# Comparative Genomics and Visualisation – Part 1

Leighton Pritchard

The James Hutton Institute

# Part 1

- What is comparative genomics?
- Levels of genome comparison
  - bulk, whole sequence, features
- A Brief History of Comparative Genomics
  - experimental comparative genomics
- Computational Comparative Genomics
  - Bulk properties
  - Whole genome comparisons
- Part 2
  - Genome feature comparisons

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.**

# What is Comparative Genomics?

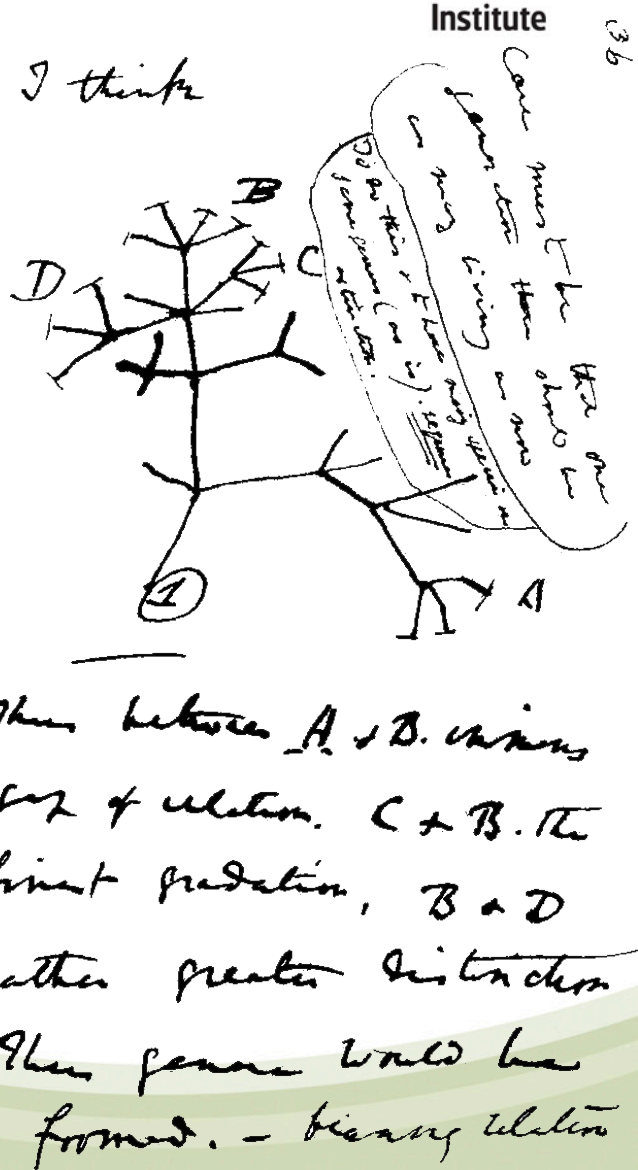
The James Hutton Institute

*"Nothing in biology makes sense, except in the light of evolution"*

**Theodosius Dobzhansky**

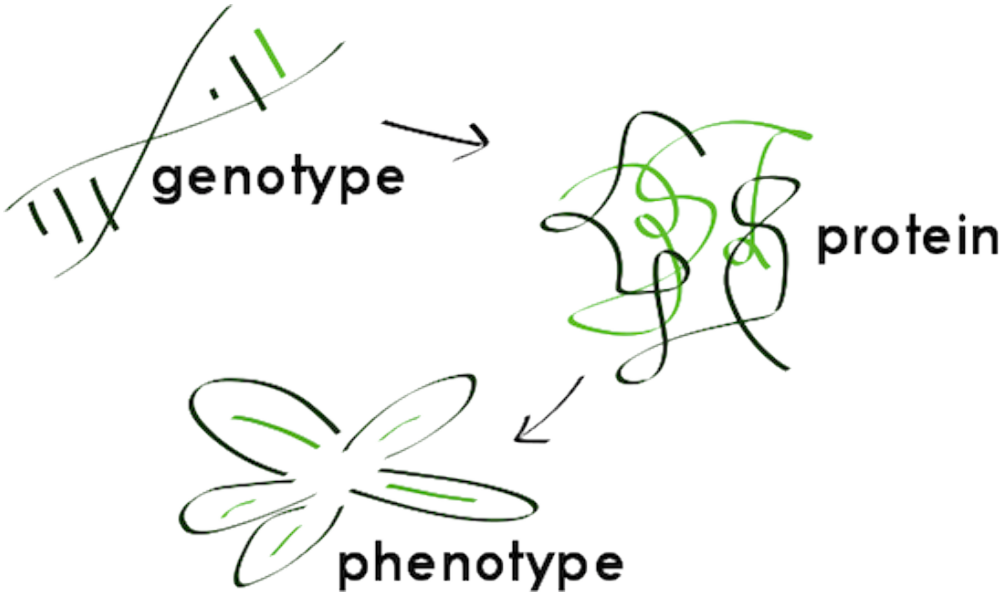
# Why Comparative Genomics?

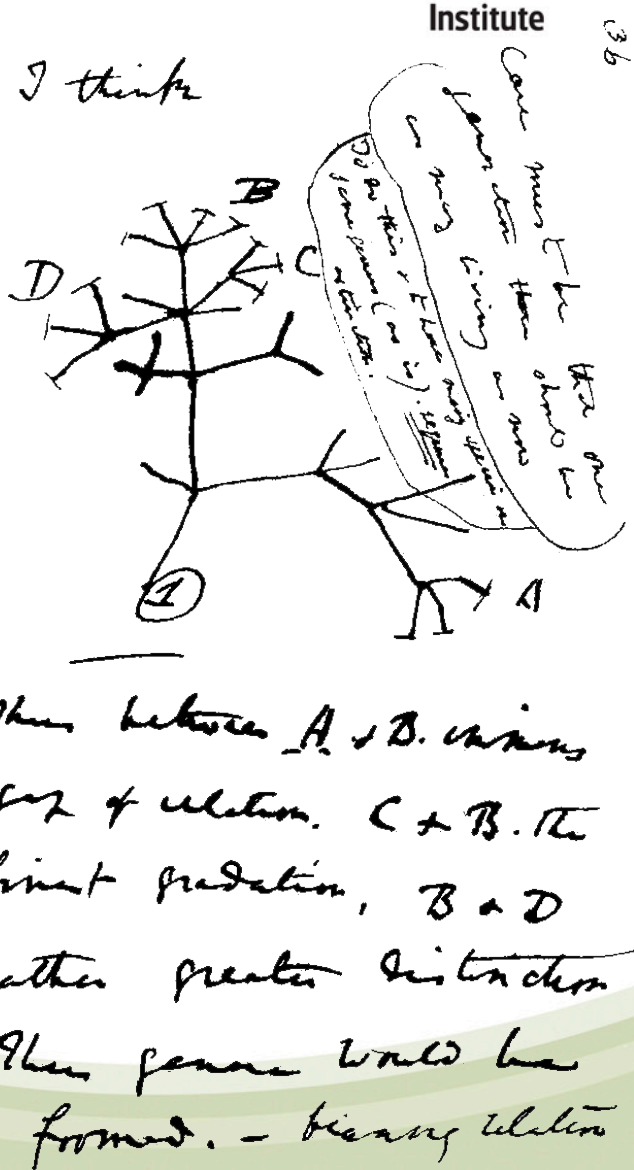
- Genomes describe heritable characteristics
- Related organisms share ancestral genomes
- Functional elements encoded in genomes are common to related organisms
- Functional understanding of model systems (*E. coli*, *A. thaliana*, *D. melanogaster*) can be transferred to non-model systems on the basis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



# Why Comparative Genomics?

The James Hutton Institute



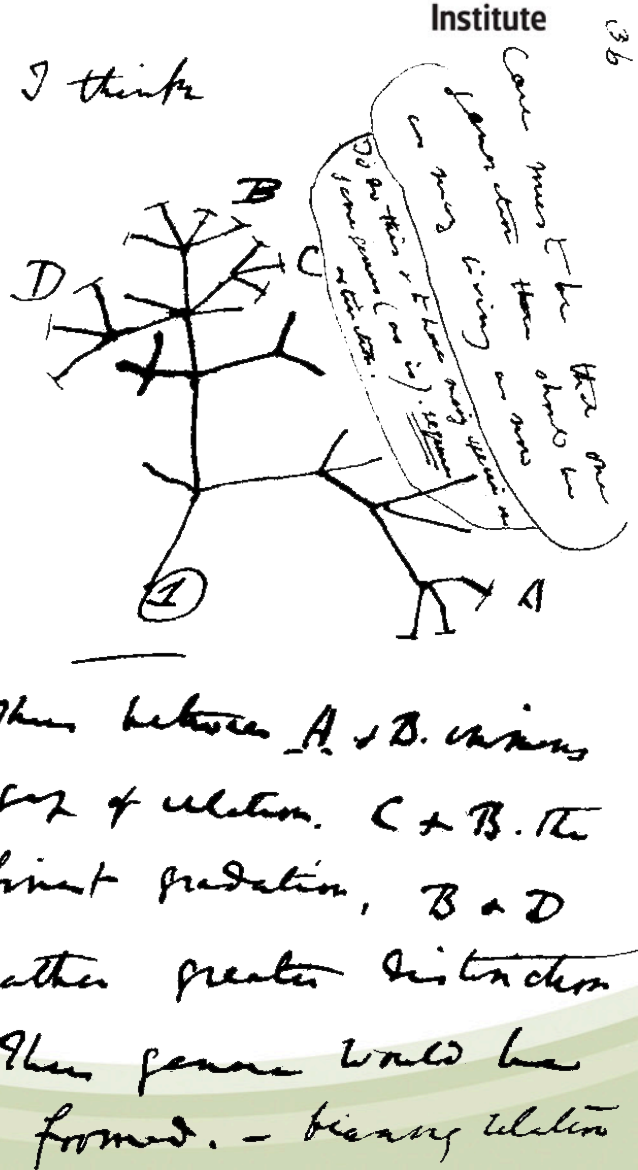




- **BUT:**
  - **Context:** epigenetics, tissue differentiation, mesoscale systems, etc.
  - **Phenotypic plasticity:** responses to temperature, stress, environment, etc.

# Why Comparative Genomics?

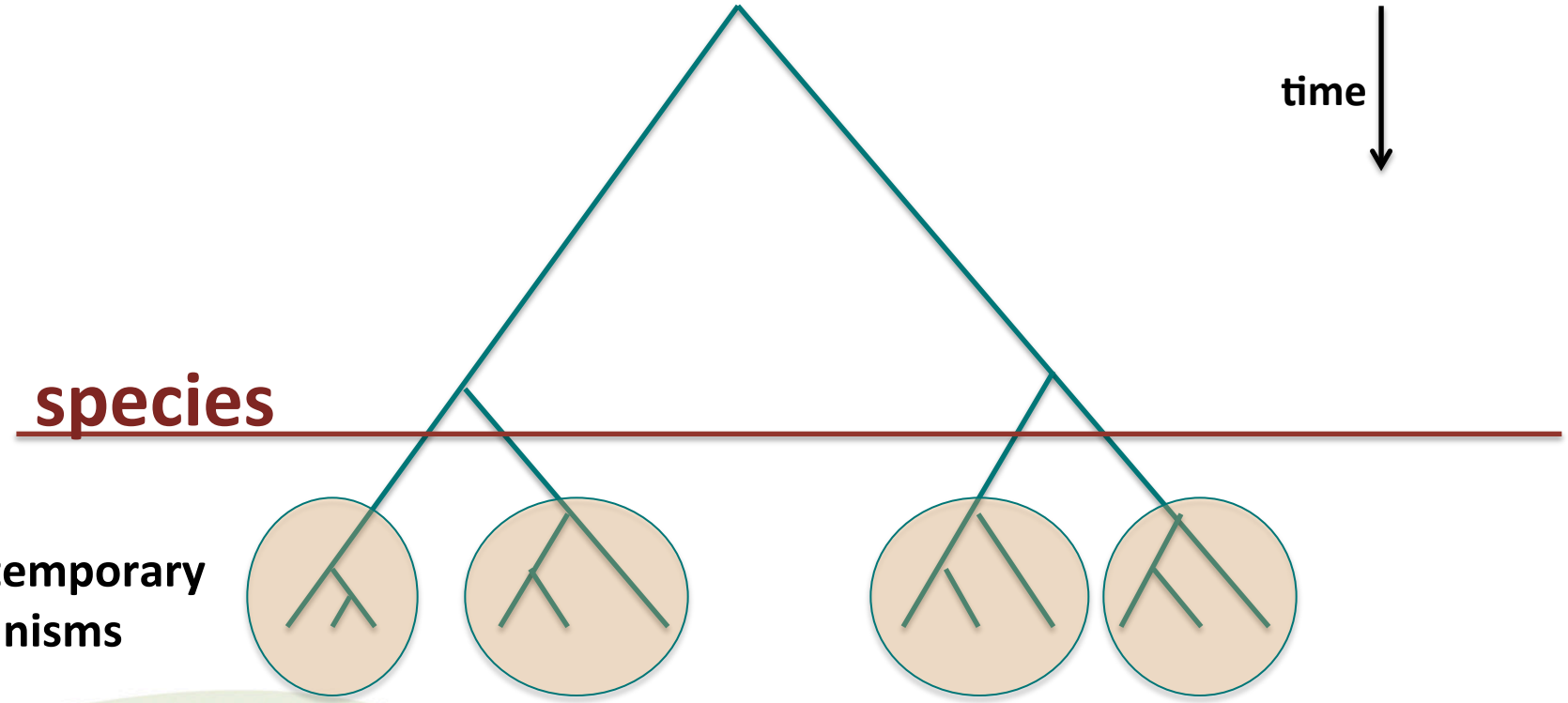
The James Hutton Institute

- 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?

The James Hutton Institute

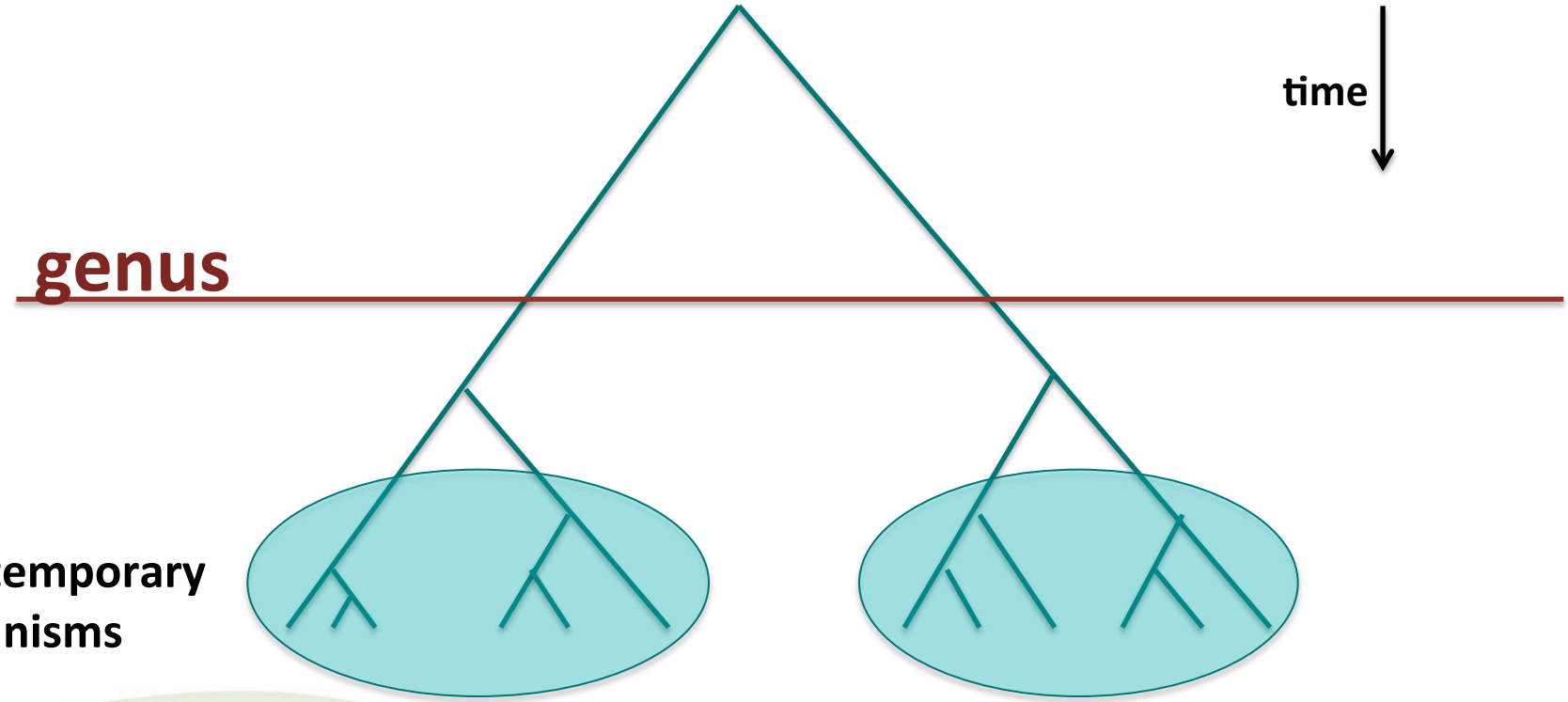



- Comparison within species (e.g. isolate-level – or even within individuals): which genome features may account for unique characteristics of organisms/ tumours? Epigenetics in an individual.

