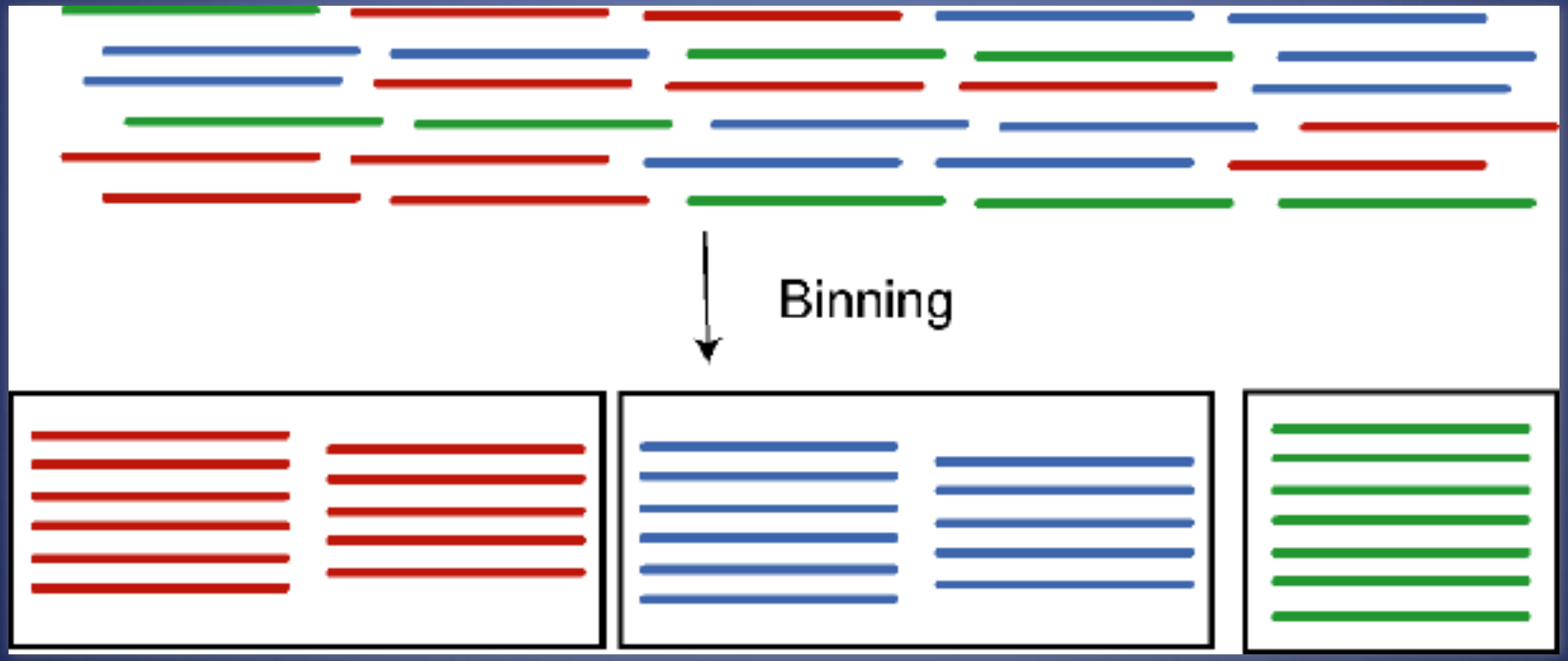
EVE 161:
Microbial Phylogenomics

Era IV:
Shotgun Metagenomics

UC Davis, Winter 2016
Instructor: Jonathan Eisen

- How do they connect “who” and “what”?

Metagenomic Binning



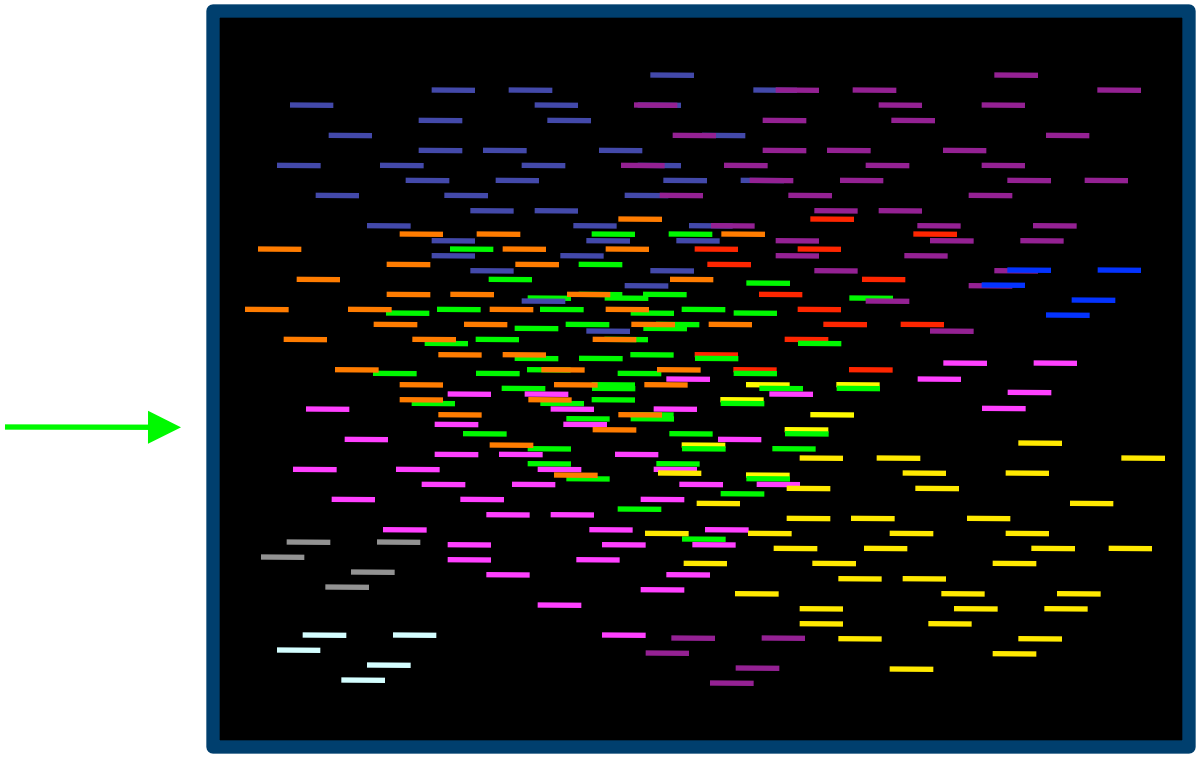
Classification of sequences by taxa

How to do Binning

- Method 1: align reads to reference genome

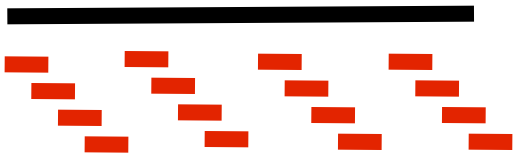
Binning challenge

A
B
C
D
E
F

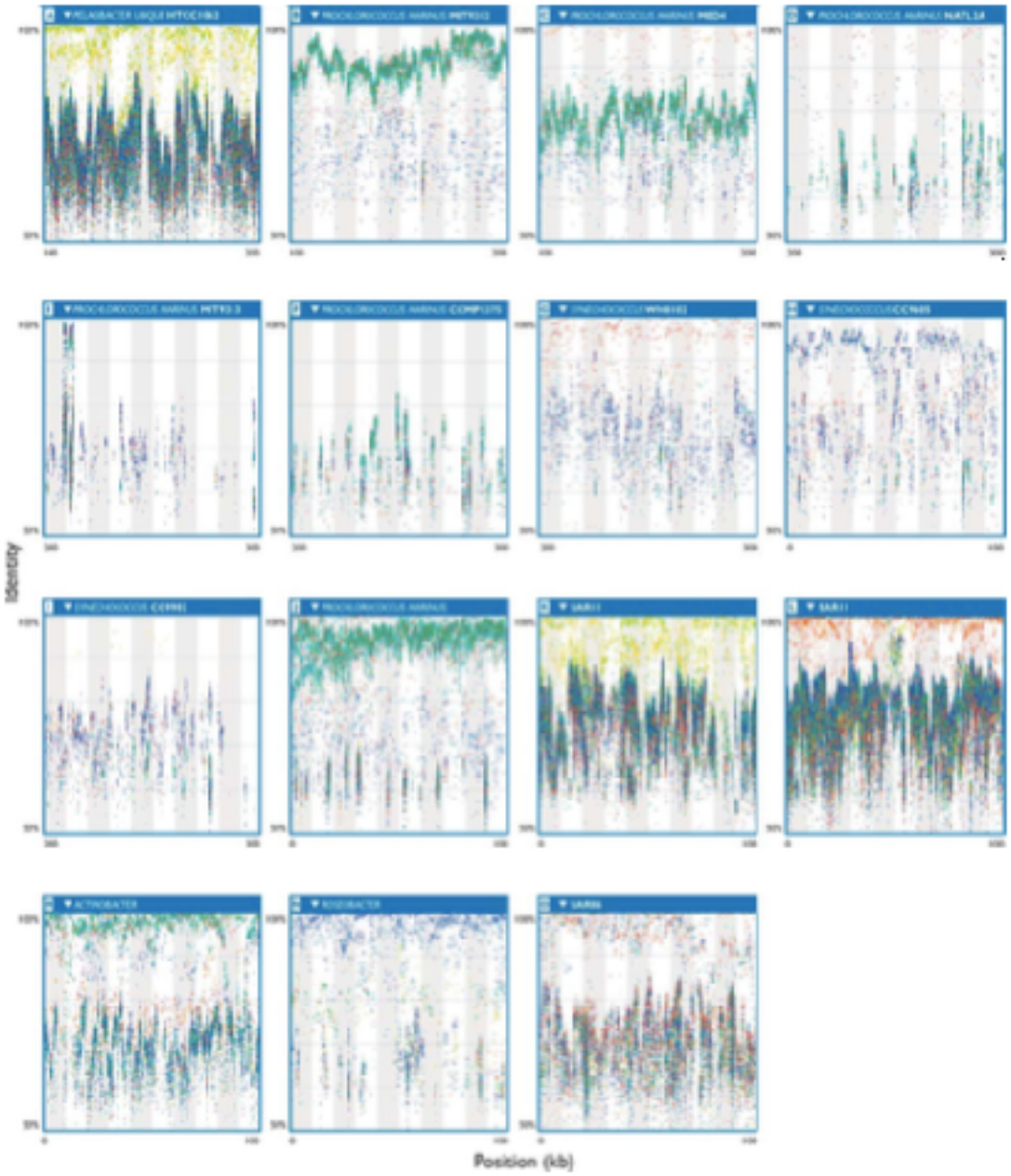


T
V
W
X
Y
Z

Best binning method: reference genomes



Fragment Recruitment



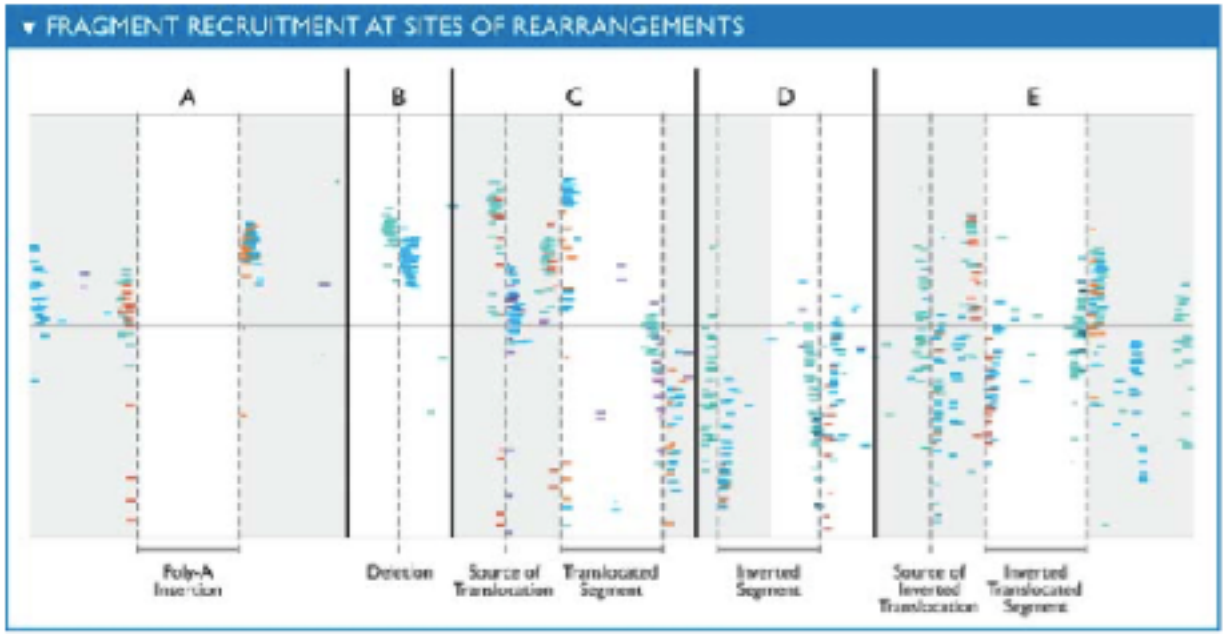


Figure 5. Fragment Recruitment at Sites of Rearrangements

Environmental sequences recruited near breaks in synteny have characteristic patterns of recruitment metadata. Indeed, each of five basic rearrangements (i.e., insertion, deletion, translocation, inversion, and inverted translocation) produced a distinct pattern when examining the recruitment metadata. Here, example recruitment plots for each type of rearrangement have been artificially generated. The "good" and "no mate" categories have been suppressed. In each case, breaks in synteny are marked by the presence of stacks of "missing" mate reads. The presence or absence of other categories distinguishes each type of rearrangement from the others.

