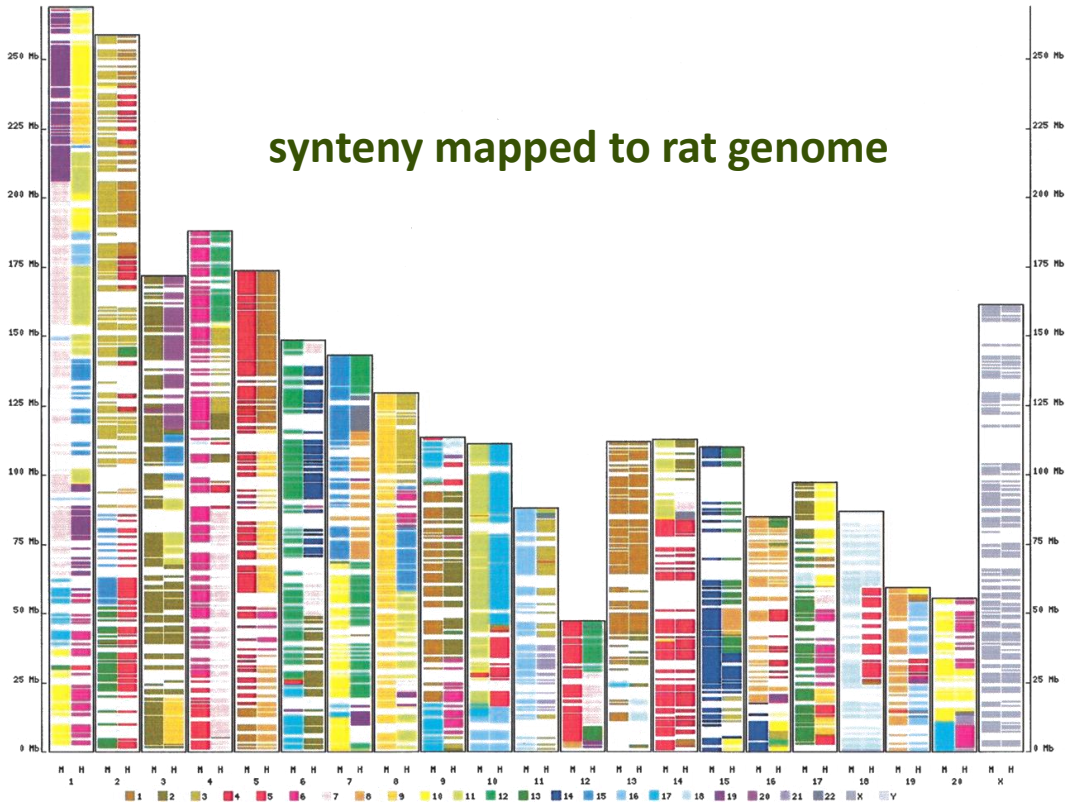


Human-Mouse-Rat Alignment

- Three-way progressive alignment, identifying:
  - Homologous (H/M/R), rodent-only (M/R) and human-mouse or human-rat (H/M, H/R) homologous regions
- Three-way synteny