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

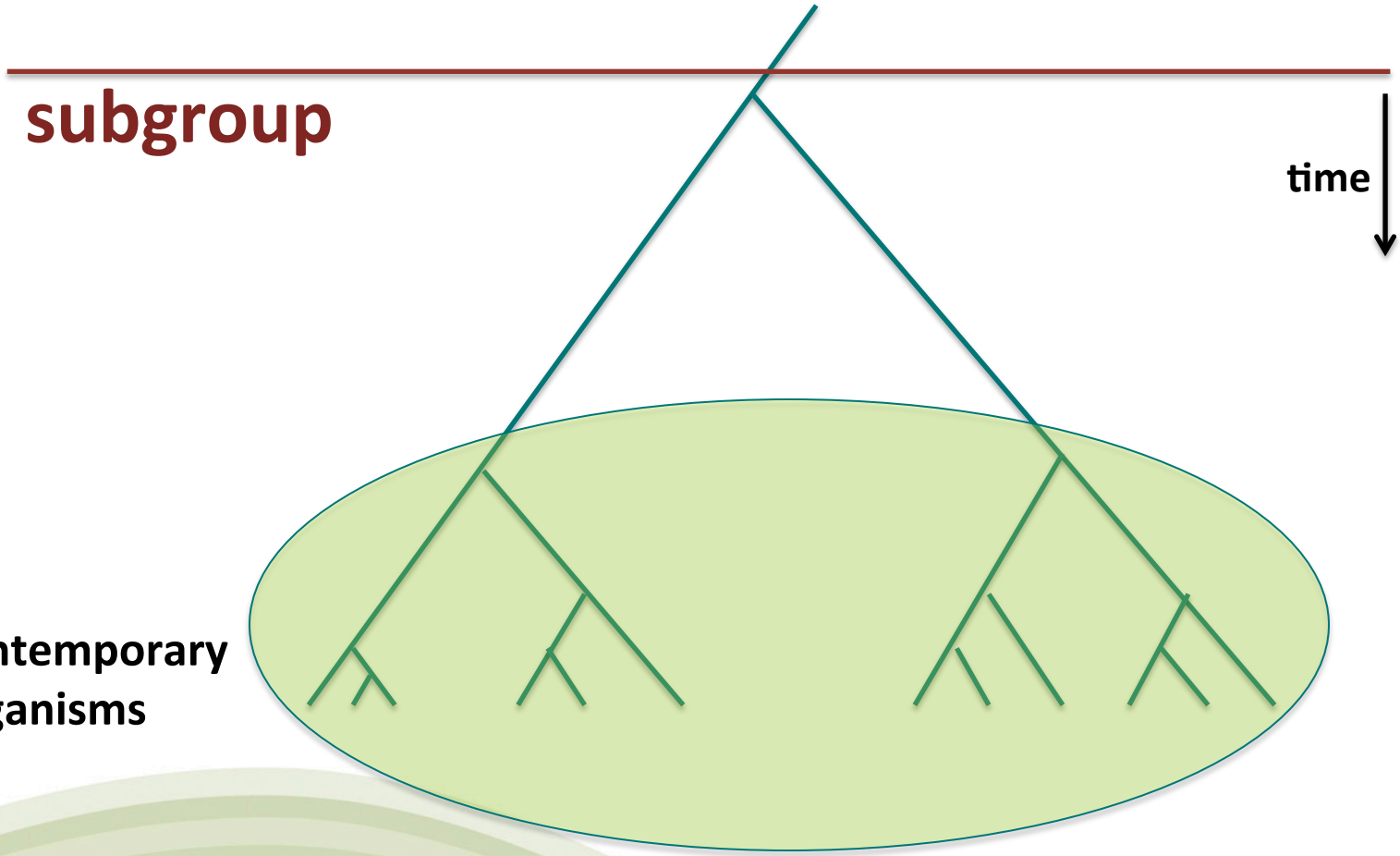
# Why Comparative Genomics?

The James Hutton Institute




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

# Why Comparative Genomics?

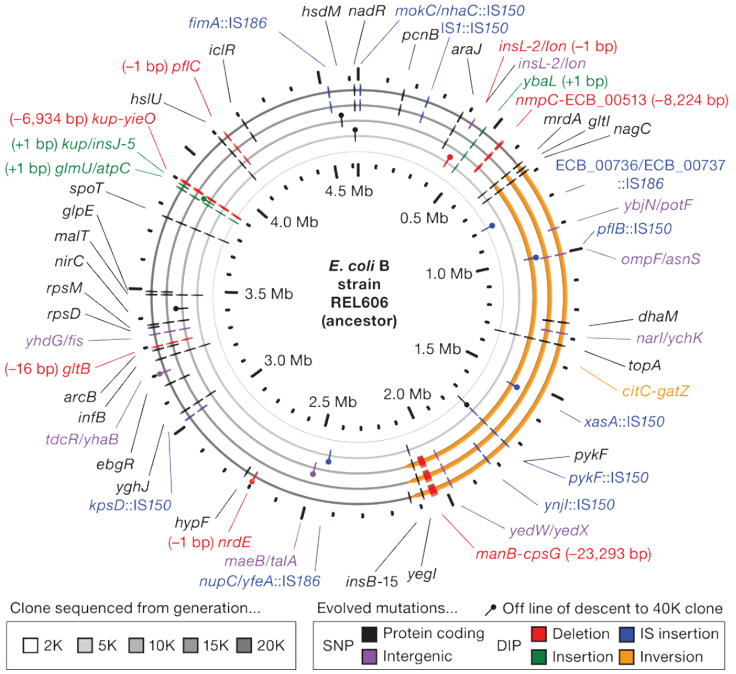


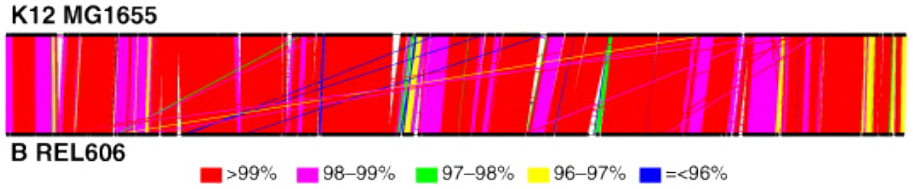

- Comparison within subgroup (e.g. genus-level): what are the core set of genome features that define a subgroup or genus?

# The *E.coli* long-term evolution experiment

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 *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

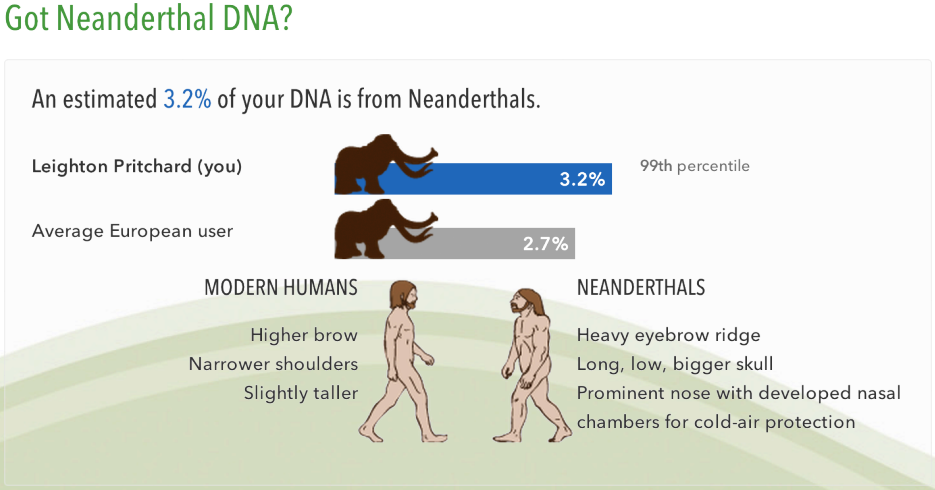
Barrick *et al.* (2009) Nature doi:10.1038/nature08480

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

# Comparative Genomics in the News

The James Hutton Institute

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

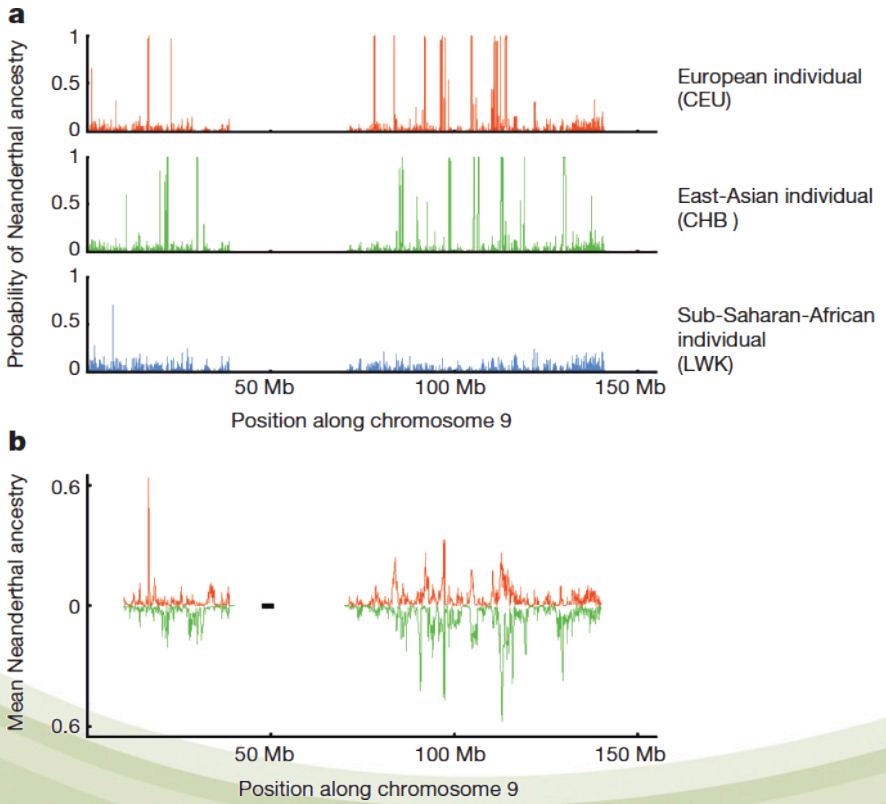



LETTER

doi:10.1038/nature12961

## The genomic landscape of Neanderthal ancestry in present-day humans

Sriram Sankararaman[1,2], Swapan Mallick[1,2], Michael Dannemann[3], Kay Prüfer[3], Janet Kelso[3], Svante Pääbo[3], Nick Patterson[1,2] & David Reich[1,2,4]




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

# Levels of Genome Comparison

The James Hutton Institute

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

# Three broad 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.

# A Brief History of Experimental Comparative Genomics

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 was not always cheap and abundant
- Practical, experimental genome comparisons were needed

# Bulk Genome Property Comparisons

The James Hutton Institute

Values calculated for individual genomes, and subsequently compared.

# Bulk Genome Properties

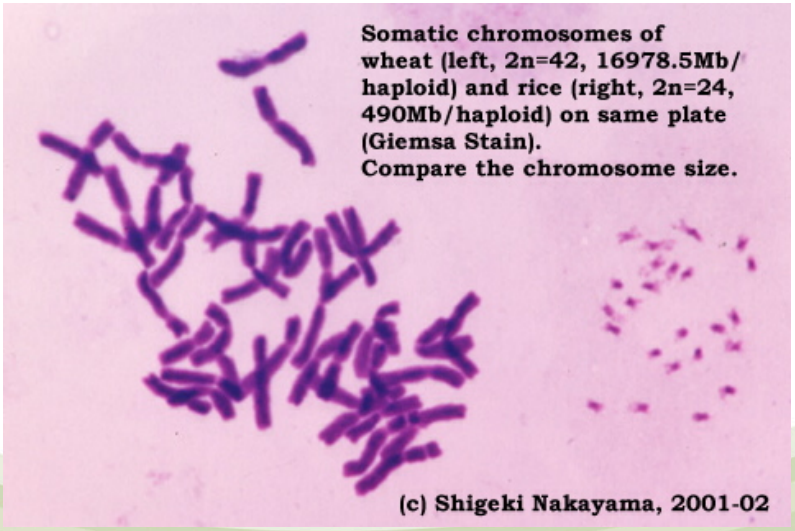
The James Hutton Institute

- 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

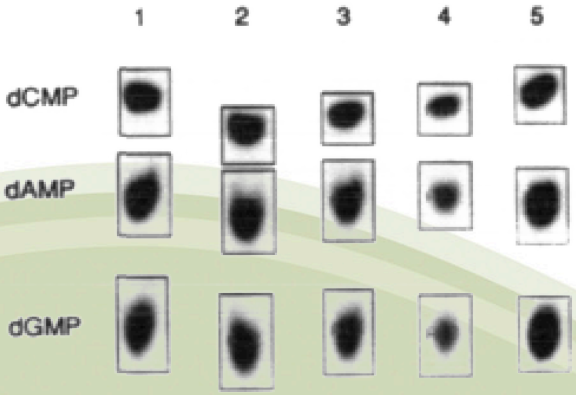



**Table 4.** Base composititon of yeast artificial chromosomes containing of human DNA inserts.

| 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 | 0 748 | 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

# Whole Genome Comparisons

The James Hutton Institute

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

# Whole Genome Comparisons

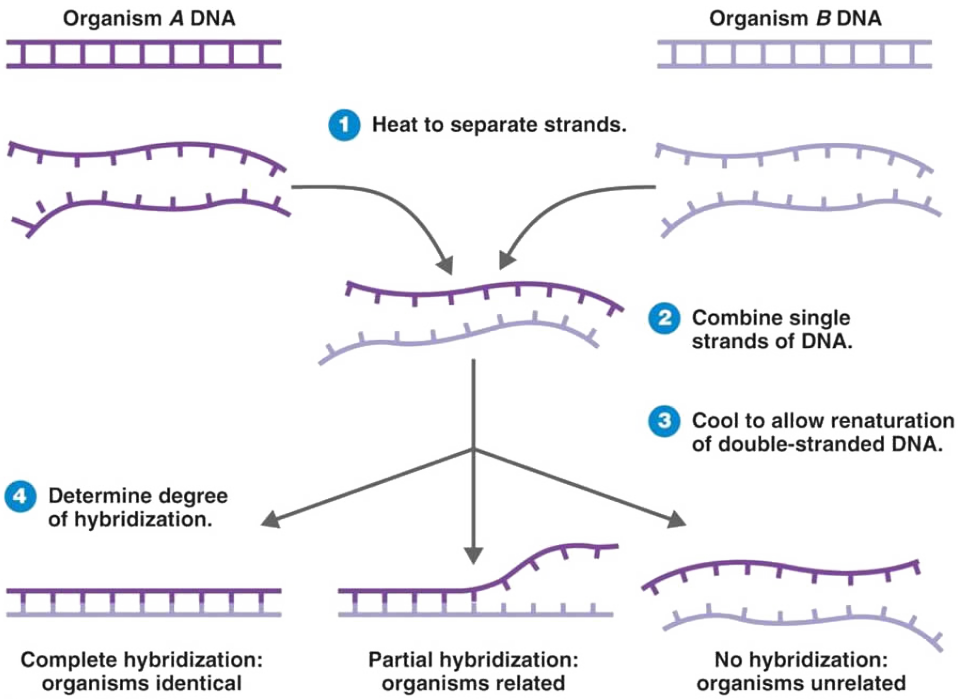
The James Hutton Institute

- Requires two genomes: “reference” and “comparator”
  - Experiment produces a comparative result, dependent on the choice of genomes
- Methods mostly based around direct or indirect DNA hybridisation
  - DNA-DNA hybridisation
  - Comparative Genomic Hybridisation (CGH)
  - Array Comparative Genomic Hybridisation (aCGH)

# DNA-DNA Hybridisation (DDH)

- Several methods based around the same principle

1. Denature organism A, B genomic DNA mixture
2. Allow to anneal – hybrids result (reassociation ≈ similarity)




Morelló-Mora & Amann (2001) *FEMS Microbiol. Rev.* doi:10.1016/S0168-6445(00)00040-1

# DNA-DNA Hybridisation (DDH)

- Several methods - same principle

1. Find homoduplex $Tm_1$
2. Denature reference, comparator gDNA + mix
3. Allow to anneal – hybrids result (reassociation ≈ similarity), find heteroduplex $Tm_2$
4. $\Delta Tm = Tm_1 - Tm_2$
5. High ΔT implies greater genomic difference (fewer H-bonds)

- Proxy for sequence similarity

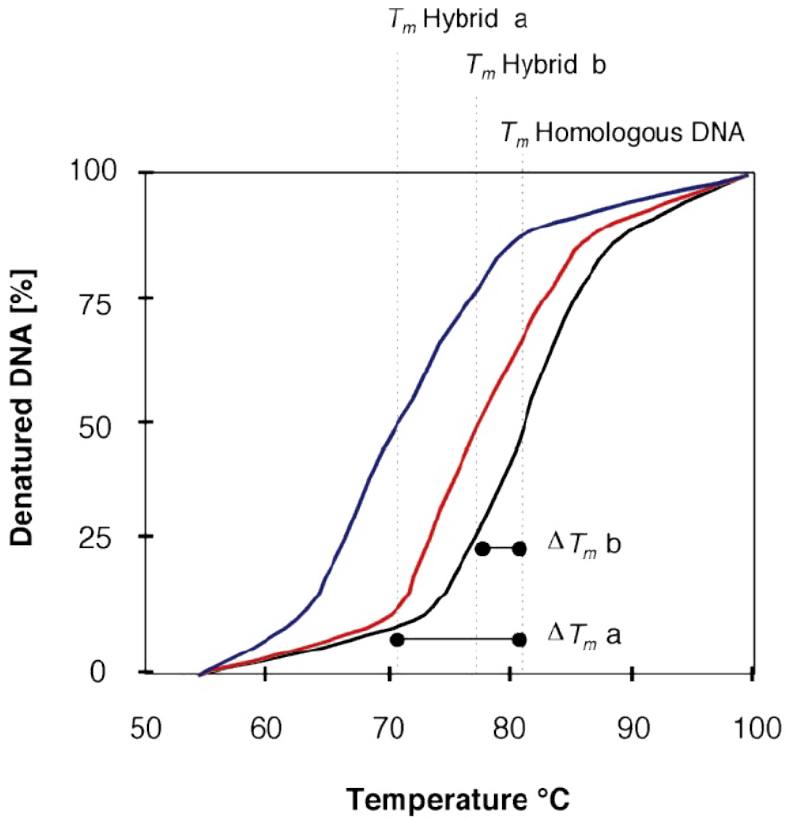



Fig. 2. Thermal denaturation curves of a homoduplex DNA and two heteroduplex DNAs.

