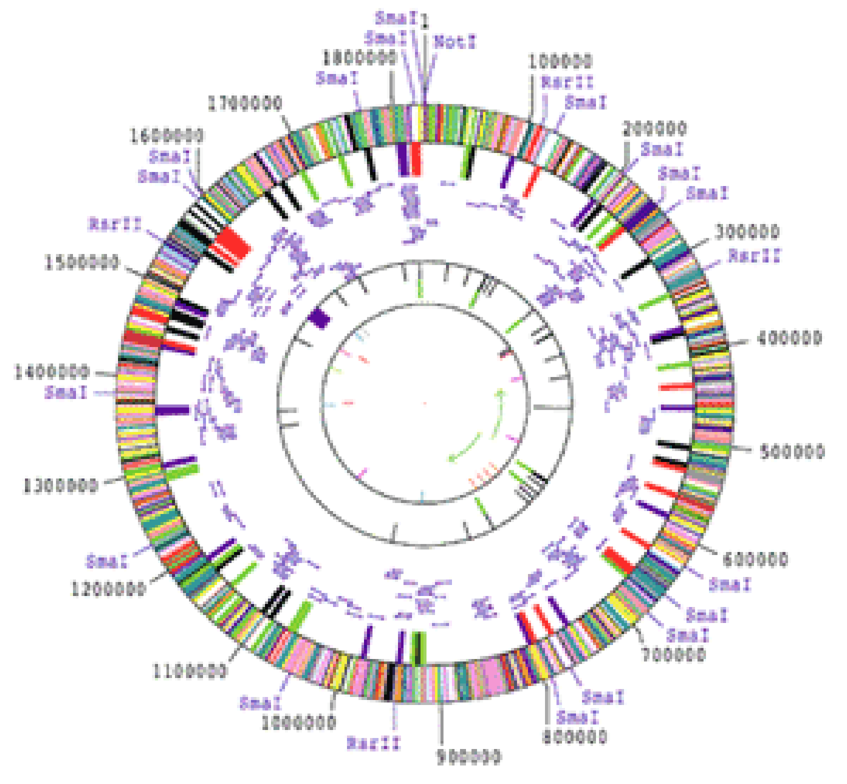
Microbial Genomics

Michael J. Stanhope,
Pop. Med. Diagnostic Sci.



Fleischmann et al. 1995. Science 269: 496

Outline

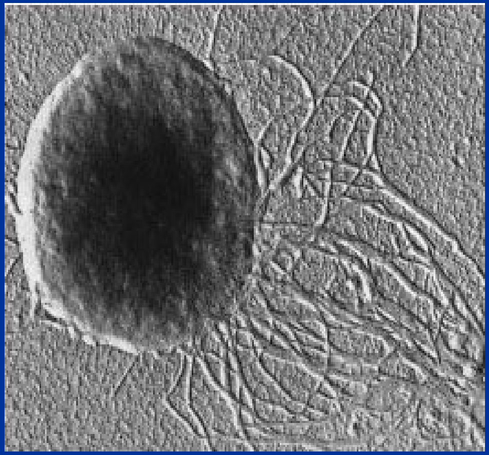
- Introduction
  - Microbial diversity
  - Universal Tree of Life
- Bacterial genome size
    - Core and pan genomes
- Horizontal Gene Transfer (HGT)
  - Mechanisms of HGT
  - Detecting HGT
- Comparative genomics of *Streptococcus*
- Comment on genome sequencing technology
- E.g. of 454 bacterial genome sequence
- Applications of microbial genomics

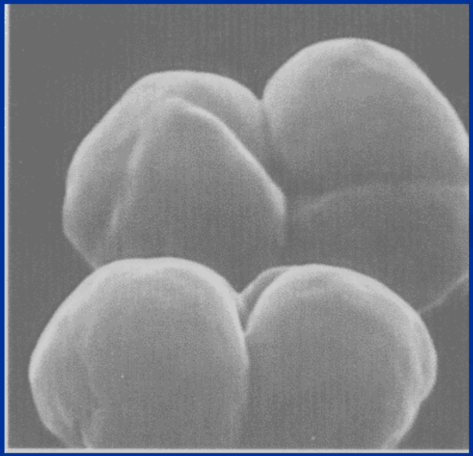
Introduction

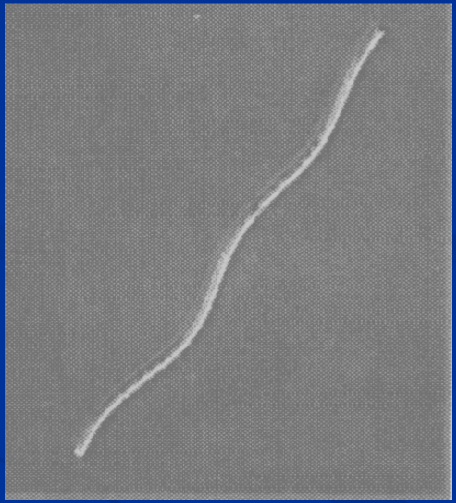
- Microbial diversity

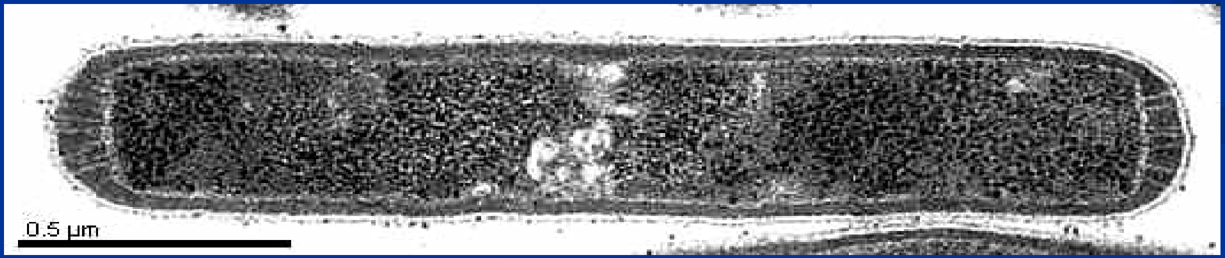
Microbial diversity

- Superficial inspection, bacteria and archaea hardly seem diverse









http://www.ucmp.berkeley.edu/archaea/archaeamm.html

Microbial diversity

- But metabolic diversity great, particularly energy generating
  - Even within a species; e.g. *E. coli*:
    - Fermentation or respiration; respire aerobically or anaerobically; glucose or lactose as sole carbon source – transforming sugar into amino acids, vitamins, nucleotides

Microbial diversity

- Energy generating metabolism in bacteria:
  - Alcohol fermentation
  - Lactic acid fermentation
  - Aerobic respiration
  - Oxygenic photosynthesis

  } present in eukaryotes & prokaryotes

  - Anaerobic degradation of carbohydrates through the Embden-Meyerhof pathway.
  - Other fermentation pathways e.g. phosphoketolase pathway
  - Anaerobic respiration
  - Lithotrophy (inorganics as source of energy)
  - Anoxygenic photosynthesis
  - Methanogenesis (H_2 as energy source and produces methane)
  - Light driven nonphotosynthetic photophosphorylation

Microbial diversity

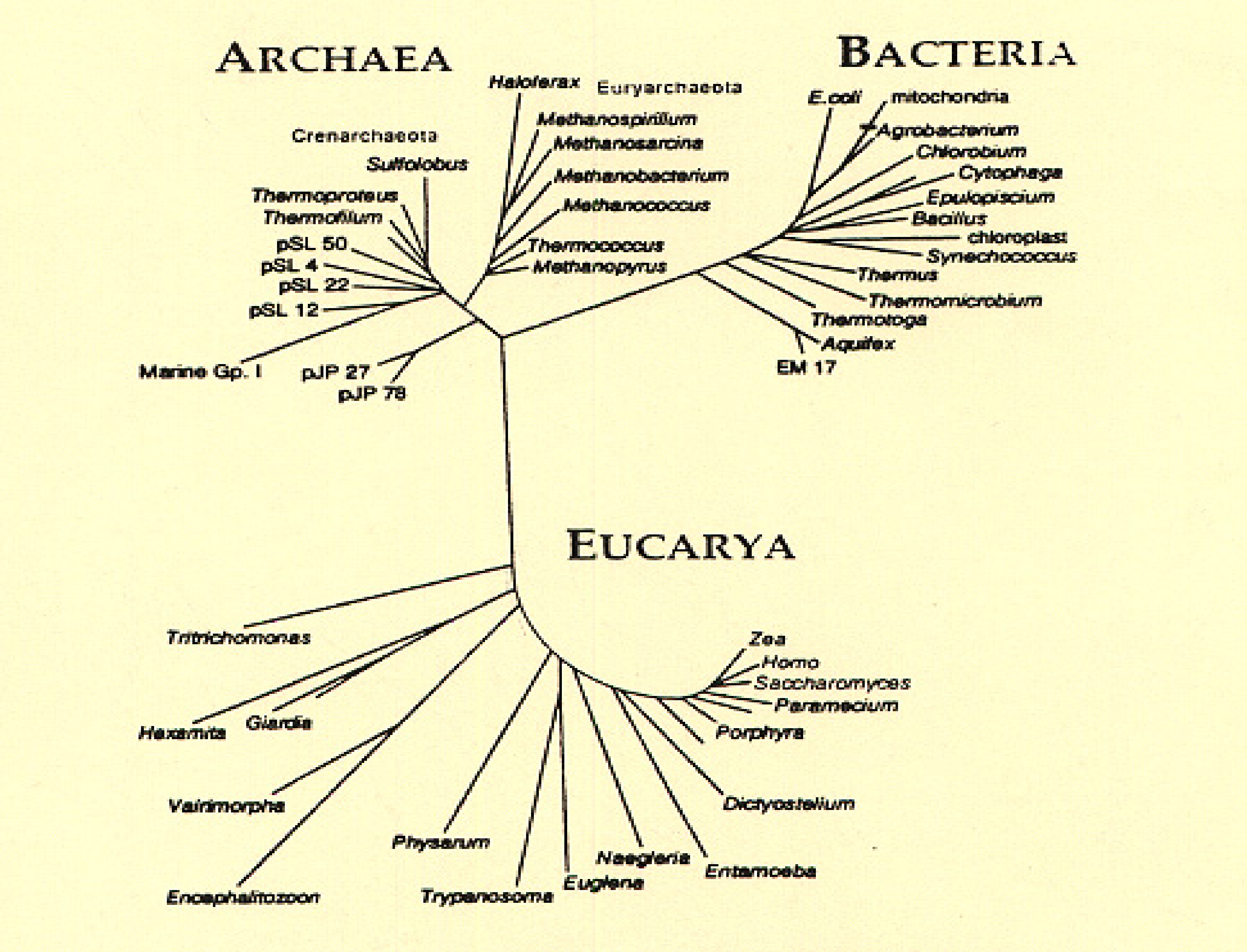
- prokaryotic cells on Earth = 6 X 10^{30}
- Prokaryotic cellular carbon = 60-100% of estimated carbon in terrestrial and marine plants.
- Abundant in environments where eukaryotes are rare
- How many species?
  - Definition of species?
    - Lack diagnostic morphological characteristics
    - Exchange genetic material in unique and unusual ways
    - Same species = 70% DNA-DNA hybridization
      - Underestimating prokaryotic diversity
  - Practical limitations in counting
    - 1% cultivable

Introduction

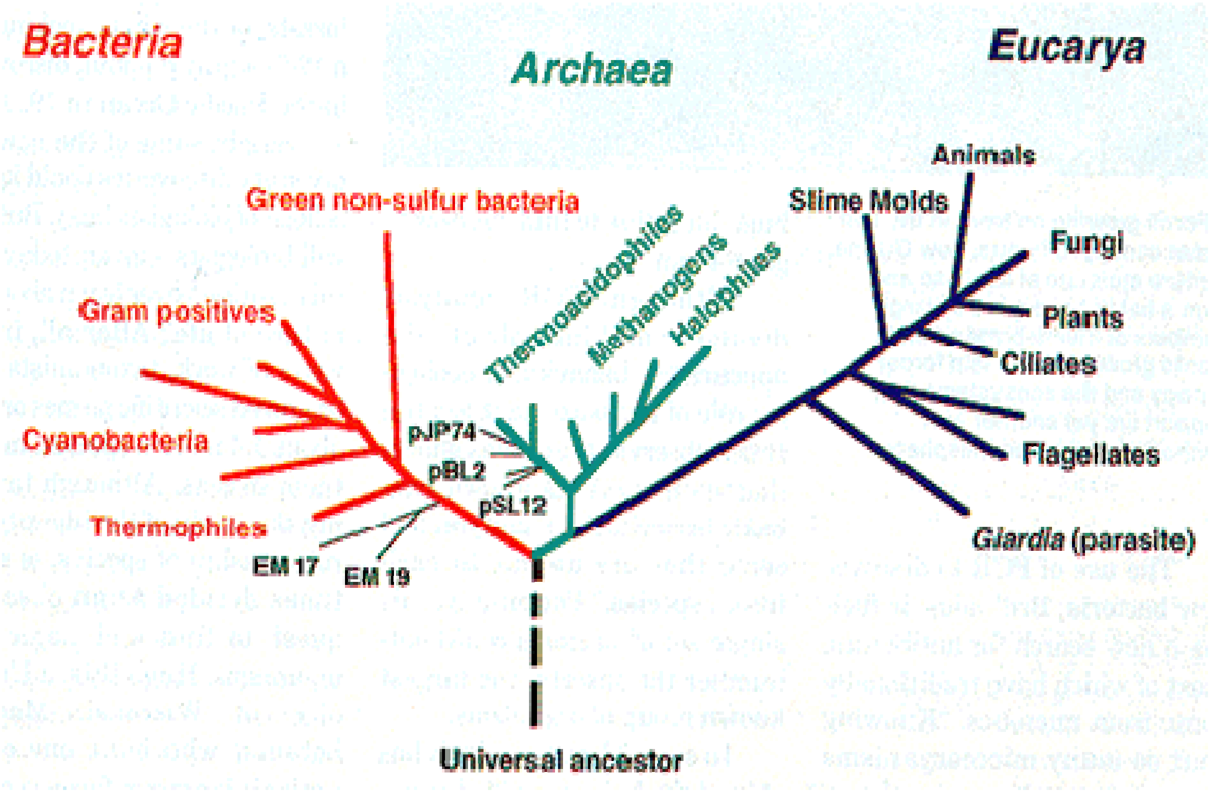
- Universal Tree

Universal Tree of Life

- 1980’s Carl Woese, phylogenetic analysis of all forms of cellular life; ssrRNA
  - Found in all cells
  - Present in thousands of copies and easy to isolate
  - Complementary to sequence of gene
  - Sequence can be compared to reveal similarity and differences
- Defined three cellular domains of life:
  - Eukaryotes
  - Eubacteria (Bacteria)
  - Archaeabacteria (Archaea)