doi:10.1371/journal.pbio.0050077.g005

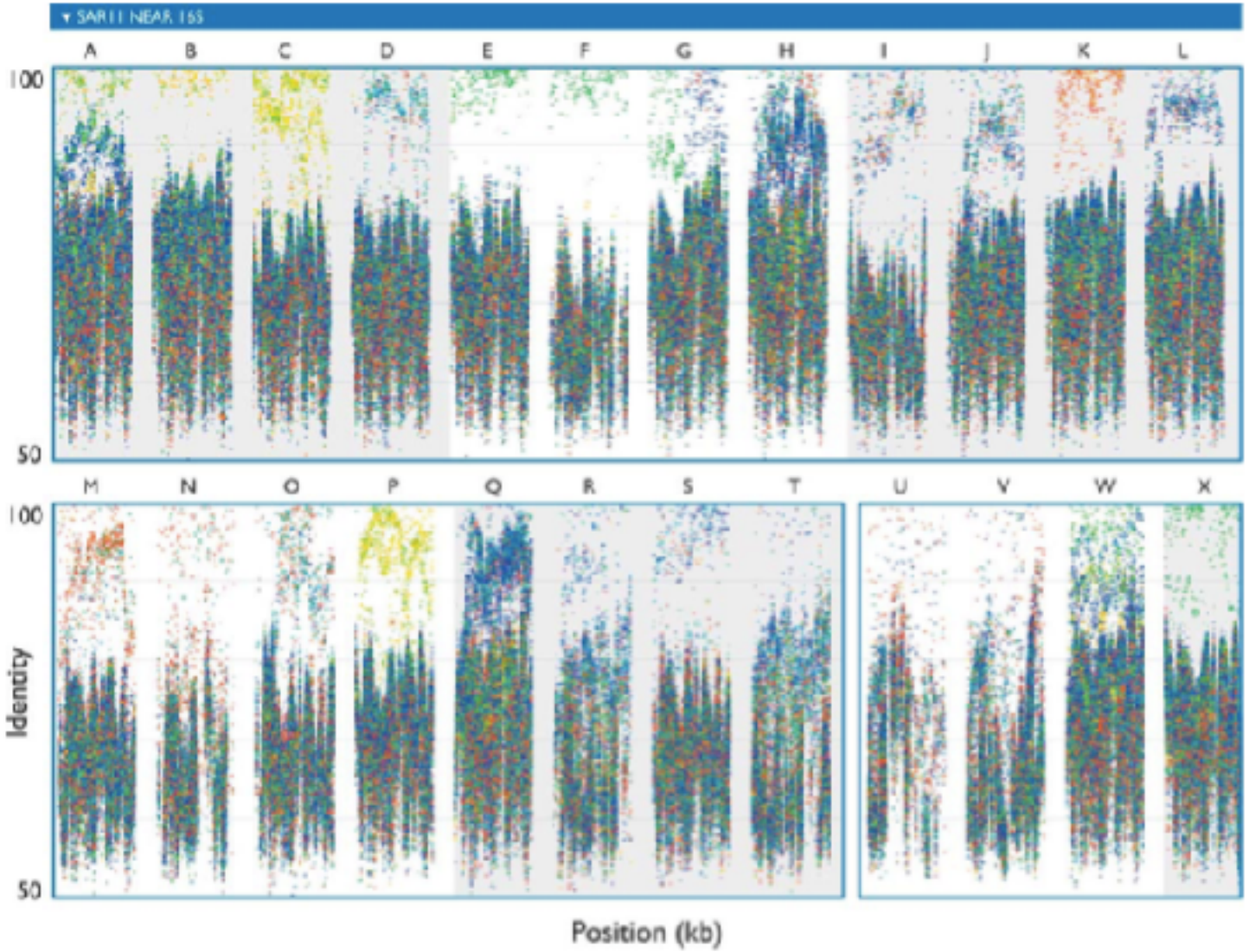


Figure 7. Fragment Recruitment Plots to 20-kb Segments of SAR11-Like Contigs Show That Many SAR11 Subtypes, with Distinct Distributions, Can Be Separated by Extreme Assembly

Each segment is constructed of a unique set of GOS sequencing reads (i.e., no read was used in more than one segment). Segments are arbitrarily labeled (A–X) for reference in Figure 8.

doi:10.1371/journal.pbio.0050077.g007

How to do Binning

- Method 1: align reads to reference genome

- How would this work?
- Advantages and disadvantages?

How to do Binning

- Method 2: Assembly

Four main challenges for *de novo* sequencing.

- Repeats.
- Low coverage.
- Errors

These introduce breaks in the construction of contigs.

- *Variation* in coverage – transcriptomes and metagenomes, as well as amplified genomic.

This challenges the assembler to distinguish between erroneous connections (e.g. repeats) and real connections.

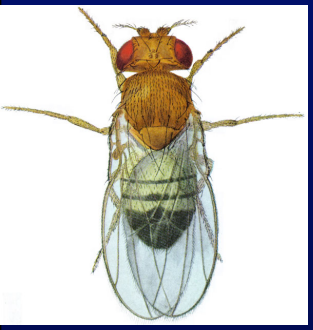
Slide from talk by Titus Brown

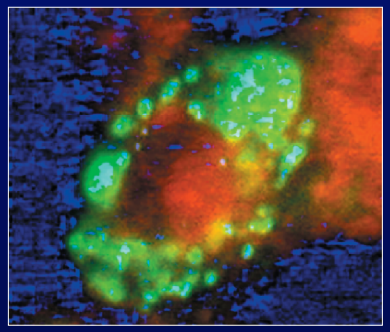
More assembly issues

- Many parameters to optimize!
- Metagenomes have variation in copy number; naïve assemblers can treat this as repetitive and eliminate it.
- Assembly requires gobs of memory (4 lanes, 60m reads => ~ 150gb RAM)
- How do we evaluate assemblies?
  - What's the best assembler?