Morelló-Mora & Amann (2001) *FEMS Microbiol. Rev.* doi:10.1016/S0168-6445(00)00040-1

# DNA-DNA Hybridisation (DDH)

- Used for taxonomic classification in prokaryotes from 1960s
- Sibley & Ahlquist redefined bird and primate phylogeny with DDH in 1980s: *Homo* shares more recent common ancestor with *Pan* than with *Gorilla* (this was previously in dispute)

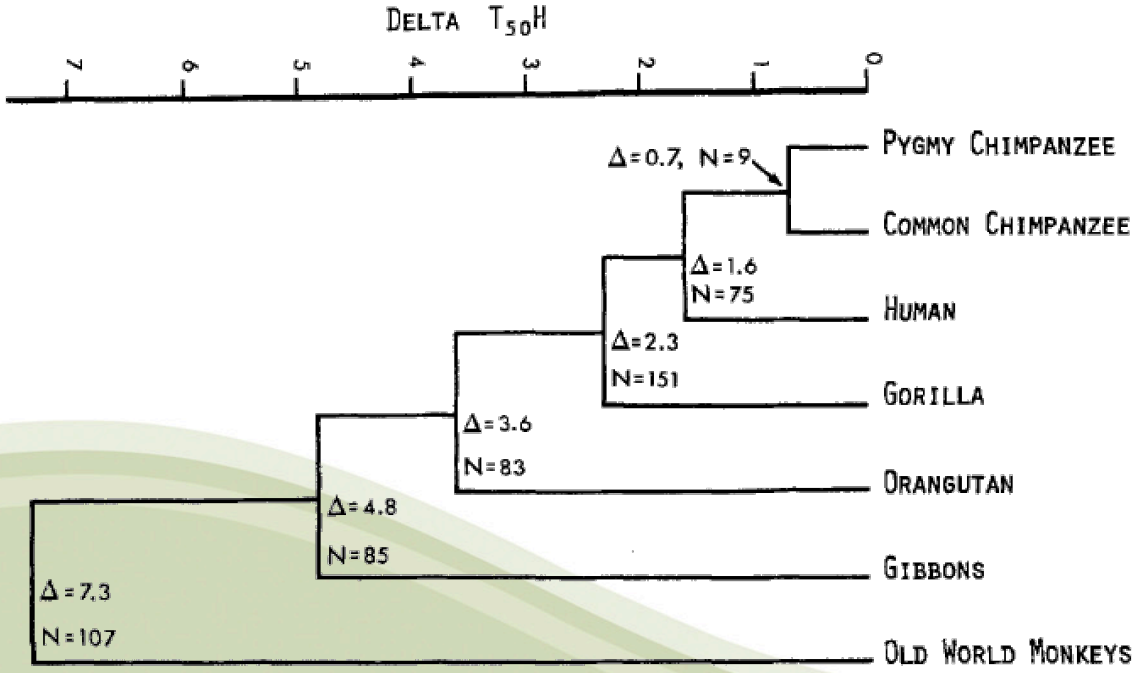


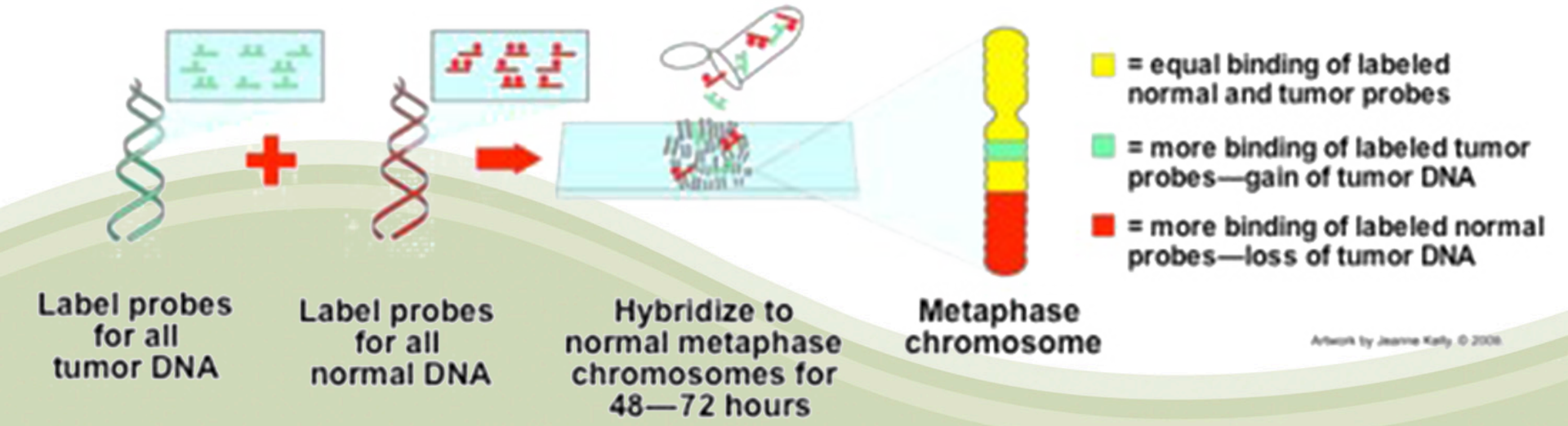

Fig. 3. Phylogeny of the hominoid primates as determined by average linkage clustering of delta $T_{50}H$ values derived from DNA–DNA hybridization

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

# Comparative Genomic Hybridisation

The James Hutton Institute

- Two genomes: "reference" and "test" are labelled (*red and green – a bad convention to choose, for visualisation*), then hybridised against a ***third*** "normal" genome
- Differences in red/green intensity mapped by microscopy correspond to relative relationship of reference and test to "normal" genome
- **Comparisons *within* species** (or individual, for tumours); **copy number variations (CNV)**
- Labour-intensive, low-resolution



# Comparative Genomic Hybridisation

- Image analysis required – intensity along medial axis.

## 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.

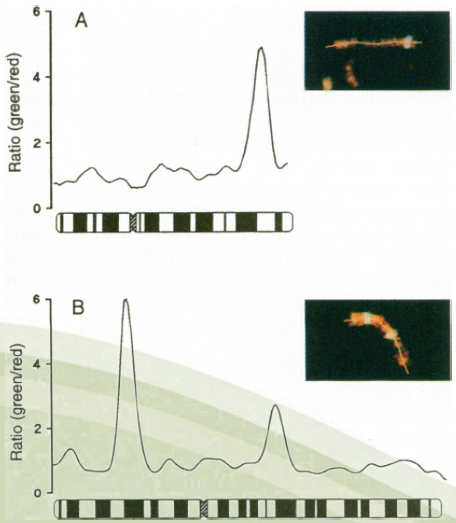
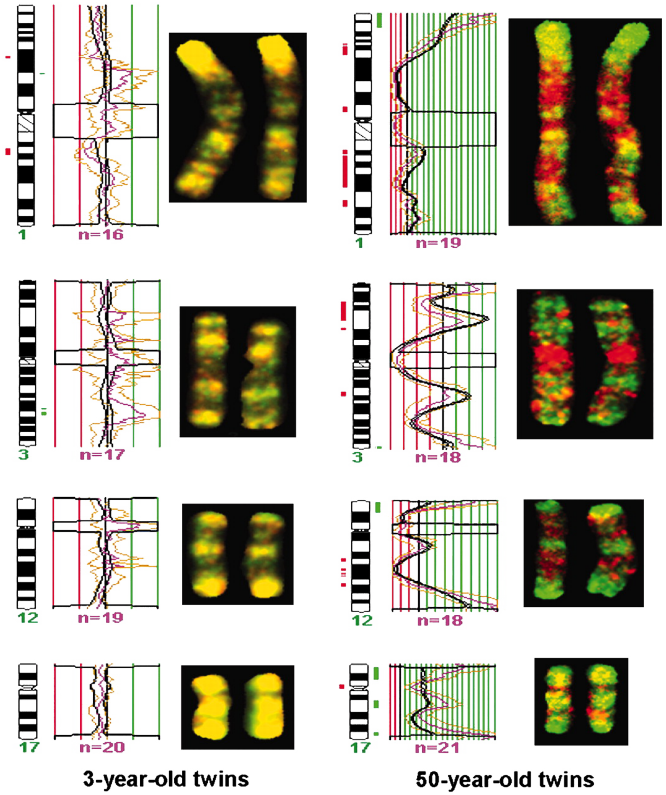



**Fig. 3.** Green-to-red fluorescence ratio profiles of chromosome 8 (**A**) and chromosome 2 (**B**) after hybridization with COLO 320HSR and NCI-H69 cell line DNAs, respectively (green). Normal reference DNA included in the hybridization is shown in red. The inserts show the overlaid green and red 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.




**Epigenetics:** hybridising methylated DNA

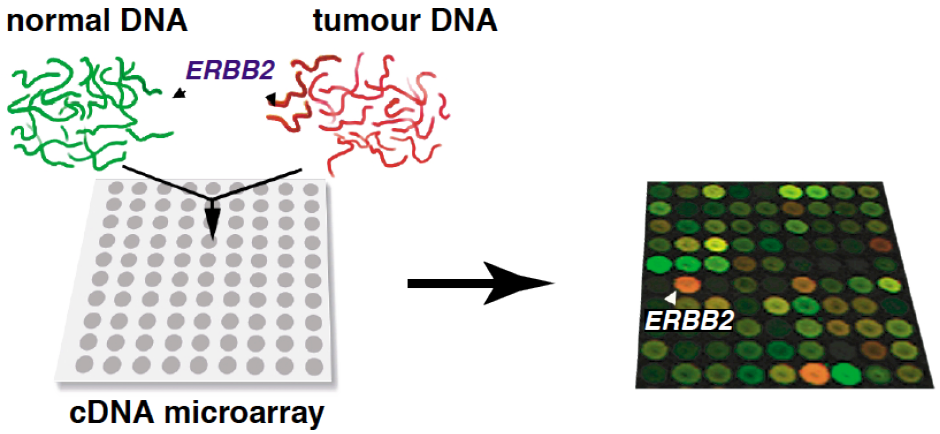
Kallioniemi *et al.* (1992) *Science* doi:10.1126/science.1359641

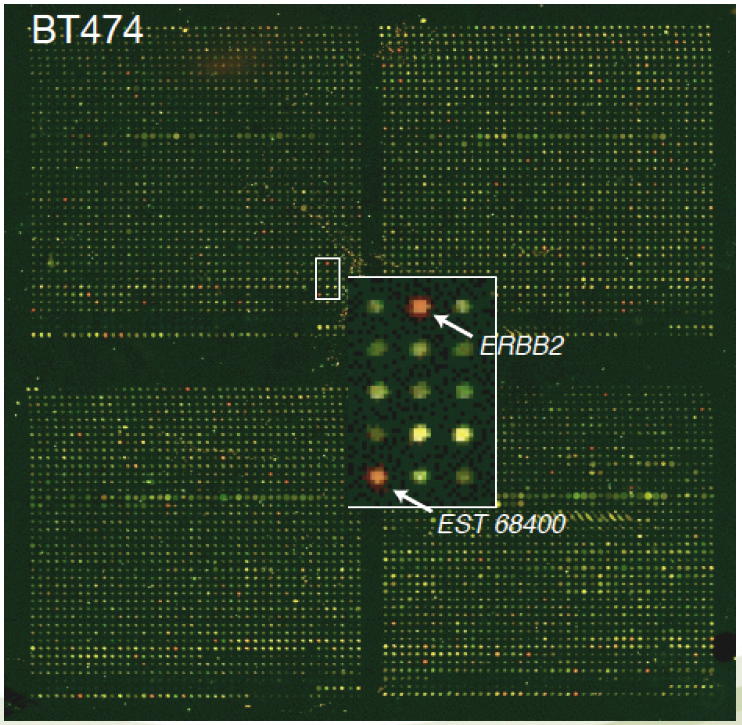
Fraga *et al.* (2005) *Proc. Natl. Acad. Sci. USA* doi:10.1073/pnas.0500398102

# Array Comparative Genomic Hybridisation

The James Hutton Institute

- Uses DNA microarrays: thousands of short DNA probes (genome fragments) immobilised on a surface
- gDNA, cDNA, etc. fluorescently-labelled and hybridised to the array




- Smaller sample sizes cf. CGH, automatable, high-throughput, high-res
- **Identifies copy number variation (CNV) and segmental duplication**

Pollack *et al.* (1999) *Nat. Genet.* doi:10.1038/12640

# Genome Feature Comparisons

The James Hutton Institute

## Comparisons on the basis of a restricted set of genome features

# 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

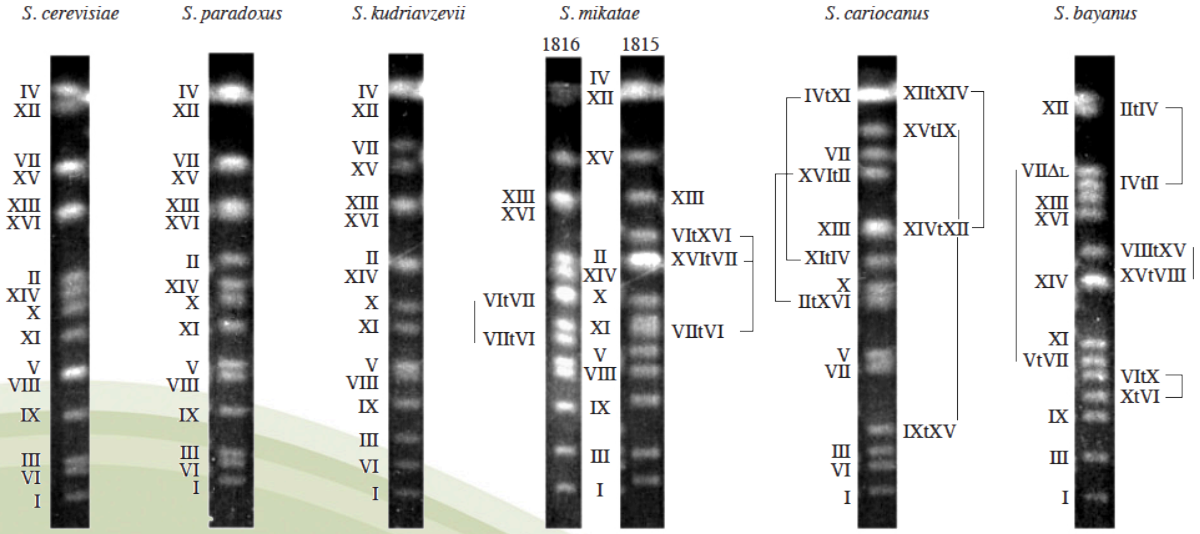
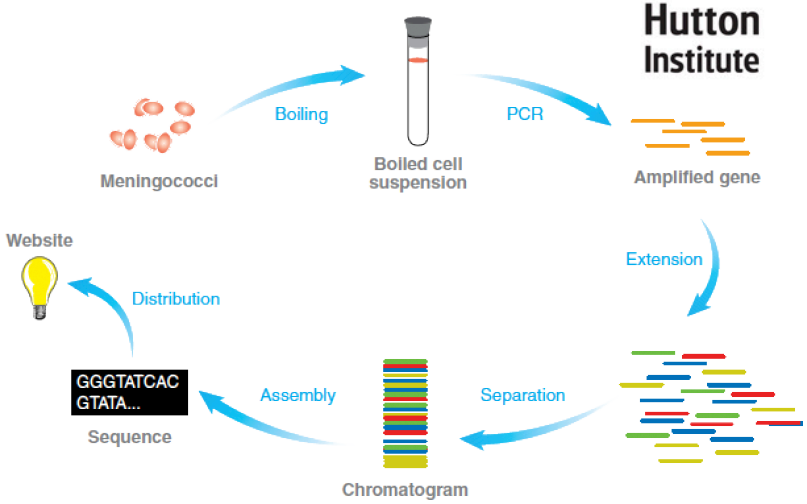



**Figure 1** Electrophoretic karyotypes of the *Saccharomyces* 'sensu stricto' species. Strains presented here are *S. cerevisiae* Y55, *S. paradoxus* N17, *S. kudriavzevii* IFO 1802, *S. mikatae* IFO 1816 and IFO 1815, *S. cariocanus* IMUFRJ 50816 and S. *bayanus* CBS 7001. Chromosomes are labelled 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

# Diagnostic PCR/MLST

- Define a set of regions (usually genes):
  - conserved enough that PCR primers can be designed to amplify the same region in multiple organisms
- and:
  - divergent enough that hybridising probes can distinguish between groups
- or:
  - sequence the amplification products
- Sequence variants given numbers
- Number profiles define groups
- Track evolution by minimum spanning trees (MST)
- http://pubmlst.org/




| | aspA | glnA | gltA | glyA | pgm | tkt | uncA |
|---|---|---|---|---|---|---|---|
| ST-45 | 4 | 7 | 10 | 4 | 1 | 7 | 1 |
| ST-2 | 4 | 7 | 51 | 4 | 1 | 7 | 1 |
| ST-25 | 4 | 7 | 10 | 1 | 1 | 7 | 1 |
| ST-94 | 4 | 7 | 10 | 1 | 1 | 1 | 1 |
| ST-68 | 4 | 7 | 10 | 4 | 32 | 7 | 6 |
| ST-88 | 4 | 7 | 10 | 4 | 36 | 28 | 1 |
| ST-243 | 4 | 7 | 10 | 3 | 42 | 7 | 1 |
| ST-109 | 2 | 7 | 10 | 4 | 1 | 7 | 5 |
| ST-116 | 26 | 7 | 10 | 5 | 1 | 7 | 21 |
| ST-203 | 17 | 7 | 10 | 30 | 1 | 7 | 4 |

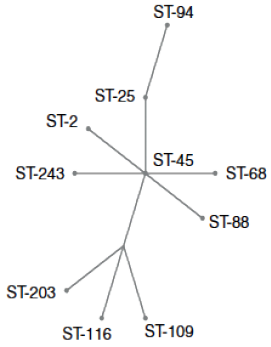



**Figure 2**

Clonal complex structures as revealed by MLST data. Clonal complexes are currently defined informally, by defining a central genotype and related genotypes, i.e., those that share up to four identical MLST loci. Identifying the central genotype depends on their frequency in samples of the population in question, their longevity, and a central position when analyzed by a variety of heuristic techniques. Once a central genotype is assigned, the MLSTdBNet software can identify all members of a clonal complex automatically. The figure shows the allelic profiles of members of the *C. jejuni* ST-45 complex visualized by split decomposition.