Pace, NR. 1997. Science 276:734



THE UNIVERSAL TREE OF LIFE

http://whyfiles.org/022critters/archaea.html

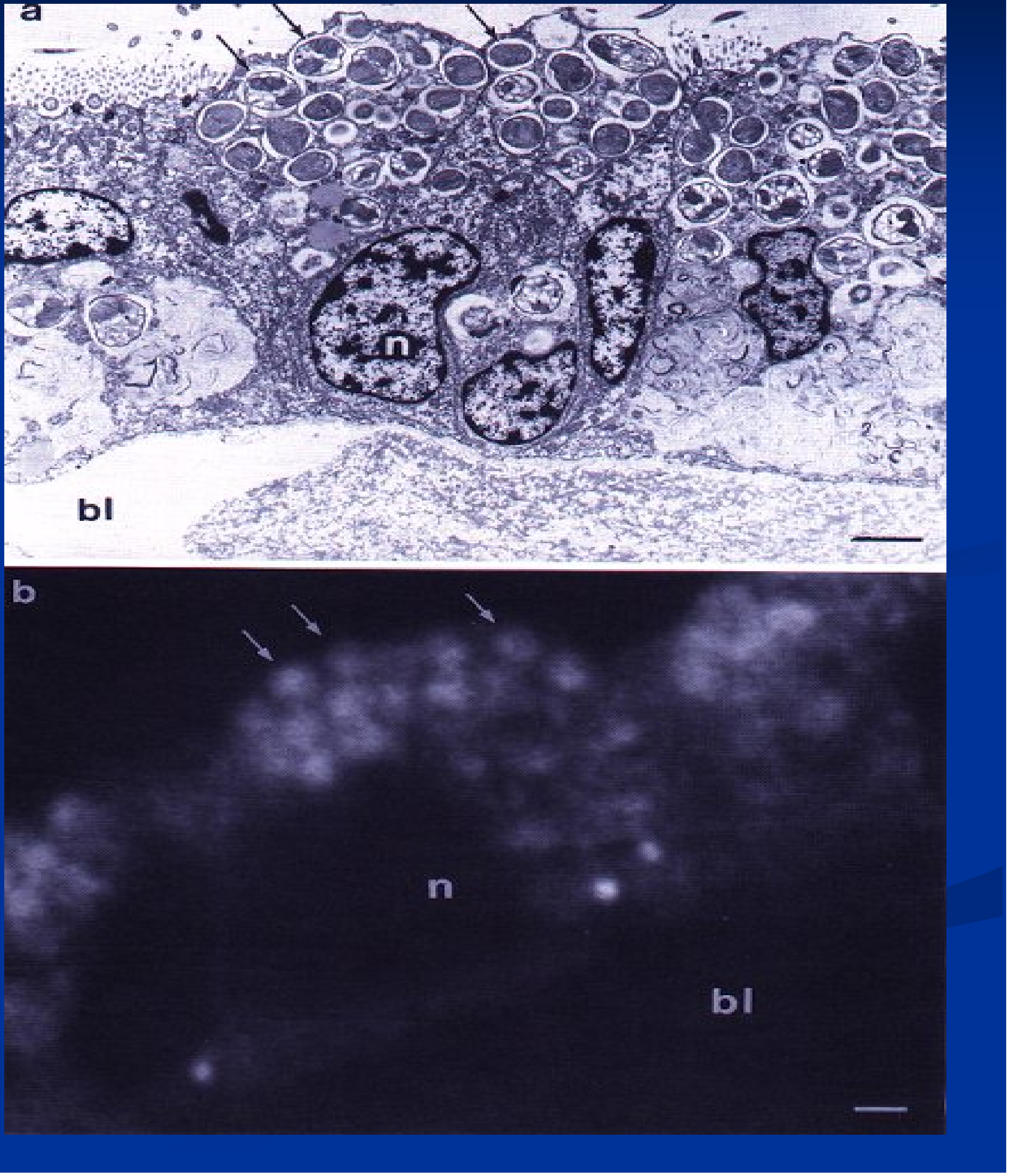
Genome Size

Genome size

- 405 complete bacterial genomes on NCBI
  - *Carsonella ruddii* (159,662) – *Burkholderia xenovorans* (9.73 Mb)
- Genome size / ecological niche
  - Smaller genomes, endocellular parasites or symbionts

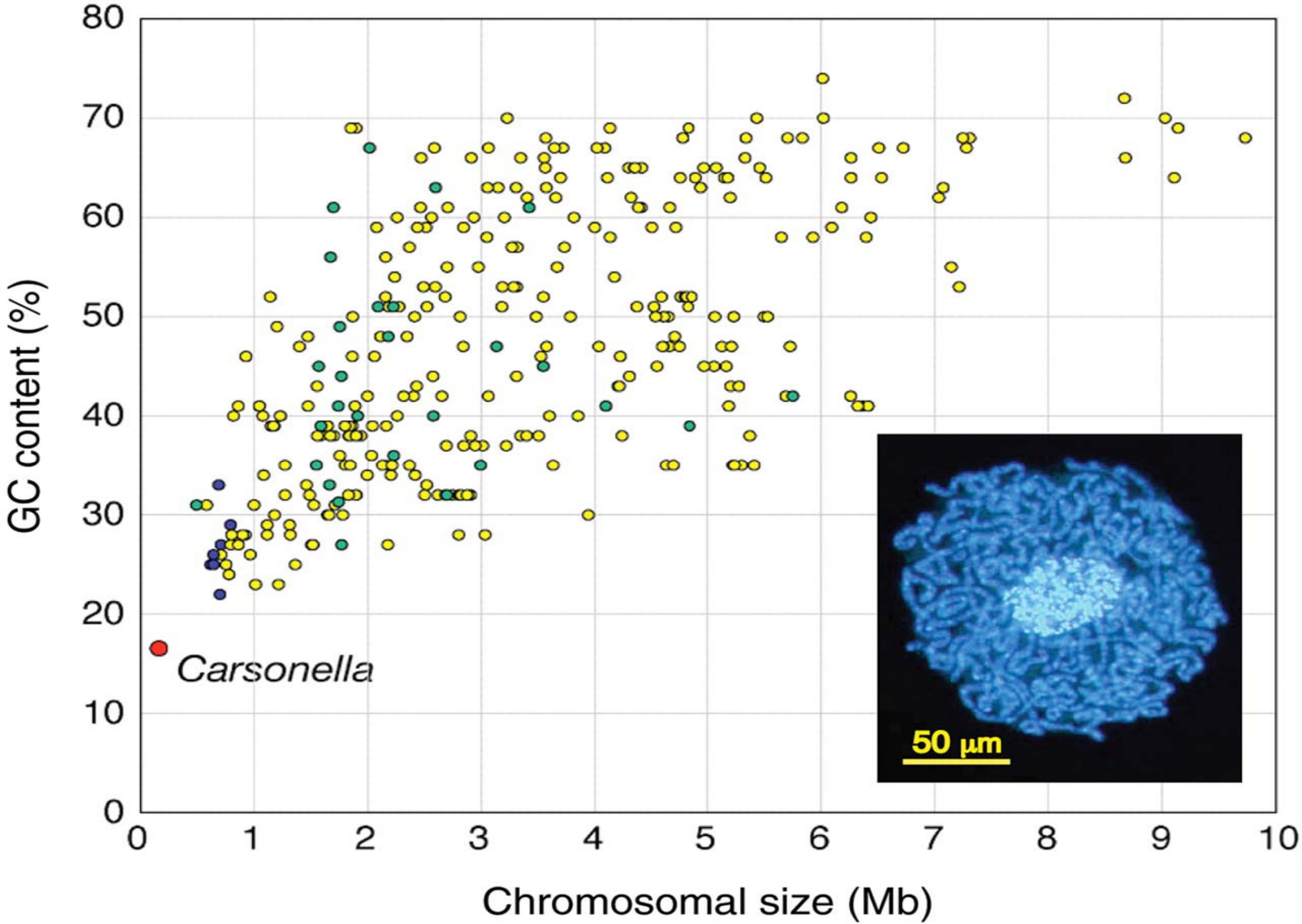
Genome size

- Mutually obligate endosymbiotic associations with animal hosts
  - bacteriocytes



Distel and Cavanaugh. 1994. J. Bact. 4:1932.

Genome size



Nakabachi et al. 2006. Science 314:267

Genome size

- Free living bacteria, genome size correlates with species metabolism & width of ecological niche
  - Pathogenic species, narrow range of hosts, small genomes; e.g. *Helicobacter, Streptococcus*
  - Anaerobic bacteria, restricted metabolism, e.g. methanogens, small genomes.
  - Aerobic organisms, and opportunistic pathogens, higher diversity of genome size; e.g. *Pseudomonas* (6 Mb)

pan and core genomes

- Core
  - Genes present in all strains
- Pan (from Greek meaning whole)
  - Dispensable genome composed of genes absent from one or more strains and genes unique to particular strains

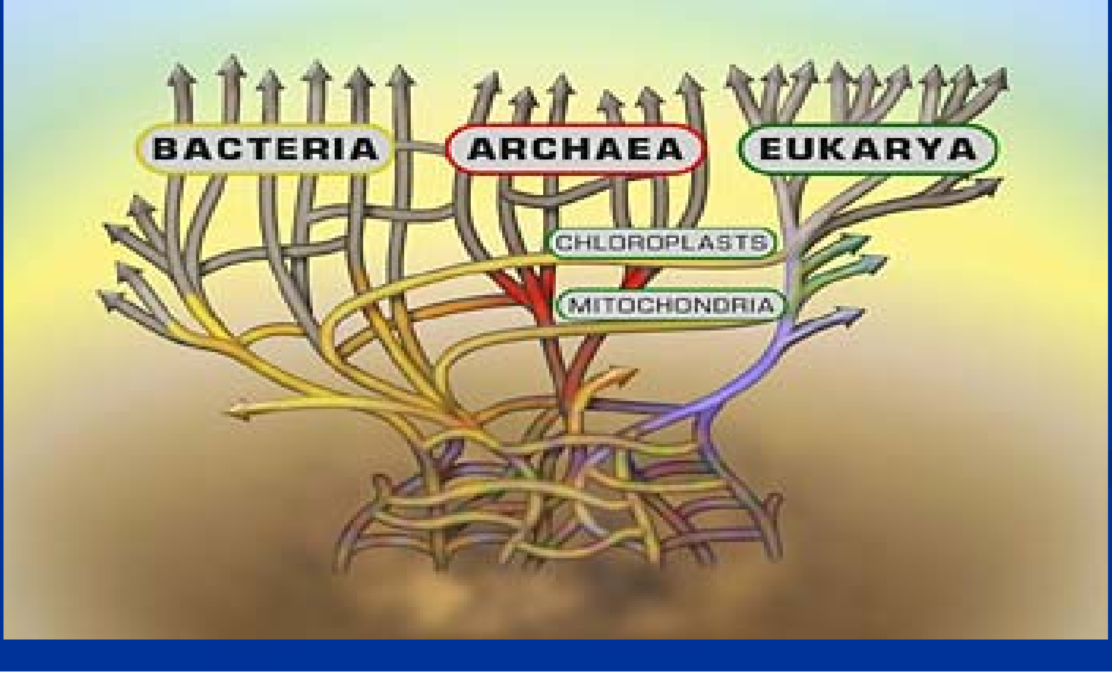
Bacteria chromosomes

- Most, single circular chromosome, but exceptions:
    - E.g. *Streptomyces*, *Borrelia*, *Agrobacterium*, linear chromosomes
    - Linear plasmids – e.g. *Klebsiella*, *Escherichia*, *Thiobacillus*
    - Linearity: enhances genomic plasticity?
    - Multichromosome spp.; e.g. some proteobacteria with free living, opportunistic lifestyle

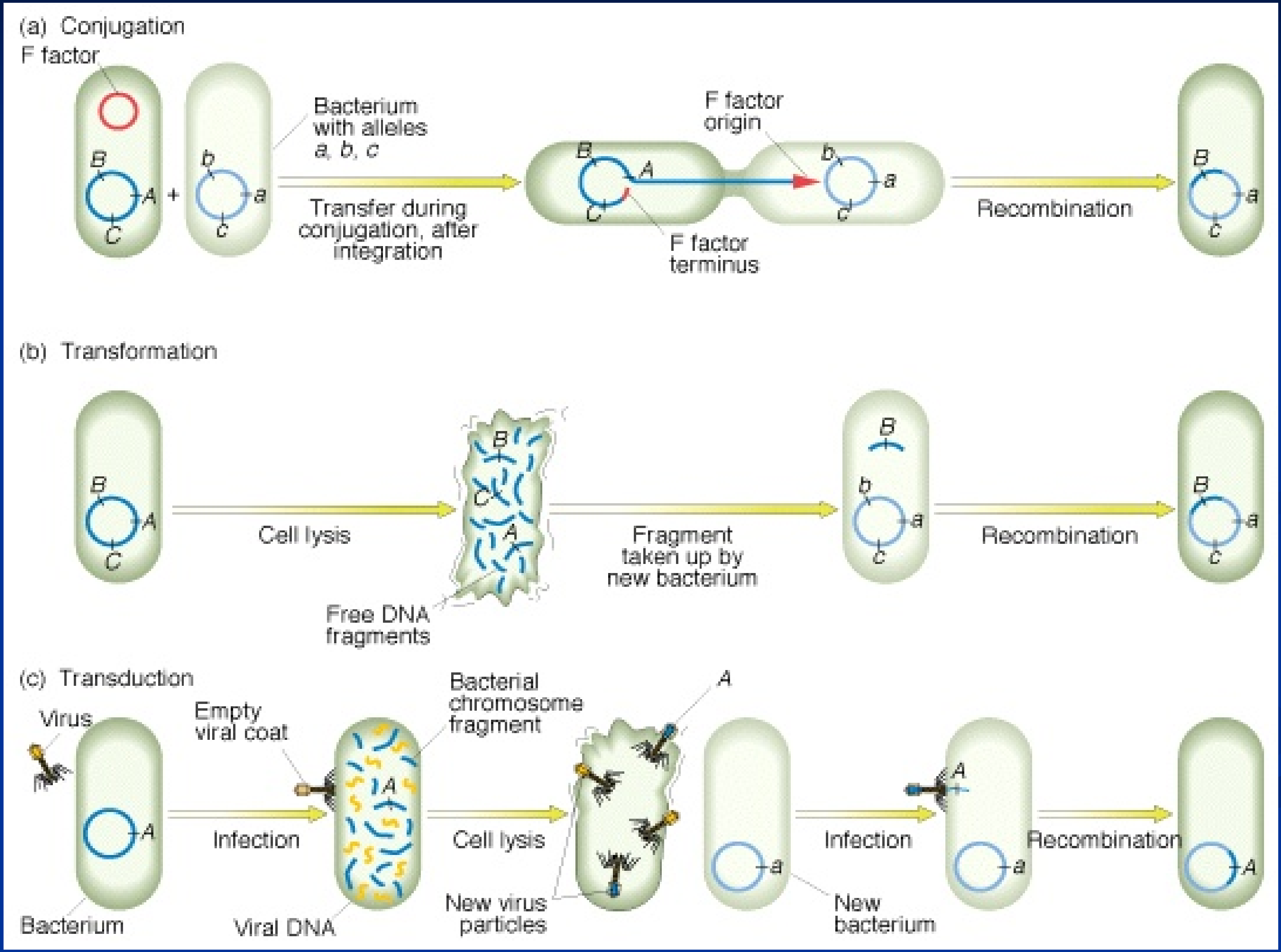
Horizontal Gene Transfer

Horizontal Gene Transfer

- Genetic exchanges between different evolutionary lineages
- 1944 Avery et al., DNA can be absorbed by microorganisms (Studies on the chemical nature of the substance inducing transformations of pneumococcal types. J. Exp. Med.79:137)
- Extent or degree is much debated