Slide from talk by Titus Brown

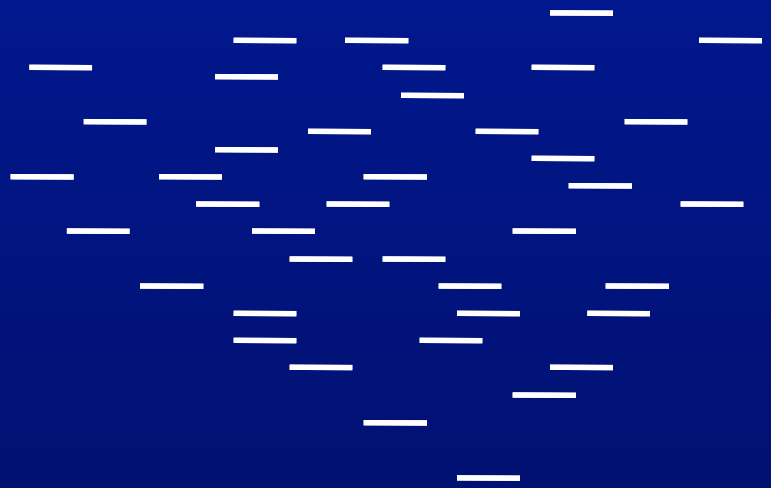
Wolbachia Metagenomic Sequencing

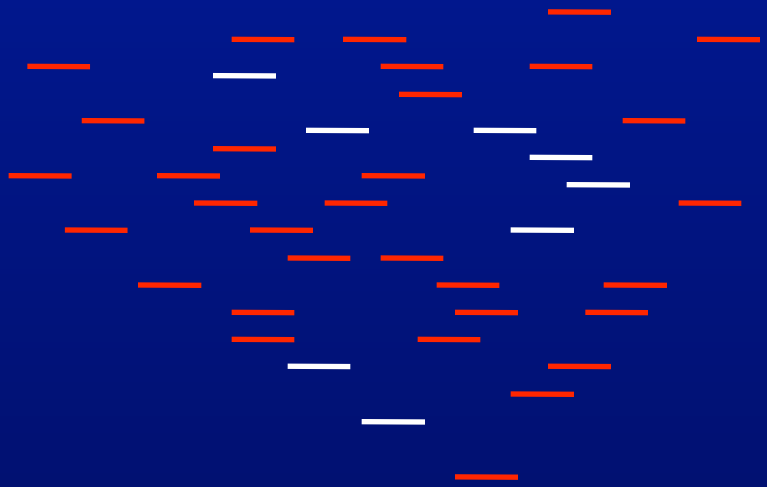


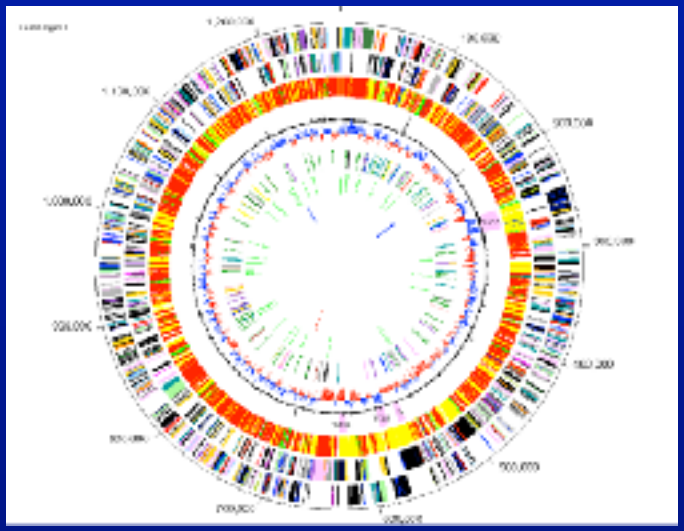


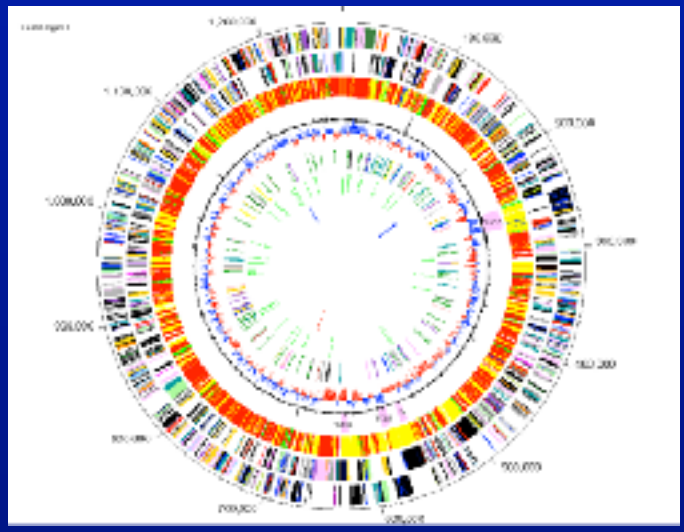
shotgun

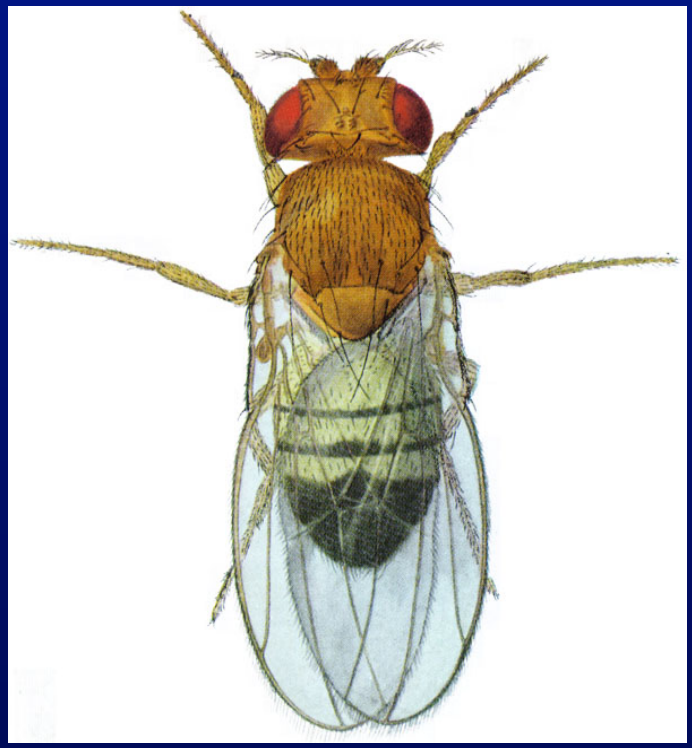
sequence



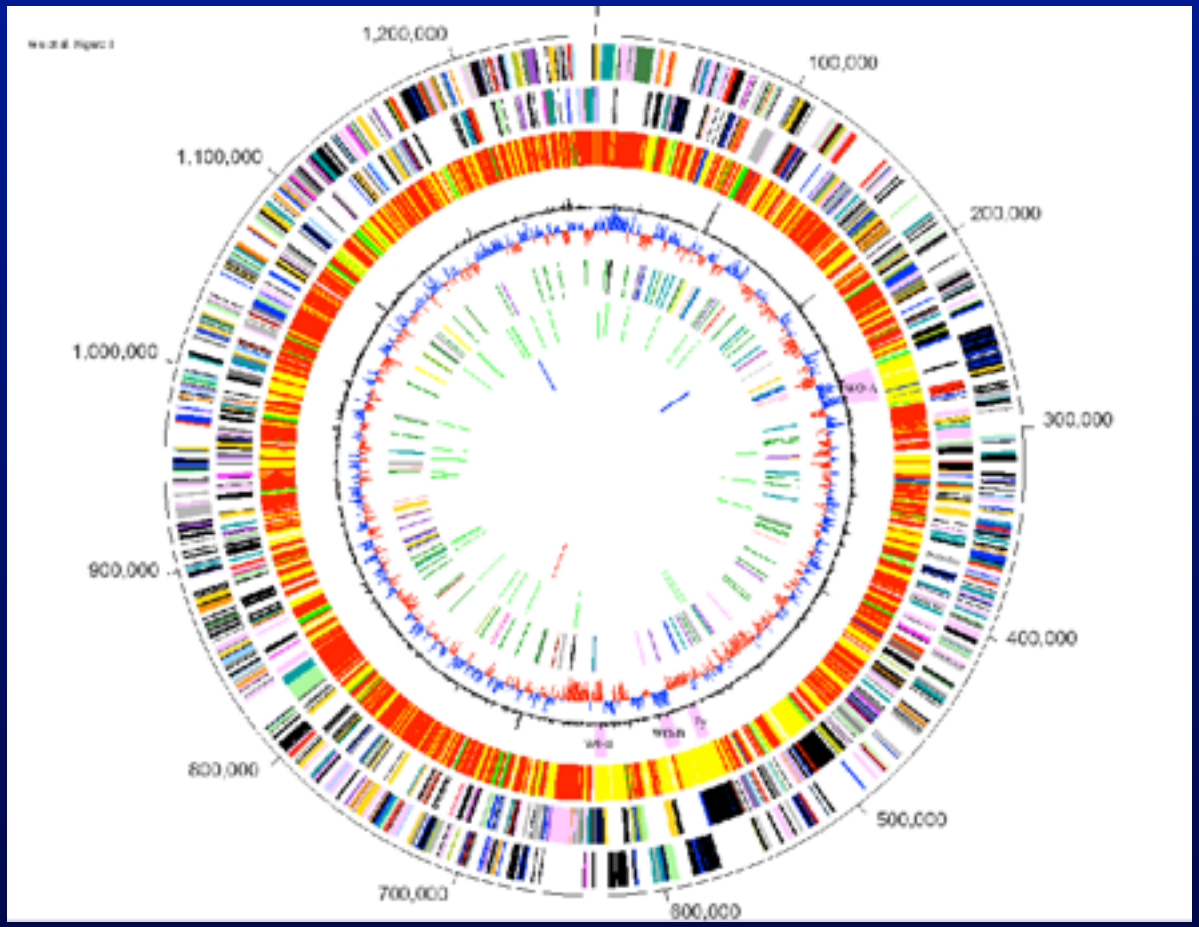








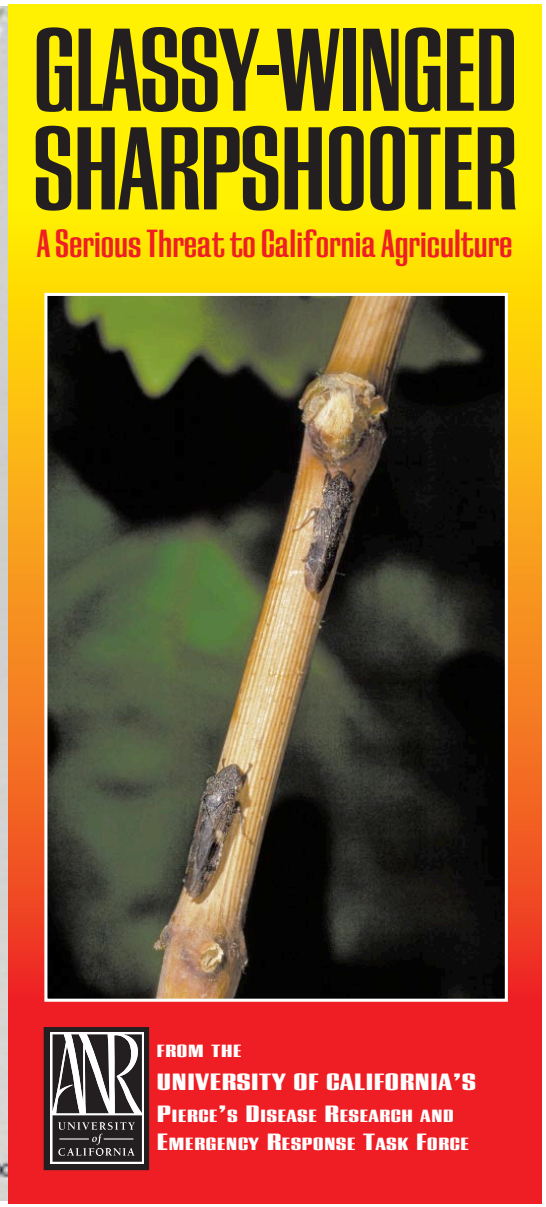
Wolbachia pipientis wMel



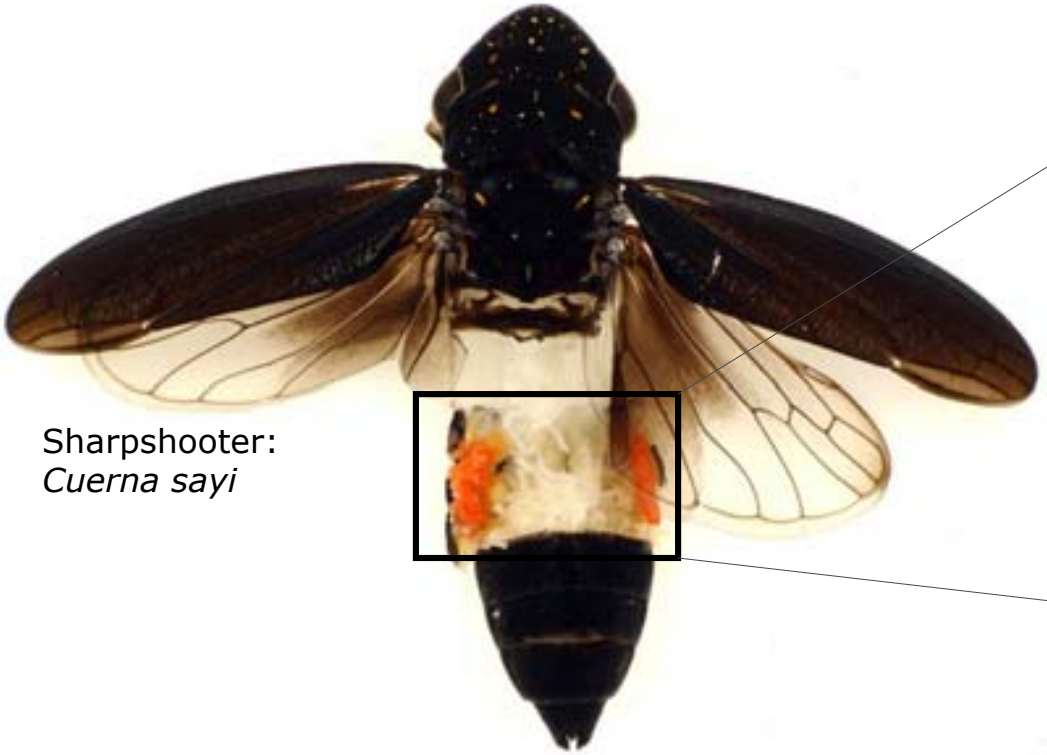
Wu et al., 2004. Collaboration between Jonathan Eisen and Scott O'Neill (Yale, U. Queensland).

Pierce's Disease

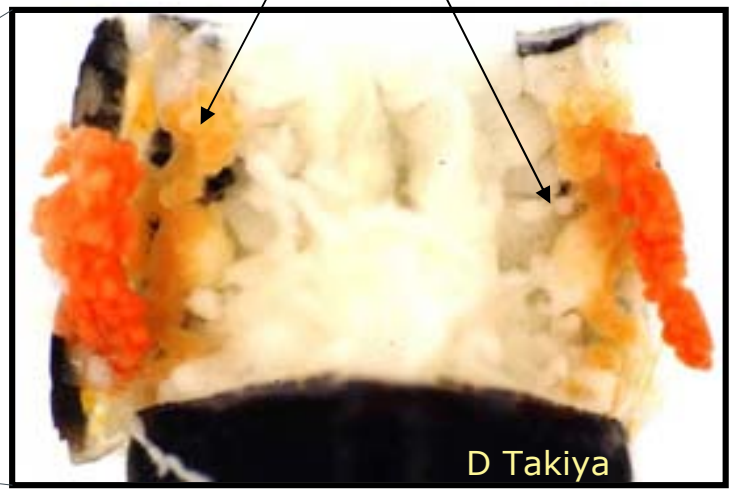
Glassy winged sharpshooter

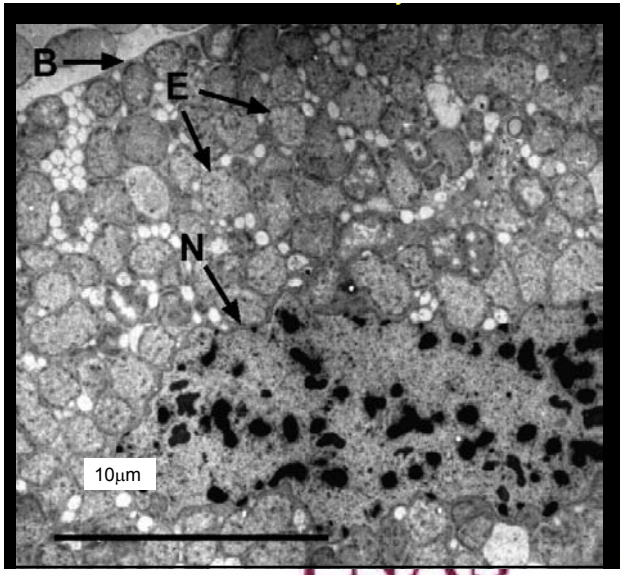


- Obligate xylem feeder
- Transmits *Xylella* between plants
- Much like mosquitoes transmit malarial pathogen
- Only animal listed as possible "bioterror" agent by US DHS



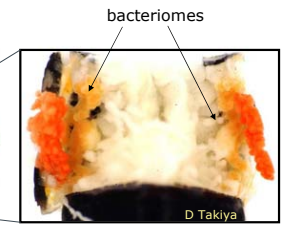
Sharpshooter:
Cuerna sayi



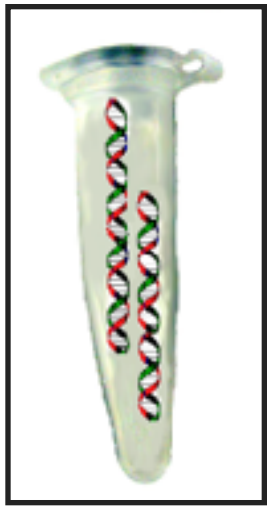


Moran N. A. PNAS 2007;104:8627-8633

PCR and phylogenetic analysis of rRNA genes

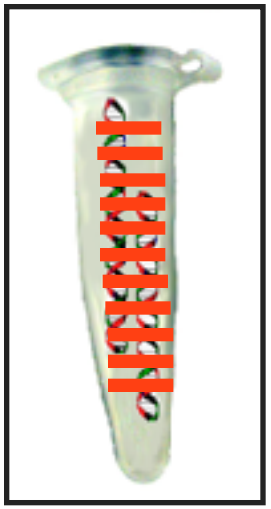


DNA extraction



PCR

Makes lots of copies of the rRNA genes in sample



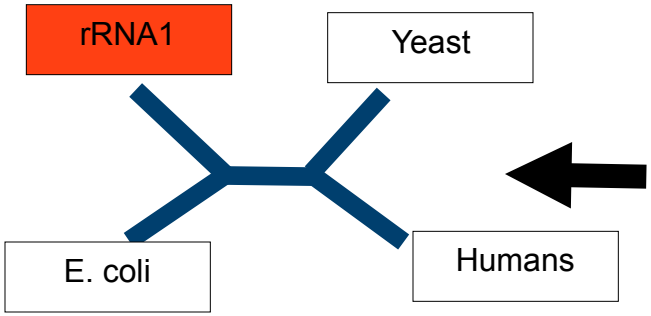
Sequence rRNA genes

rRNA1
5' ...TACAGTATAGGTG
GAGCTAGCGATCGAT
CGA... 3'