**Maiden *et al.* (2006) *Ann. Rev. Microbiol.* doi:10.1146/annurev.micro.59.030804.121325**

# Array Comparative Genomic Hybridisation

The James Hutton Institute

- aCGH can also be **applied *across* species** for **classification/diagnostics**:
  - Microarray probes represent genes from one or more organisms
  - "Off-species" gDNA fragmented, labelled, and hybridised
  - Hybridisation ≈ sequence similarity ≈ gene presence
- Heatmap of 217 *Staphylococcus aureus* isolates on **7-strain array**.
  - columns=isolates
  - yellow/red=gene present
  - blue/white/grey=gene absent
- Lower bars coloured by lineage and host (green=cattle, blue=horse, purple=human)


Sung *et al.* (2008) *Microbiol.* doi:10.1099/mic.0.2007/015289-0

# But This Happened…

- High-throughput sequencing

# …And Then It Rained Sequence Data

The James Hutton Institute

- Modern high-throughput sequencing (454, Illumina) completely changed the landscape.
- Complete, (mainly) accurate sequence data much cheaper, enabling:
  - more precise sequence comparison
  - novel analyses, insights and visualisations
  - Genomic & exomic comparisons
- 19/2/2014 at GOLD:
  - 3,011 “finished” genomes
  - 9,891 “permanent draft” genomes
- 19/2/2014 at NCBI WGS:
  - 17,023 whole genome projects

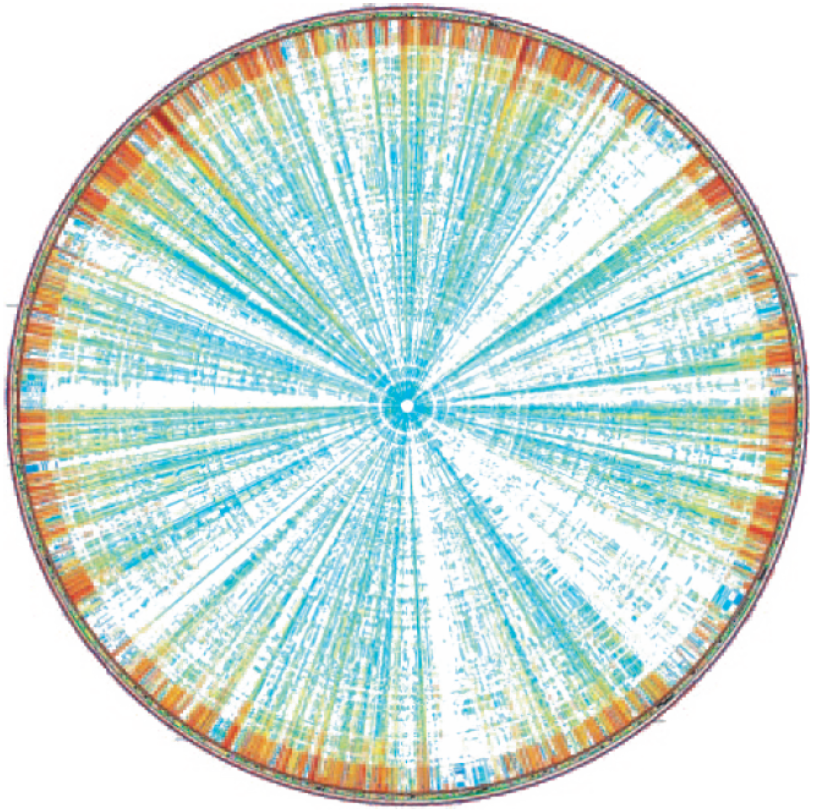
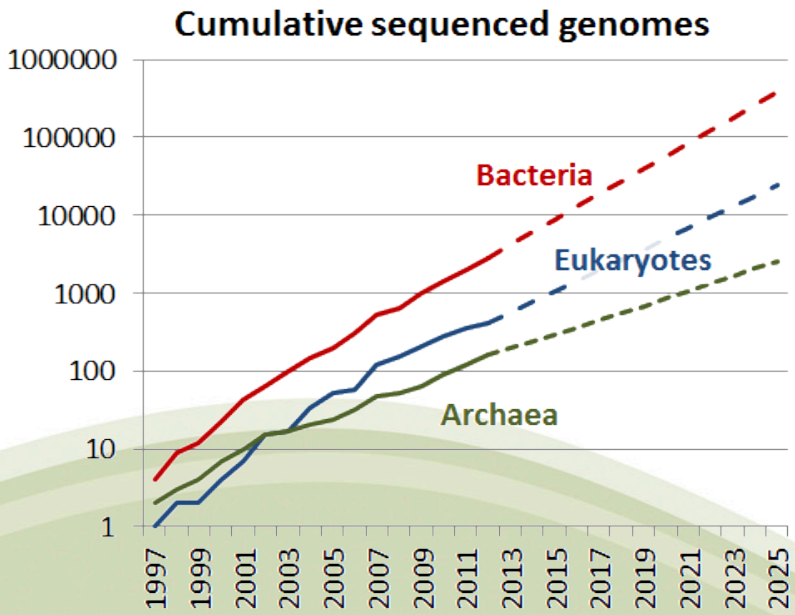
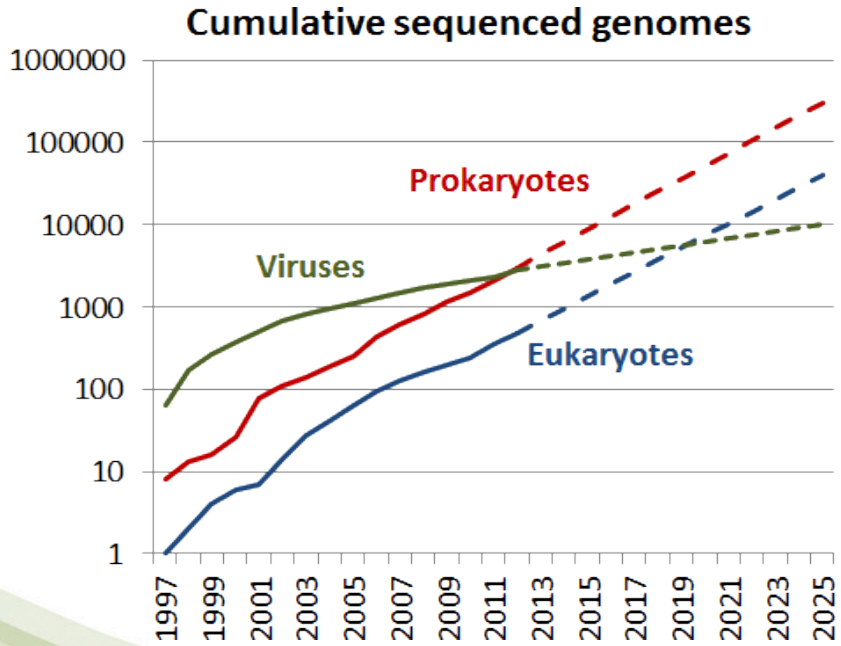


**Fig. 1.** Circular diagram rendered by GenomeDiagram of the reciprocal best hit comparison of all coding sequences from *Eca* against 229 bacterial genomes. Individual hits are coloured in a graduated scheme from dark red to light blue in order of decreasing sequence similarity. Successive inner rings represent distinct bacterial genomes in order of decreasing average similarity of coding sequence. The image summarizes over 1 000 000 data points.

# …And Then It Rained Sequence Data

- In 2012, GOLD added 3736 genomes, NCBI added 4585
- Mostly prokaryotes (archaea and bacteria)
- We're a little ahead of Su's (Scripps, La Jolla) projections




**Figures and code from: http://sulab.org/2013/06/sequenced-genomes-per-year/**

# Computational Comparative Genomics

The James Hutton Institute

**Massively enabled by high-throughput sequencing, much more powerful and precise.**

# Three broad 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 (rearrangements), 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.

# Bulk Genome Property Comparisons

The James Hutton Institute

Values calculated for individual genomes, and subsequently compared.