Mechanisms of HGT

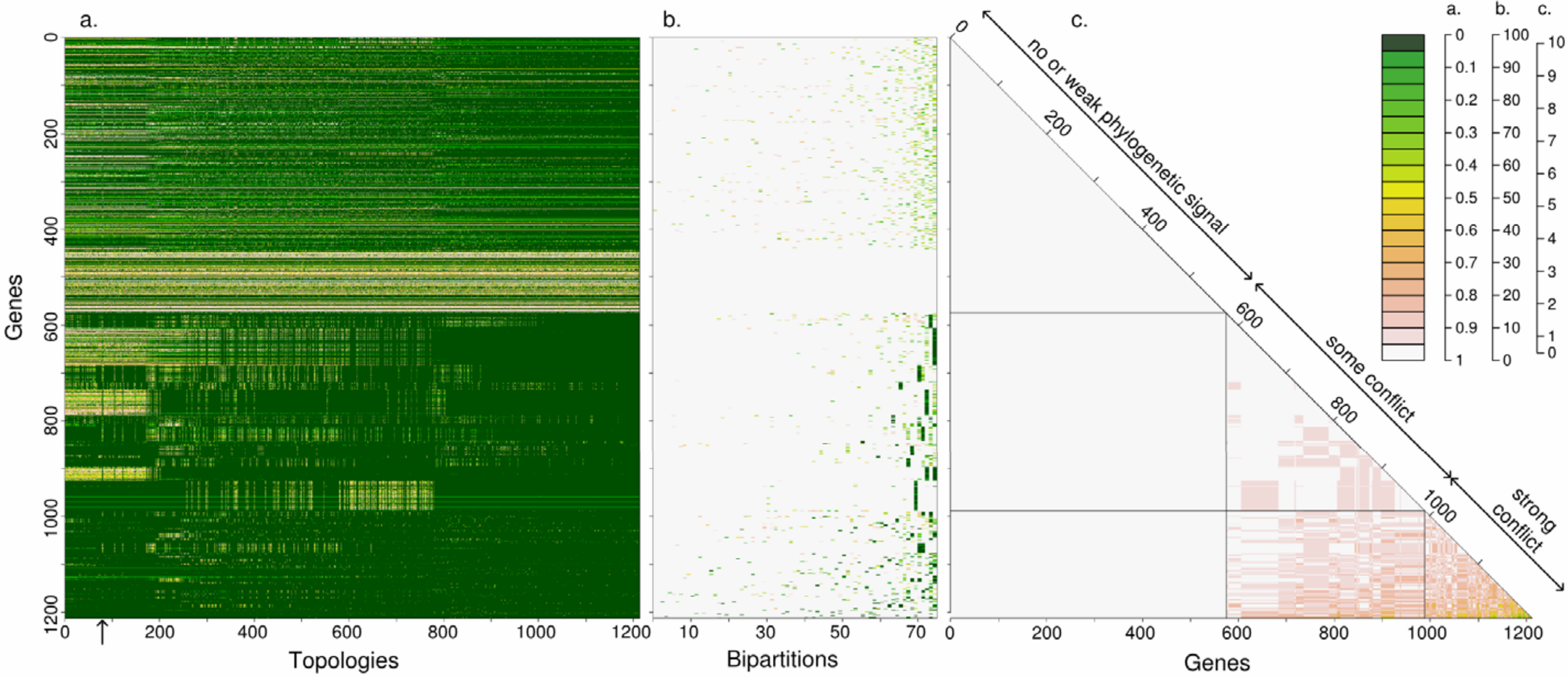


http://fig.cox.miami.edu/Faculty/Dana/bacfun.jpg

Detecting HGT

- Phylogenetics
  - Gene tree that differs significantly from species tree
  - Compare all gene trees; gene trees that are significantly different from majority are putative LGT

Detecting HGT



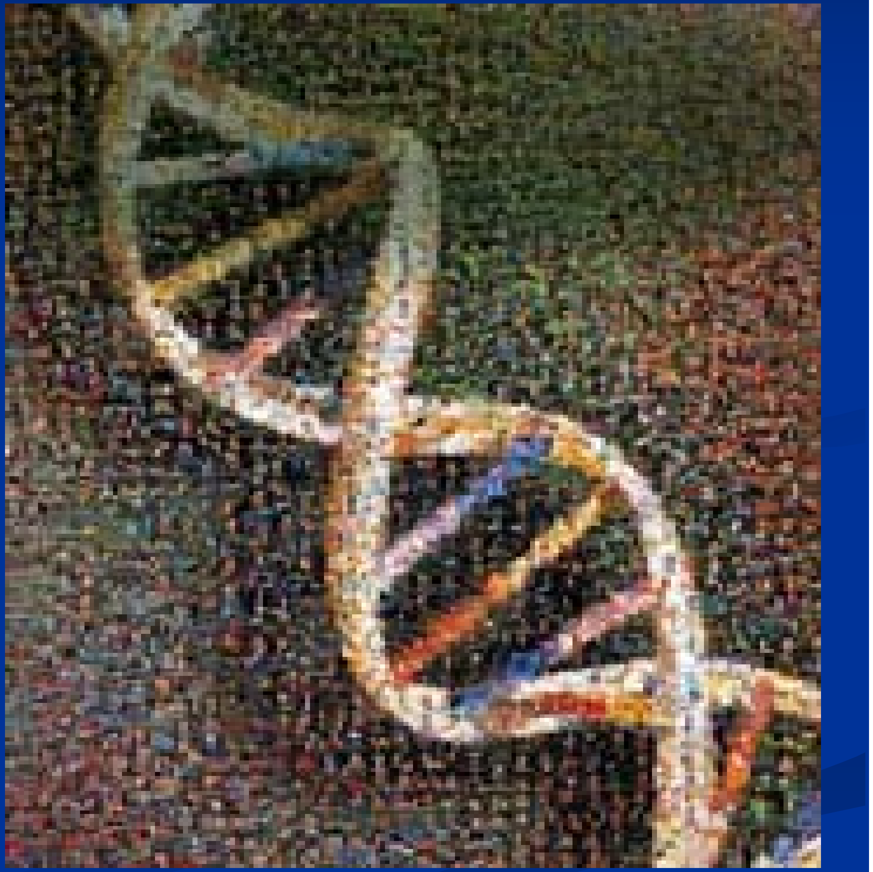
Detecting HGT

- Best sequence match detection (BLAST)
  - Rapid, but of limited use, since sequence similarity not necessarily correlated with evolutionary history.

Bacteria to Vertebrate Horizontal Gene Transfer??

"Hundreds of human genes appear likely to have resulted from horizontal transfer from bacteria at some point in the vertebrate lineage."

International Human Genome Sequencing Consortium.
Initial Sequencing and Analysis of the Human Genome.
***Nature* 409, 860 (2001).**



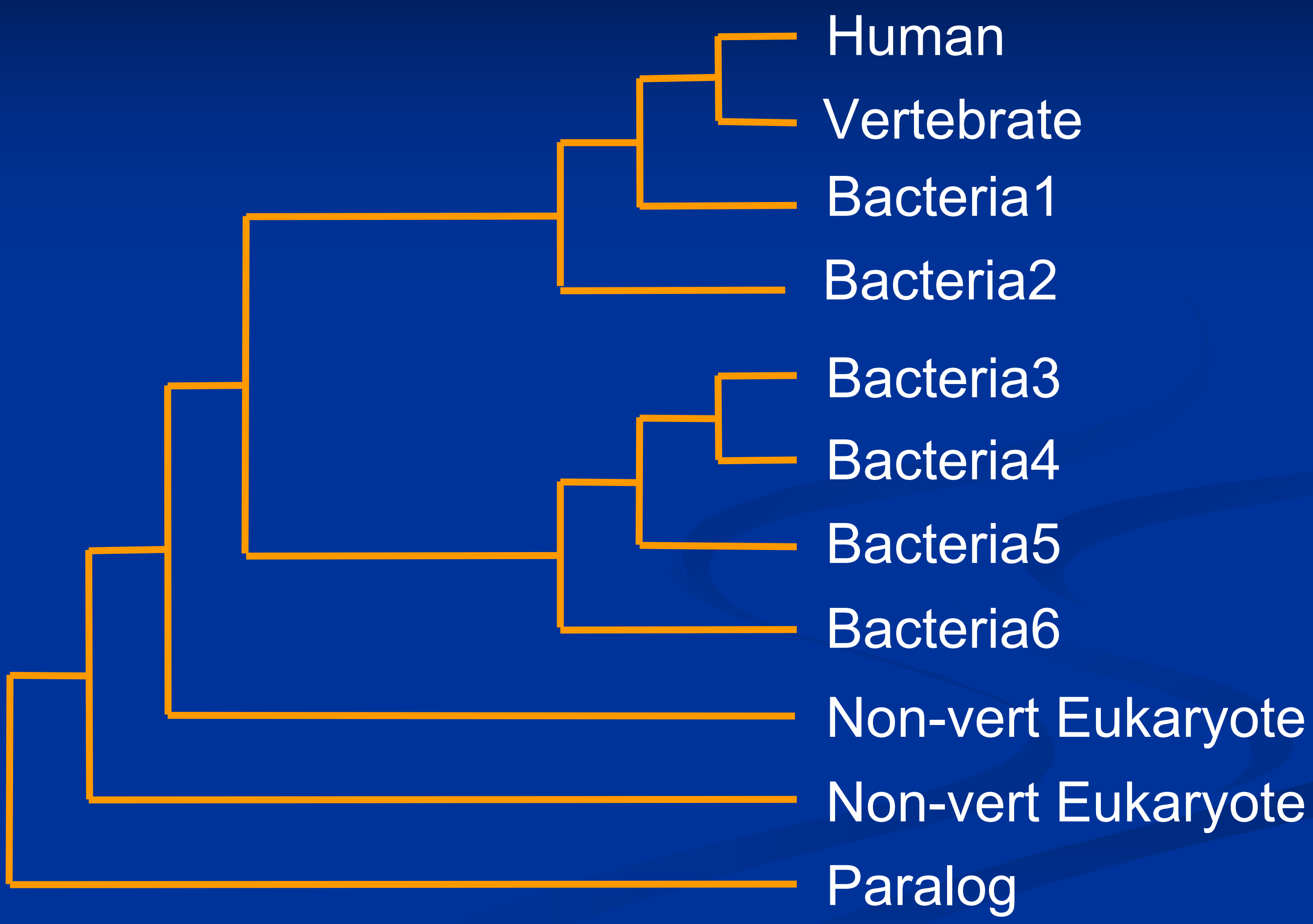
Bacteria to Vertebrate HGT -- Implications (*If True*)

- HGT bacterial genes became fixed in vertebrates through insertion into germ cells (because somatic cell HGT genes would be lost within a generation).
- Foreign bacterial genes can co-opt vertebrate regulatory regions and transcription factors.
- Humans could accumulate foreign, perhaps deleterious, genes from bacterial infections and/or GM foods.

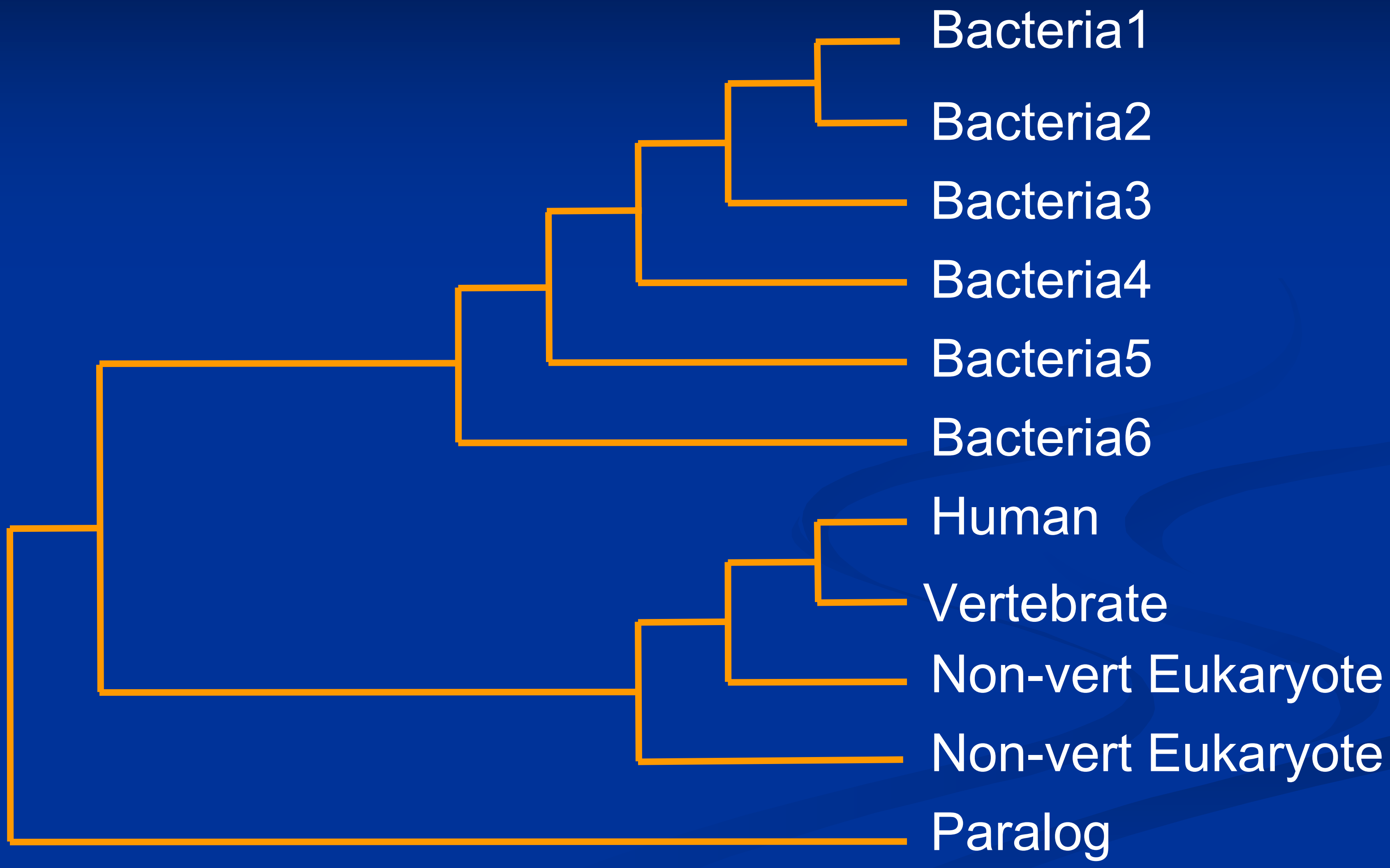
Nature Human Genome Issue

- International Human Genome Sequencing Consortium (IHGSC)
  - 113 genes that are likely examples of bacteria to vertebrate HGT (horizontal gene transfer).
    - Conclusion based on BLASTP alignment scores. Best sequence match detection (BLAST)

Phylogenetic evidence in support of bacteria - vertebrate HGT



Phylogenetic evidence rejecting bacteria - vertebrate HGT



Why did the IHGSC conclude bacteria to vertebrate HGT?

- Equated BLAST ranking with evolutionary relatedness.

Detecting HGT

- Nucleotide compositional analysis
  - Based on premise that DNA fragments obtained through HGT retain sequence characteristics of donor genome
  - Advantage is it only requires genome sequence from 1 spp.