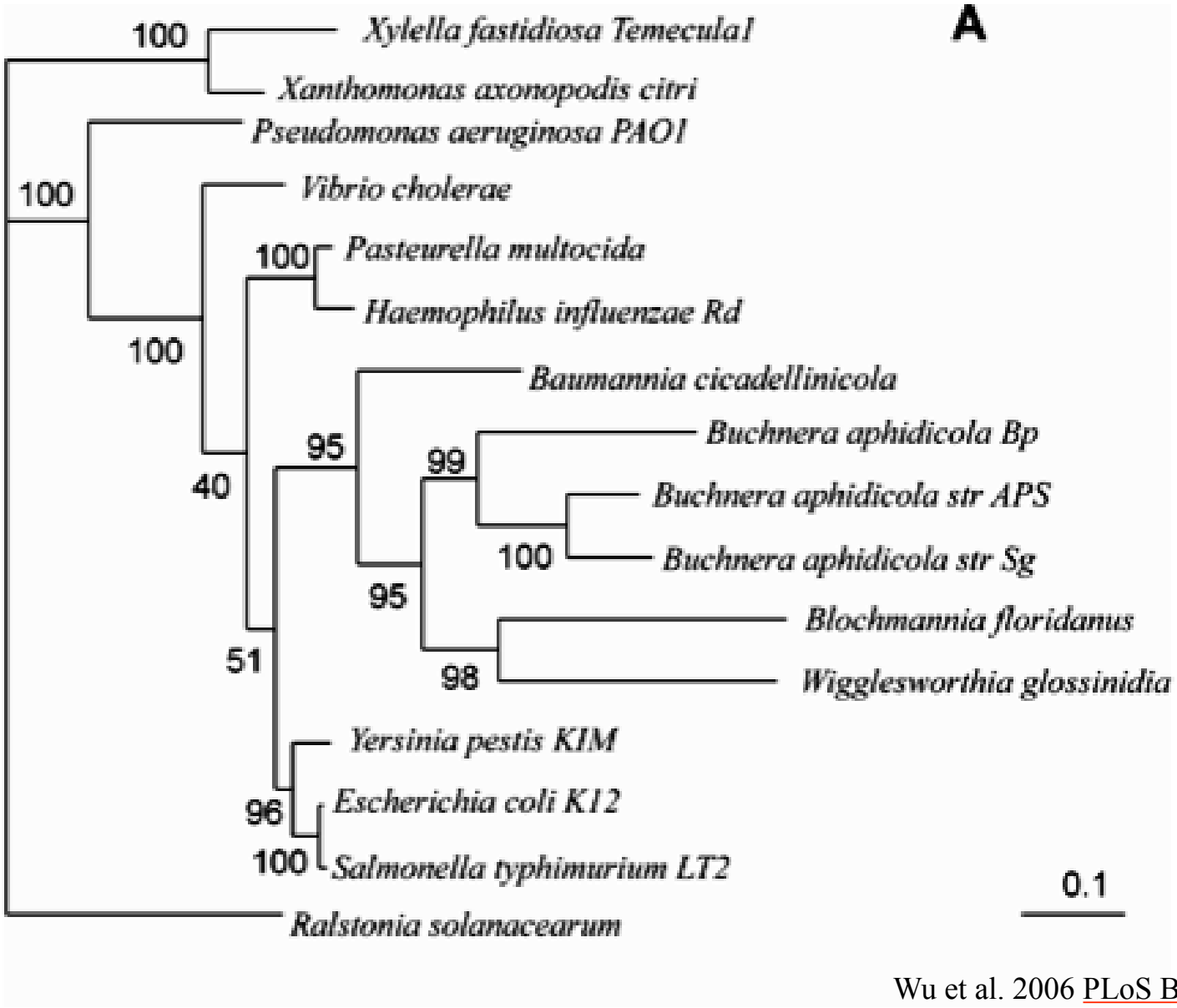
Sequence alignment = Data matrix

rRNA1	A	C	A	C	A	C
Yeast	T	A	C	A	G	T
E. coli	A	G	A	C	A	G
Humans	T	A	T	A	G	T

Phylogenetic tree

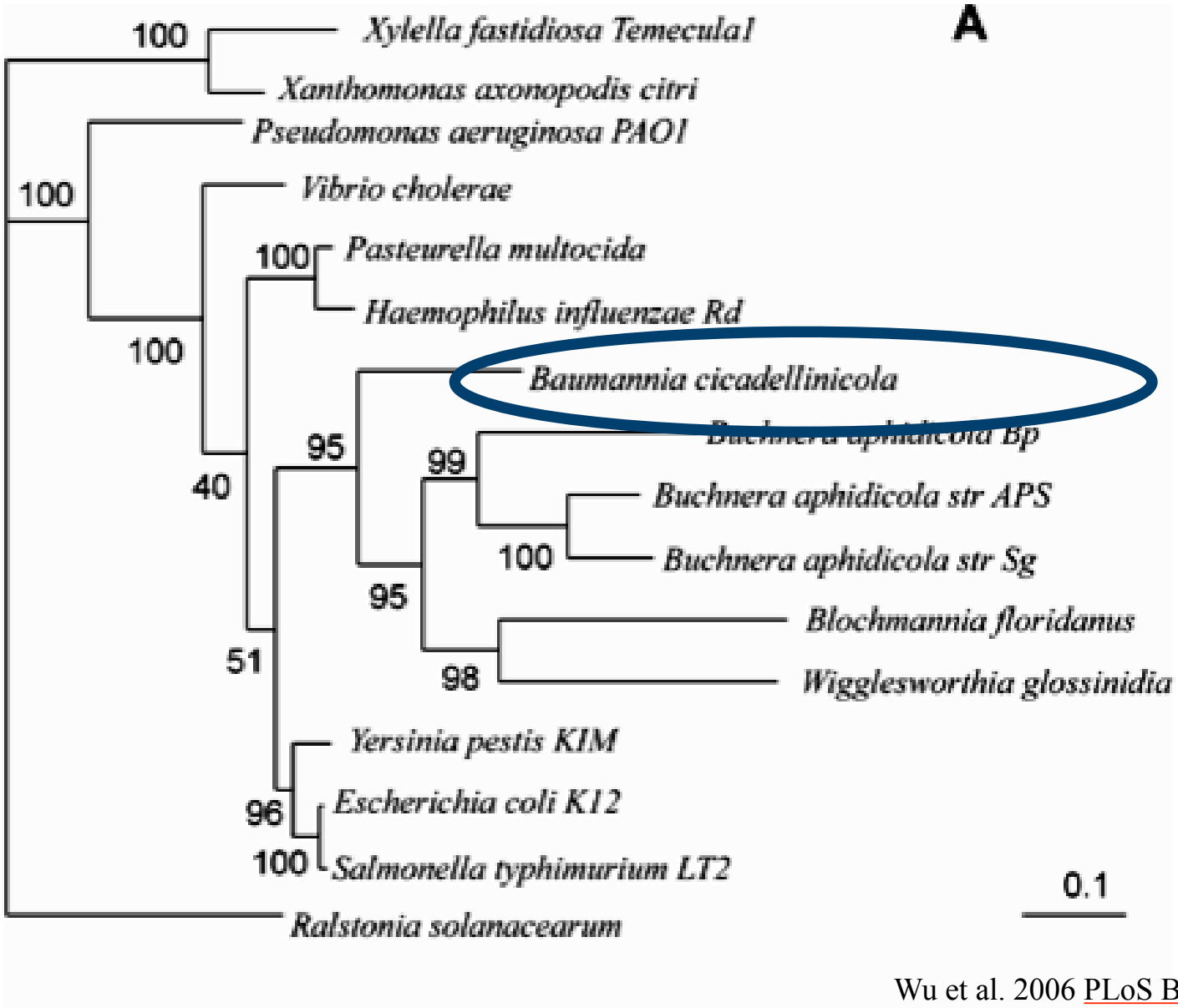


Baumania is close relative of Buchnera symbionts of aphids



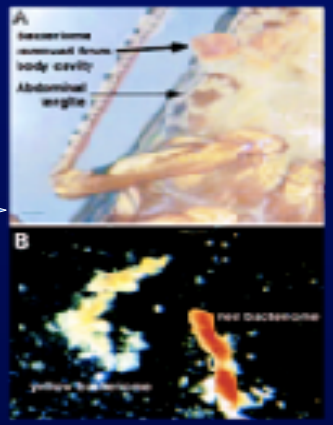
Wu et al. 2006 PLoS Biology 4: e188.

Baumania is close relative of Buchnera symbionts of aphids

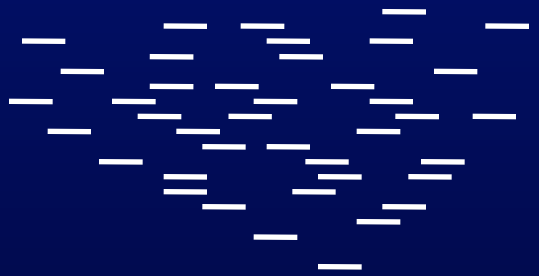


Wu et al. 2006 PLoS Biology 4: e188.

Sharpshooter Shotgun Sequencing

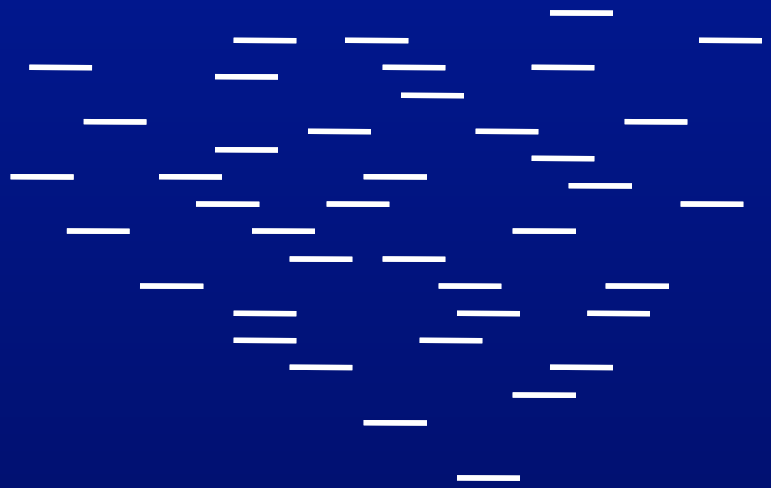


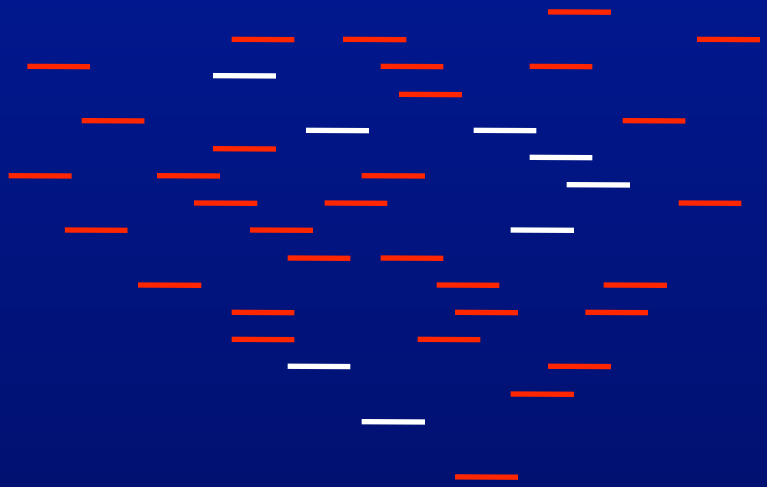
shotgun

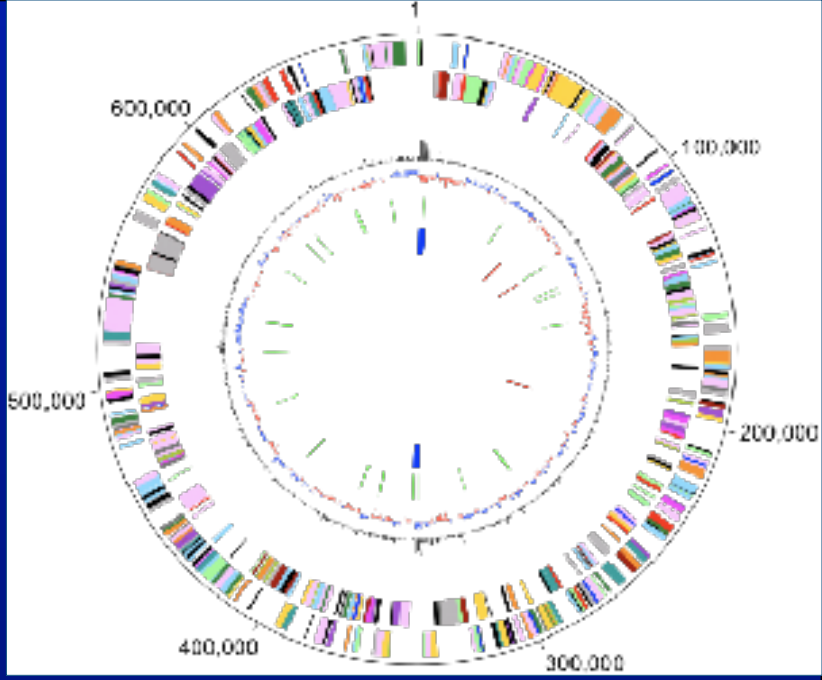


Collaboration with Nancy Moran's lab

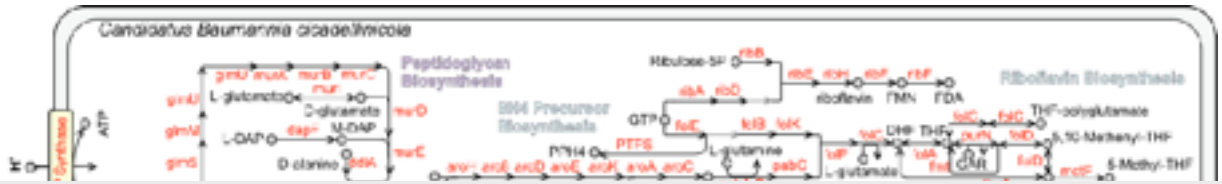
Wu et al. 2006 PLoS Biology 4: e188



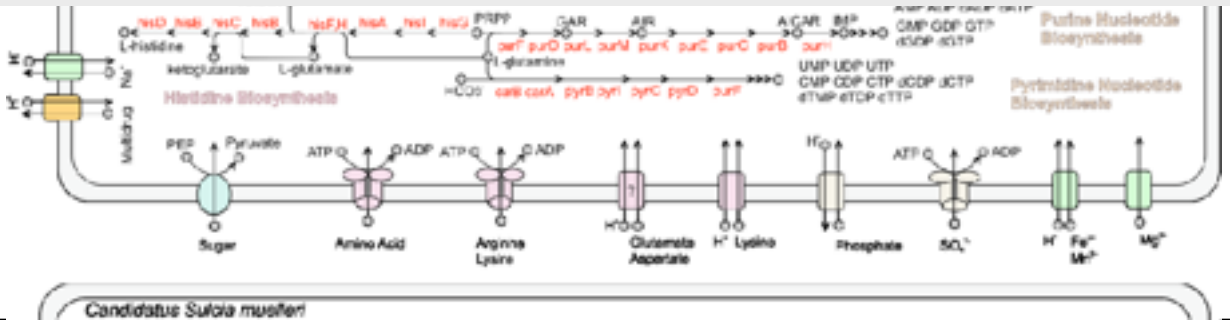



1
100,000
200,000
300,000
400,000
500,000
600,000

Baumannia is a Vitamin and Cofactor Producing Machine



VITAMIN AND COFACTOR PRODUCING MACHINE



Wu et al. 2006 PLoS Biology 4: e188.

How to do Binning

- Method 2: Assembly in More Complex Ecosystems

- How would this work?
- Advantages and disadvantages?

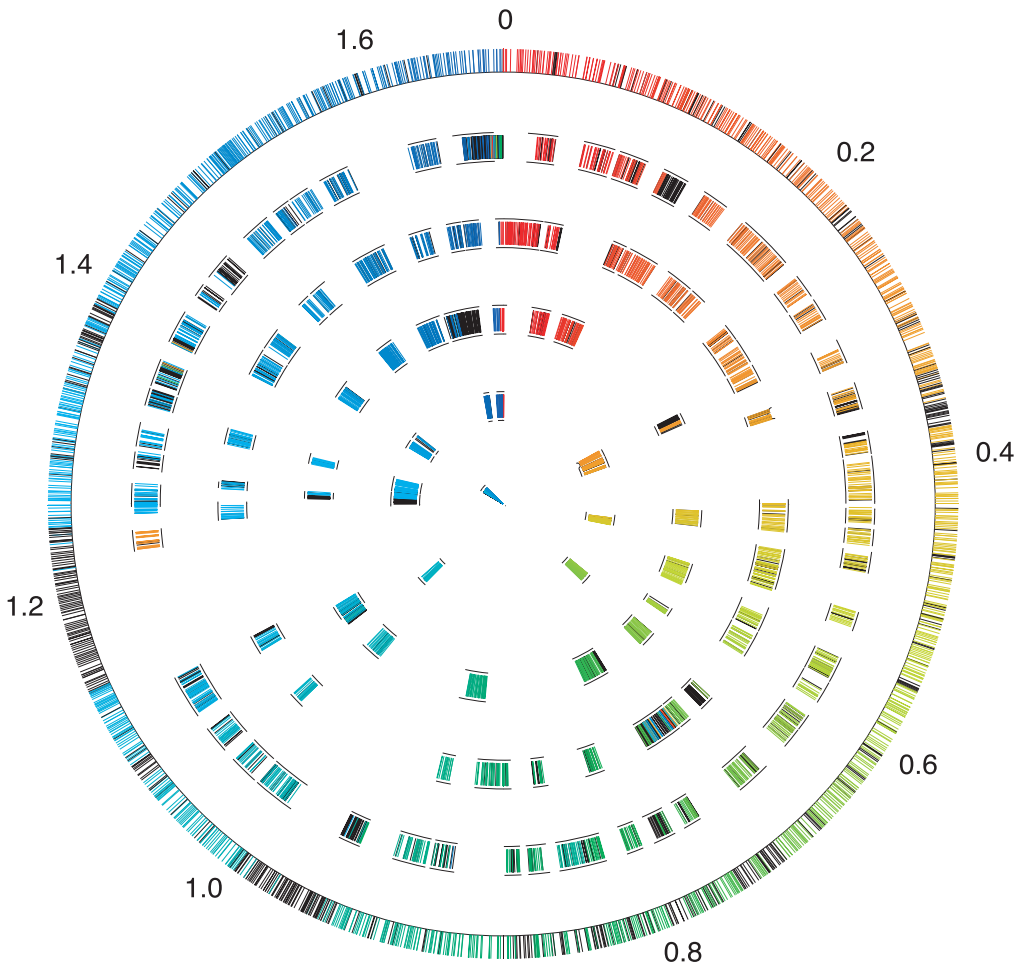


Fig. 2. Gene conservation among closely related *Prochlorococcus*. The outermost concentric circle of the diagram depicts the competed genomic sequence of *Prochlorococcus marinus* MED4 (*11*). Fragments from environmental sequencing were compared to this completed *Prochlorococcus* genome and are shown in the inner concentric circles and were given boxed outlines. Genes for the outermost circle have been assigned psuedospectrum colors based on the position of those genes along the chromosome, where genes nearer to the start of the genome are colored in red, and genes nearer to the end of the genome are colored in blue. Fragments from environmental sequencing were subjected to an analysis that identifies conserved gene order between those fragments and the completed *Prochlorococcus* MED4 genome. Genes on the environmental genome segments that exhibited conserved gene order are colored with the same color assignments as the *Prochlorococcus* MED4 chromosome. Colored regions on the environmental segments exhibiting color differences from the adjacent outermost concentric circle are the result of conserved gene order with other MED4 regions and probably represent chromosomal rearrangements. Genes that did not exhibit conserved gene order are colored in black.