# Nucleotide Frequencies/Genome Size

The James Hutton Institute

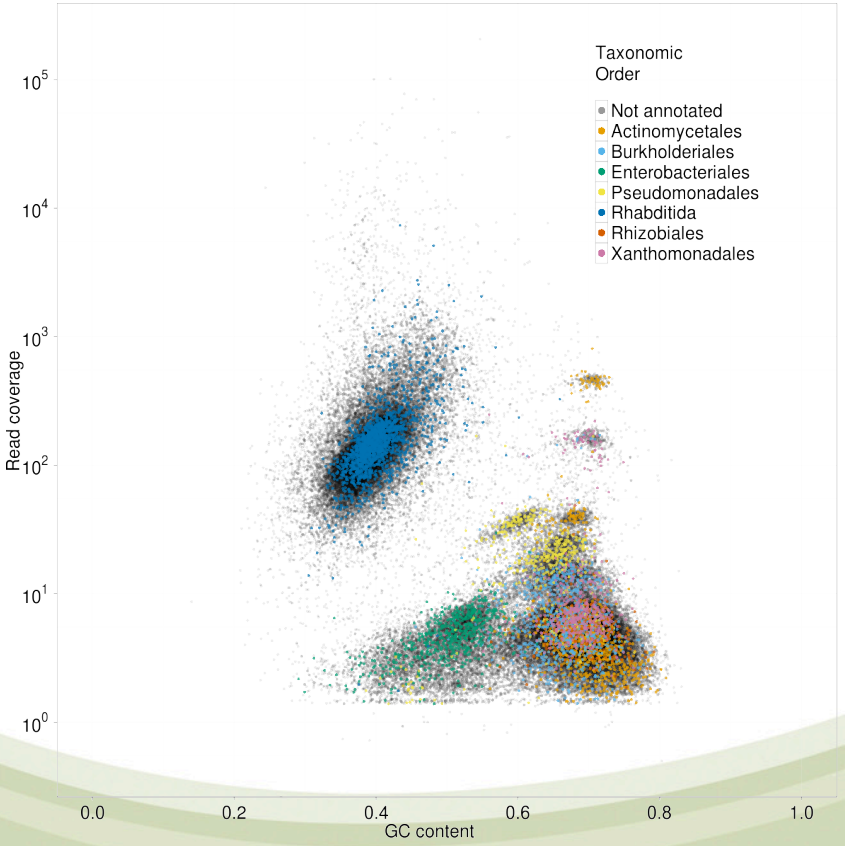
- Very easy to calculate from complete or draft genome sequence
  - (or in a region of genome sequence)

```
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
- **[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

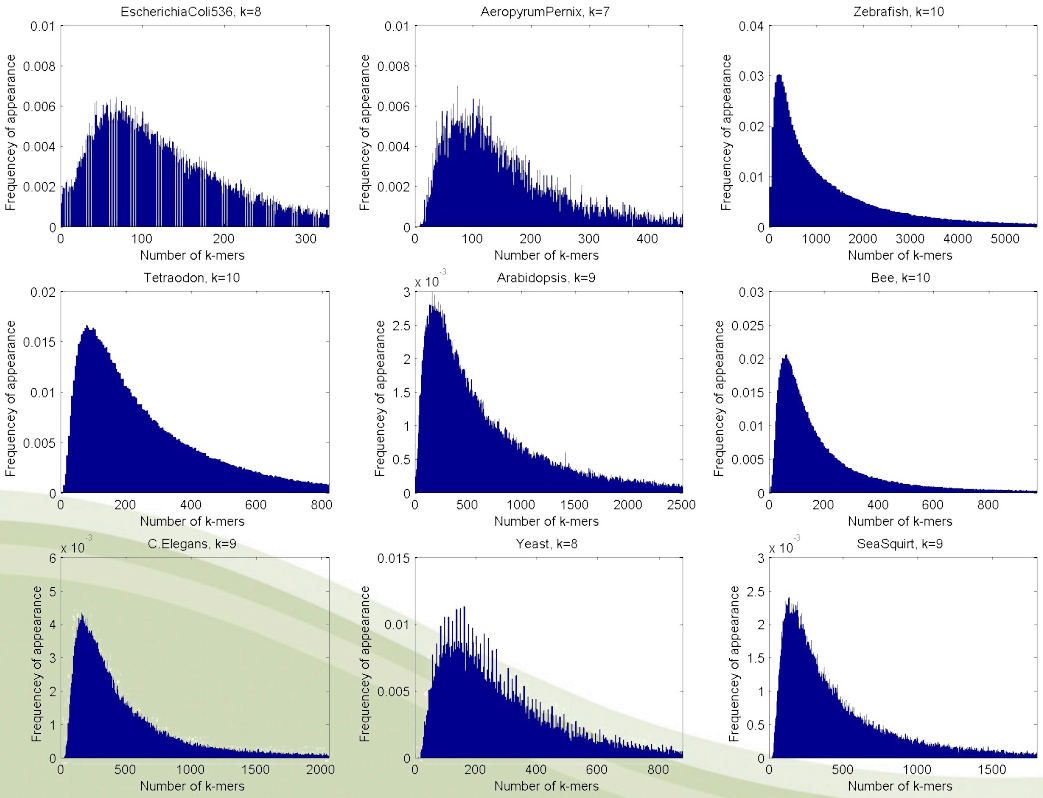
The James Hutton Institute

- **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

The James Hutton Institute

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

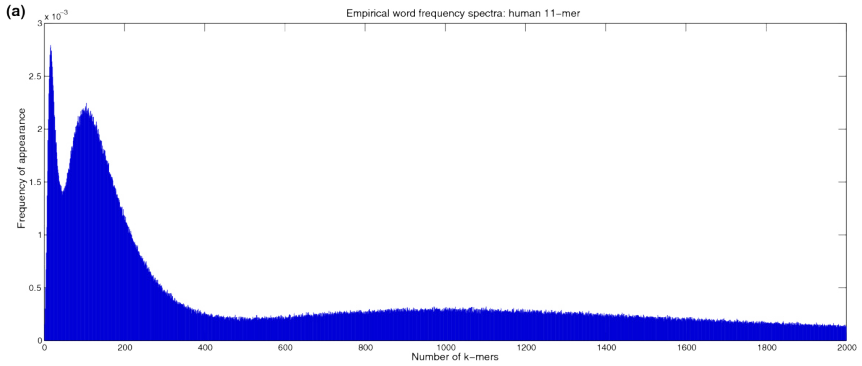


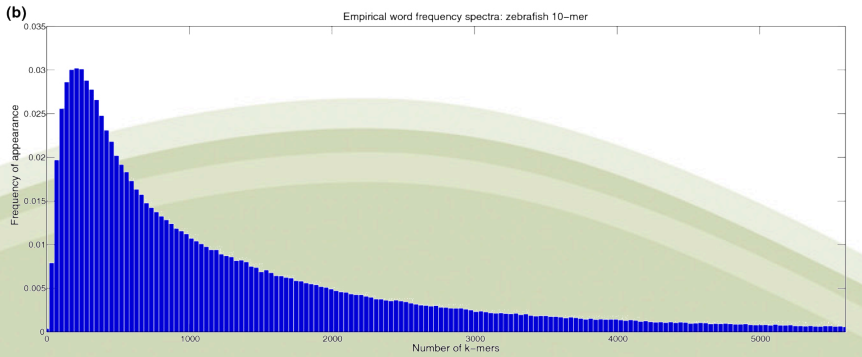

Chor *et al.* (2009) *Genome Biol.* doi:10.1186/gb-2009-10-10-r108

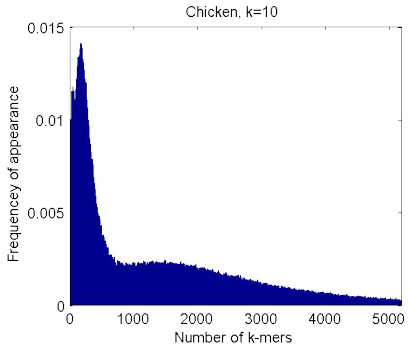
# *k*-mer Spectra

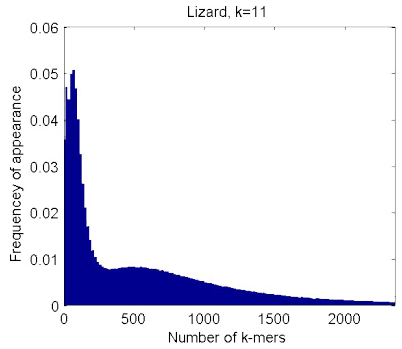
The James Hutton Institute

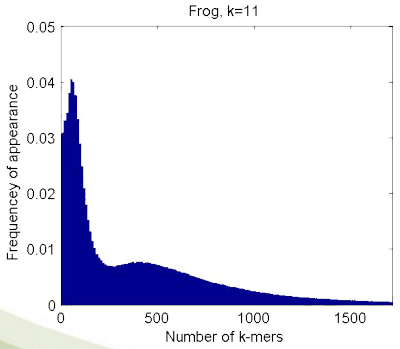
- ***k*-mer spectrum:**
  - All mammals tested (and some other) species have a multimodal *k*-mer spectrum
  - Genomic regions differ in this property

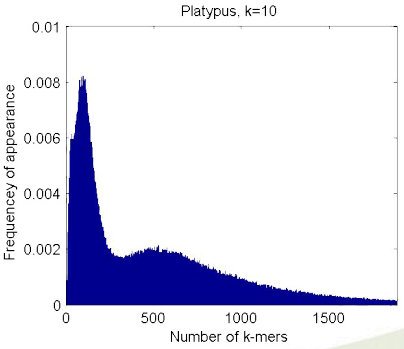


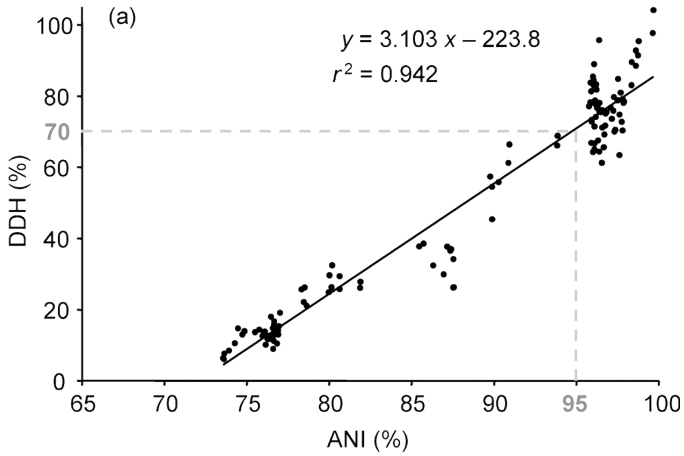



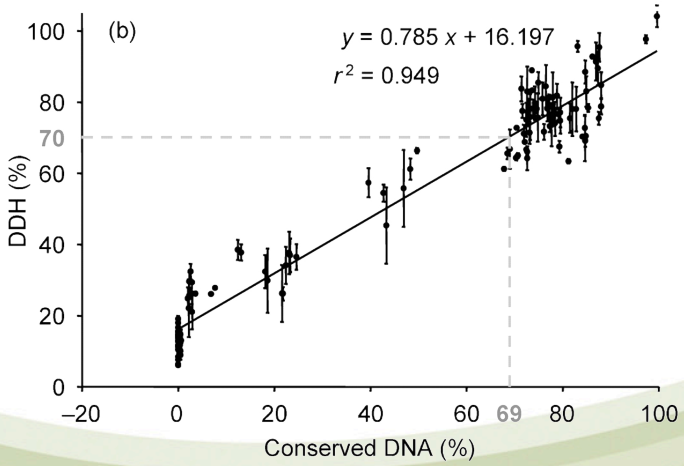




Chor *et al.* (2009) *Genome Biol.* doi:10.1186/gb-2009-10-10-r108

# 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)

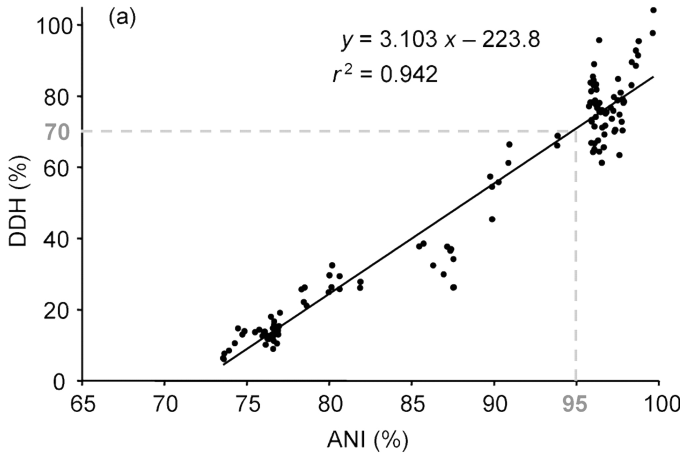


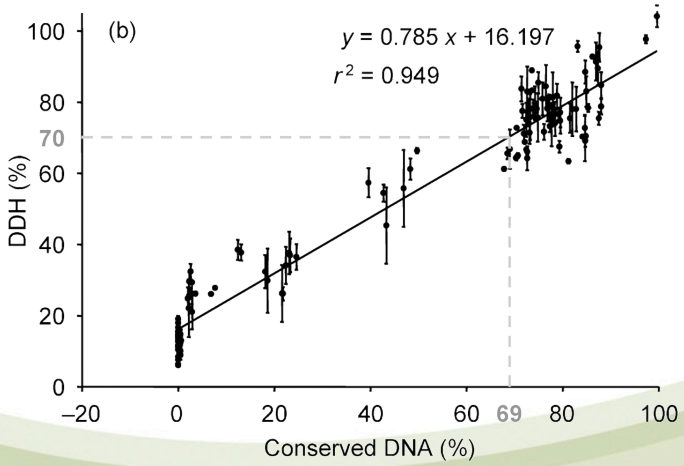




Goris *et al.* (2007) *Int. J. System. Evol. Biol.* doi:10.1099/ijs.0.64483-0

# 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)
- 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



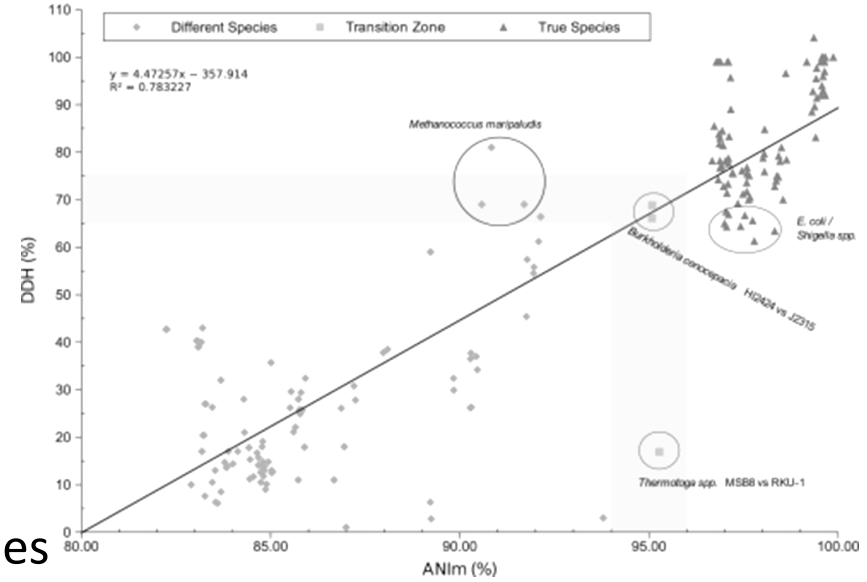




Goris *et al.* (2007) *Int. J. System. Evol. Biol.* doi:10.1099/ijs.0.64483-0

# Average Nucleotide Identity (ANI)

The James Hutton Institute

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

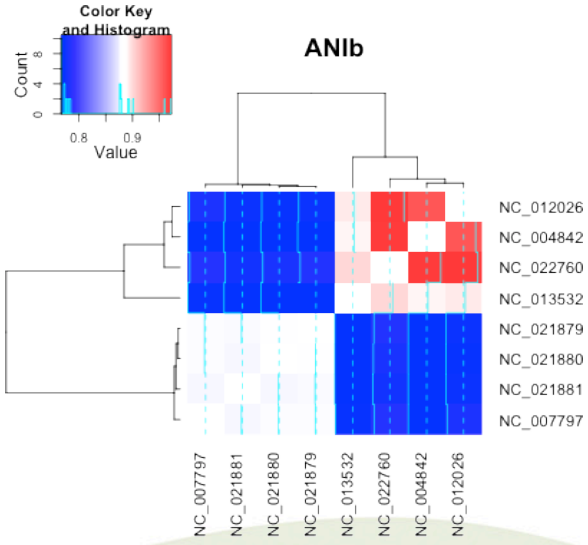


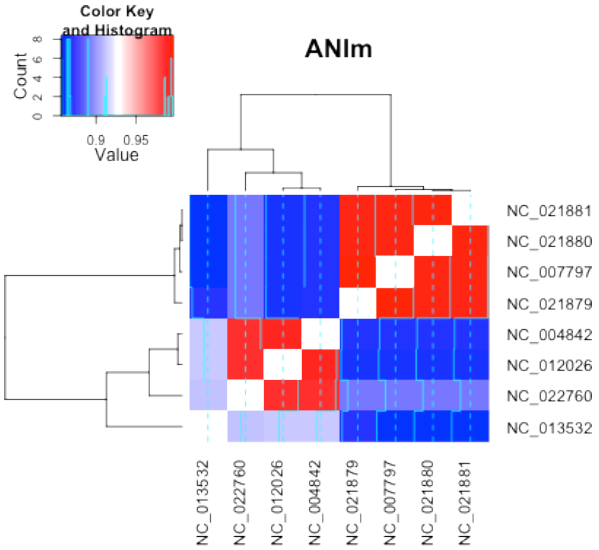

- ANIm and TETRA introduced (2009)
  1. Align sequences using NUCmer
  2. ANI = mean %identity of matches
- TETRA:
  1. Calculate tetranucleotide frequencies
  2. Determine each tetramer deviation from expectation (Z-score)
  3. TETRA = Pearson correlation coefficient of tetramer Z-scores

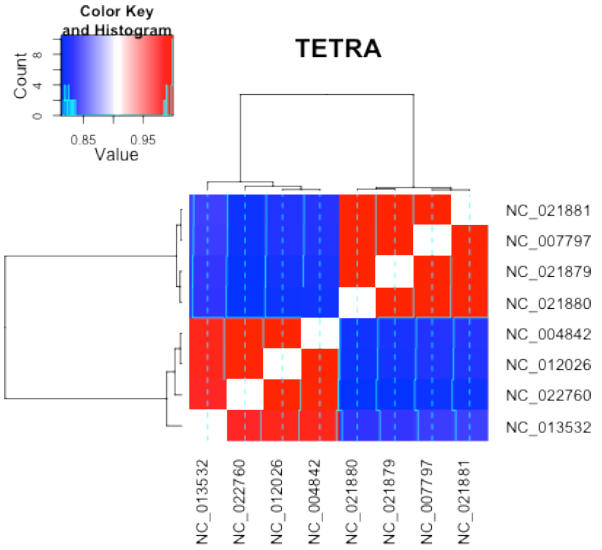

Richter & Rosselló-Móra (2009) *Proc. Natl. Acad. Sci. USA* doi:10.1073/pnas.0906412106

# 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)

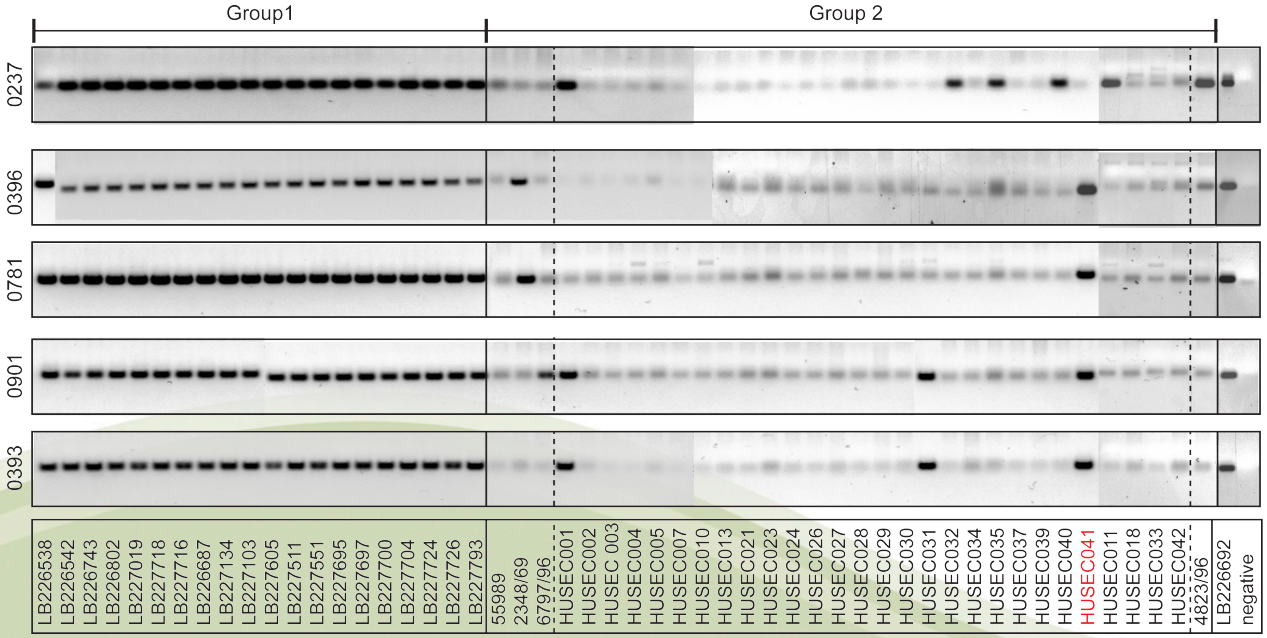
The James Hutton Institute

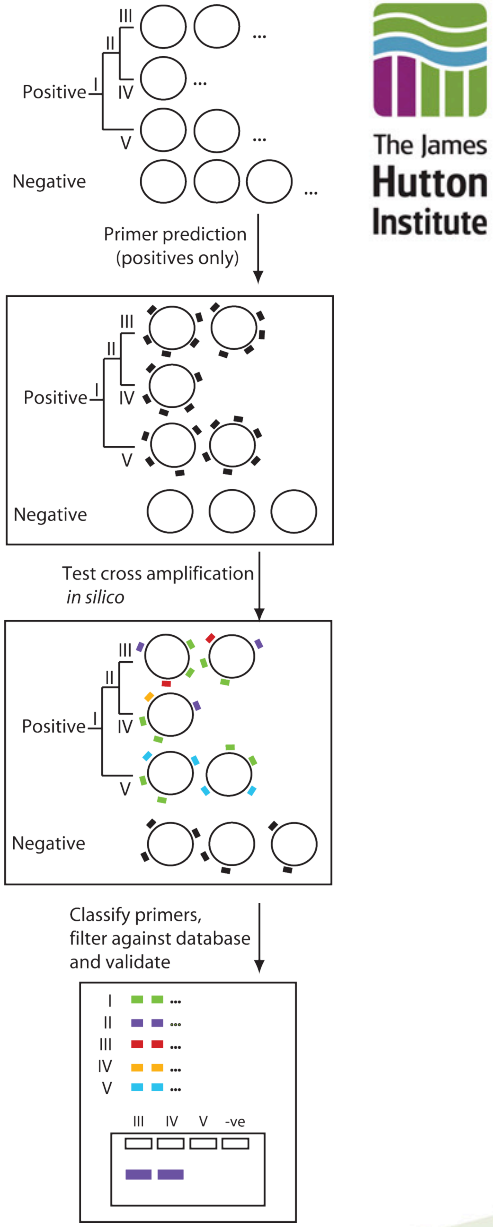
- Jspecies (http://www.imedea.uib.es/jspecies/)
  - WebStart
  - `java -jar -Xms1024m -Xmx1024m jspecies1.2.1.jar`
- Python script
  - `scripts/calculate_ani.py`
- **[ACTIVITY]**
  - `average_nucleotide_identity/README.md` Markdown

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

# Diagnostic PCR/MLST

- PCR/MLST still cheap
  - (but for how much longer?)
- Use whole genomes to identify unique/ diagnostic regions for PCR/MLST




The James Hutton Institute


Slezak *et al.* (2003) *Brief. Bioinf.* doi:10.1093/bib/4.2.133

Pritchard *et al.* (2012) *PLoS One* doi:10.1371/journal.pone.0034498

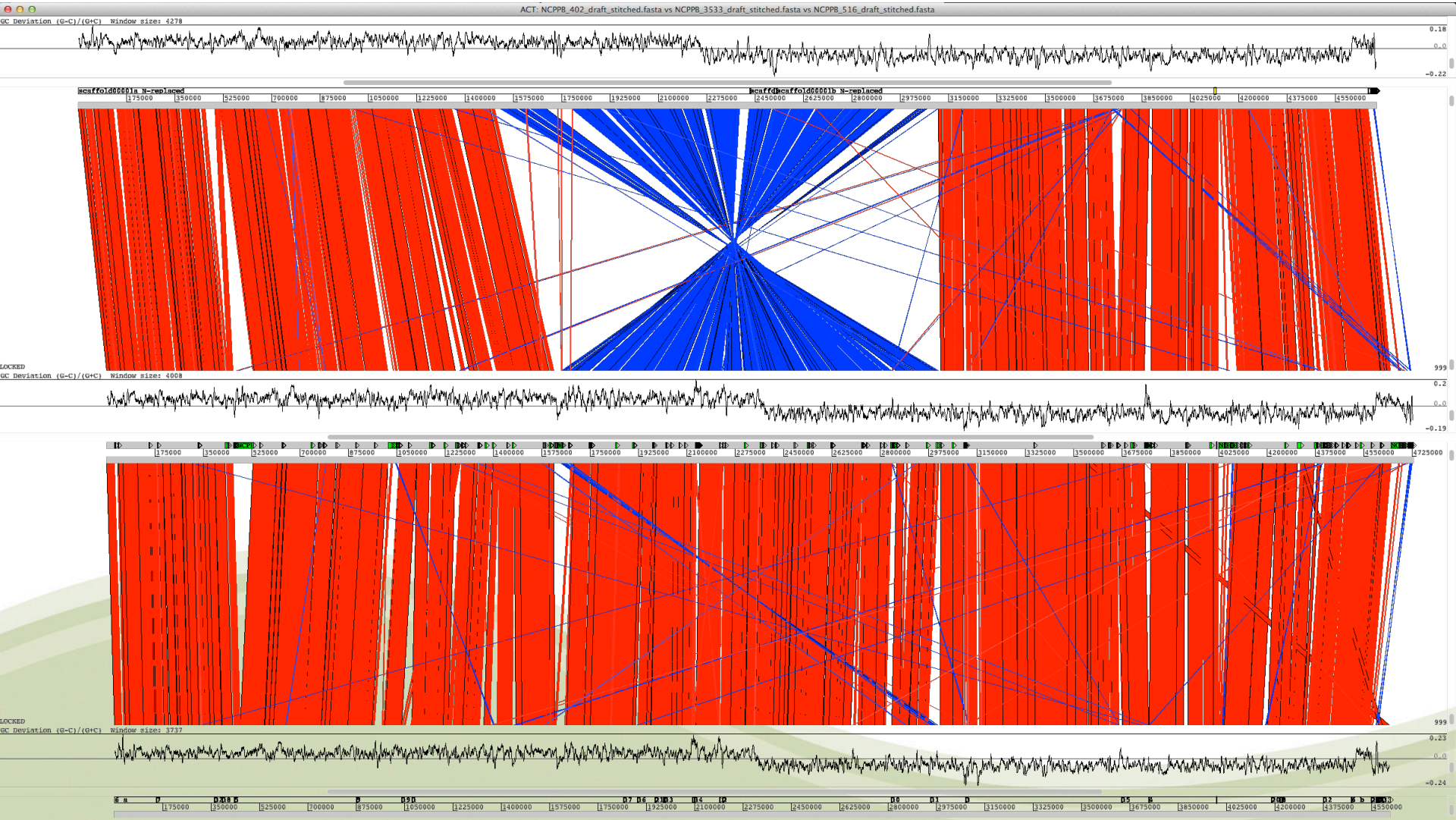
# Whole Genome Sequence Comparisons

The James Hutton Institute

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

# Whole Genome Alignment

The James Hutton Institute



# Whole Genome Alignment

The James Hutton Institute

- 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

The James Hutton Institute

- 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…

# Whole Genome Alignment

The James Hutton Institute

- **BLAT**
  - BLAT is broadly similar to BLAST
  - Main differences:
    - optimised to find only exact or near-exact matches, for speed
    - indexes the subject genome, retains the index and scans the query