Comparative Genomics of *Streptococcus*

Streptococcus genomes

- 26 genomes (public) from 6 spp
  - *S. pneumoniae* (2), *S. agalactiae* (8), *S. pyogenes* (11), *S. thermophilus* (3), *S. mutans* (1), *S. suis* (1)

Adaptive potential of bacteria

- 1. Darwinian or positive selection, favoring the fixation of advantageous mutations
- 2. acquisition of new genetic material by lateral DNA exchange
- 3. gene regulation

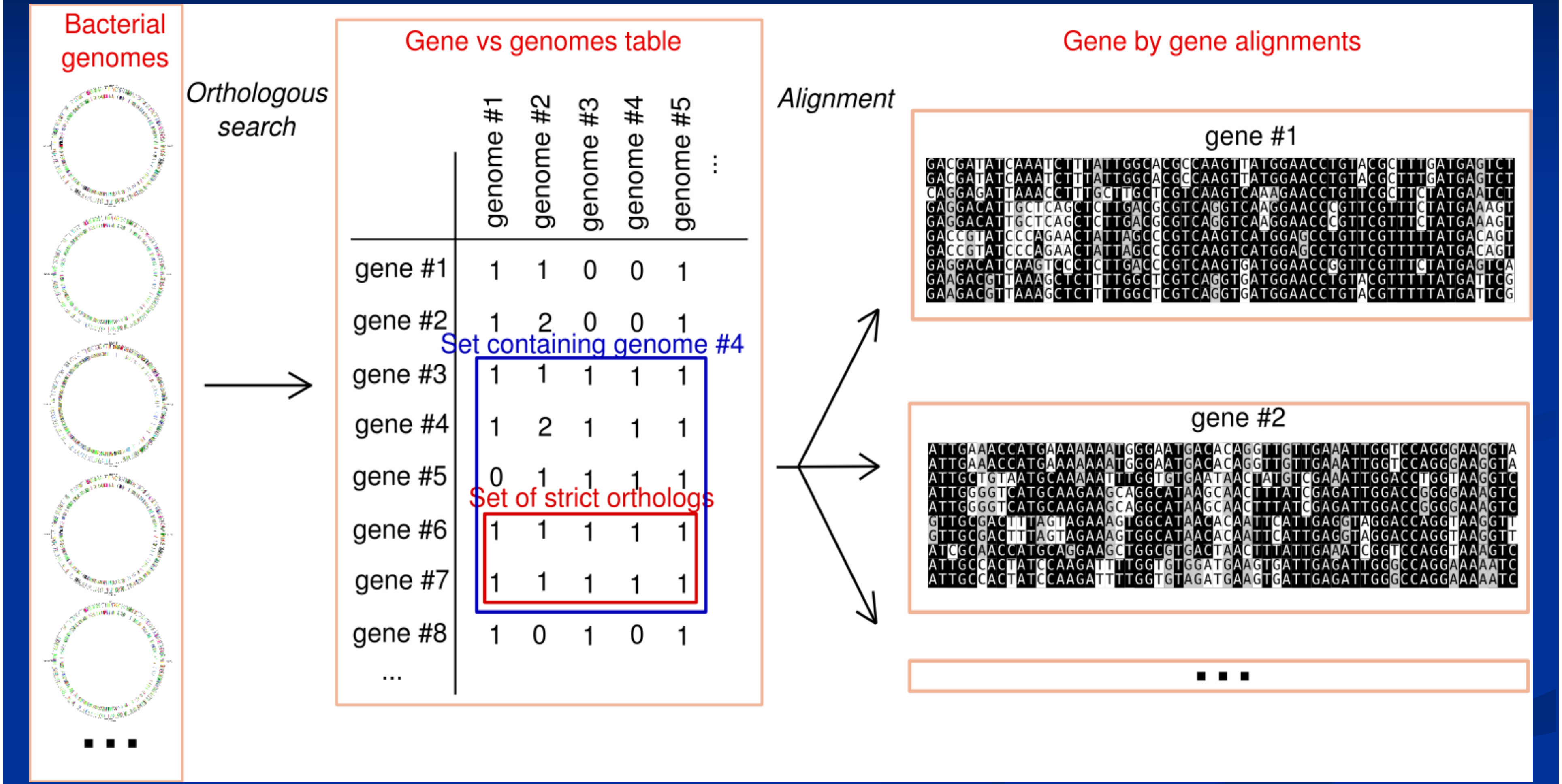
Core genome

- LGT of bacterial genomes, possibly key factor in adaptation
  - Nonetheless, core genome, possibly relatively LGT free
- Focus on adaptation often centered on species specific loci
  - Selection pressure on core genome not explored

Molecular selection

- Powerful statistical methods for detecting adaptive molecular evolution (Yang and Nielsen)
    - Nonsynonymous substitution rate elevated above the synonymous rate as evidence for positive selection
        - Fixation of advantageous mutations, driven by NS =>evolutionary innovations
- **Our goal**: assess positive selection pressure across core genome components of *Streptococcus*, while concomitantly assessing levels of recombination within core genome

Pipeline (part 1)



Gene vs genomes table

	genome #1	genome #2	genome #3	genome #4	genome #5	...
gene #1	1	1	0	0	1	
gene #2	1	2	0	0	1	
gene #3	1	1	1	1	1	
gene #4	1	2	1	1	1	
gene #5	0	1	1	1	1	
gene #6	1	1	1	1	1	
gene #7	1	1	1	1	1	
gene #8	1	0	1	0	1	
...						

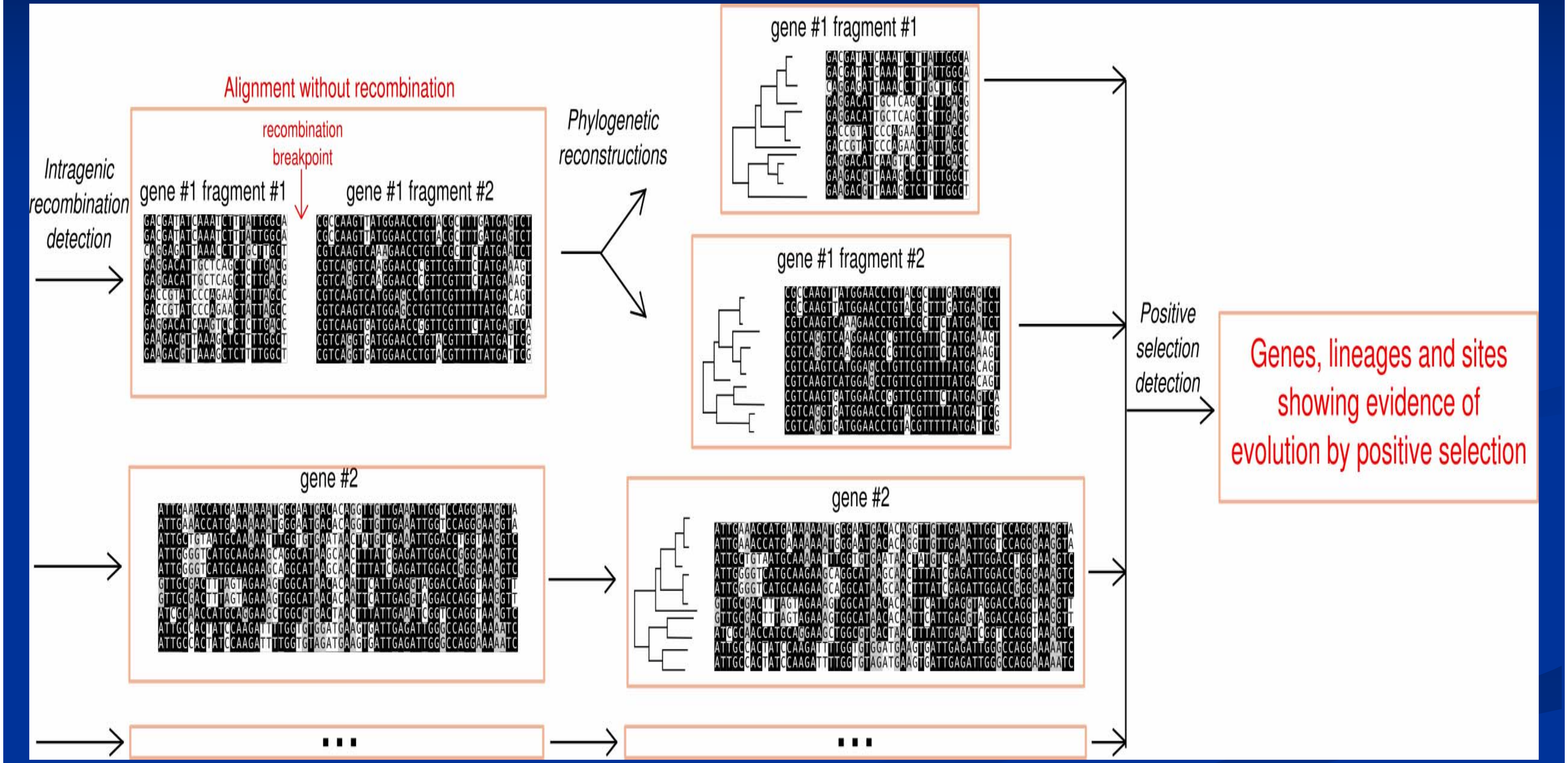
gene #1

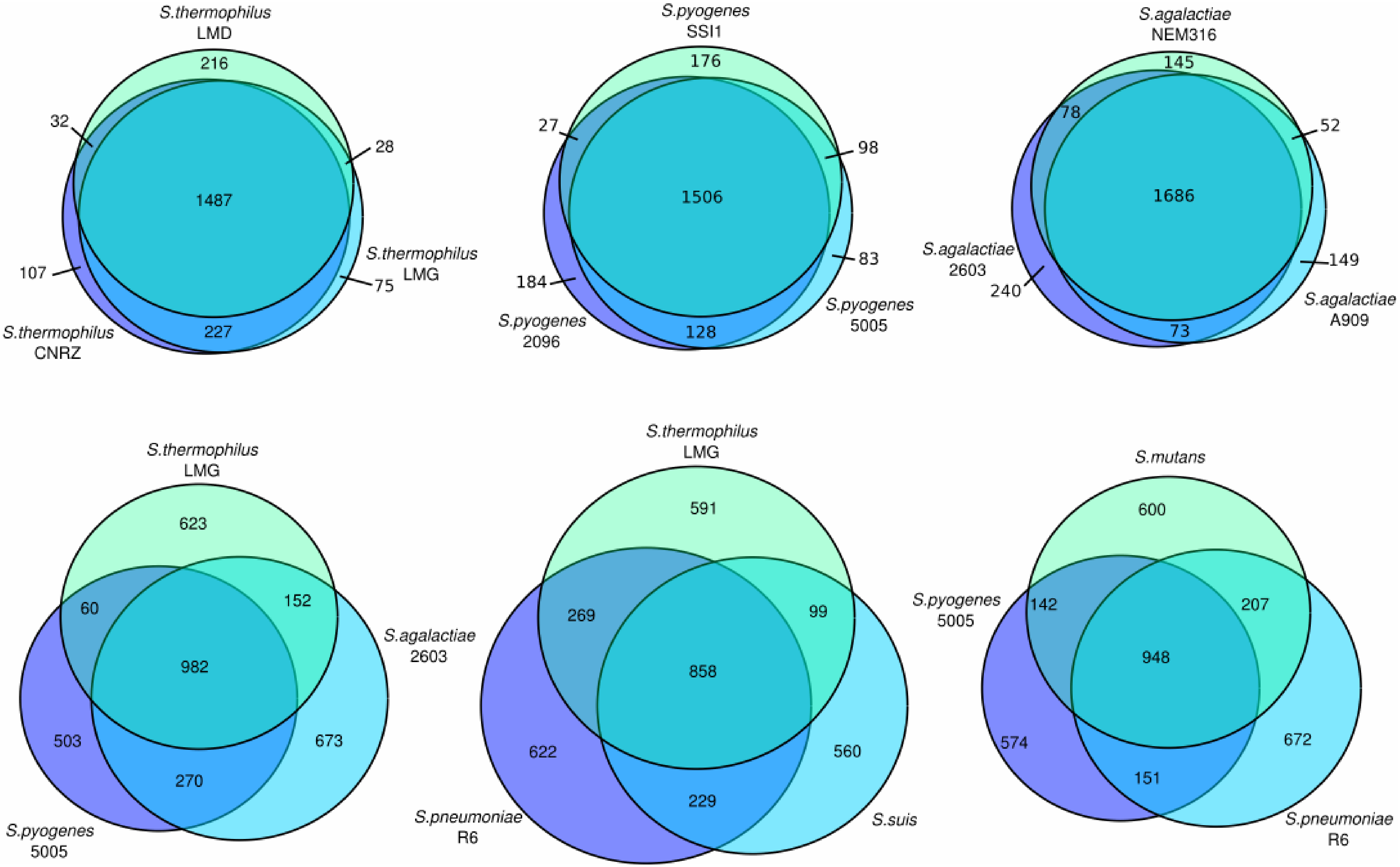
```
GACGATATCAAATCTTTATTGGCACGCCAAGTTATGGAACCTGTACGCTTTGATGAGTCT
GACGATATCAAATCTTTATTGGCACGCCAAGTTATGGAACCTGTACGCTTTGATGAGTCT
CAGGAGATTAAACCTTTGCTTGCTCGTCAAGTCAAAGAACCTGTTCGCTTCTATGAATCT
GAGGACATTGCTCAGCTCTTGACGCGTCAGGTCAAGGAACCCGTTCGTTTCTATGAAAGT
GAGGACATTGCTCAGCTCTTGACGCGTCAGGTCAAGGAACCCGTTCGTTTCTATGAAAGT
GACCGTATCCCAGAACTATTAGCCCGTCAAGTCATGGAGCCTGTTCGTTTTTATGACAGT
GACCGTATCCCAGAACTATTAGCCCGTCAAGTCATGGAGCCTGTTCGTTTTTATGACAGT
GAGGACATCAAGTCCCTCTTGACCCGTCAAGTGATGGAACCGGTTCGTTTCTATGAGTCA
GAAGACGTTAAAGCTCTTTTGGCTCGTCAGGTGATGGAACCTGTACGTTTTTATGATTCG
GAAGACGTTAAAGCTCTTTTGGCTCGTCAGGTGATGGAACCTGTACGTTTTTATGATTCG
```

gene #2

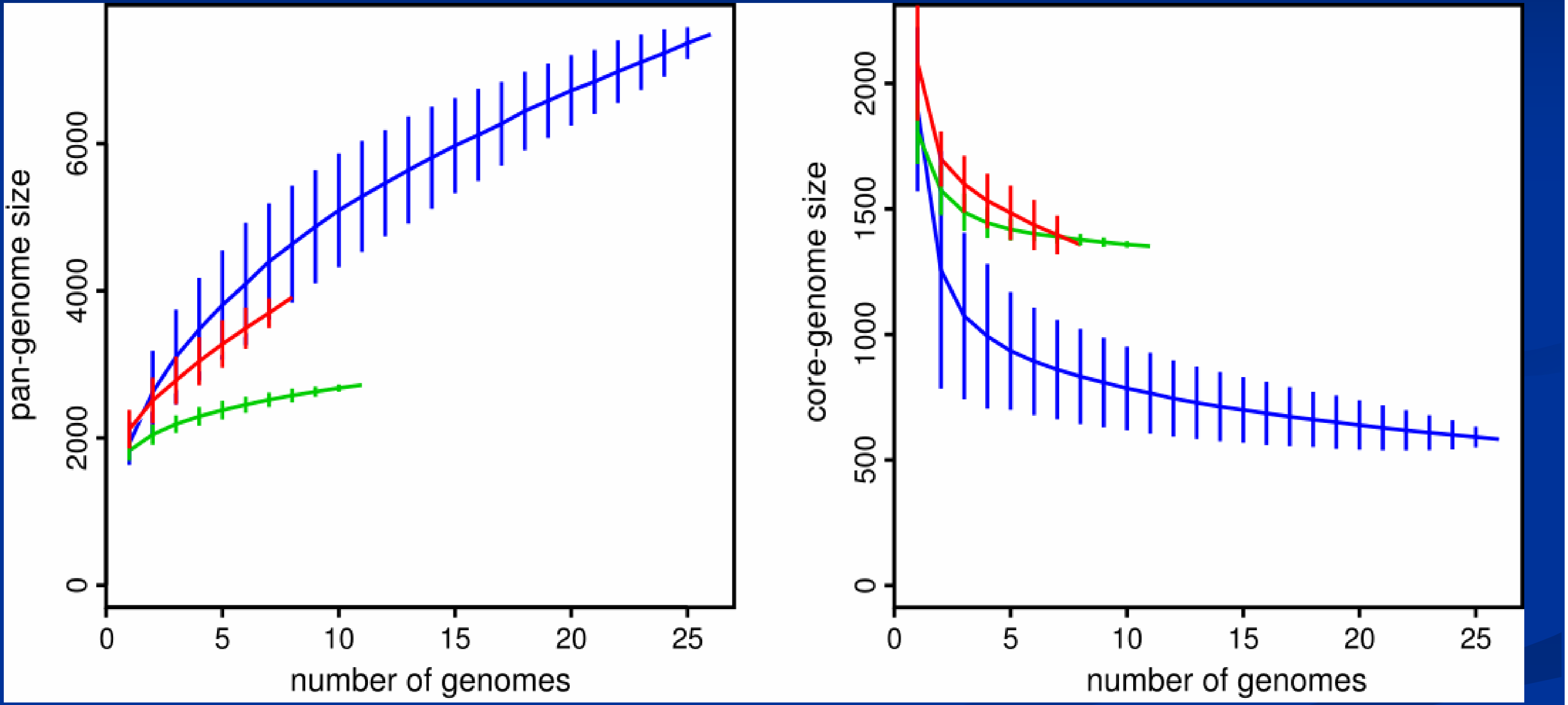
```
ATTGAAACCATGAAAAAAATGGGAATGACACAGGTTGTTGAAATTGGTCCAGGGAAGGTA
ATTGAAACCATGAAAAAAATGGGAATGACACAGGTTGTTGAAATTGGTCCAGGGAAGGTA
ATTGCTGTAATGCAAAAAATTTGGTGTGAATAACTATGTCGAAATTGGACCTGGTAAGGTC
ATTGGGGTCATGCAAGAAGCAGGCATAAGCAACTTTATCGAGATTGGACCGGGGAAAGTC
ATTGGGGTCATGCAAGAAGCAGGCATAAGCAACTTTATCGAGATTGGACCGGGGAAAGTC
GTTGCGACTTTAGTAGAAAGTGGCATAACACAATTCATTGAGGTAGGACCAGGTAAGGTT
GTTGCGACTTTAGTAGAAAGTGGCATAACACAATTCATTGAGGTAGGACCAGGTAAGGTT
ATCGCAACCATGCAGGAAGCTGGCGTGACTAACTTTATTGAAATCGGTCCAGGTAAAGTC
ATTGCCACTATCCAAGATTTTGGTGTGGATGAAGTGATTGAGATTGGGCCAGGAAAAATC
ATTGCCACTATCCAAGATTTTGGTGTAGATGAAGTGATTGAGATTGGGCCAGGAAAAATC
```