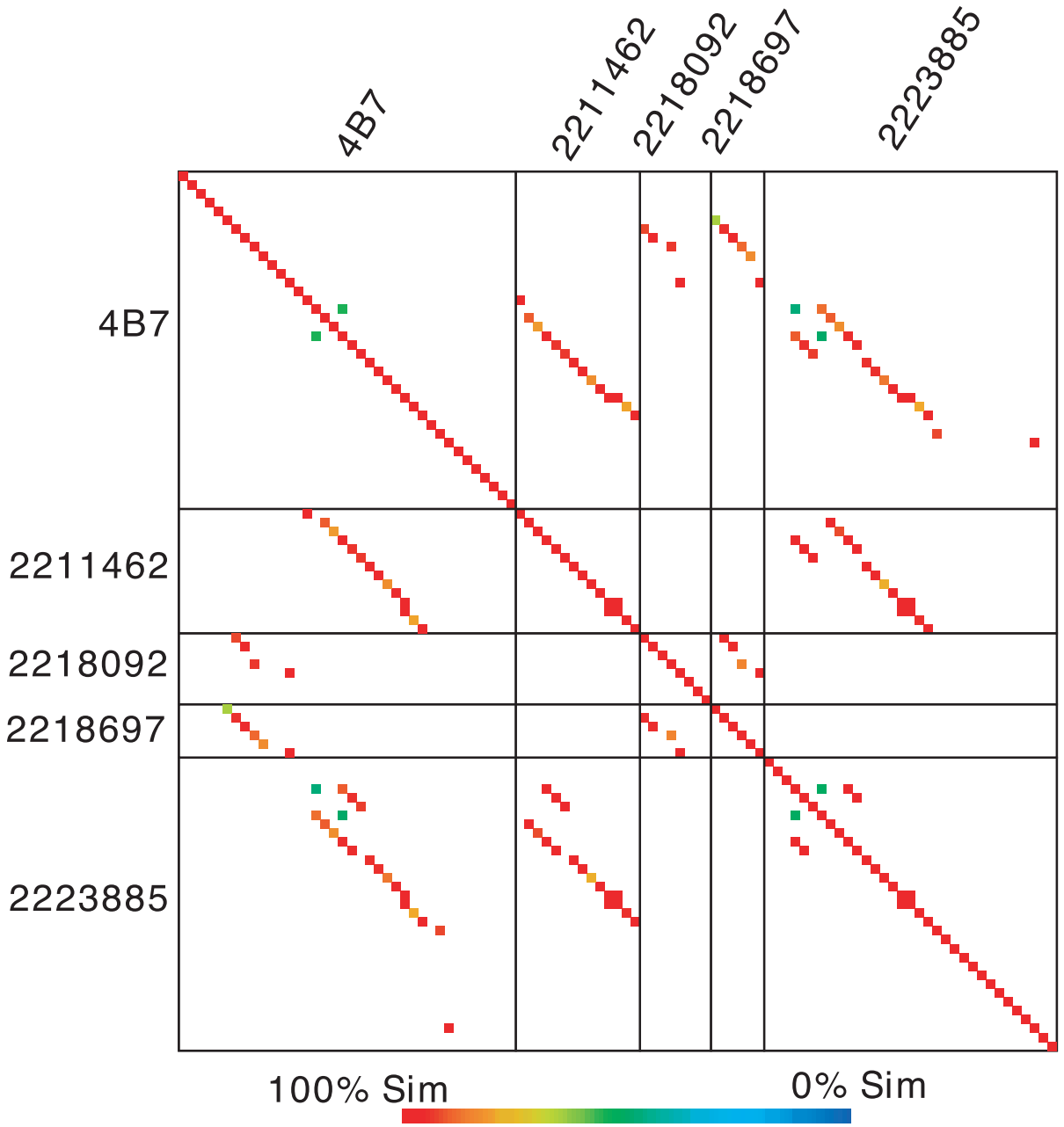


Fig. 3. Comparison of Sargasso Sea scaffolds to Crenarchaeal clone 4B7. Predicted proteins from 4B7 and the scaffolds showing significant homology to 4B7 by tBLASTx are arrayed in positional order along the *x* and *y* axes. Colored boxes represent BLASTp matches scoring at least 25% similarity and with an e value of better than 1e-5. Black vertical and horizontal lines delineate scaffold borders.

- Studying Sar86 and other marine plankton
- Note - published one of first genomic studies of uncultured microbes - in 1996

JOURNAL OF BACTERIOLOGY, Feb. 1996, p. 591–599 Vol. 178, No. 3
0021-9193/96/$04.00+0

Characterization of Uncultivated Prokaryotes: Isolation and Analysis of a 40-Kilobase-Pair Genome Fragment from a Planktonic Marine Archaeon

JEFFEREY L. STEIN,[1]* TERENCE L. MARSH,[2] KE YING WU,[3] HIROAKI SHIZUYA,[4] AND EDWARD F. DELONG[3]*

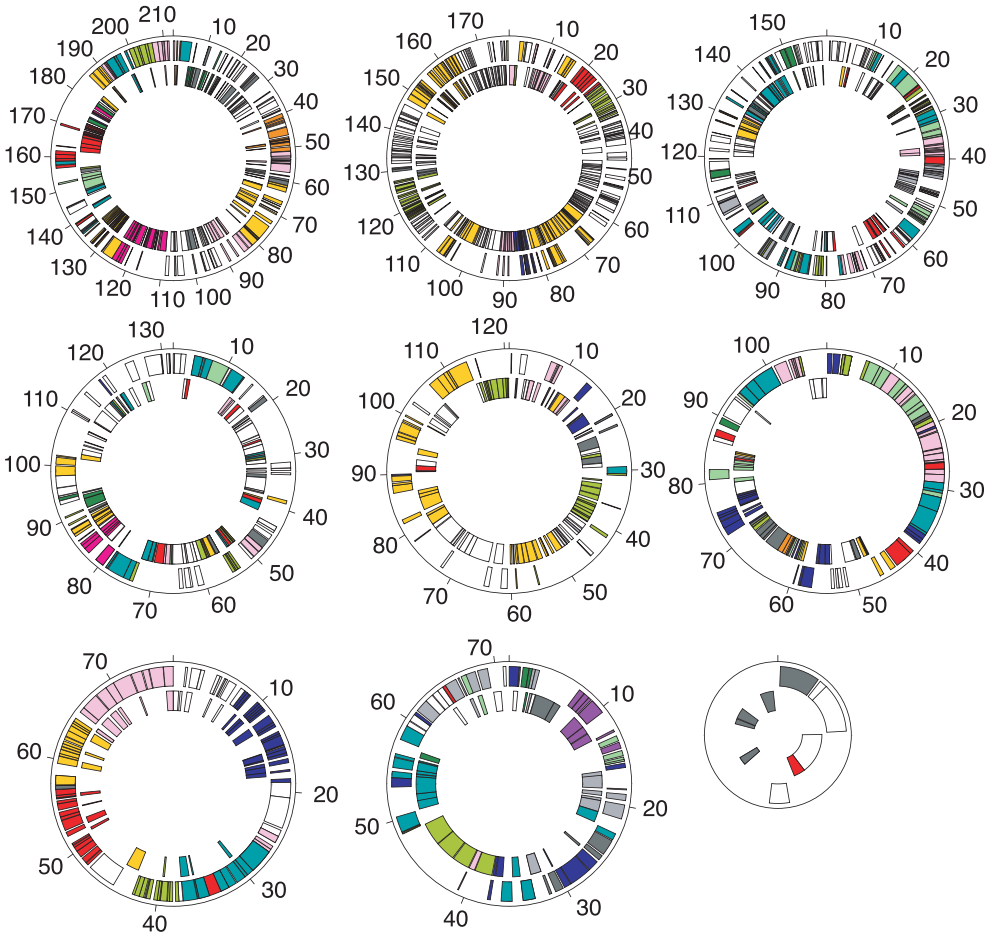
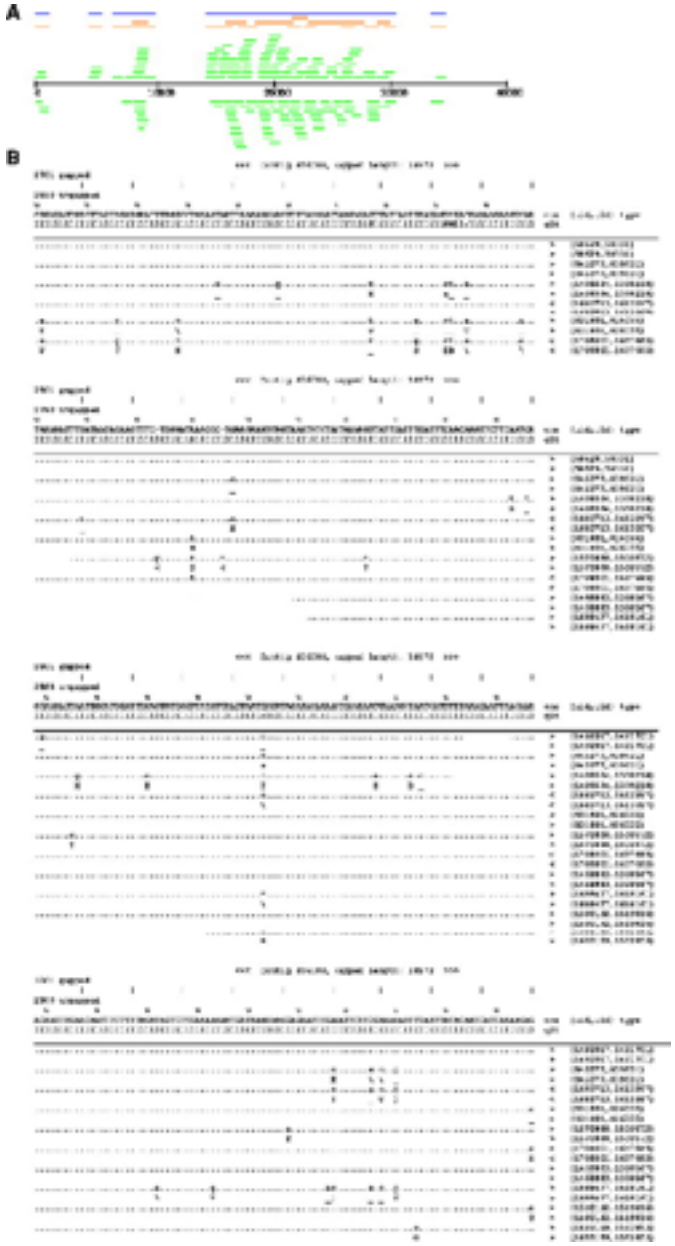
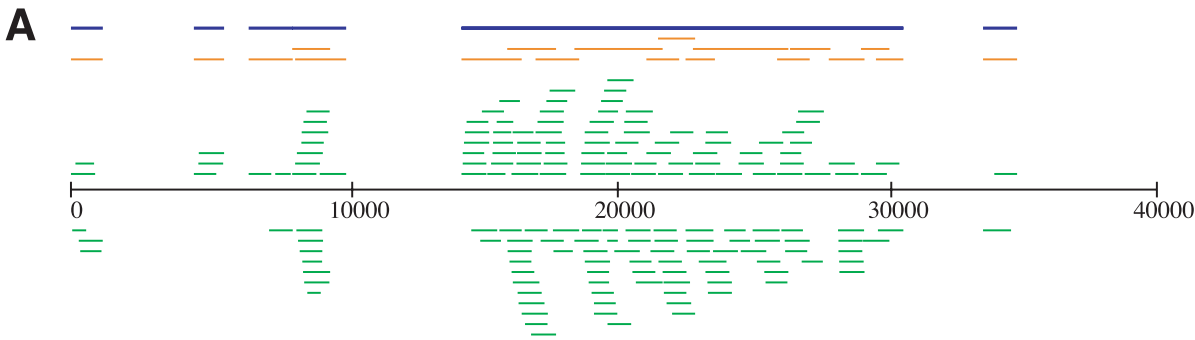


Fig. 4. Circular diagrams of nine complete megaplasmids. Genes encoded in the forward direction are shown in the outer concentric circle; reverse coding genes are shown in the inner concentric circle. The genes have been given role category assignment and colored accordingly: amino acid biosynthesis, violet; biosynthesis of cofactors, prosthetic groups, and carriers, light blue; cell envelope, light green; cellular processes, red; central intermediary metabolism, brown; DNA metabolism, gold; energy metabolism, light gray; fatty acid and phospholipid metabolism, magenta; protein fate and protein synthesis, pink; purines, pyrimidines, nucleosides, and nucleotides, orange; regulatory functions and signal transduction, olive; transcription, dark green; transport and binding proteins, blue-green; genes with no known homology to other proteins and genes with homology to genes with no known function, white; genes of unknown function, gray; Tick marks are placed on 10-kb intervals.