    - connects homologous match regions into a single alignment (BLAST reports them separately)
    - reports mRNA match intron-exon boundaries exactly (BLAST tends to extend)
  - **Advantages**: fast; exact exon boundaries; UCSC integration
  - **Disadvantages**: does not find more remote/very divergent matches

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

# Whole Genome Alignment

The James Hutton Institute

- **megaBLAST**
  - Optimised for speed over BLASTN (see http://www.ncbi.nlm.nih.gov/blast/Why.shtml):
    - genome-level searches
    - queries on large sequence sets
    - long alignments of very similar sequence (sequencing errors/SNPs)
  - Uses Zhang *et al.* (2000) greedy algorithm
  - Concatenates queries to improve performance ("query packing")
    - **NOTE: this is good practice for large query sets!**
  - Two modes: megaBLAST, and discontinuous megaBLAST (dc-megablast)
    - **dc-megablast** intended for more divergent sequences

**Zhang *et al.* (2000) *J. Comp. Biol.* 7(1-2) 203-14**
**Korf *et al.* (2003) "BLAST", O'Reilly & Associates, Sebastopol, CA**

# Whole Genome Alignment

The James Hutton Institute

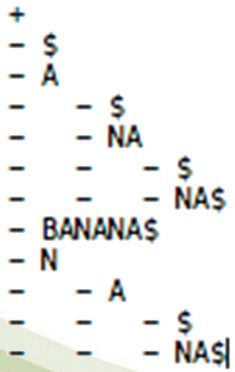
- **MUMmer**
  - Uses suffix trees for pattern matching: very fast even for large sequences
    - Finds *maximal exact matches*
    - Memory use depends only on reference sequence size

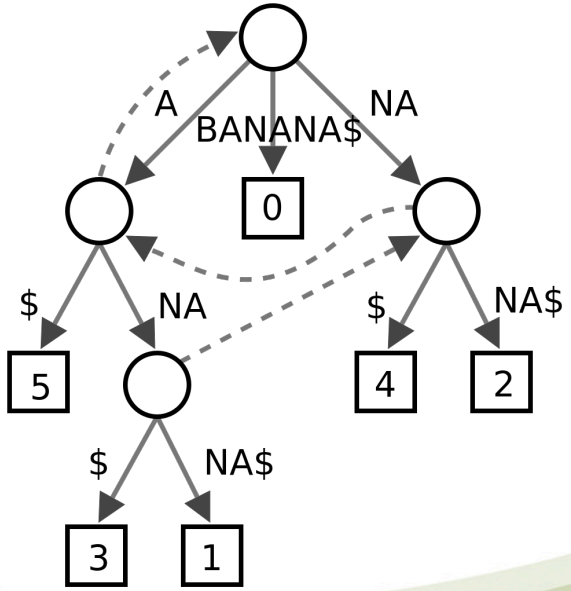
Kurtz *et al.* (2004) *Genome Biol.* doi:10.1186/gb-2004-5-2-r12

# Whole Genome Alignment

The James Hutton Institute

- **MUMmer**
  - Uses suffix trees for pattern matching: very fast even for large sequences
    - Finds *maximal exact matches*
    - Memory use depends only on reference sequence size
- Suffix Tree:
  - Can be constructed and searched in *O(n)* time
  - Useful algorithms are nontrivial
  - BANANA$
    - B followed by ANANA$ only
    - A followed by $, NA$, NANA$
    - N followed by A$, ANA$




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

# Whole Genome Alignment

The James Hutton Institute

- **MUMmer**
  - Process:
    - 1) Identify a non-overlapping subset of maximal exact matches: often *Maximum Unique Matches* (MUMs - though not always unique)
    - 2) Cluster into *alignment anchors*
    - 3) Extend between anchors to produce a final gapped alignment
  - Very flexible approach: a suite of programs (`mummer`, `nucmer`, `promer`, …)
    - nucleotide and "conceptual protein" (more sensitive) alignments
    - used for genome comparisons, assembly scaffolding, repeat detection, etc.
    - forms the basis for other aligners/assemblers, e.g. Mugsy, AMOS

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

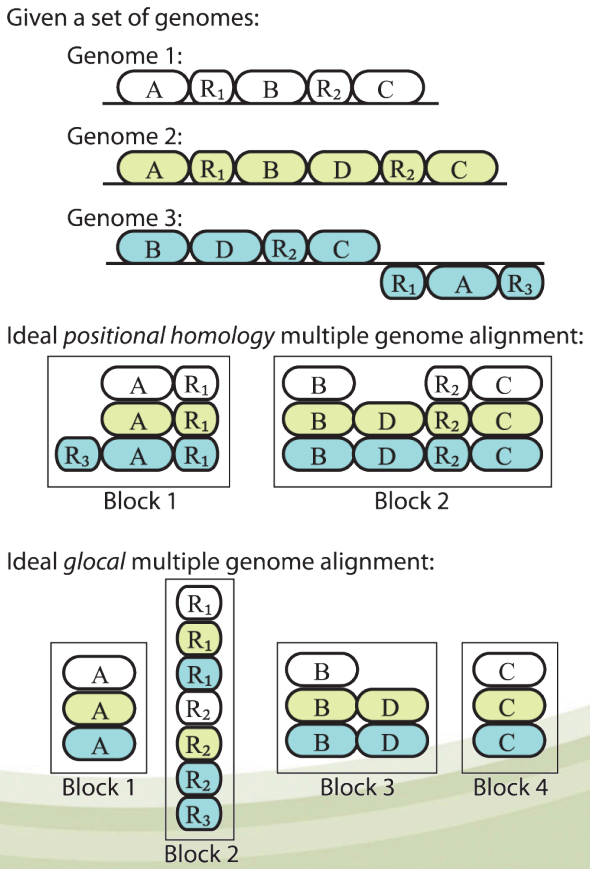
# Whole Genome Alignment

The James Hutton Institute

- [ACTIVITY]
  - `whole_genome_alignments_A.md` **Markdown**
  - https://github.com/widdowquinn/Teaching/blob/master/Comparative_Genomics_and_Visualisation/Part_1/whole_genome_alignment/whole_genome_alignments_A.md

# 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*



# Multiple Genome Alignment

The James Hutton Institute

- **LAGAN:** rapid alignment of two homologous genome sequences
  - Generate local alignments (*anchors*, B)
  - Construct rough global map (maximal-scoring ordered subset, C)
    - Join anchors that lie within a threshold distance, the same way
  - Compute global alignment by dynamic programming (D)

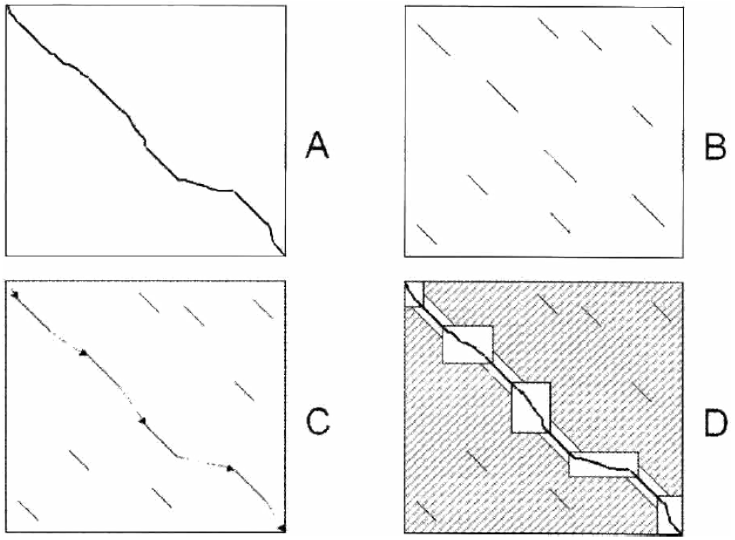



**Figure 1** The LAGAN algorithm. (*A*) A global alignment between two sequences is a path between the top-left and the bottom-right corner of their alignment matrix. (*B*) LAGAN first finds all local alignments between the two sequences. (*C*) LAGAN computes a maximal-scoring ordered subset of the alignments, 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.

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