Pipeline (part 2)


Intragenic recombination detection
Alignment without recombination
recombination breakpoint
gene #1 fragment #1
gene #1 fragment #2
gene #2
...
Phylogenetic reconstructions
gene #1 fragment #1
gene #1 fragment #2
gene #2
...
Positive selection detection
Genes, lineages and sites showing evidence of evolution by positive selection

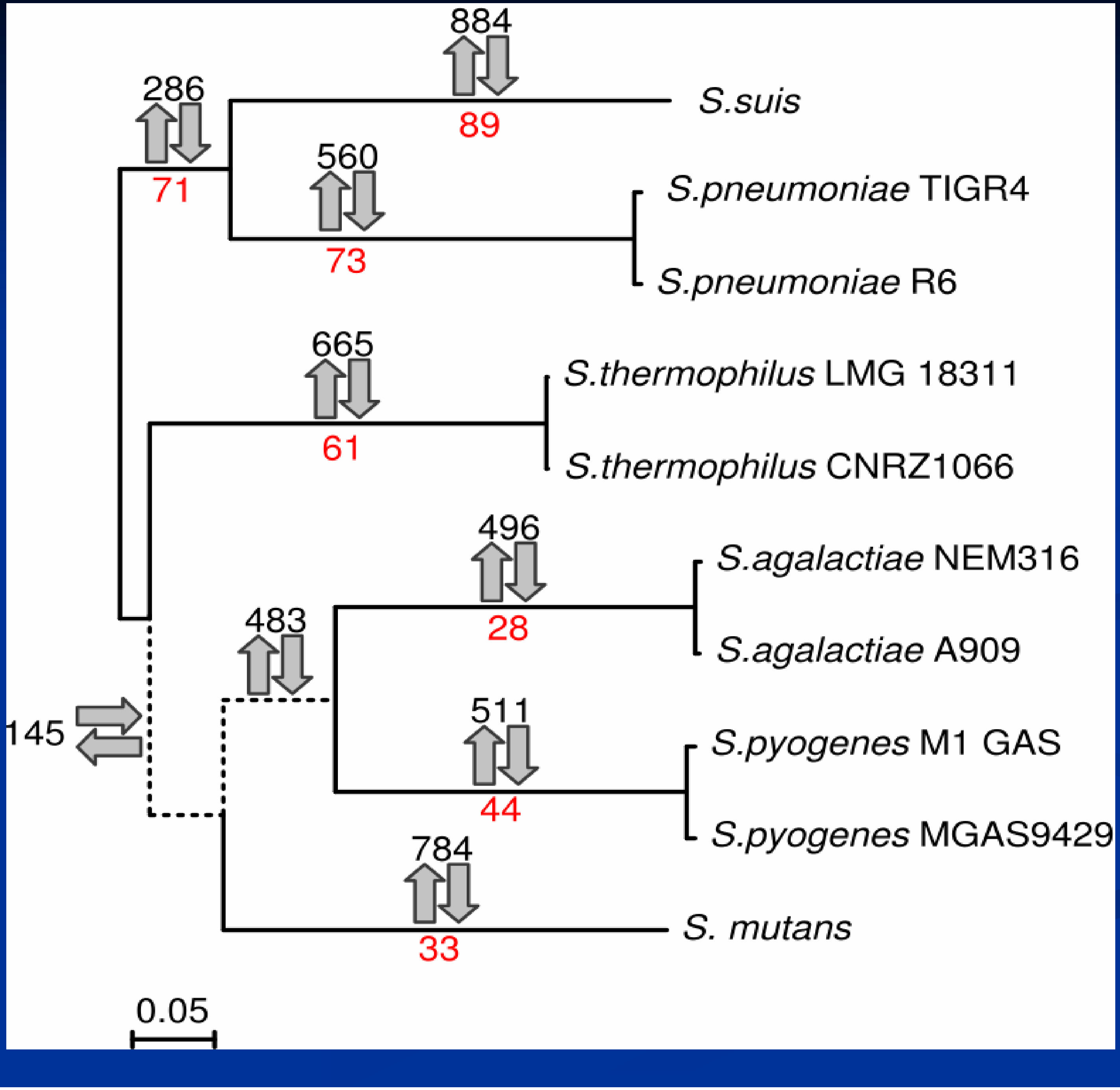

S.thermophilus LMD
216
32
28
1487
S.thermophilus LMG
107
75
227
S.thermophilus CNRZ
S.pyogenes SSI1
176
27
98
1506
83
184
128
S.pyogenes 2096
S.pyogenes 5005
S.agalactiae NEM316
145
78
52
1686
S.agalactiae 2603
149
240
73
S.agalactiae A909
S.thermophilus LMG
623
60
152
982
S.agalactiae 2603
503
673
270
S.pyogenes 5005
S.thermophilus LMG
591
269
99
858
622
560
229
S.pneumoniae R6
S.suis
S.mutans
600
S.pyogenes 5005
142
207
948
574
672
151
S.pneumoniae R6

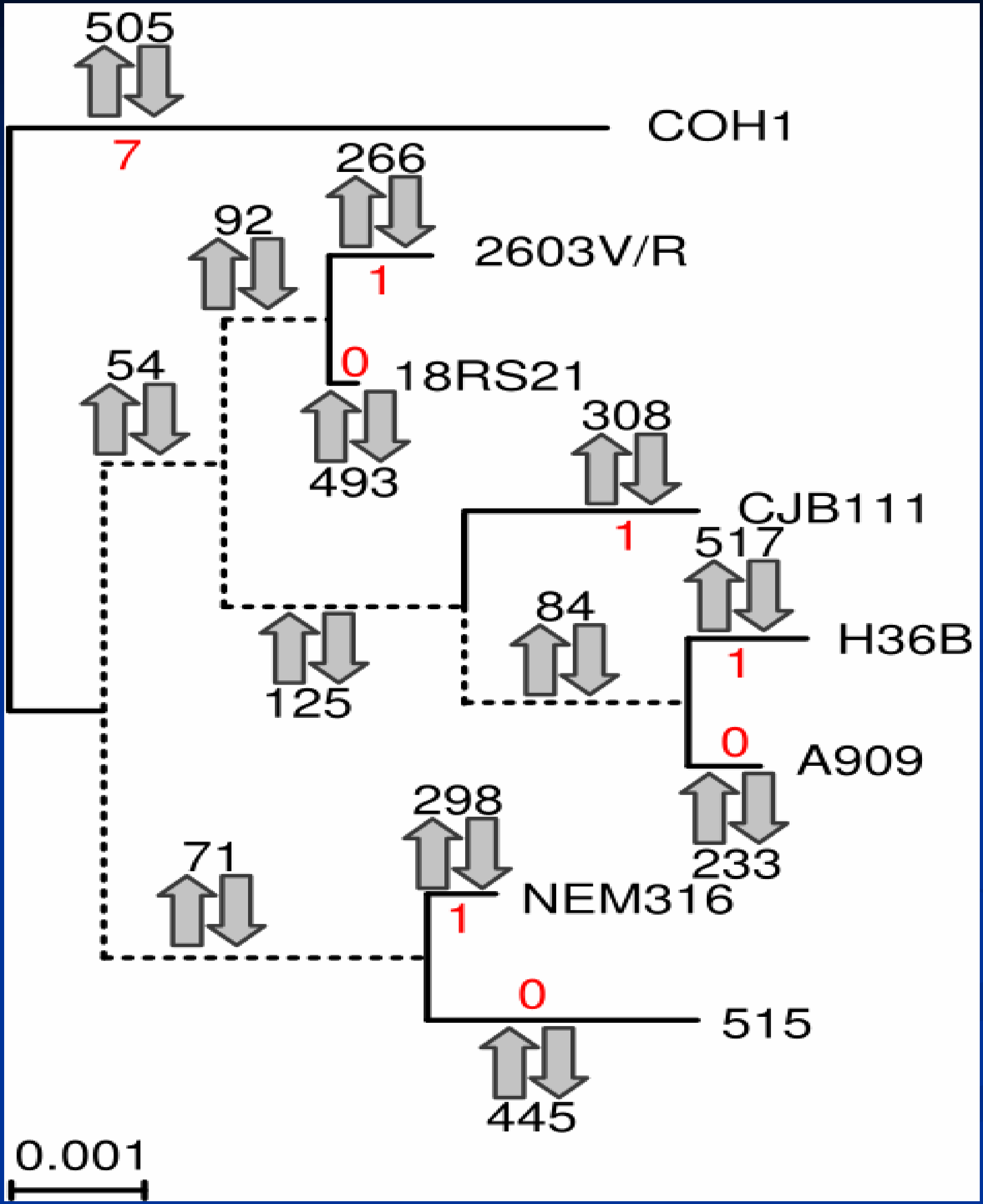
Estimated pan & core genome sizes

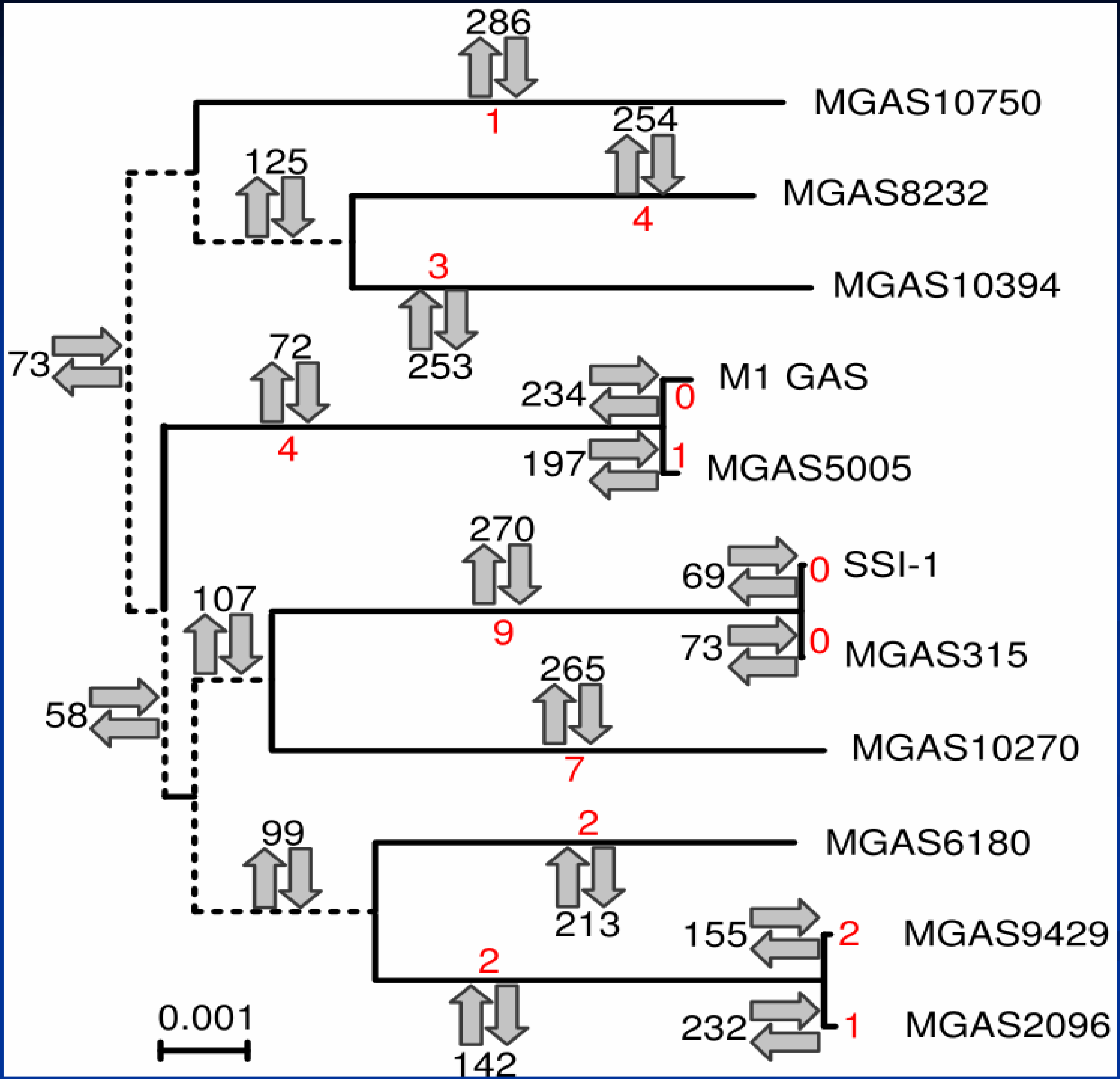


S. agalactiae
S. pyogenes

Alignable core genome
size for interspecific analysis = 260


884
89
S.suis
286
71
560
73
S.pneumoniae TIGR4
S.pneumoniae R6
665
61
S.thermophilus LMG 18311
S.thermophilus CNRZ1066
496
28
S.agalactiae NEM316
S.agalactiae A909
483
145
511
44
S.pyogenes M1 GAS
S.pyogenes MGAS9429
784
33
S. mutans
0.05


505
COH1
7
266
2603V/R
1
92
0
18RS21
54
493
308
CJB111
1
517
84
H36B
1
125
0
A909
233
298
71
NEM316
1
0
515
445
0.001


286
1
MGAS10750
254
125
MGAS8232
4
3
MGAS10394
253
73
72
234
M1 GAS
0
4
197
1
MGAS5005
270
69
0
SSI-1
107
9
73
0
MGAS315
265
58
MGAS10270
7
99
2
MGAS6180
213
155
2
MGAS9429
2
0.001
232
1
MGAS2096
142

Genes under positive selection: between species

Lineage	nbr analyzed	nbr under PS	% under PS
S. mutans	260	33	12.69
S. pneumoniae	260	73	28.08
S. suis	260	**89**	**34.23**
S. thermophilus	260	61	23.46
S. agalactiae	260	28	10.77
S. pyogenes	260	44	16.92
(S. pneumoniae, S. suis)	221	71	32.13