A
B


A
0
10000
20000
30000
40000

MS 1093857: Environmental Genome Shotgun Sequencing of the Sargasso Sea
Venter et al., revised

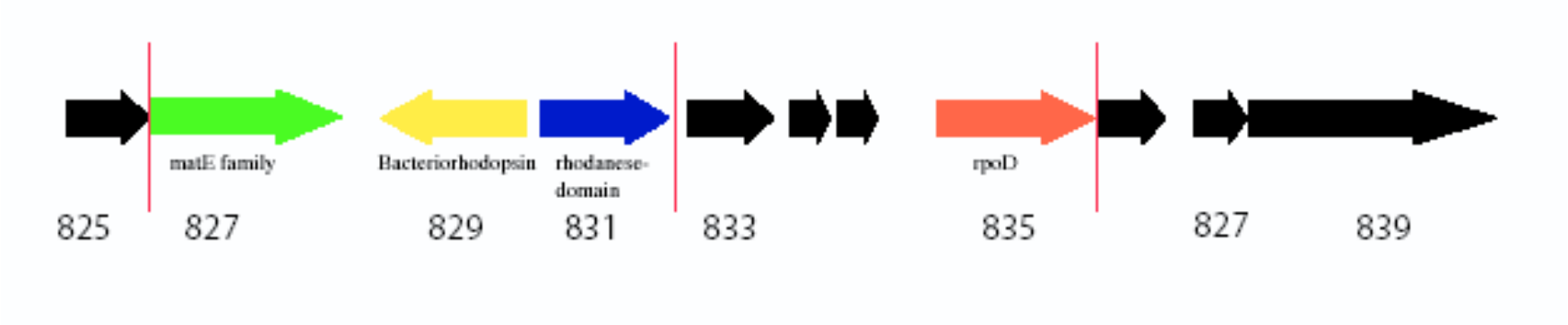


Figure S10. Scaffold 2217664, containing the gene encoding Proteorhodopsin. Genes are colored using color assignments described in Fig. 2, and contig boundaries are indicated with red vertical lines. In this scaffold, rhodopsin is associated with a DNA-directed RNA polymerase, sigma subunit (rpoD) originating in the CFB group.

How to do Binning

- Method 2: Assembly in More Complex Ecosystems

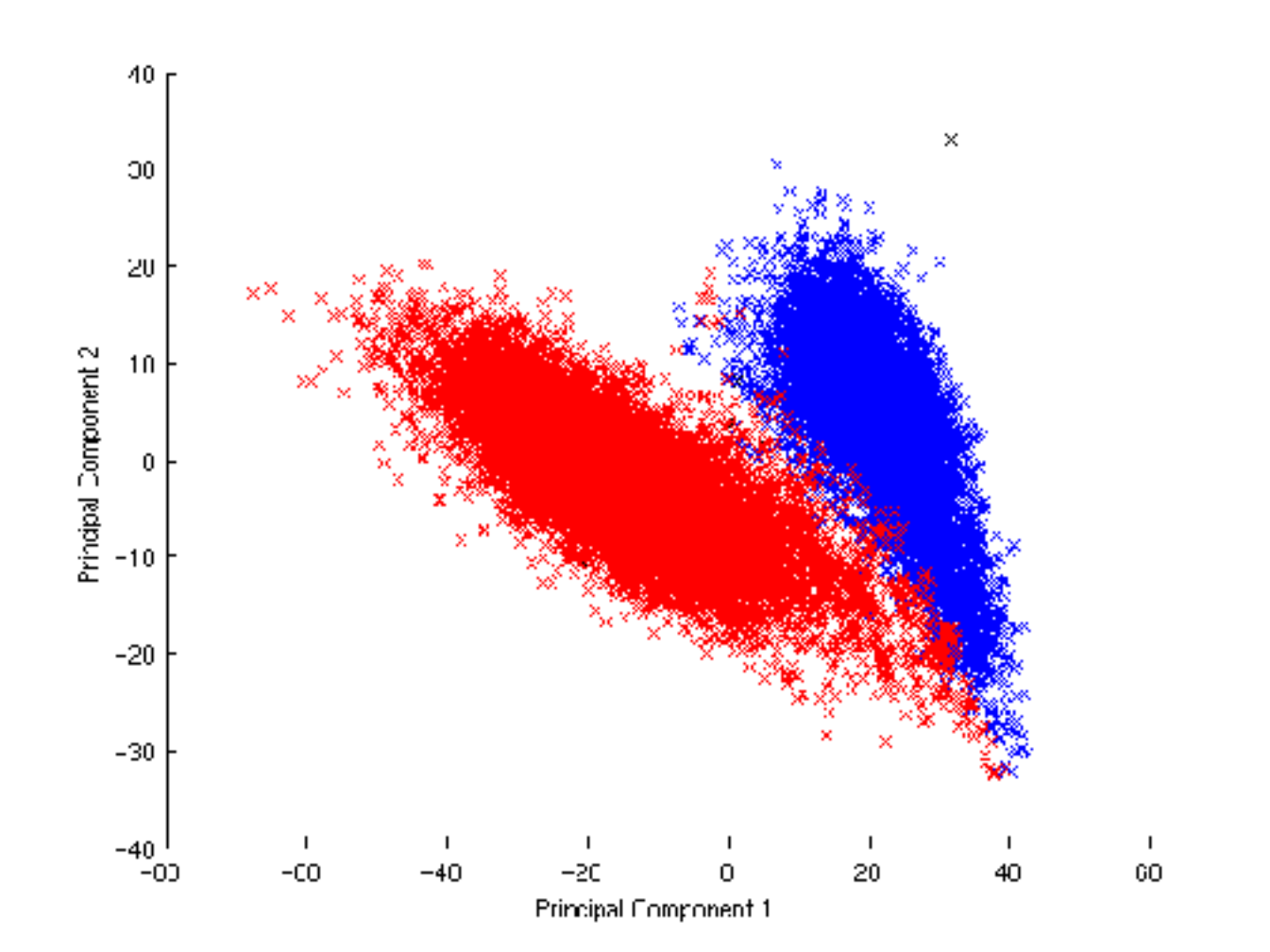
- How would this work?
- Advantages and disadvantages?

How to do Binning

- Method 3: Composition of Reads / Contigs

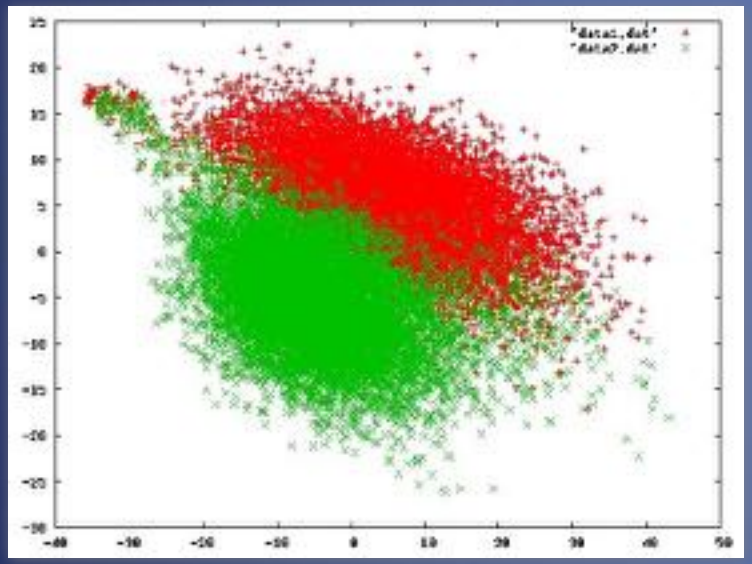
- How would this work?
- Advantages and disadvantages?

PCA separates species

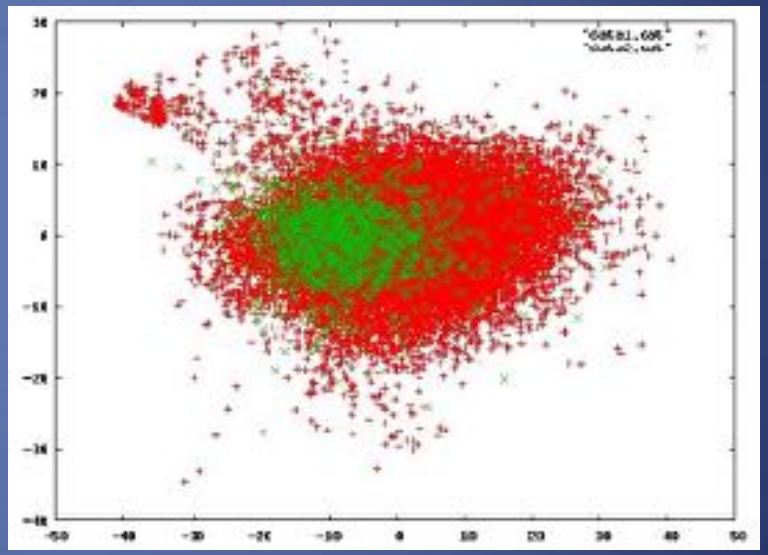


Gluconobacter oxydans[65% GC] and *Rhodospirillum rubrum*[61% GC]

Effect of Skewed Relative Abundance



Abundance 1:1



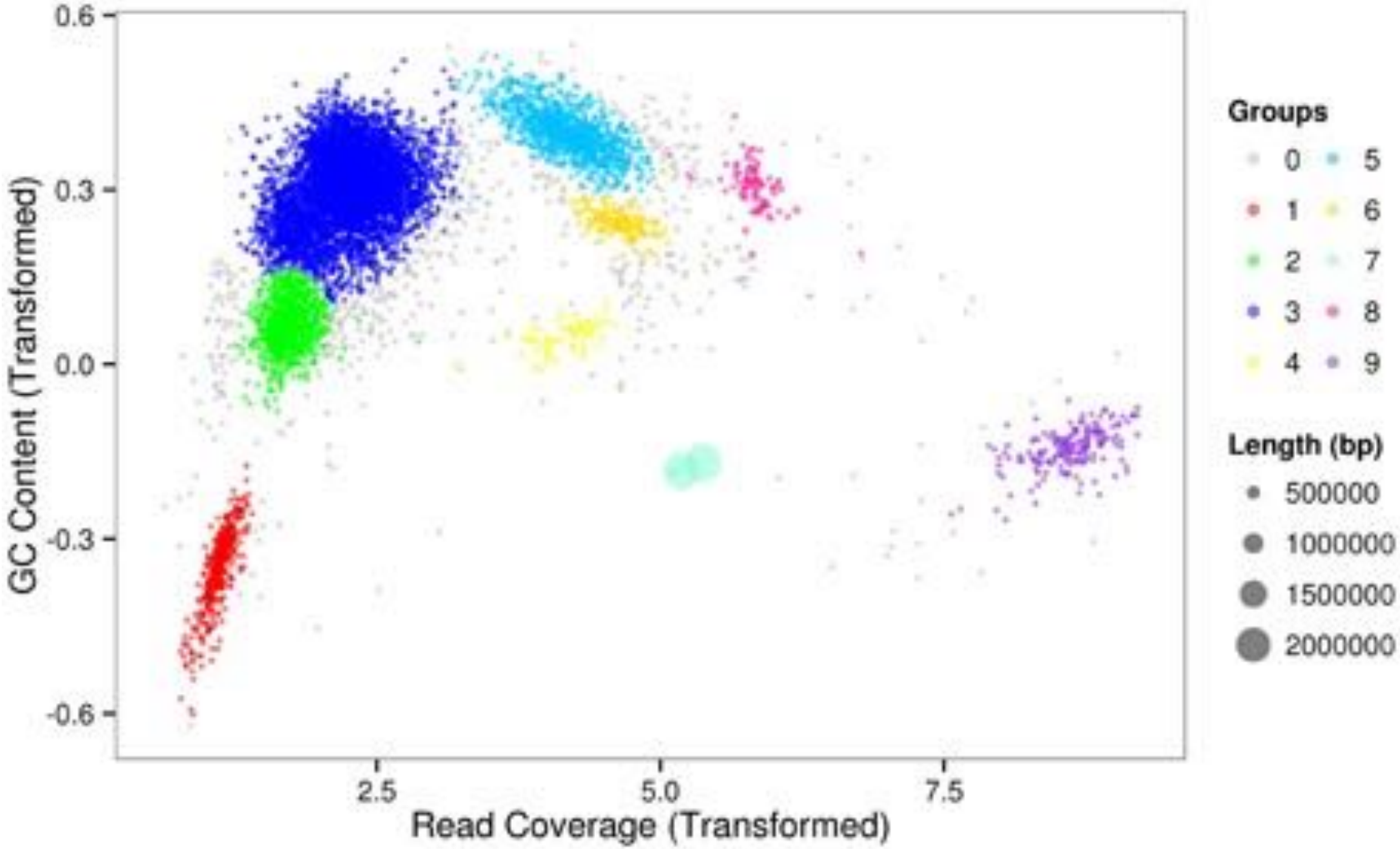
Abundance 20:1

B. anthracis **and** ***L. monogocytes***

How to do Binning

- Method 4: Coverage of Reads / Contigs

- How would this work?
- Advantages and disadvantages?