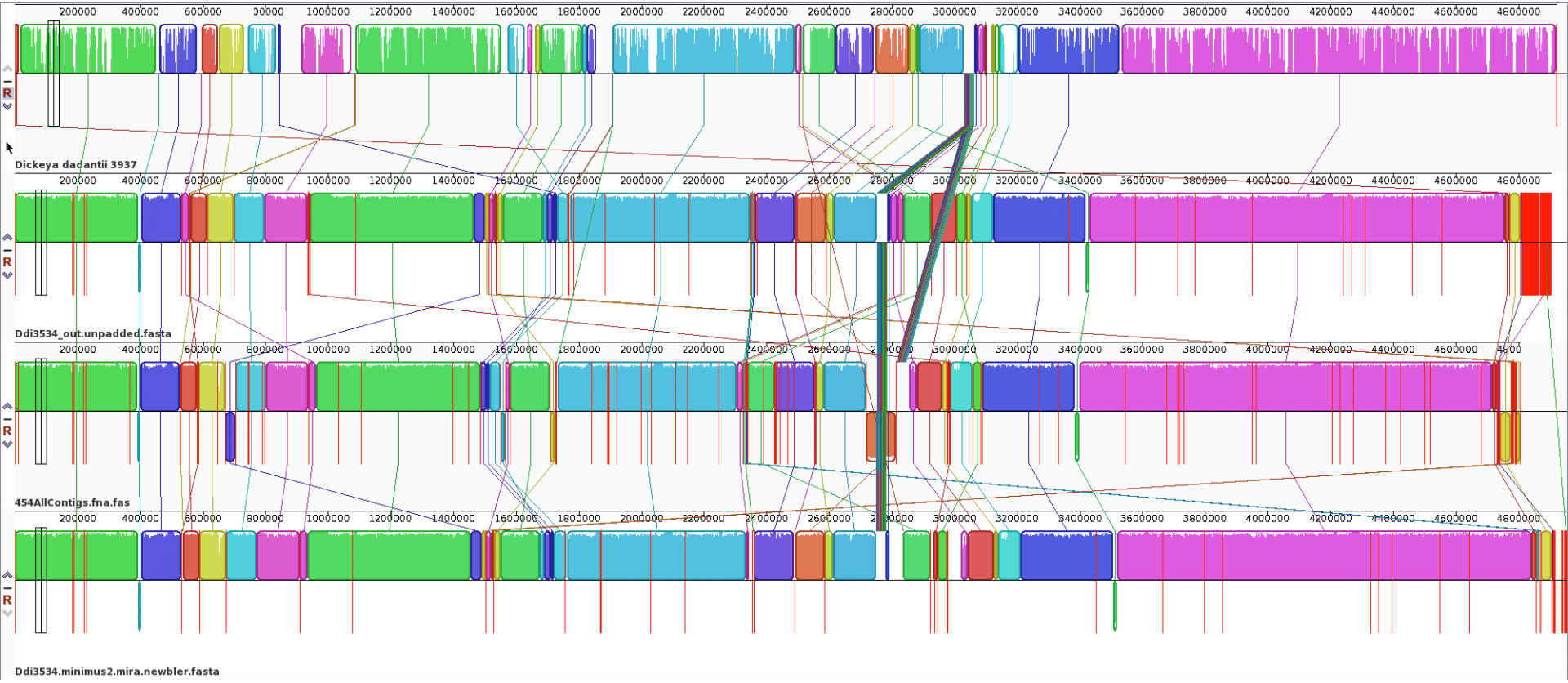


Initial alignments by BLAT
Syntenous regions aligned with LAGAN



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

Draft Genome Alignment



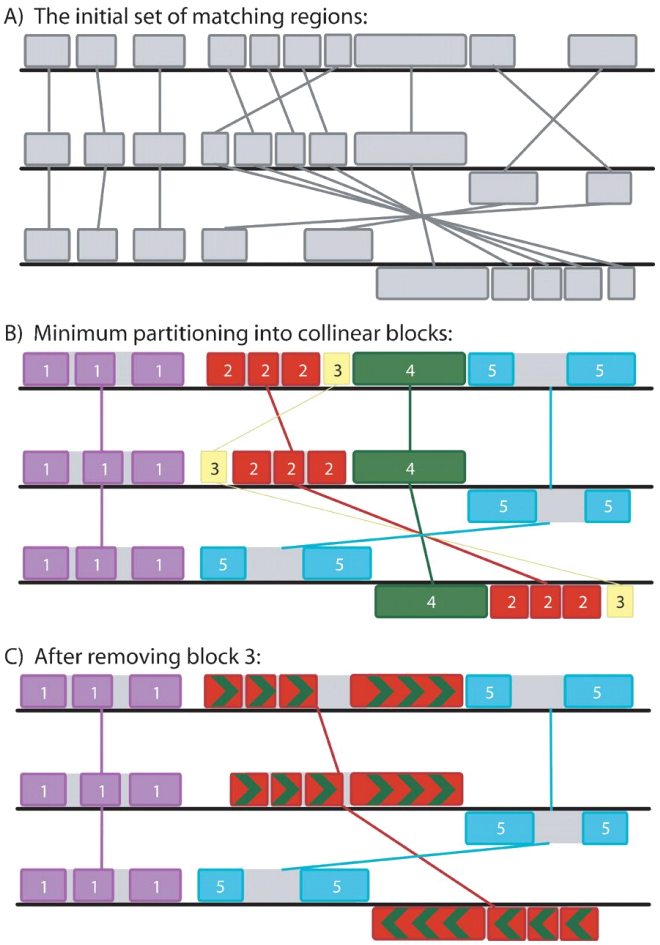
Draft Genome Alignment

- Whole genome alignments useful for scaffolding assemblies
  - High-throughput sequence assemblies come in fragments (contigs)
  - Contigs can sometimes be ordered if paired reads or long read technologies are used
  - Can also align to a known reference genome
- MUMmer
  - Can use NUCmer or, for more distant relations, PROmer
- Mauve/Progressive Mauve
  - http://gel.ahabs.wisc.edu/mauve/

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

Mauve

- **Mauve**'s alignment algorithm
  1. Find local alignments (multi-MUMs – seed & extend)
  2. Construct phylogenetic guide tree from multi-MUMs