# Multiple Genome Alignment

The James Hutton Institute

- **MLAGAN:** multiple genome alignment of *k* genomes in *k-1* alignment steps, using a phylogenetic tree (CLUSTAL-like):
  - Make rough global maps between each pair of sequences (step C in LAGAN)
  - Progressive multiple alignment with anchors (iterated)
    1. Perform global alignment between closest pair of sequences with LAGAN: alignments are "*multi-sequences*"
    2. Find rough global maps of this *multi-sequence* to all other *multi-sequences*.

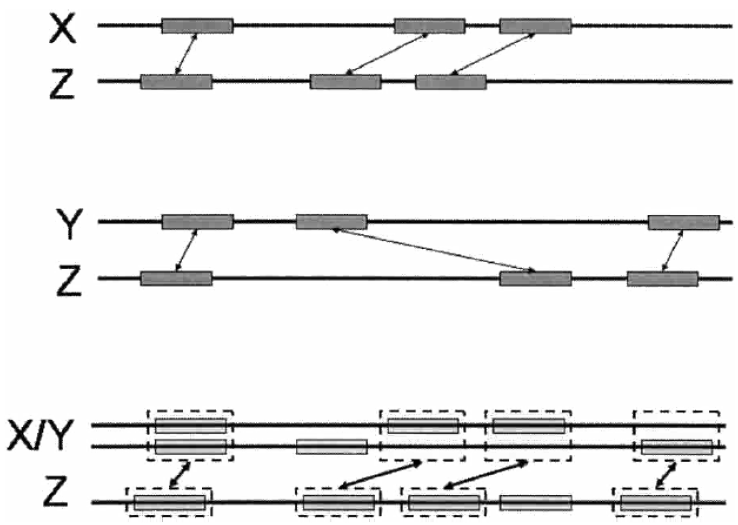


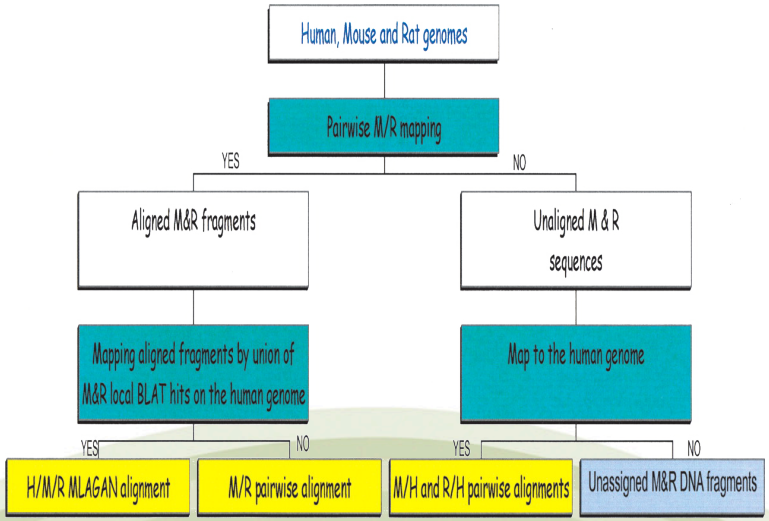

**Figure 5** Generation of anchors during progressive alignment. Multi-sequence X/Y is aligned to sequence Z. Anchors between X and Z (*top*) and anchors between Y and Z (*middle*) are remapped to co-ordinates in the X/Y multi-sequence, and given a new score. Then, the Longest Increasing Subsequence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

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

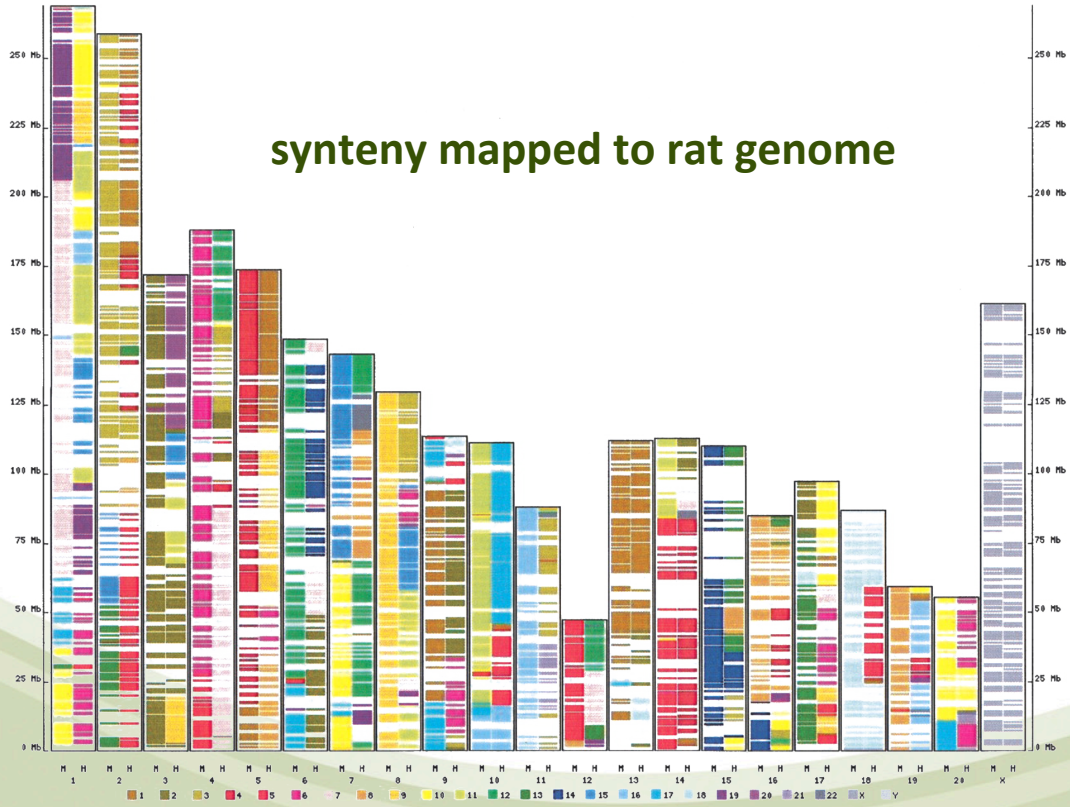
# Human-Mouse-Rat Alignment

The James Hutton Institute

- 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



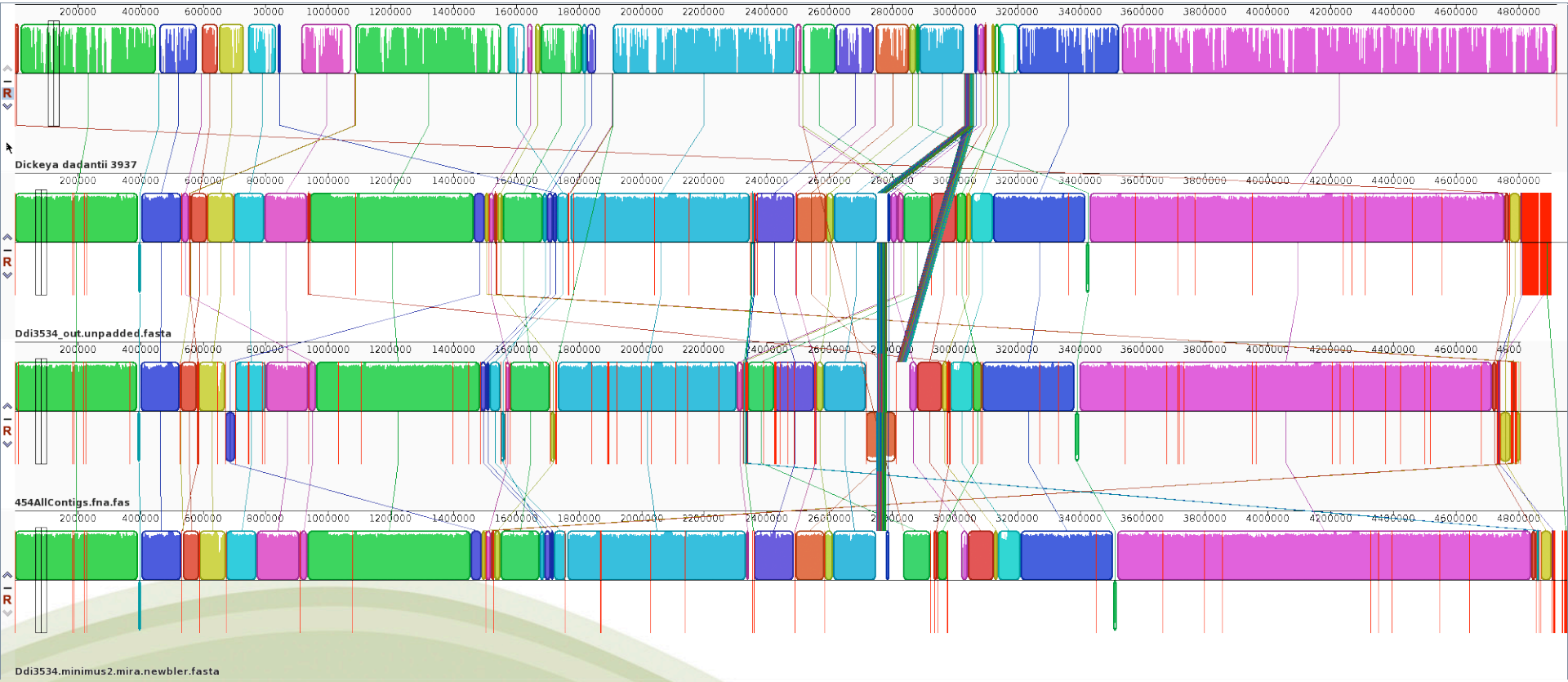

**Initial alignments by BLAT**
**Syntenous regions aligned with LAGAN**




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

# Draft Genome Alignment

The James Hutton Institute



# Draft Genome Alignment

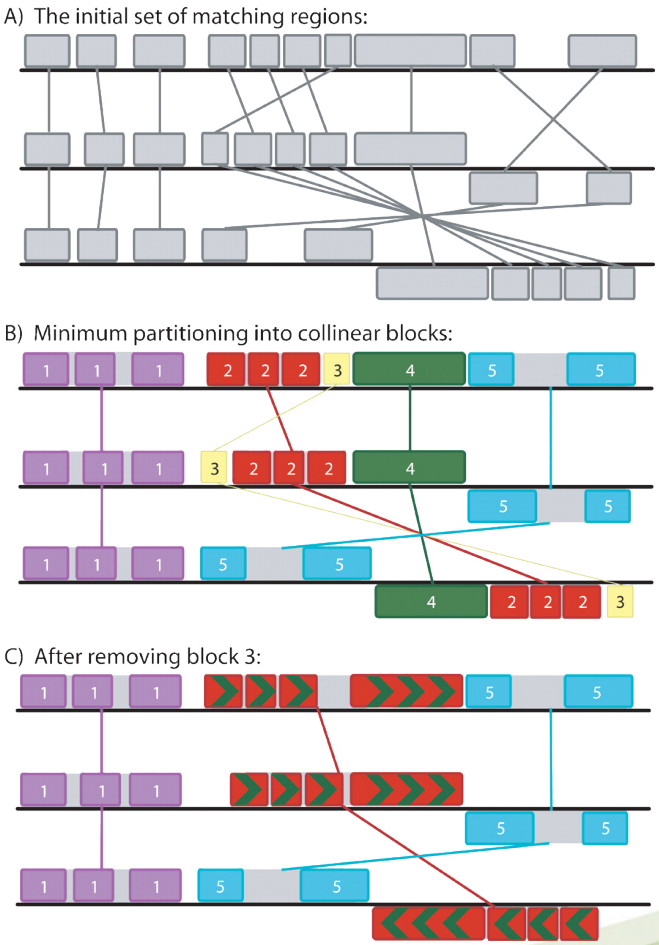
The James Hutton Institute

- 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

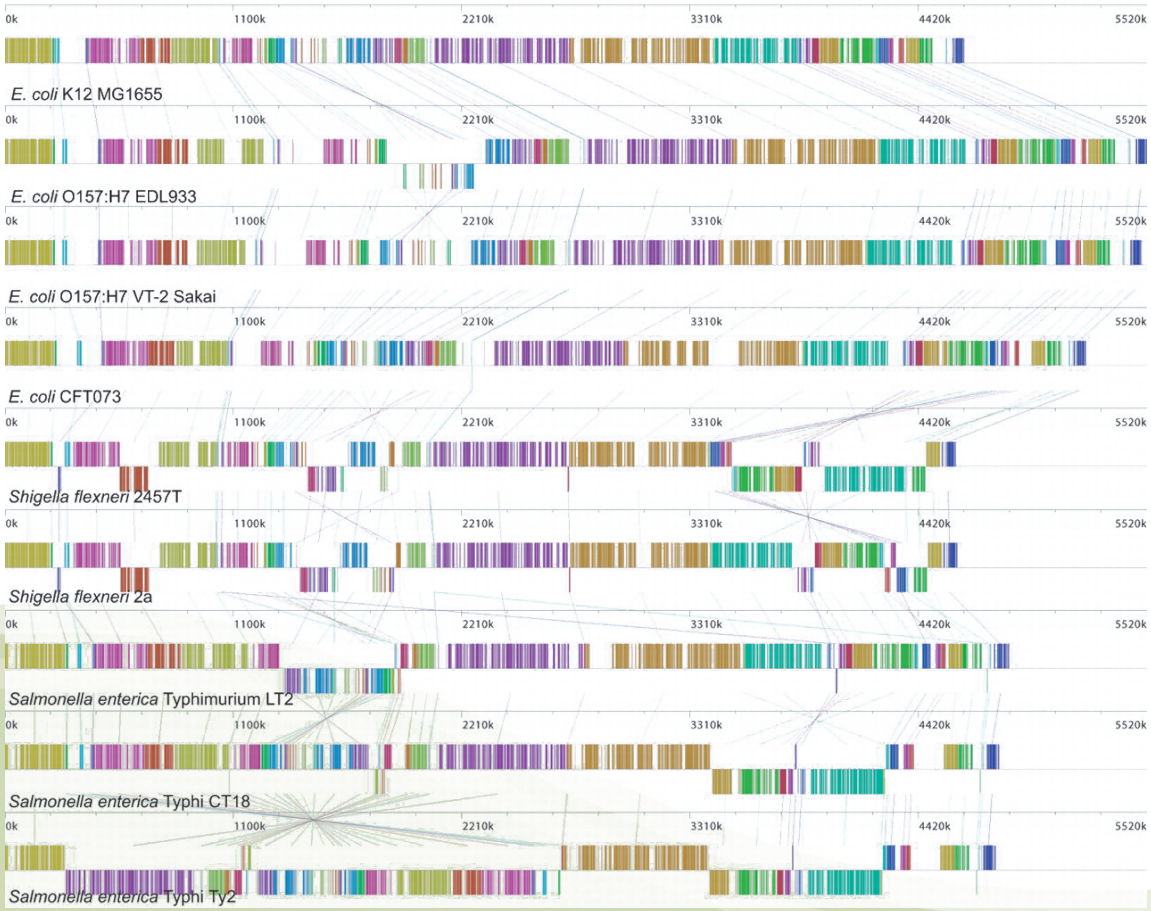



- Mauve Contig Mover (MCM) for ordering contigs

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

# Mauve

- **Mauve** alignment of LCBs in nine enterobacterial genomes
  - Rearrangement of homologous backbone sequence




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

# Draft Genome Alignment

The James Hutton Institute

- **[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

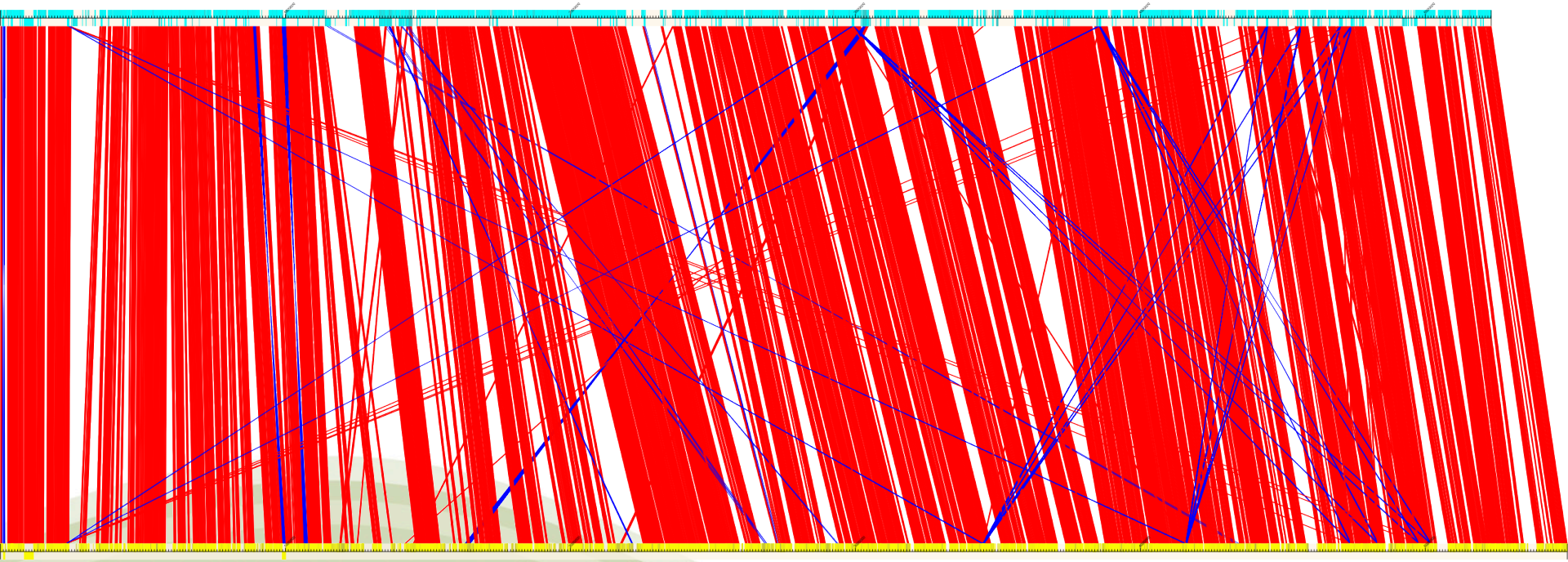
# Collinearity and Synteny

The James Hutton Institute

- 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

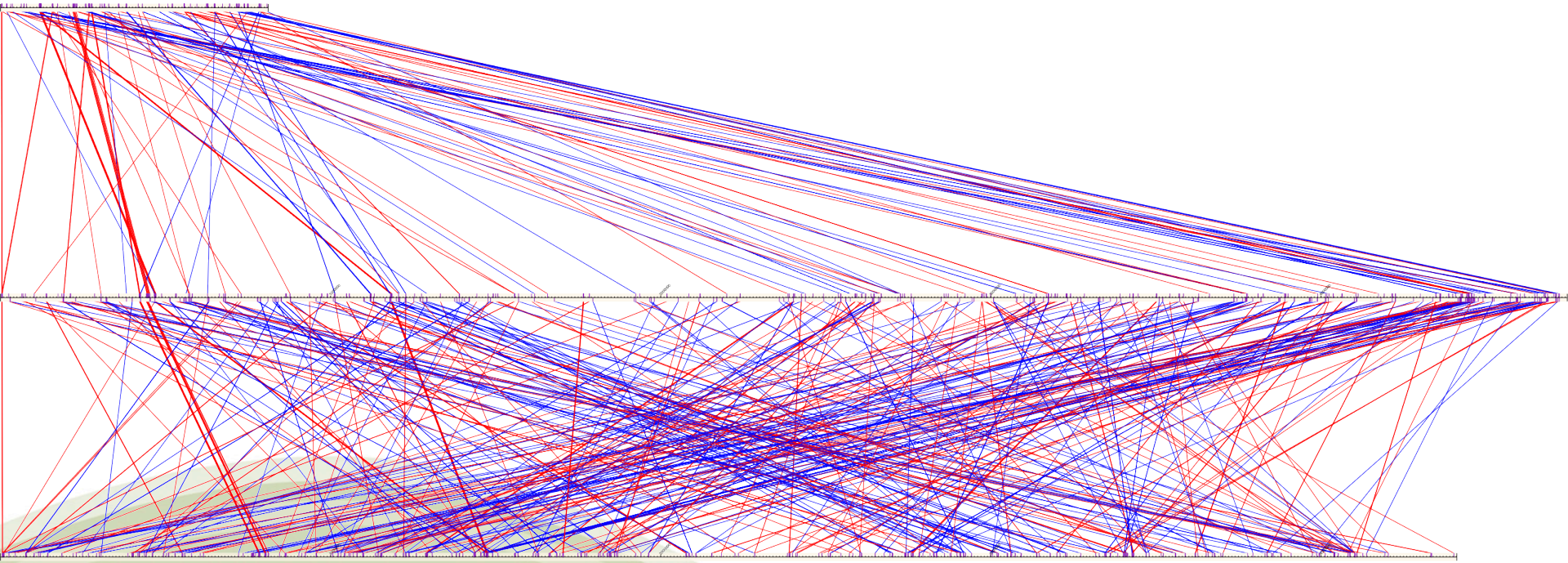
# Syntenous

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



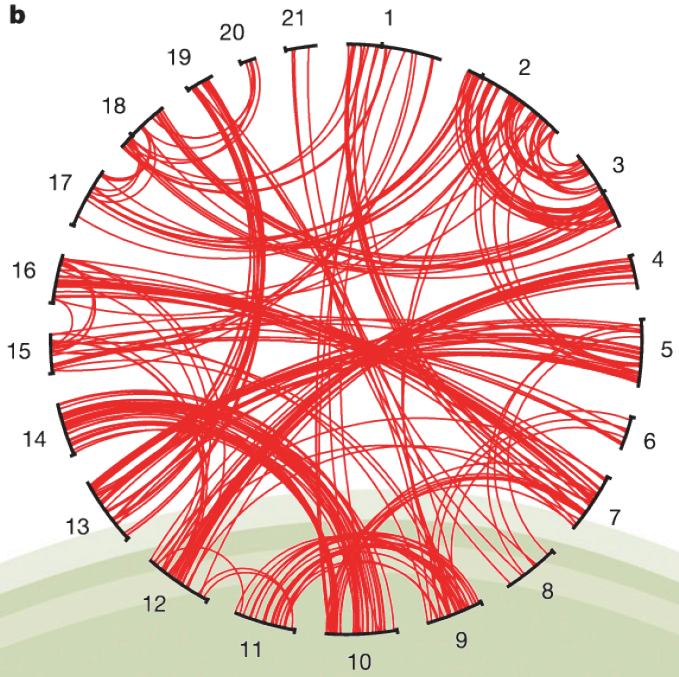
# Nonsyntenous

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

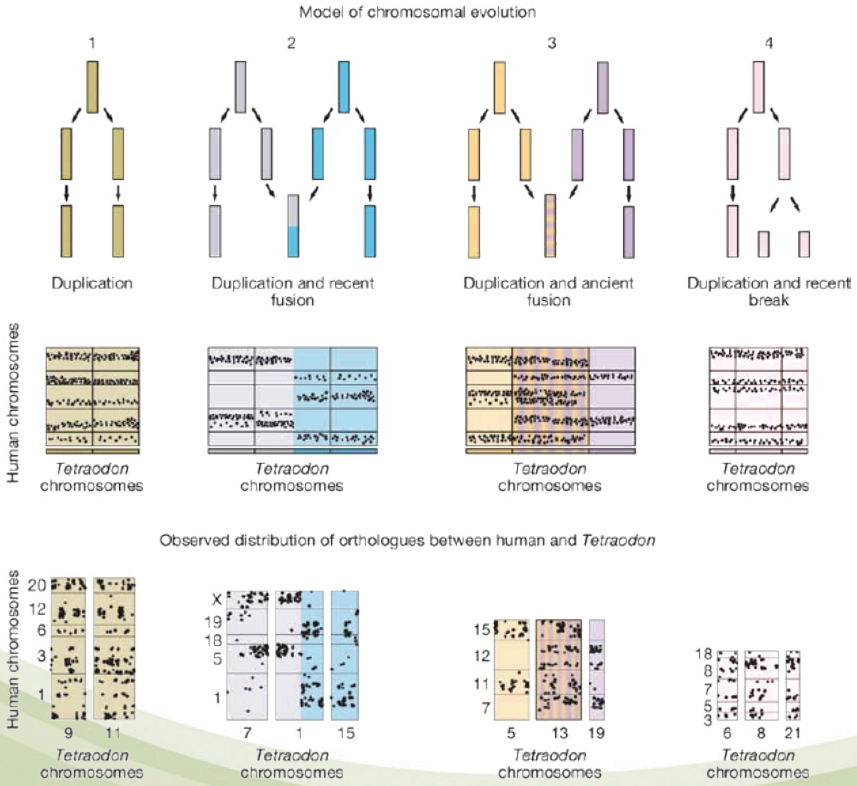


# Whole Genome Duplication

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



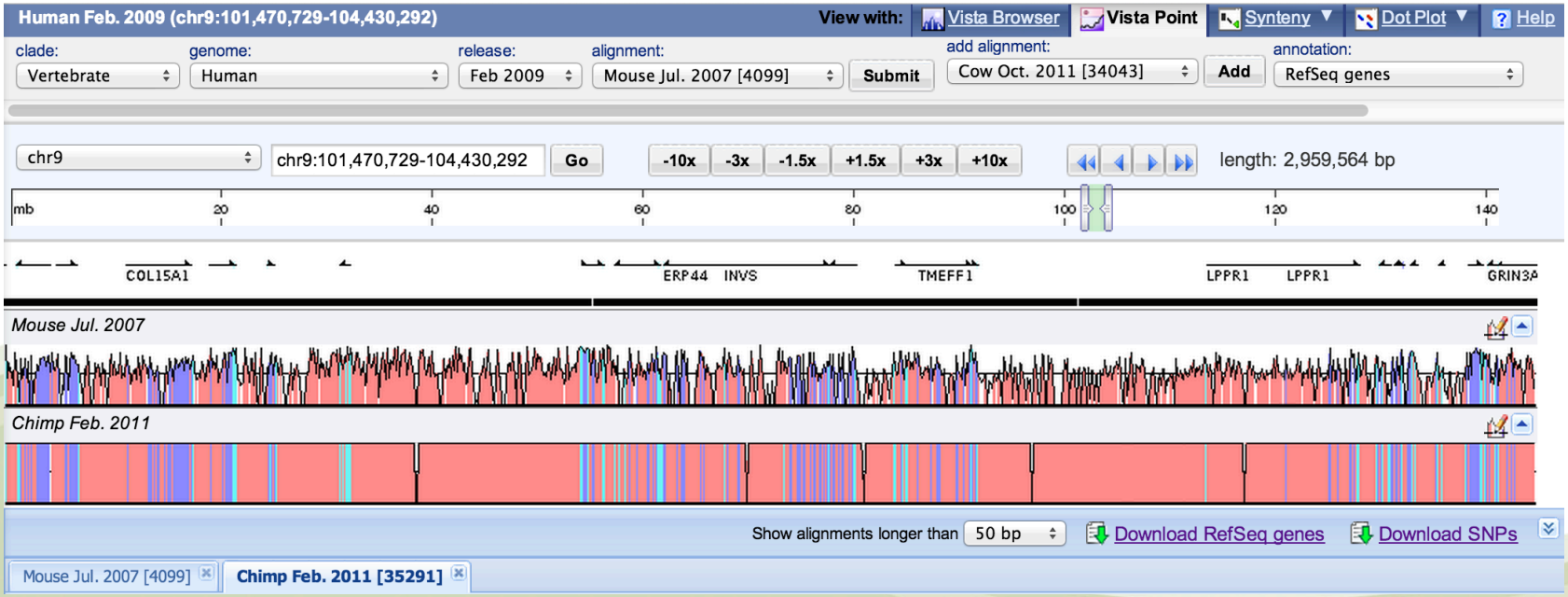

Duplicated genes (ExoFish) on 21 chromosomes




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

# VISTA, mVISTA, VISTA-Point

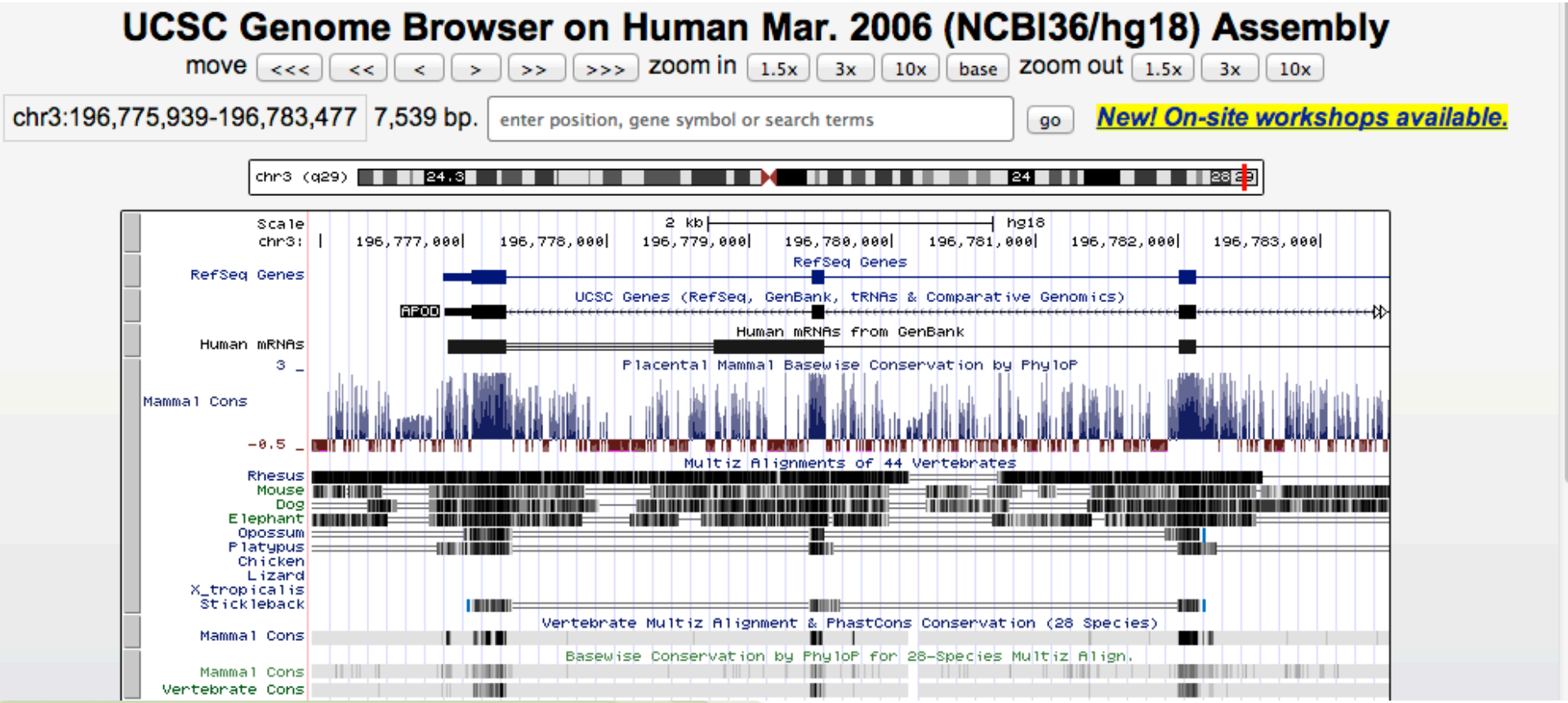
- Alignment/visualisation tools:
  - http://genome.lbl.gov/vista/index.shtml
- mVISTA: align and compare submitted sequences (up to 2Mbp)
- VISTA-Point: visualise precomputed alignments




Frazer *et al.* (2004) *Nucl. Acids Res.* doi:10.1093/nar/gkh458

# UCSC

- http://genome.ucsc.edu/




- Many vertebrate/invertebrate model genomes

Kent *et al.* (2002) *Genome Res.* doi:10.1101/gr.229102

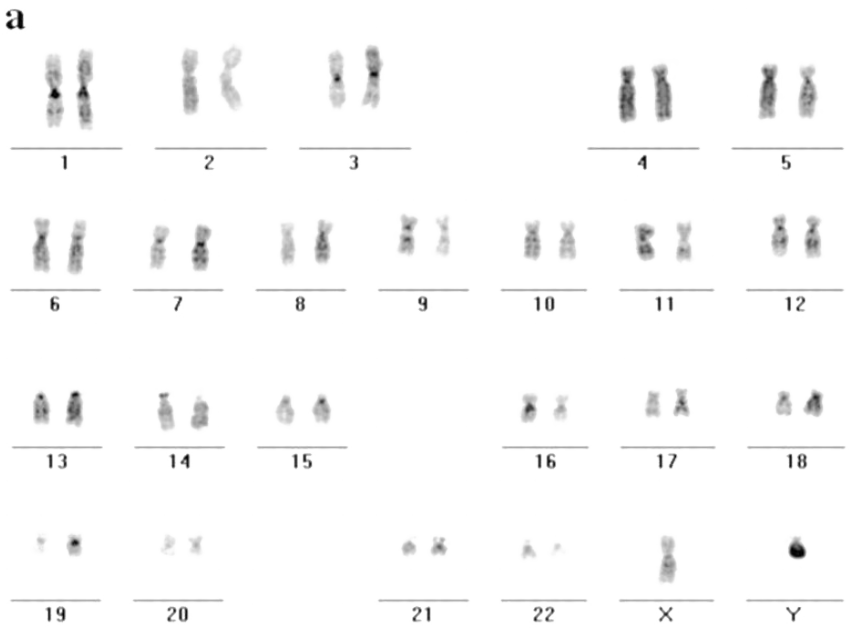
# Conclusion

- 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
- Coming in **Part 2**: genomic functional elements

The James Hutton Institute

# Nucleotide Content

- A, C, G, T composition
- Varies between, and within genomes
  - staining varies across genomes, due to variation in GC content
- *"isochores"*: regions with little internal GC variation (homogeneous)
  - long a point of discussion – difficult to define
- In humans:
  - L1, L2 isochores: low GC (≲41%)