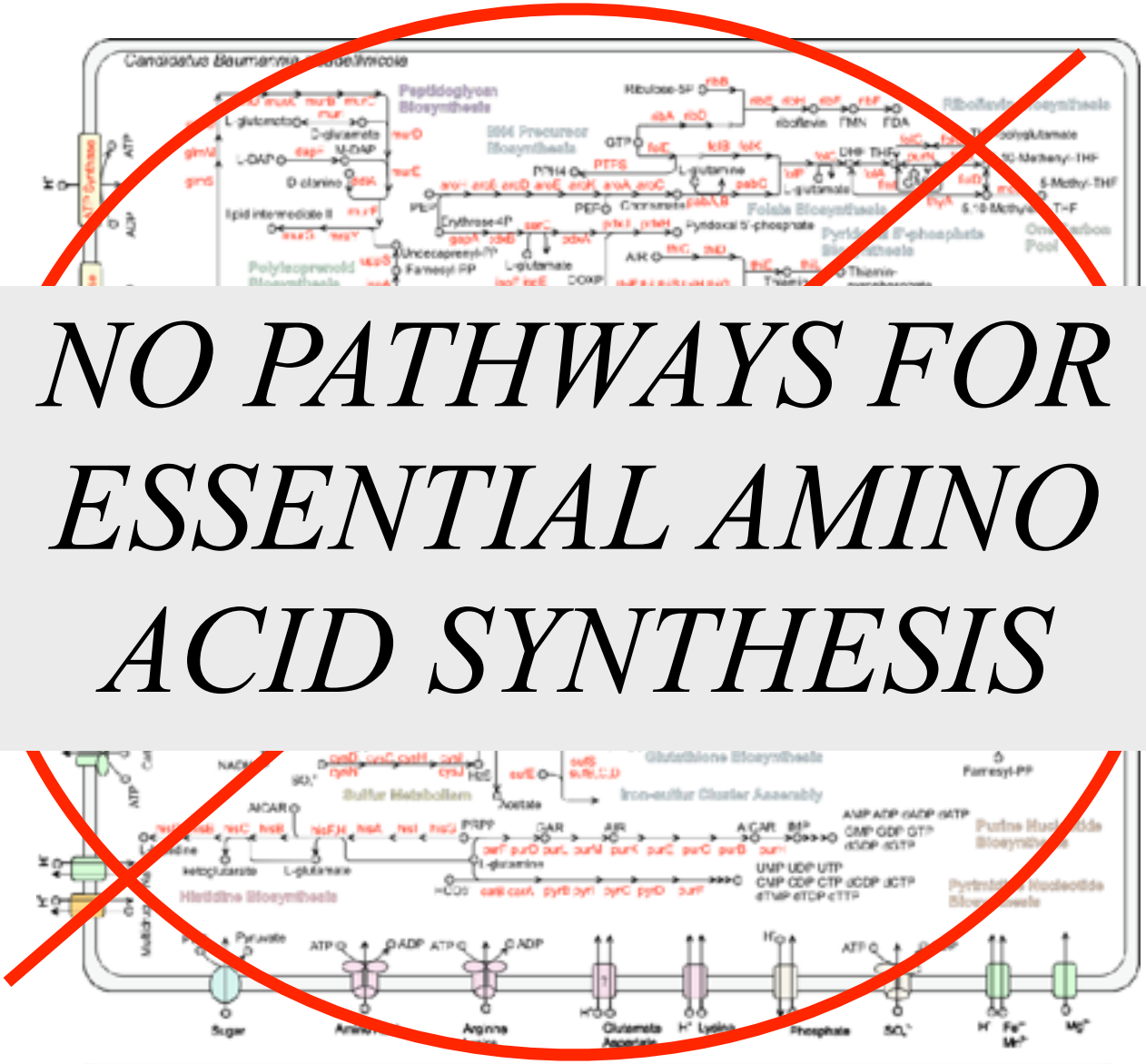

0.6
0.3
0.0
-0.3
-0.6
GC Content (Transformed)
2.5
5.0
7.5
Read Coverage (Transformed)
Groups
0
1
2
3
4
5
6
7
8
9
Length (bp)
500000
1000000
1500000
2000000

How to do Binning

- Method 5: Phylogenetic Analysis of Reads / Contigs

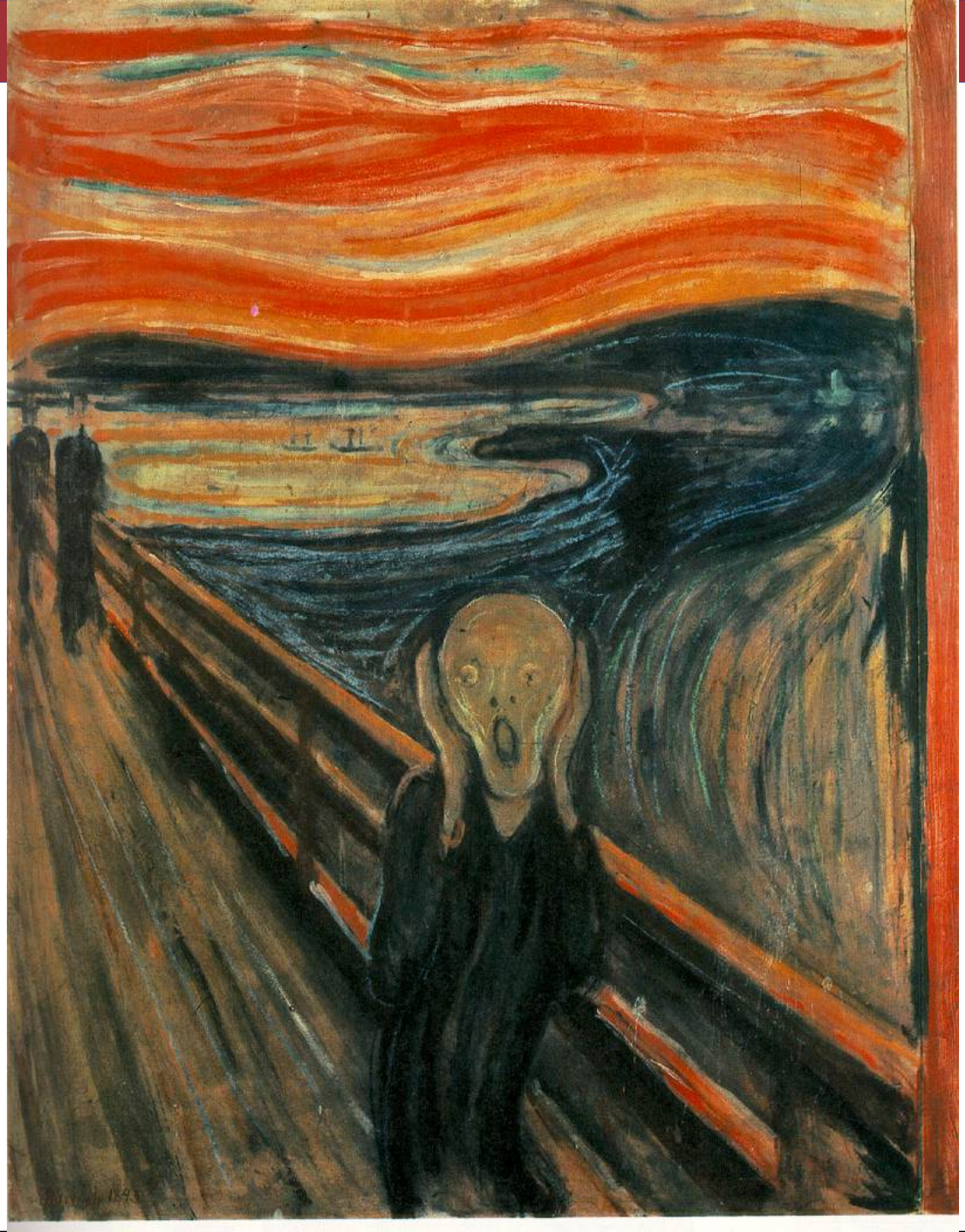
- How would this work?
- Advantages and disadvantages?

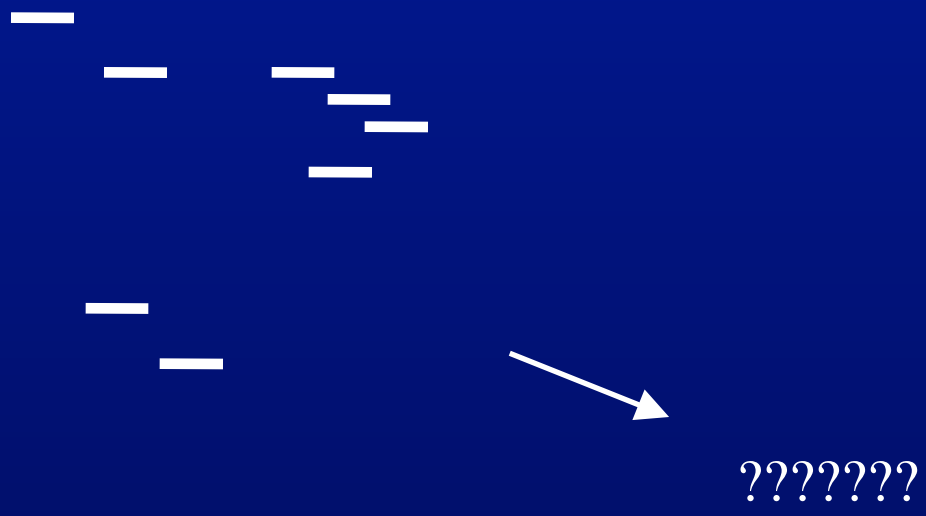
Baumannia is a Vitamin and Cofactor Producing Machine

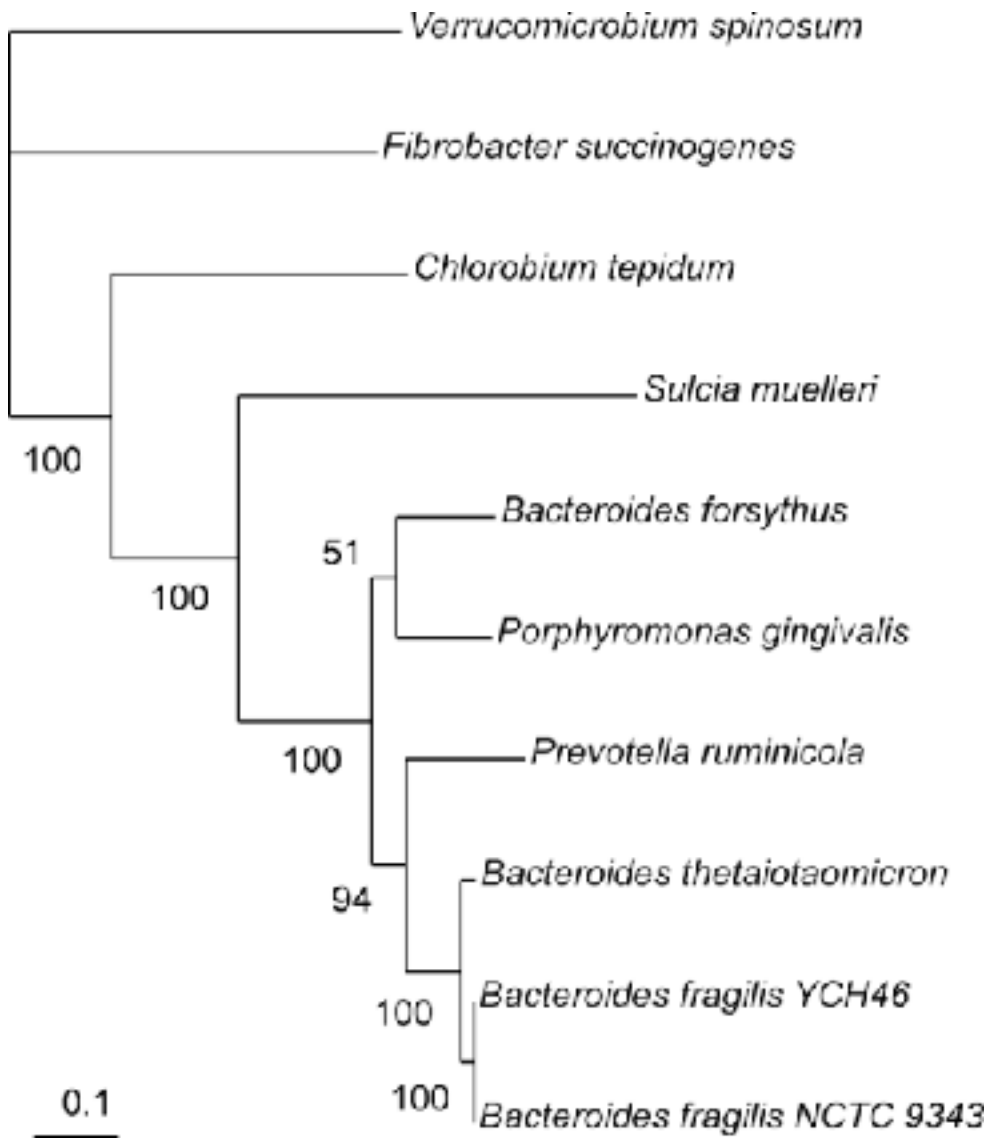


NO PATHWAYS FOR ESSENTIAL AMINO ACID SYNTHESIS

Wu et al. 2006 PLoS Biology 4: e188.

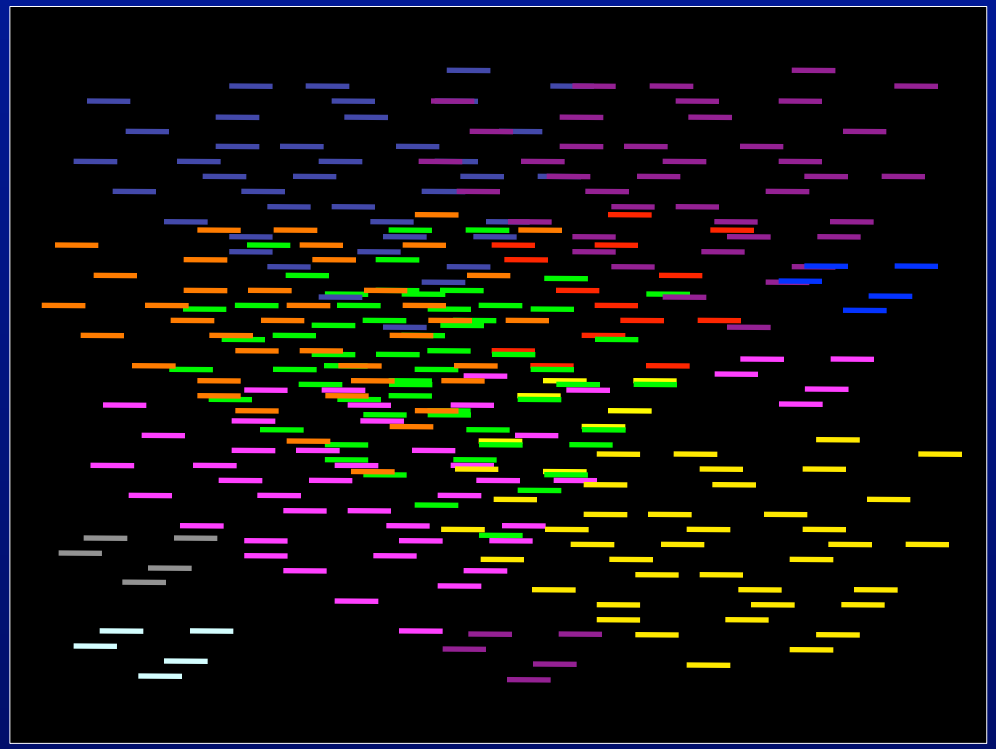





Verrucomicrobium spinosum
Fibrobacter succinogenes
Chlorobium tepidum
Sulcia muelleri
100
Bacteroides forsythus
51
100
Porphyromonas gingivalis
Prevotella ruminicola
100
Bacteroides thetaiotaomicron
94
Bacteroides fragilis YCH46
100
100
Bacteroides fragilis NCTC 9343
0.1