  3. Select subset of multi-MUMs as anchors.
     - Partition anchors into Local Collinear Blocks (LCBs) – *consistently-ordered subsets*
  4. Perform recursive anchoring to identify further anchors
  5. Perform progressive alignment (similar to CLUSTAL), against guide tree



- Mauve Contig Mover (MCM) for ordering contigs

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

Draft Genome Alignment

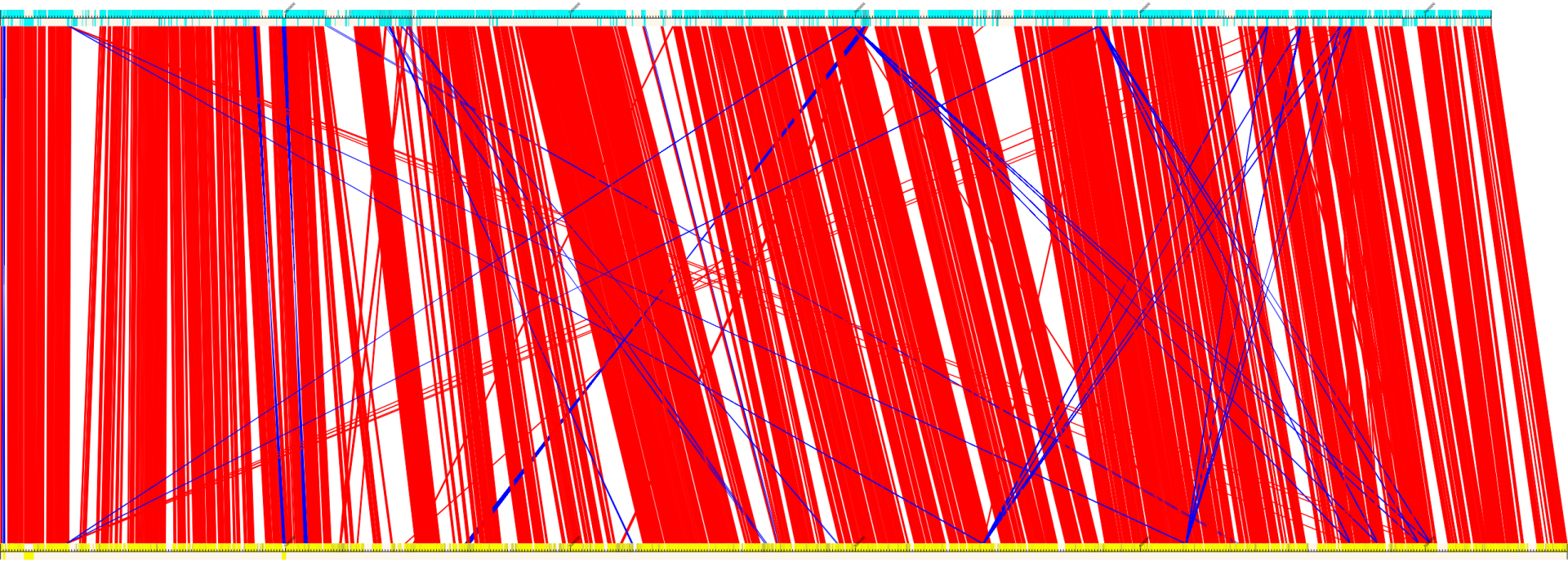
- **[OPTIONAL ACTIVITY] (useful for exercise)**
  - Alignment and reordering of draft genome contigs
  - **`whole_genome_alignments_B.md`** Markdown
  - https://github.com/widdowquinn/Teaching/blob/master/Comparative_Genomics_and_Visualisation/Part_1/whole_genome_alignment/whole_genome_alignments_B.md
- **[ACTIVITY]**
  - Visualisation of whole genome alignment with Biopython
  - **`biopython_visualisation`** iPython notebook