  - H1, H2, H3 isochores: high GC (≳41%)
  - Imprecise bulk measurement

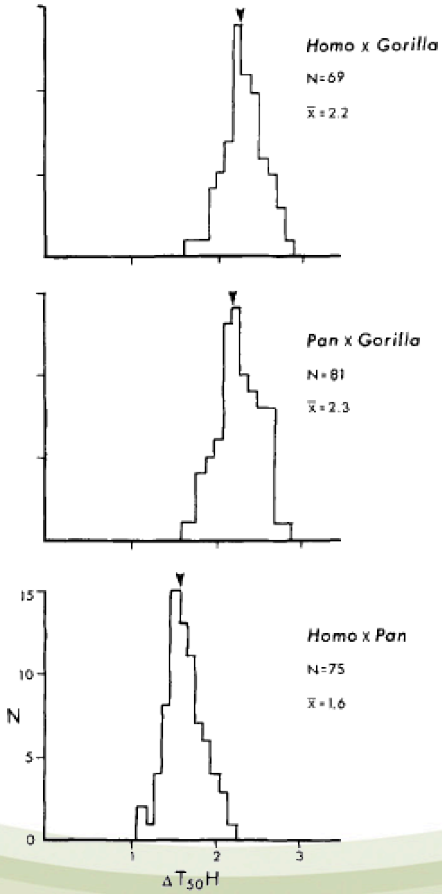



hybridisation of H3 isochore to human genome

Sadoni *et al.* (1999) *J. Cell Biol*. doi:10.1083/jcb.146.6.1211

# DNA-DNA Hybridisation (DDH)

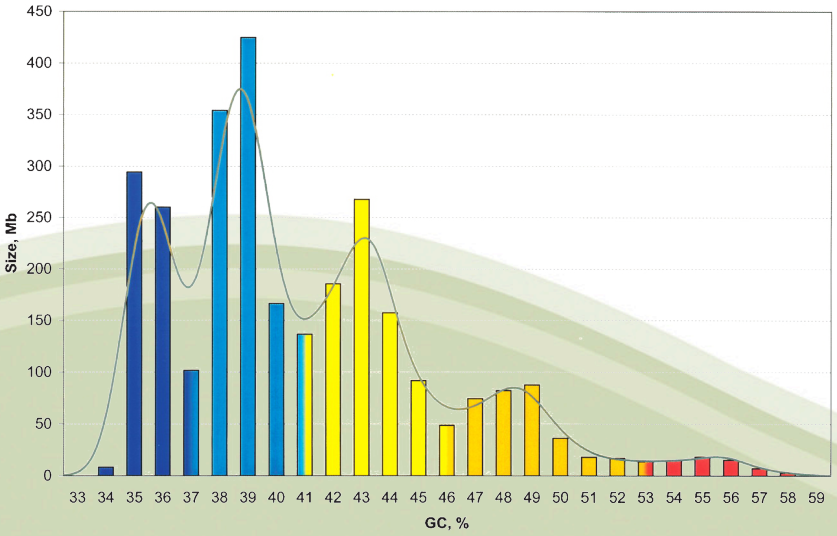
- Used for taxonomic classification in prokaryotes from 1960s
- Sibley & Ahlquist redefined bird and primate phylogeny with DDH in 1980s:
- Not without controversy:
  - Suggestions of data manipulation (see here)
  - Close evolutionary relationships difficult to resolve due to paralogy (more on paralogy later…)
- Still hanging on as a *de facto* "gold standard" in microbiological taxonomic classification.

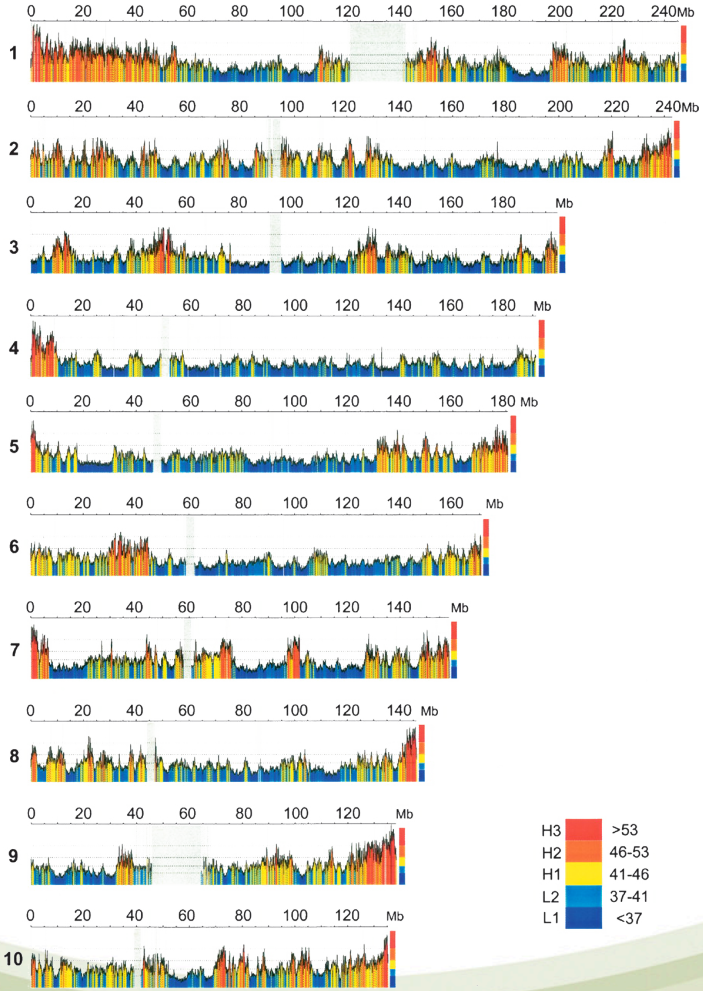



**Sibley & Ahlquist (1987) *J. Mol. Evol.* doi:10.1007/BF02111285**

# Finding isochores

- Isochores: homogeneous regions of %GC content
  - Easy to find with windowed (100kbp) %GC calculation, from sequenced genomes.
  - 3200 isochores characterised in the human genome, consistent with 5 levels (L1, L2, H1, H2, H3) found by staining/hybridisation.




Costantini *et al.* (2006) *Genome Res.* doi:10.1101/gr.4910606

# Comparative Genomic Hybridisation

The James Hutton Institute

- Two genomes: "reference" and "test" labelled (red and green), then hybridised against a "normal" genome

**BMC Cancer**

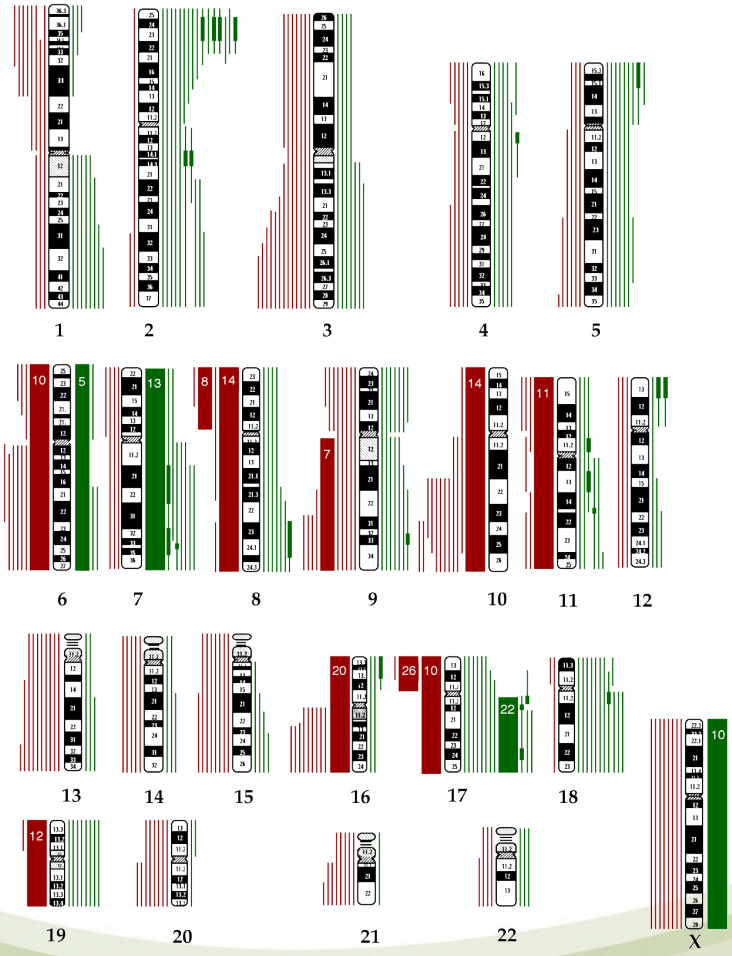
BioMed Central

Research article

Open Access

**Medulloblastoma outcome is adversely associated with overexpression of EEF1D, RPL30, and RPS20 on the long arm of chromosome 8**

Massimiliano De Bortoli[1], Robert C Castellino[1], Xin-Yan Lu[2], Jeffrey Deyo[3], Lisa Marie Sturla[4], Adekunle M Adesina[6], Laszlo Perlaky[1], Scott L Pomeroy[5], Ching C Lau[1], Tsz-Kwong Man[1], Pulivarthi H Rao[1] and John YH Kim*[1]




- **semiquantitative:**
- **Red**: ***loss*** (<2 copies) in tumour
- **Green**: ***gain*** (3-4 copies) in tumour
- Amplifications (>4 copies) in **BOLD**
- Cases with the same Copy Number Aberration (CNA) are numbered

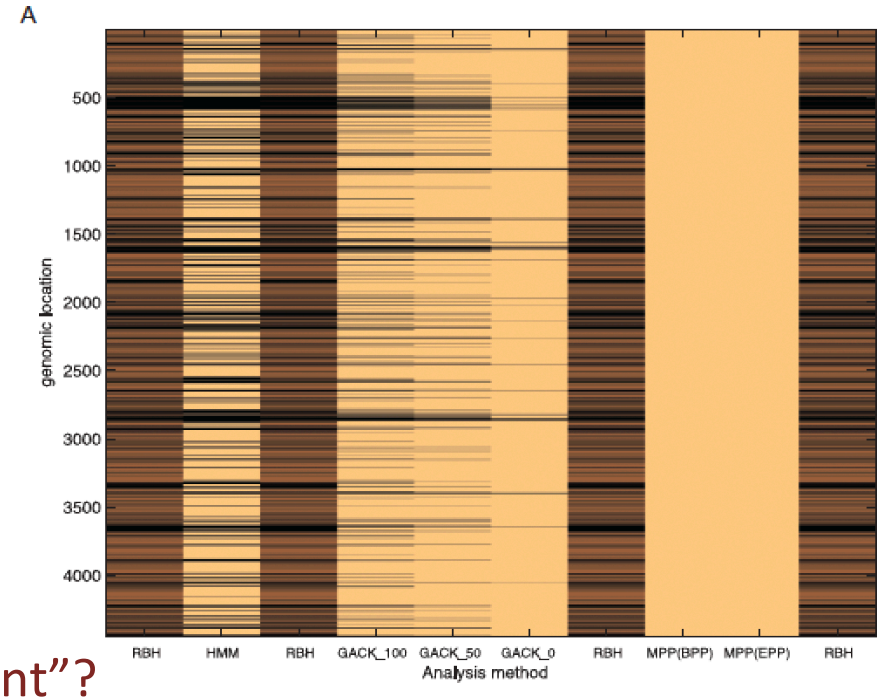
**De Bortoli *et al.* (2006) *BMC Cancer* doi:10.1186/1471-2407-6-223**

# Array Comparative Genomic Hybridisation

- Early approaches took a threshold score (present/absent)
- Later approaches used known reference genome sequence context (HMMs, synteny) to improve presence/absence calls

| Analysis Method | Correct Prediction Rate | Positive Prediction Rate | Sensitivity | Count |
|---|---|---|---|---|
| *Pba*1043:*Dda*3937 | | | | 1630 |
| HMM (*Pba*1043:*Dda*3937) | 0.7796 | 0.7752 | 0.5607 | 1179 |
| GACK (0% EPP) | 0.6512 | 0.9642 | 0.0497 | 84 |
| GACK (50% EPP) | 0.7011 | 0.9360 | 0.1975 | 344 |
| GACK (100%EPP) | 0.7355 | 0.8807 | 0.3214 | 595 |
| MPP (BPP) | 0.6337 | 0.0000 | 0.0000 | 0 |
| MPP (EPP) | 0.6337 | 0.0000 | 0.0000 | 0 |
| *Lactococcus lactis* | | | | 379 |
| HMM (*Lactococcus*) | 0.8404 | 0.7253 | 0.1741 | 91 |
| GACK (0% EPP) | 0.8210 | 0.0000 | 0.0000 | 0 |
| GACK (50% EPP) | 0.7616 | 0.2418 | 0.1557 | 244 |
| GACK (100%EPP) | 0.7616 | 0.2418 | 0.1557 | 244 |
| MPP (BPP) | 0.8220 | 1.0000 | 0.0053 | 2 |
| MPP (EPP) | 0.8210 | 0.5000 | 0.0053 | 4 |

doi:10.1371/journal.pcbi.1000473.t002

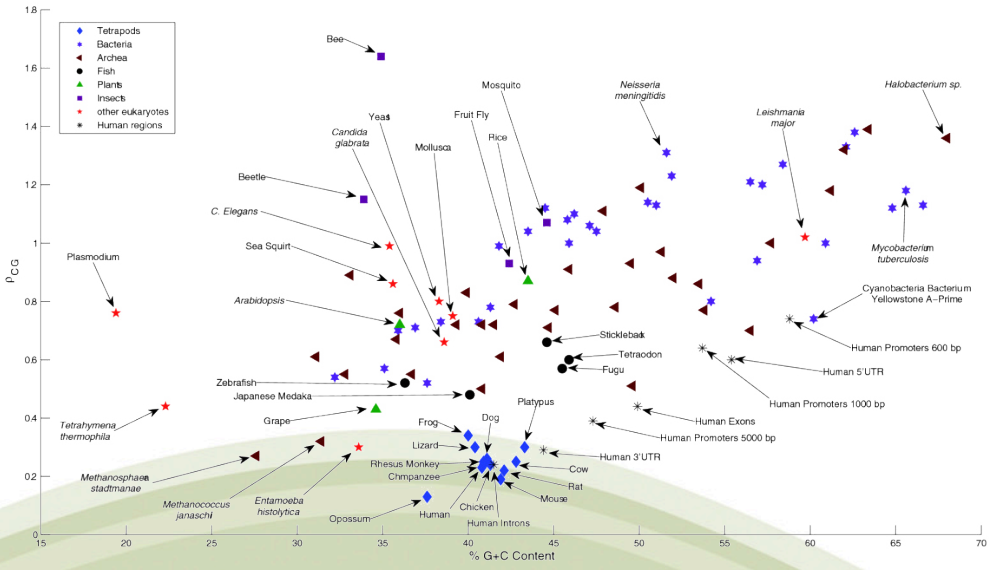


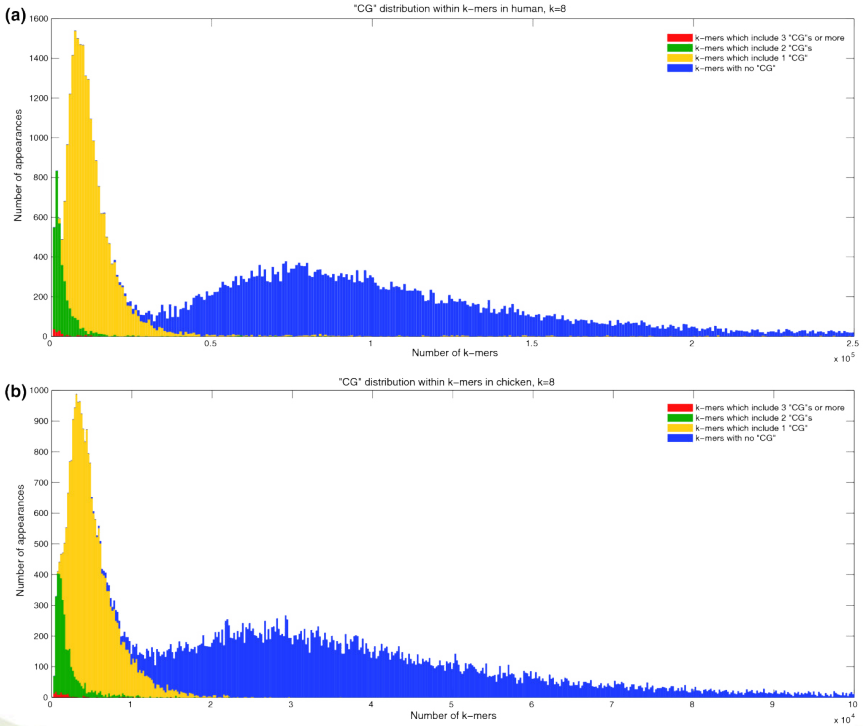

- No hybridisation = “absent” or“divergent”?
- **Not nearly as good as sequencing directly!**

**Pritchard *et al.* (2009) *PLoS Comp. Biol.* doi:10.1371/journal.pcbi.1000473**

# *k*-mer Spectra

- ***k*-mer spectrum:**
  - CpG suppression (CGs are uncommon in vertebrate genomes), but (by simulation) *only* when in combination with a particular %GC, explains multimodality




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