Binning challenge

A
B
C
D
E
F
G

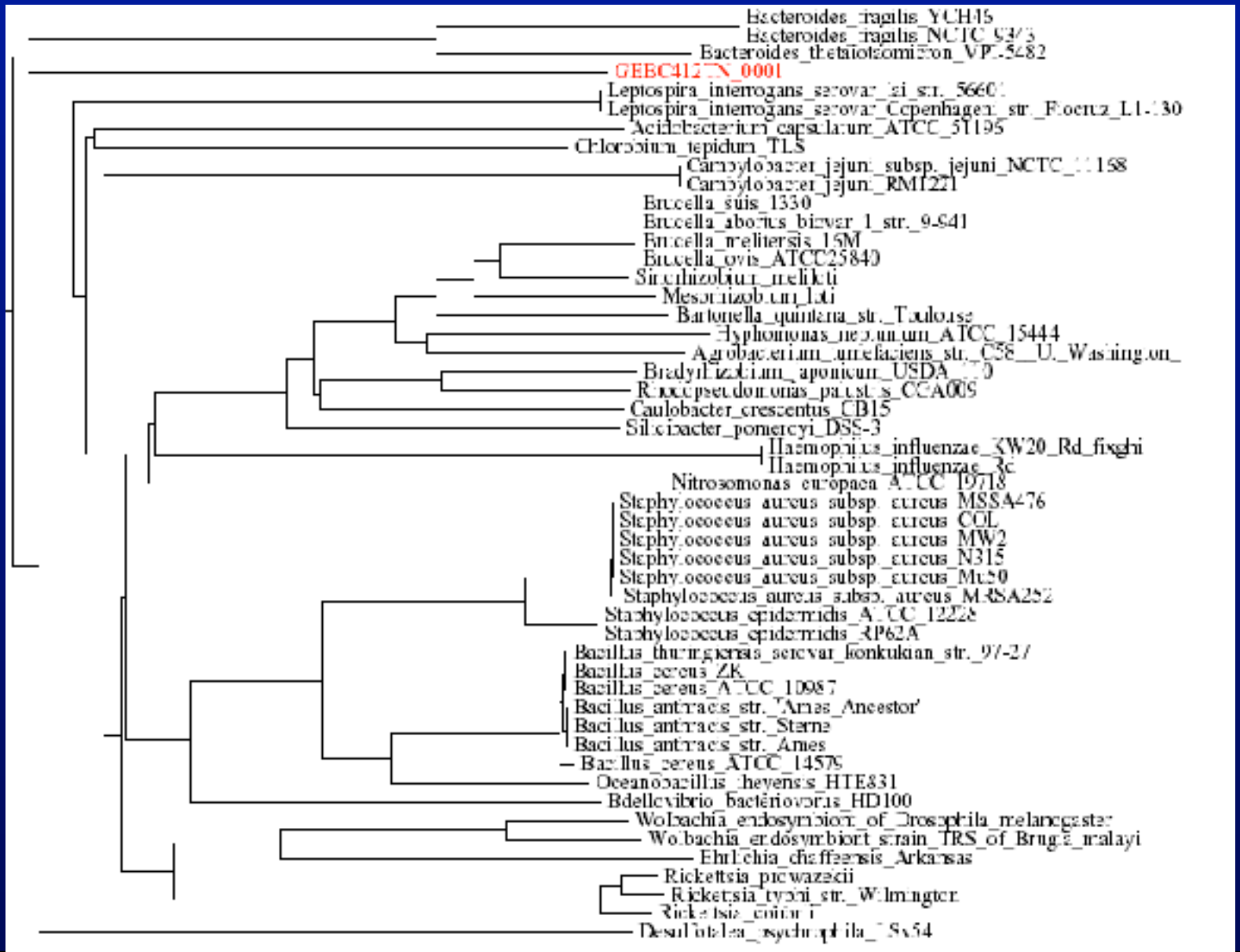


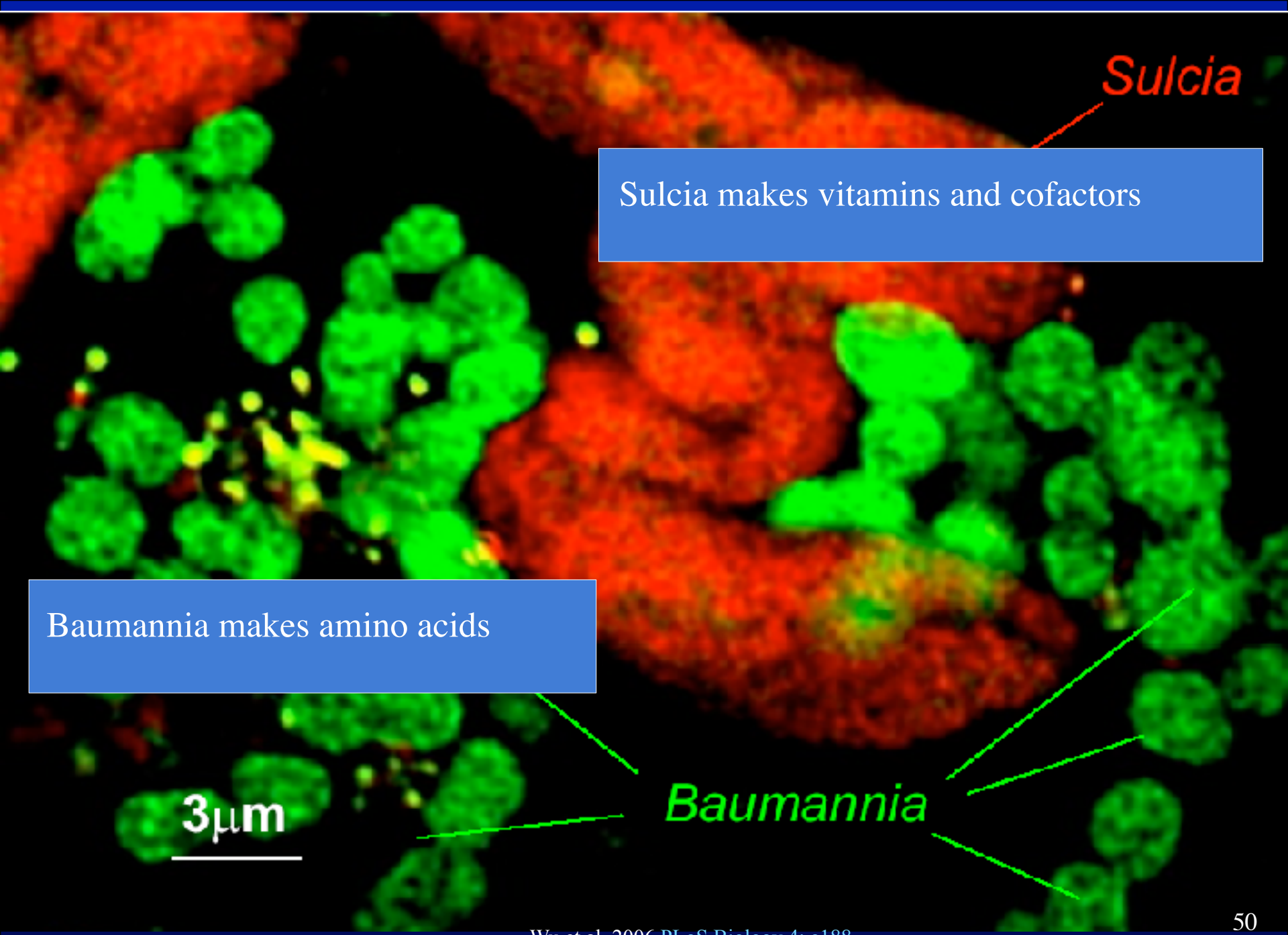
T
U
V
W
X
Y
Z

No reference genome? What do you do?

Phylogeny ....

CFB Phyla


Bacteroides_fragilis_YCH46
Bacteroides_fragilis_NCTC_9343
Bacteroides_thetaiotaomicron_VPI-5482
GEBC412_N_0001
Leptospira_interrogans_serovar_lai_str._56601
Leptospira_interrogans_serovar_Copenhageni_str._Fiocruz_L1-130
Acidobacterium_capsulatum_ATCC_51196
Chlorobium_tepidum_TLS
Campylobacter_jejuni_subsp._jejuni_NCTC_11168
Campylobacter_jejuni_RM1221
Brucella_suis_1330
Brucella_abortus_biovar_1_str._9-941
Brucella_melitensis_16M
Brucella_ovis_ATCC25840
Sinorhizobium_meliloti
Mesorhizobium_loti
Bartonella_quintana_str._Toulouse
Xylella_fastidiosa_Temecula1
Agrobacterium_tumefaciens_str._C58_(U._Washington)
Bradyrhizobium_japonicum_USDA_110
Rhodopseudomonas_palustris_CGA009
Caulobacter_crescentus_CB15
Silicibacter_pomeroyi_DSS-3
Haemophilus_influenzae_KW20_Rd_fixghi
Haemophilus_influenzae_3c
Nitrosomonas_europaea_ATCC_19718
Staphylococcus_aureus_subsp._aureus_MSSA476
Staphylococcus_aureus_subsp._aureus_COL
Staphylococcus_aureus_subsp._aureus_MW2
Staphylococcus_aureus_subsp._aureus_N315
Staphylococcus_aureus_subsp._aureus_Mu50
Staphylococcus_aureus_subsp._aureus_MRSA252
Staphylococcus_epidermidis_ATCC_12228
Staphylococcus_epidermidis_RP62A
Bacillus_thuringiensis_serovar_konkukian_str._97-27
Bacillus_cereus_ZK
Bacillus_cereus_ATCC_10987
Bacillus_anthracis_str._'Ames_Ancestor'
Bacillus_anthracis_str._Sterne
Bacillus_anthracis_str._Ames
Bacillus_cereus_ATCC_14579
Oceanobacillus_iheyensis_HTE831
Bdellovibrio_bacteriovorus_HD100
Wolbachia_endosymbiont_of_Drosophila_melanogaster
Wolbachia_endosymbiont_strain_TRS_of_Brugia_malayi
Ehrlichia_chaffeensis_Arkansas
Rickettsia_prowazekii
Rickettsia_typhi_str._Wilmington
Rickettsia_conorii
Desulfotalea_psychrophila_LSv54



Wu et al. 2006 PLoS Biology 4: e188

How to do Binning

- Method 6: Long Reads

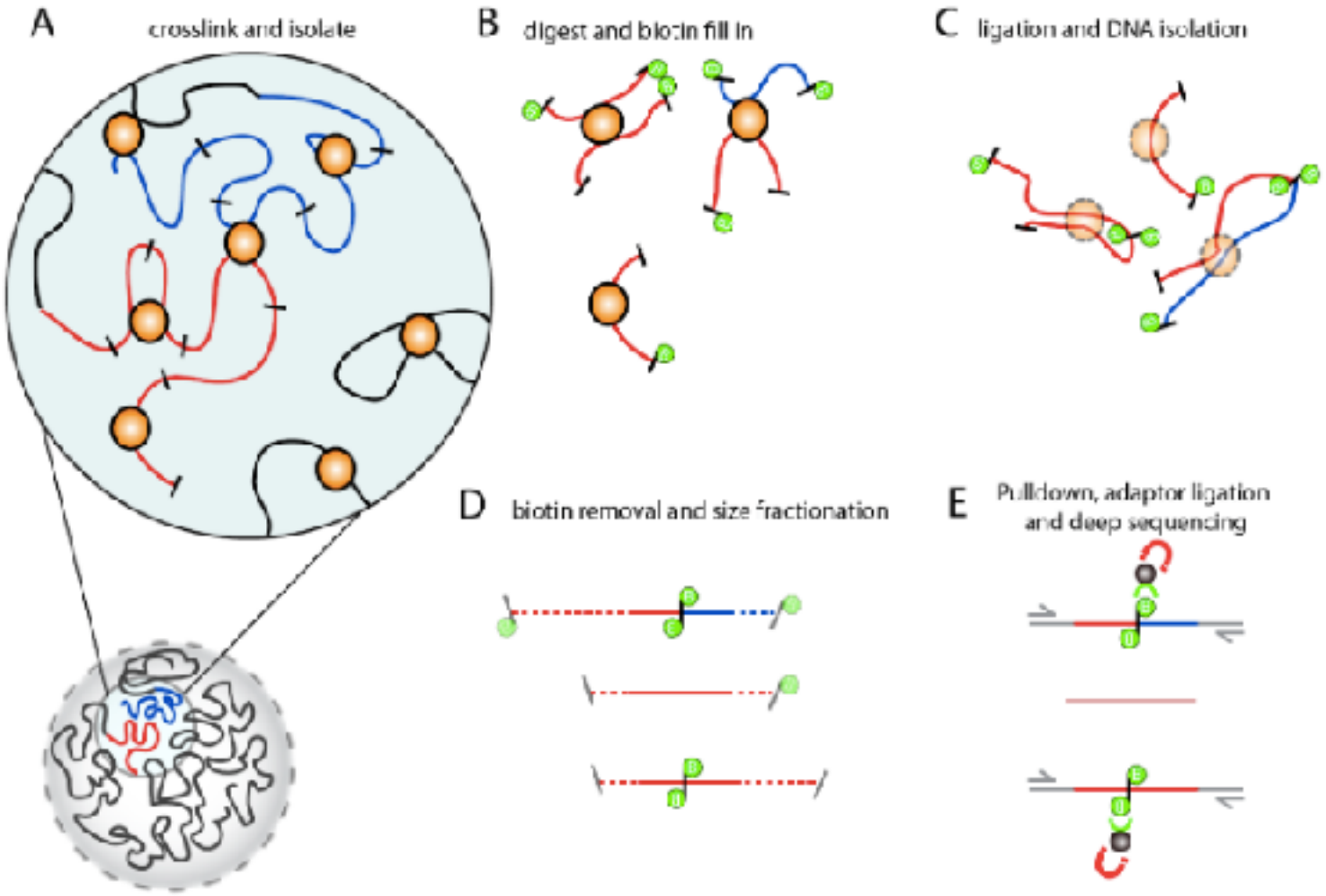
- How would this work?
- Advantages and disadvantages?

How to do Binning

- Method 7: Cross-linking

- How would this work?
- Advantages and disadvantages?

HiC



From Belton JM1, McCord RP, Gibcus JH, Naumova N, Zhan Y, Dekker J. Methods. 2012 Nov;58(3):268-76. doi: 10.1016/j.ymeth.2012.05.001. Hi-C: a comprehensive technique to capture the conformation of genomes.

How to do Binning

- Method 8: Experiments …

- How would this work?
- Advantages and disadvantages?