Collinearity and Synteny

- Rearrangements may occur post-speciation
- Different species still exhibit conservation of sequence similarity and order
  - Two elements are *collinear* if they lie in the same linear sequence
  - Two elements are *syntenous* (*syntenic*) if:
    - (*orig.*) they lie on the same chromosome
    - (*mod.*) conservation of blocks of order within the same chromosome
- Signs of evolutionary constraints, including synteny, may indicate functional genome regions
- More about this in **Part 2**, related to genome features

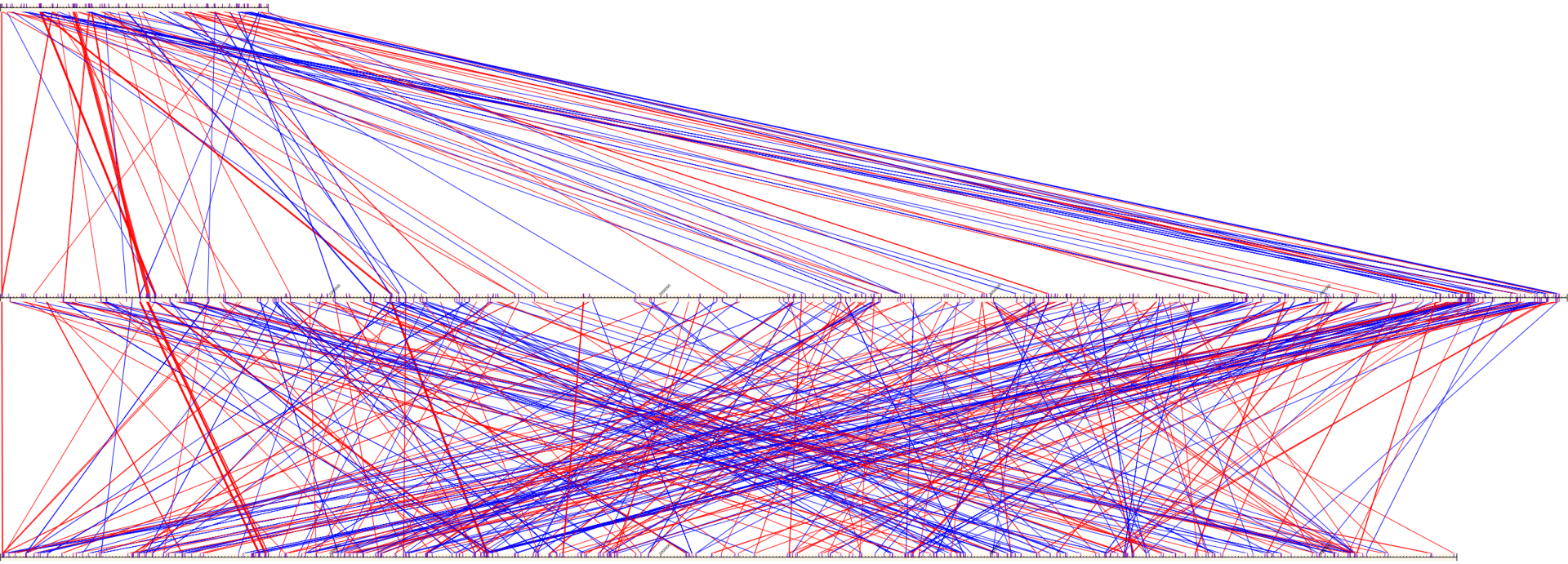
Syntenous

- **`example1.png`** from **`biopython_visualisation`** activity



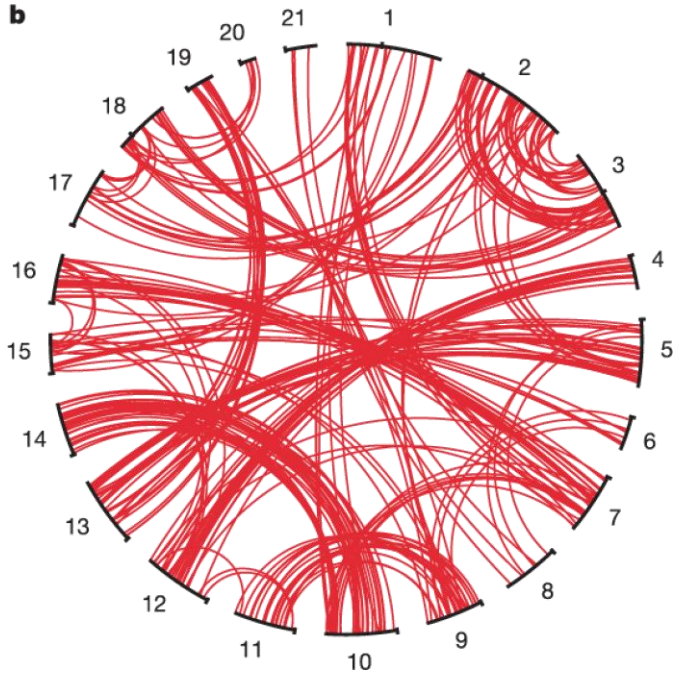
Nonsyntenous