Genes under positive selection: *S. agalactiae*

Lineage	nbr analyzed	nbr under PS	% under PS
COH1	1212	7	0.58
18RS21	1212	0	0.00
NEM316	1212	1	0.08
H36B	1212	1	0.08
A909	1212	0	0.00
2603V/R	1212	1	0.08
CJB111	1212	1	0.08
515	1212	0	0.00

Genes under positive selection: *S. pyogenes*

Lineage	nbr analyzed	nbr under PS	% under PS
MGAS10270	1297	7	0.54
MGAS10394	1297	3	0.23
MGAS10750	1297	1	0.08
MGAS2096	1297	1	0.08
MGAS315	1297	0	0.00
MGAS5005	1297	1	0.08
MGAS6180	1297	2	0.15
MGAS8232	1297	4	0.31
MGAS9429	1297	2	0.15
M1 GAS	1297	0	0.00
SSI-1	1297	0	0.00
(MGAS9429, MGAS2096)	925	2	0.22
(MGAS5005, M1 GAS)	978	4	0.41
(SSI-1, MGAS315)	983	**9**	0.92

Recombination

Data set	SPI (strong phylogenetic incongruence)	PHI (intragenic method)	PHI ∩ MaxChi ∩ NSS (set of intragenic methods)	SPI ∩ PHI	SPI U intragenic set
interspecific	26 (10%)	54 (21%)	35 (14%)	11 (4%)	53 (20%)
S. pyogenes	**434 (33%)**	284 (22%)	168 (13%)	186 (14%)	477 (37%)
S. agalactiae	222 (18%)	34 (3%)	7 (1%)	18 (1%)	223 (18%)

Recombination and positive selection

Data set	Genes under PS	PS + recombinant	PS + SPI	PS + intragenic
interspecific	175	43 (25%)	20 (8%)	29 (11%)
S. agalactiae	10	4 (40%)	4 (40%)	0
S. pyogenes	32	25 (78%)	21 (65%)	17 (53%)

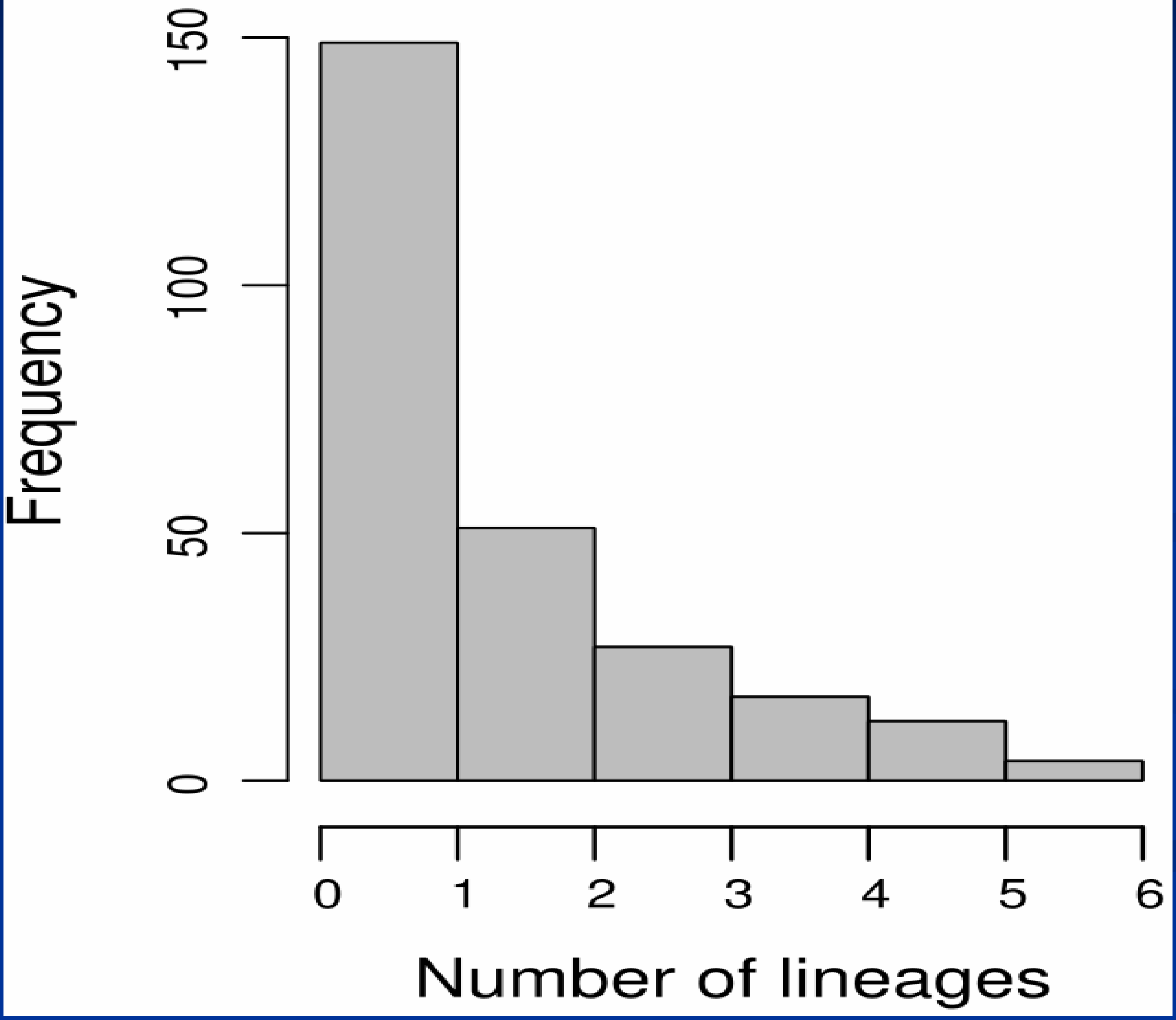
Pan genome and recombination

- Habitat differences for *S. pyogenes* and *S. agalactiae*
  - Reduced gene pool environment for *S. pyogenes*, could result in smaller pan genome and potentially more homologous recombination

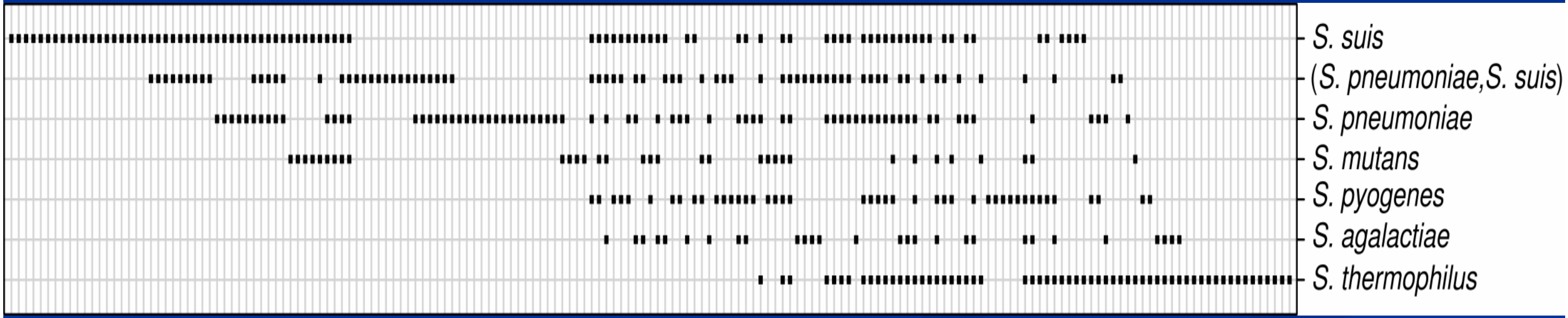
Statistical analysis of PS data

- Significant affect of lineage (ANOVA; p<0.0001):
  - Majority of pairwise multiple comparisons significantly different
- Significant affect of biochemical category (p<0.0001)
  - Amino acid biosynthesis; Biosynthesis of cofactors, prosthetic groups, and carriers; Cell envelope; Cellular processes; Central intermediary metabolism; **DNA metabolism**; Energy metabolism; Fatty acid and phospholipid metabolism; Hypothetical proteins; Protein fate; Protein synthesis; Purines, pyrimidines, nucleosides, and nucleotides; Regulatory functions; Signal transduction; **Transcription**; Transport and binding proteins; Unknown function
- Significant interaction between lineage and biochemical category (p=0.003)
  - (*S. pneumoniae, S. suis*) DNA metabolism, Transcription, Protein fate

Genes selected per lineage



Genes selected on each lineage



19 unique loci for *S. suis*; 15 for *S. thermophilus*; 14 for *S. pneumoniae*

Lineages with unusual selection pressure

- *S. suis*
  - Both gene gain and loss and PS; suggesting evolutionary flexibility – host jumping?
- *S. agalactiae*, COH1
  - Significantly associated with neonatal disease, and of recent bovine ancestry
- *S. pyogenes*, M3 serotype
  - M3 cause more cases of invasive disease, higher rate of lethal infections, epidemic tendencies
- *S. thermophilus*, LMD-9
  - ?

Streptococcus comparative genomics tentative conclusions

- Considerable recombination and positive selection pressure in *Streptococcus* core genome
- Several loci identified for *S. agalactiae* and *S. pyogenes* that could be linked to the specific pathogenic features of these strains
- Identification and cataloguing of these loci, serve as an evolutionary short-cut for laboratory mutation experiments, to assess specific functional significance of these genes.

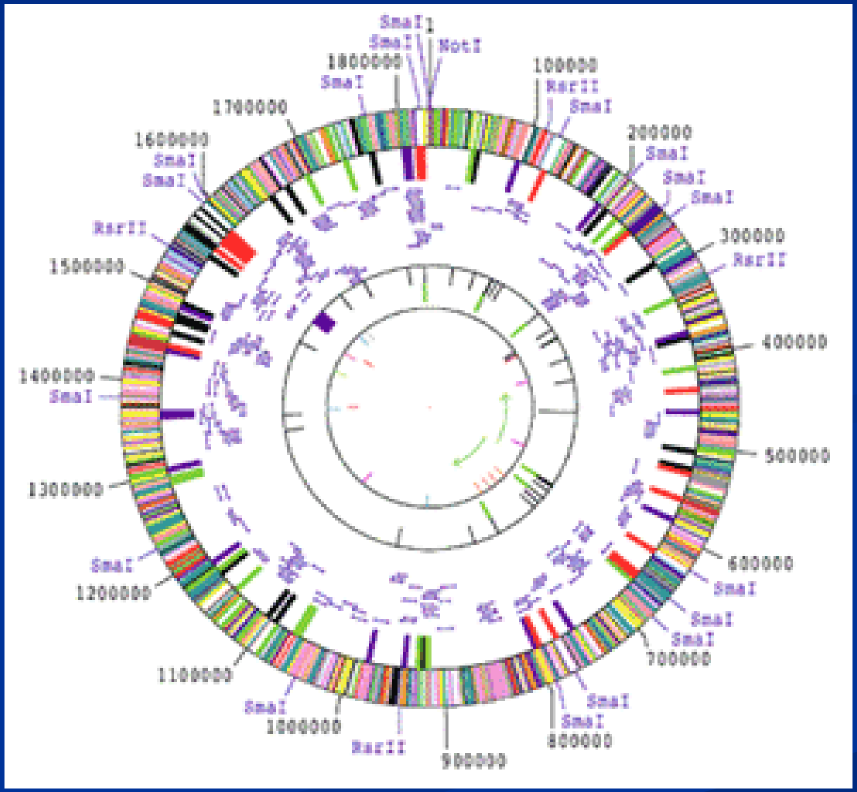
Sequencing Technology

First “shot gun” microbial genome sequence

Haemophilus influenzae 1.8 Mb

Library of plasmid clones, 1600-2000 bp fragments; sequences of these clones with their many overlaps represent the raw data entered into computer programs (e.g. TIGR assembler) which assemble the genome;

remaining gaps closed with other strategies (e.g. long range PCR)



Fleischmann et al. 1995. Science 269: 496

Race for the $1000 genome

- First to produce $1000 human genome
  - J. Craig Venter Science Foundation: $500,000
  - X Prize Foundation: $5 million
- 2004; NIH; $70 million grant program

Next generation of sequencers

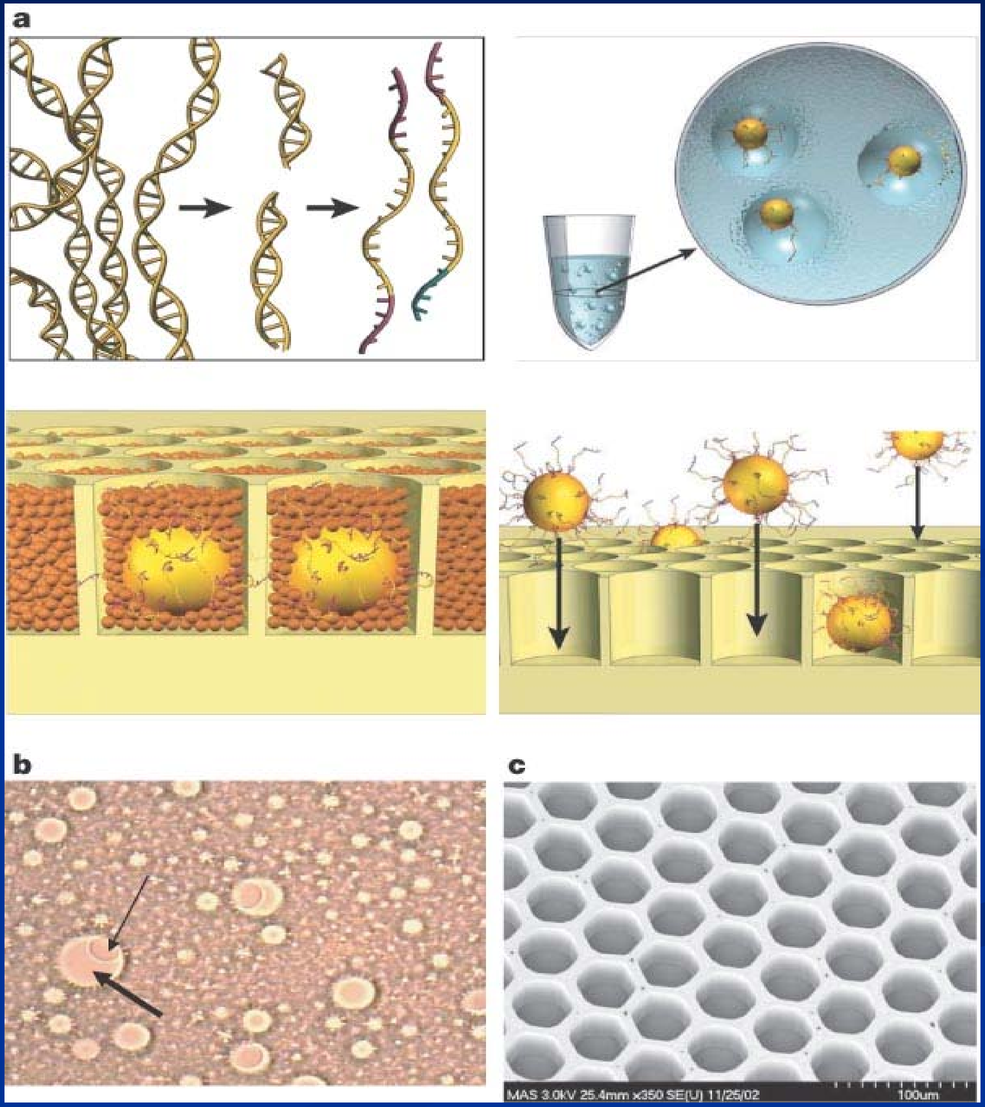
Searching for Cheaper Genome Sequencers

Company	Format	Read Length (bases)	Expected Throughput Mb (million bases)/day
454 Life Sciences	Parallel bead array	100	96
Agencourt Bioscience	Sequencing by ligation	50	200
Applied Biosystems	Capillary electrophoresis	1000	3-4
LI-COR Biosciences	Electronic microchip	20,000	14,000
Microchip Biotechnologies	Parallel bead array	850-1000	7
Network Biosystems	Biochip	800+	5
NimbleGen Systems	Map and survey microarray	30	100
Solexa	Parallel microchip	35	500
VisiGen Biotechnologies	Single-molecule array	NA	1000