- **`example2.png`** from **`biopython_visualisation`** activity

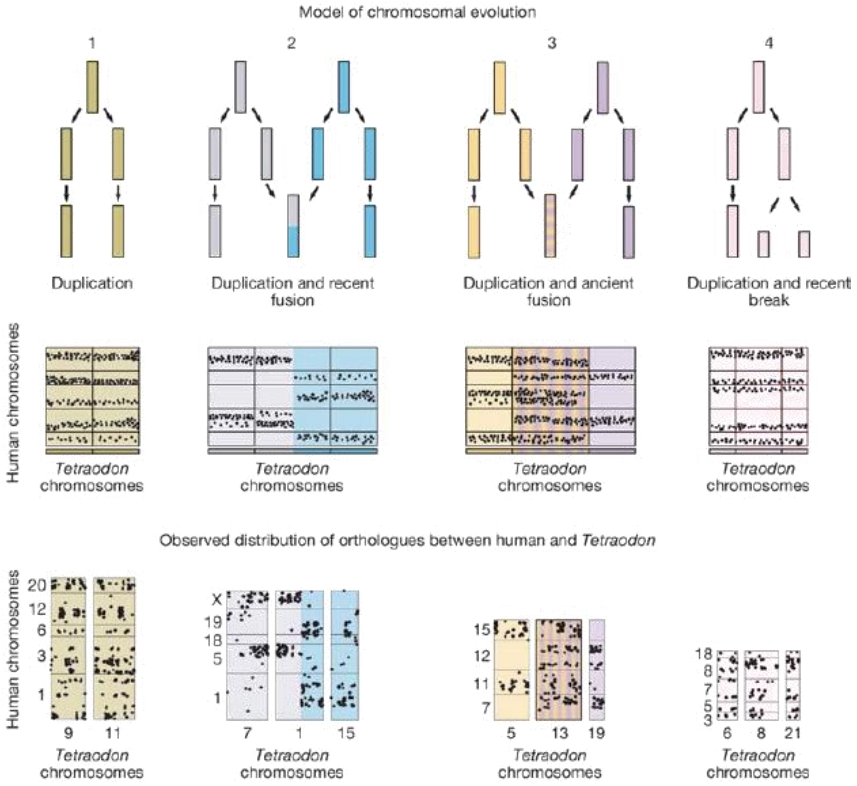


Whole Genome Duplication

- Puffer fish *Tetraodon nigroviridis* (smallest known vertebrate genome)
  - Whole-genome duplication, subsequent to divergence from mammals.
  - Ancestral vertebrate genome inferred to have 12 chromosomes.



Duplicated genes (ExoFish) on 21 chromosomes



Jaillon *et al.* (2004) *Nature* doi:10.1038/nature03025

Conclusion

- Physical and computational genome comparisons:
  - Similar biological questions -> similar concepts
- Lots of sequence data in modern biology
- Conservation ≈ evolutionary constraint
- Many choices of algorithms/analysis software
- Many choices of visualisation software/tools