from: Service, RF 2006. Science 311:1544

454 sequencing

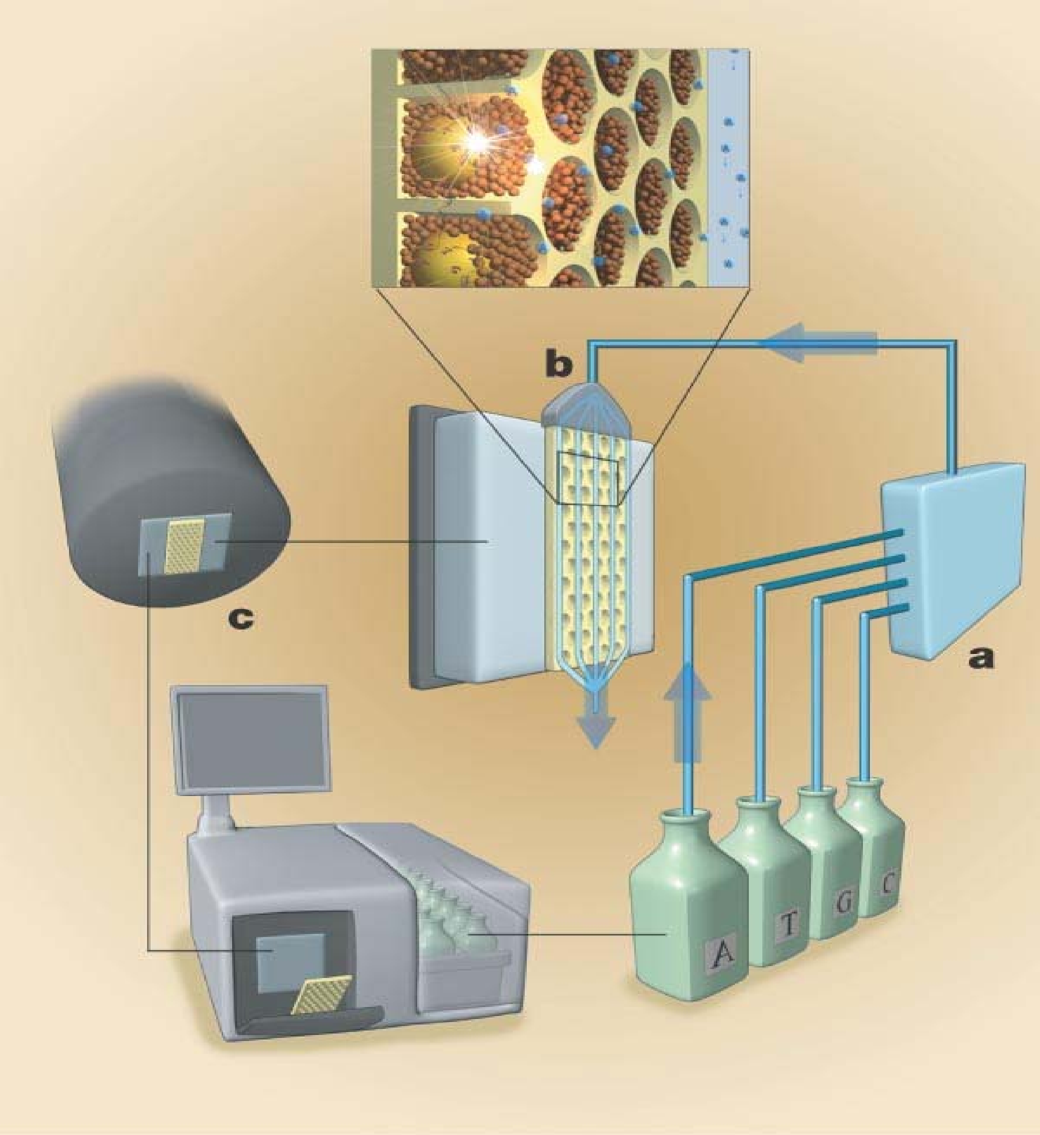
- Sequencing by synthesis (tracks bases as they are added); pyrosequencing
- 300-500 bp pieces, denatured;
- link one strand to plastic bead
- copy using emulsion PCR
- beads are separated on a fibre-optic plate containing approx. 1.6 million wells;
- add sequencing reagents



from: Margulies et al. 2005 Nature 437:376

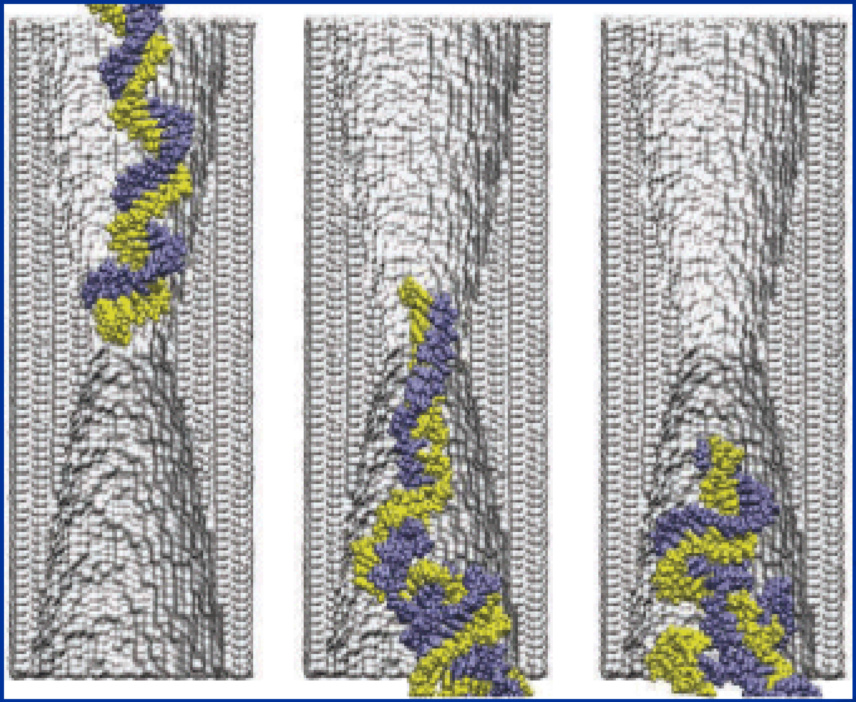
454 sequencing

- nucleotides added release pyrophophate, prompting luciferase & flash of light
- correlating flashes
from each cell with nucleotides presented in flow through, computer tracks sequence growth



from: Margulies et al. 2005 Nature 437:376

Nanopore sequencing



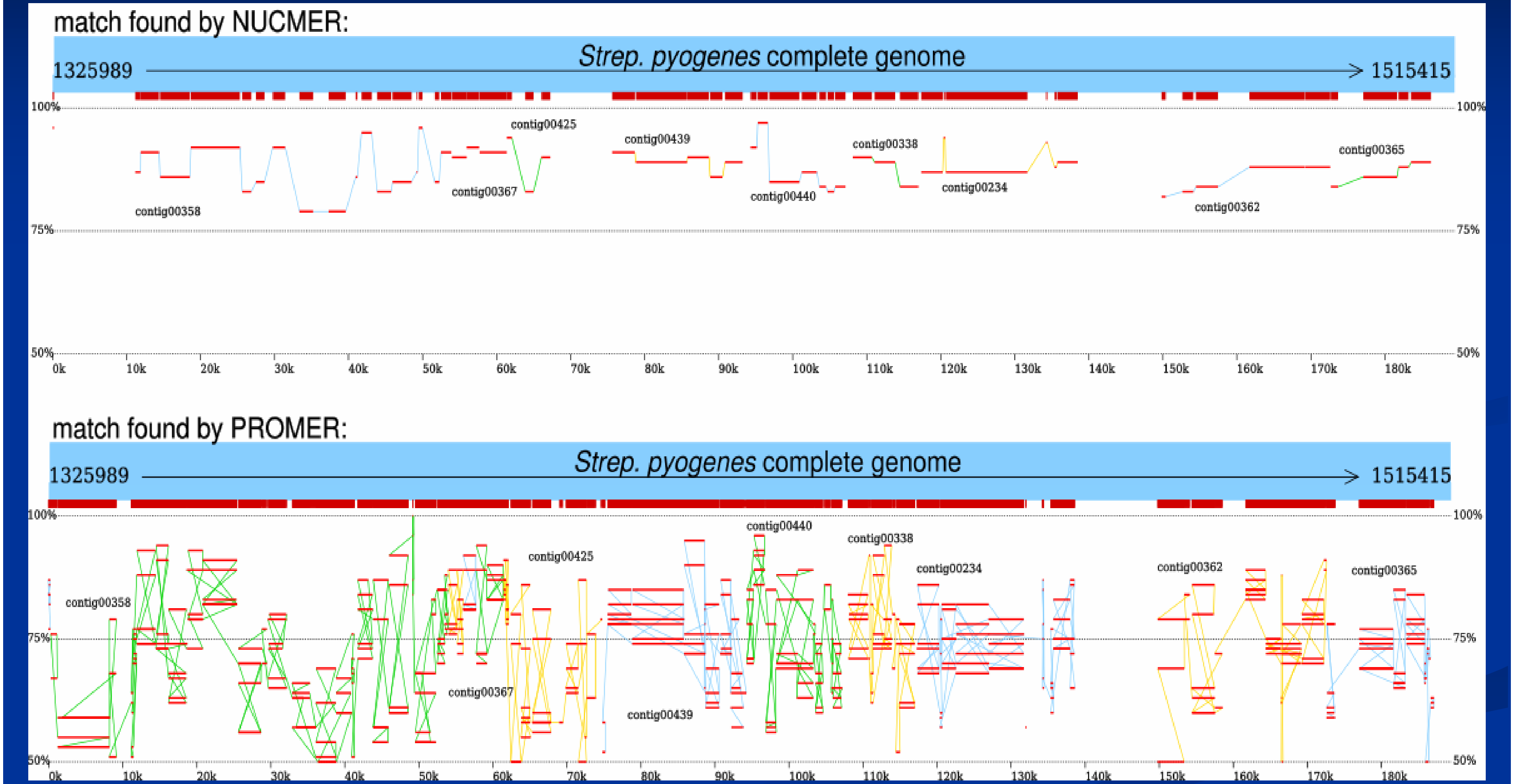
from: Service, RF 2006. Science 311:1544

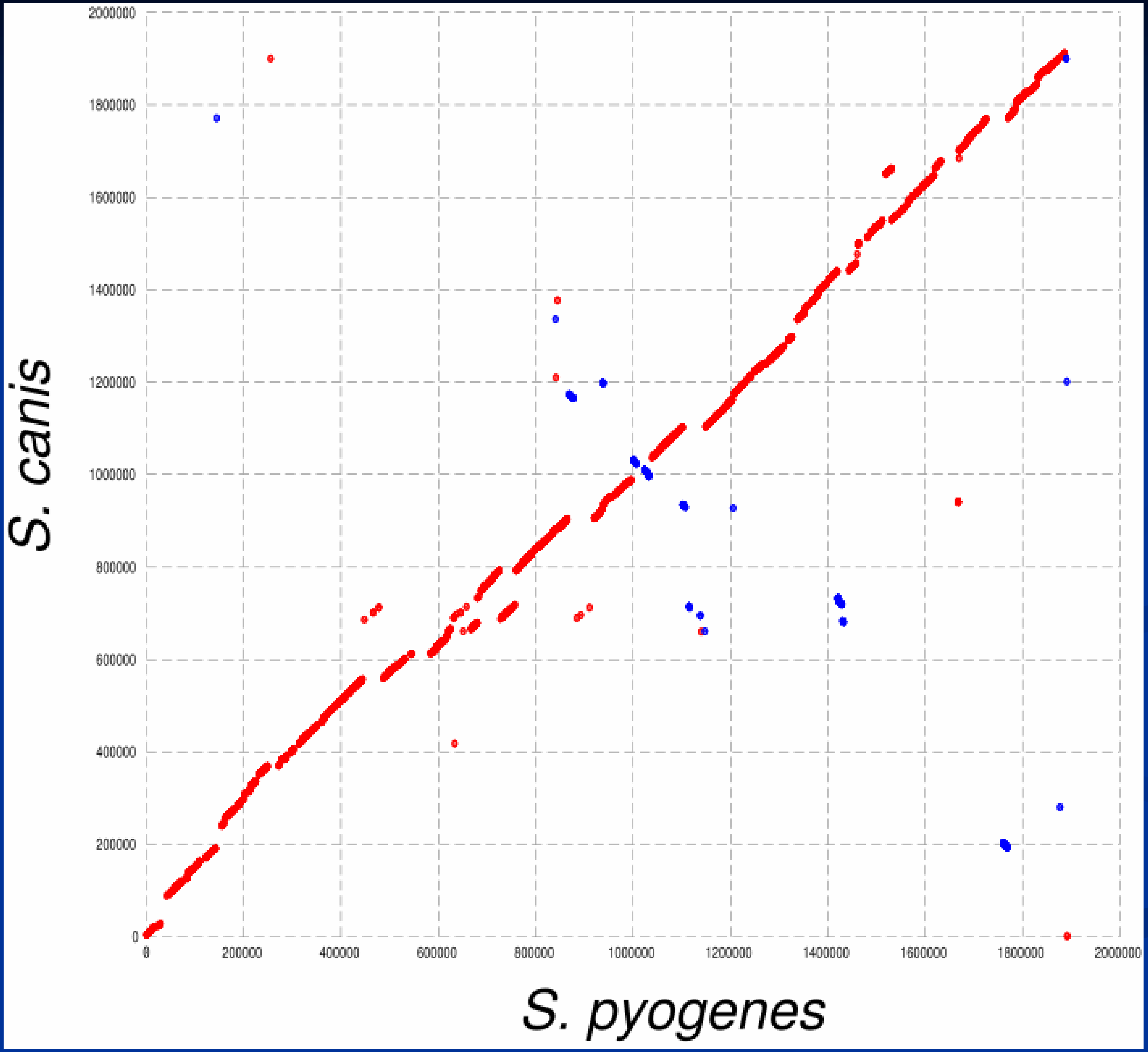
Example of 454 Bacteria Genome Sequence

Streptococcus canis

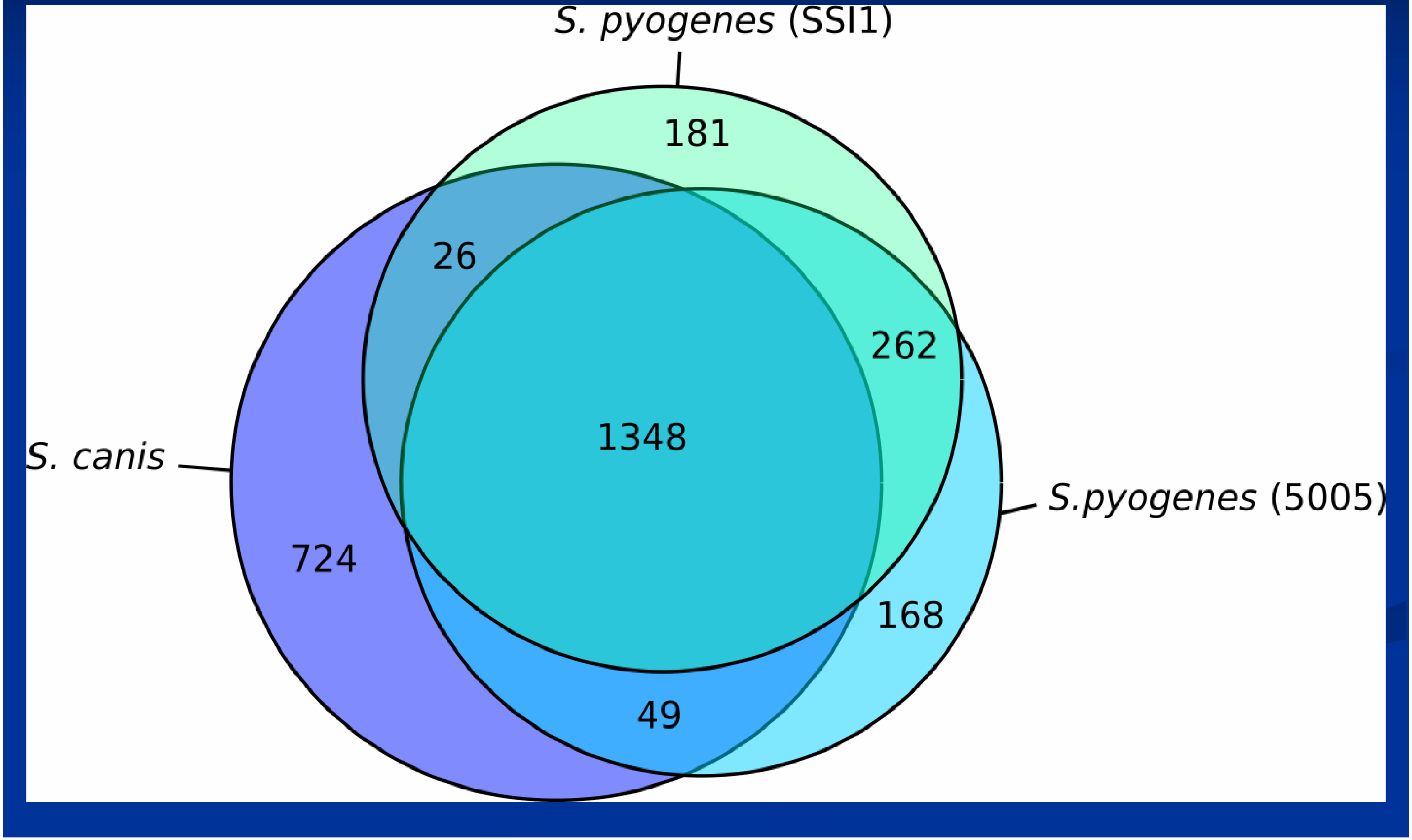
- Genome sequence data for putative sister groups to major pathogens often not available
  - e.g. *S. pyogenes*; putative sister group *S. canis*
- *S. canis* from 454 Life Sciences
  - 103 cotigs, 2,191,310 bp, 98.5% coverage, 39.6% GC
  - 100% of the bases with Q40+ rating (99.99% accuracy)

canis / pyogenes genome wide alignments




2000000
1800000
1600000
1400000
1200000
1000000
800000
600000
400000
200000
0
S. canis
0
200000
400000
600000
800000
1000000
1200000
1400000
1600000
1800000
2000000
S. pyogenes

canis / *pyogenes* genomic content comparison



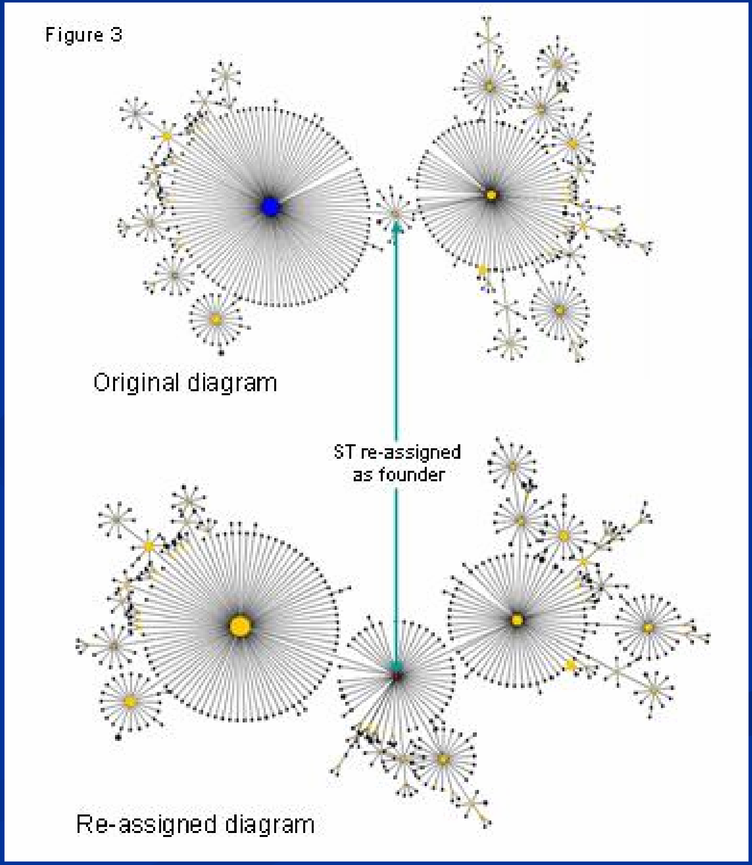
Applications of Microbial Genomics

Comparative genomics and drug discovery

- Genes need prioritization
- Drug development against a single bacterial species usually impractical
- Gene products, with orthologs in humans, may lack selectivity
    - => compare genomes, find potential drug targets shared by clinically important range of taxa, & absent or divergent from human host

Molecular Epidemiology

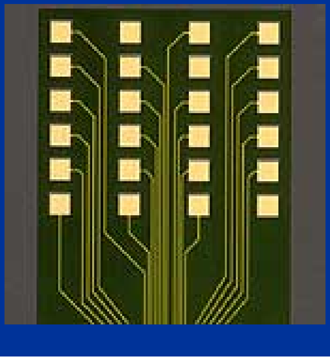
- MLST = multi locus sequence typing; sequence of portions of 7 (or more) housekeeping genes; combination of alleles = sequence type (ST); closely related STs (differ by one or two alleles) = clonal complex

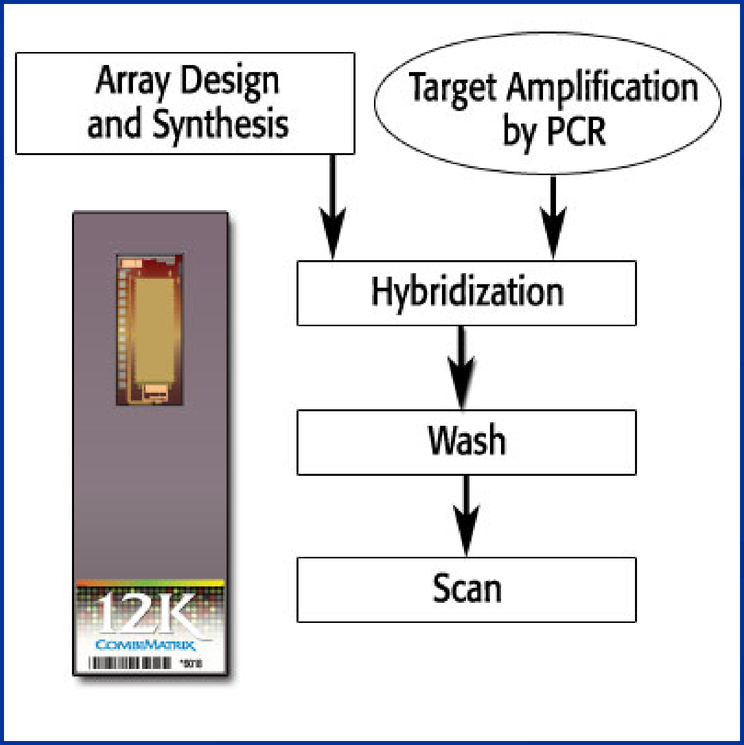


http://eburst.mlst.net/6.asp

Microarray gene / presence absence

- Genome sequence allows gene presence / absence detection across strains using microarrays
  - E.g. Combimatrix 4 X 2K microarrays





Array Design and Synthesis

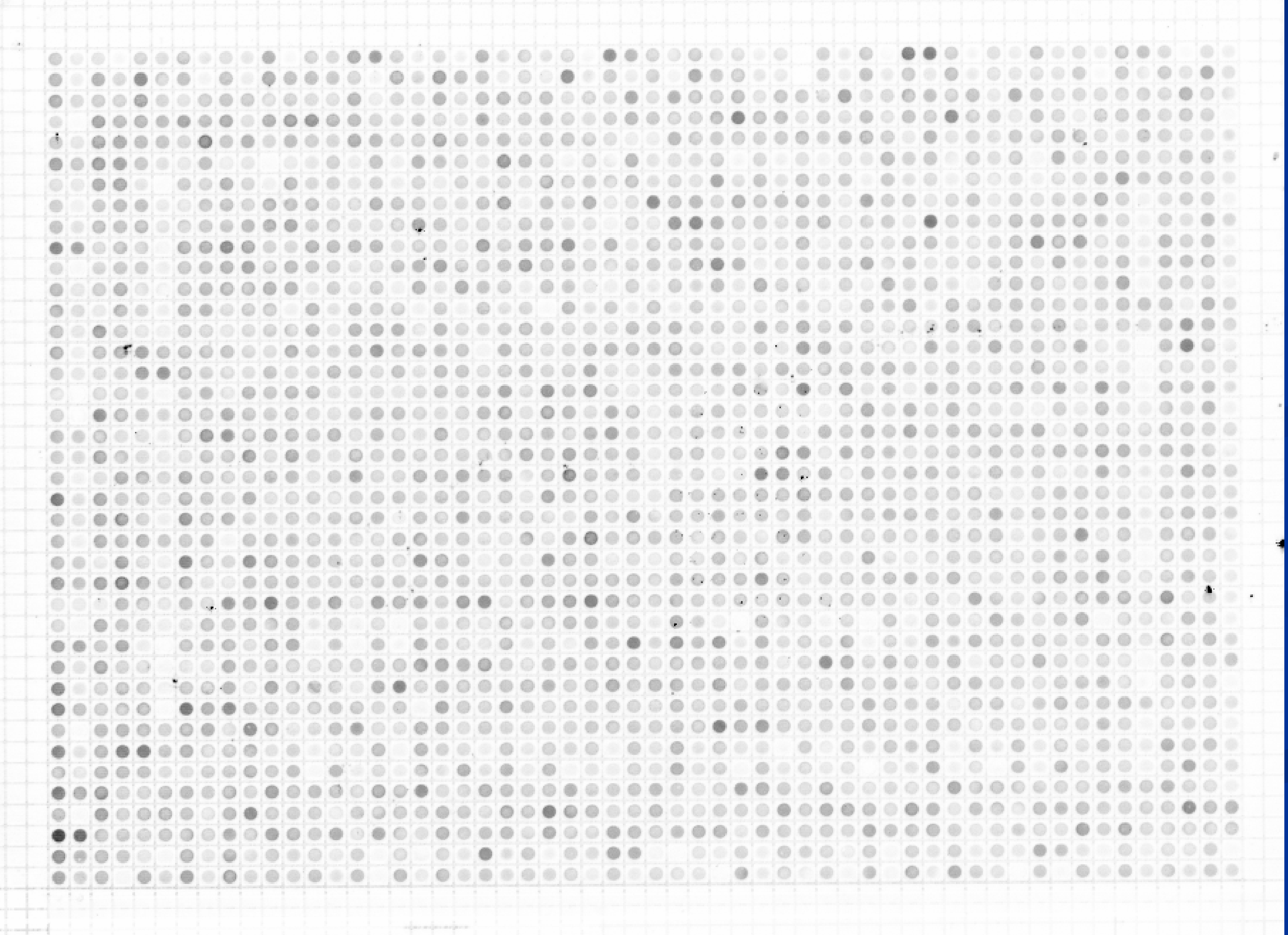
Target Amplification by PCR

Hybridization

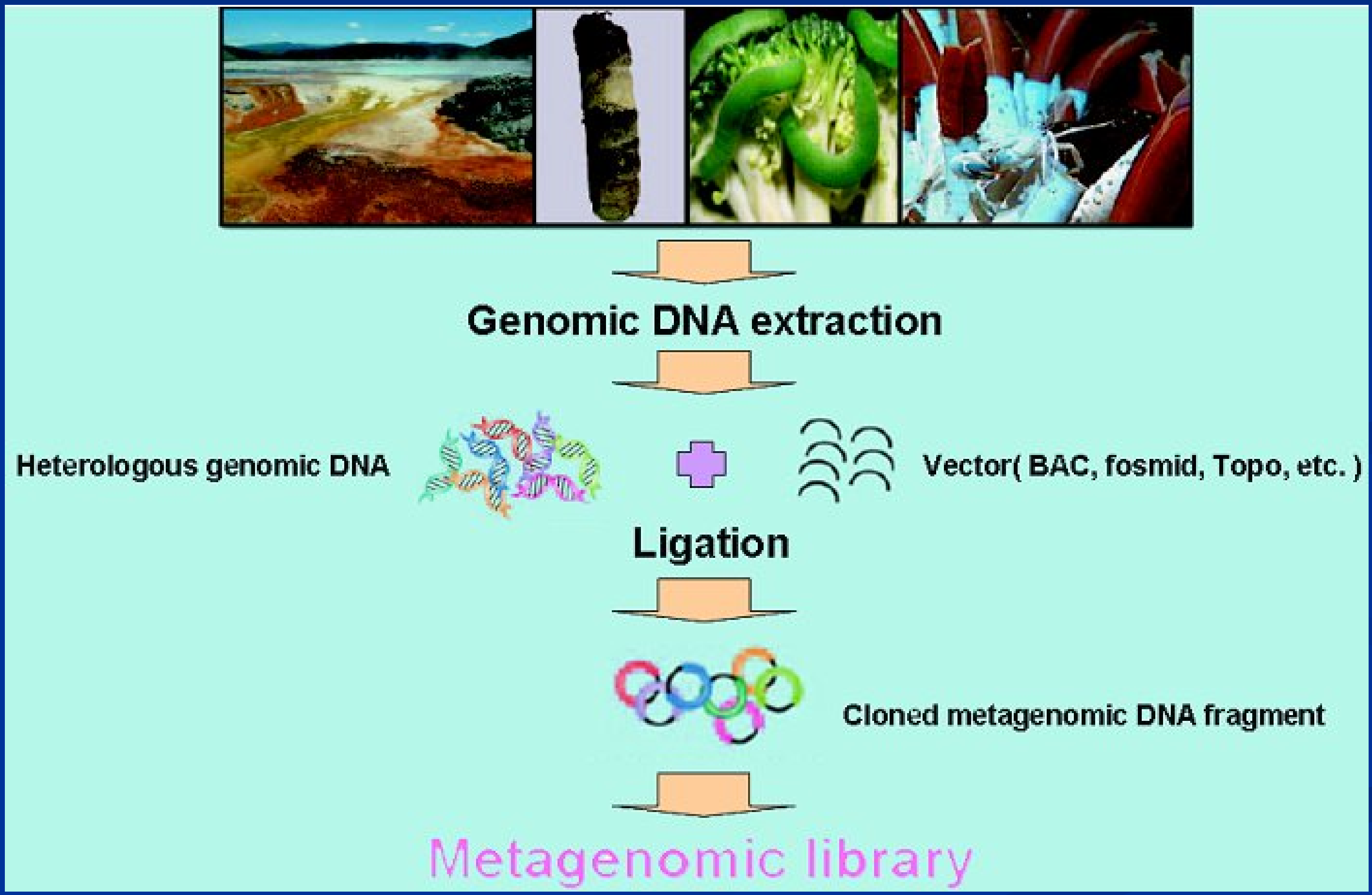
Wash

Scan

Gene / presence absence hybridization



Metagenomics



Genomic DNA extraction

Heterologous genomic DNA

Vector(BAC, fosmid, Topo, etc.)

Ligation

Cloned metagenomic DNA fragment

Metagenomic library

http://chunlab.snu.ac.kr/meta.htm