EVE 161:
Microbial Phylogenomics

Class #10:
Genome Sequencing

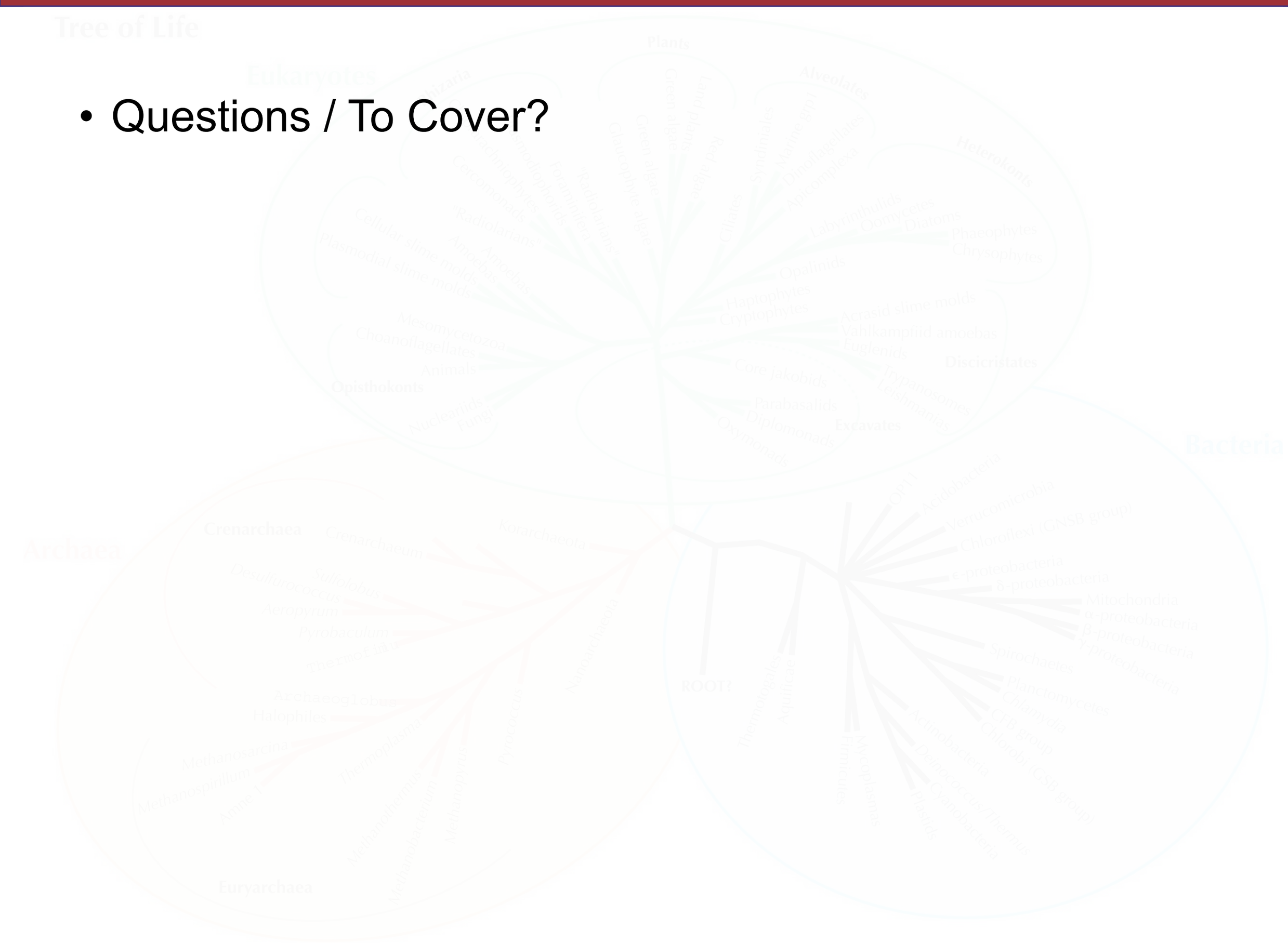
UC Davis, Winter 2018
Instructor: Jonathan Eisen
Teaching Assistant: Cassie Ettinger

Whole-Genome Random Sequencing and Assembly of *Haemophilus influenzae* Rd

Robert D. Fleischmann, Mark D. Adams, Owen White, Rebecca A. Clayton, Ewen F. Kirkness, Anthony R. Kerlavage, Carol J. Bult, Jean-Francois Tomb, Brian A. Dougherty, Joseph M. Merrick, Keith McKenney, Granger Sutton, Will FitzHugh, Chris Fields,* Jeannine D. Gocayne, John Scott, Robert Shirley, Li-Ing Liu, Anna Glodek, Jenny M. Kelley, Janice F. Weidman, Cheryl A. Phillips, Tracy Spriggs, Eva Hedblom, Matthew D. Cotton, Teresa R. Utterback, Michael C. Hanna, David T. Nguyen, Deborah M. Saudek, Rhonda C. Brandon, Leah D. Fine, Janice L. Fritchman, Joyce L. Fuhrmann, N. S. M. Geoghagen, Cheryl L. Gnehm, Lisa A. McDonald, Keith V. Small, Claire M. Fraser, Hamilton O. Smith, J. Craig Venter†

- Presenters?

- Questions / To Cover?



Whole-Genome Random Sequencing and Assembly of *Haemophilus influenzae* Rd

Robert D. Fleischmann, Mark D. Adams, Owen White, Rebecca A. Clayton, Ewen F. Kirkness, Anthony R. Kerlavage, Carol J. Bult, Jean-Francois Tomb, Brian A. Dougherty, Joseph M. Merrick, Keith McKenney, Granger Sutton, Will FitzHugh, Chris Fields,* Jeannine D. Gocayne, John Scott, Robert Shirley, Li-Ing Liu, Anna Glodek, Jenny M. Kelley, Janice F. Weidman, Cheryl A. Phillips, Tracy Spriggs, Eva Hedblom, Matthew D. Cotton, Teresa R. Utterback, Michael C. Hanna, David T. Nguyen, Deborah M. Saudek, Rhonda C. Brandon, Leah D. Fine, Janice L. Fritchman, Joyce L. Fuhrmann, N. S. M. Geoghagen, Cheryl L. Gnehm, Lisa A. McDonald, Keith V. Small, Claire M. Fraser, Hamilton O. Smith, J. Craig Venter†

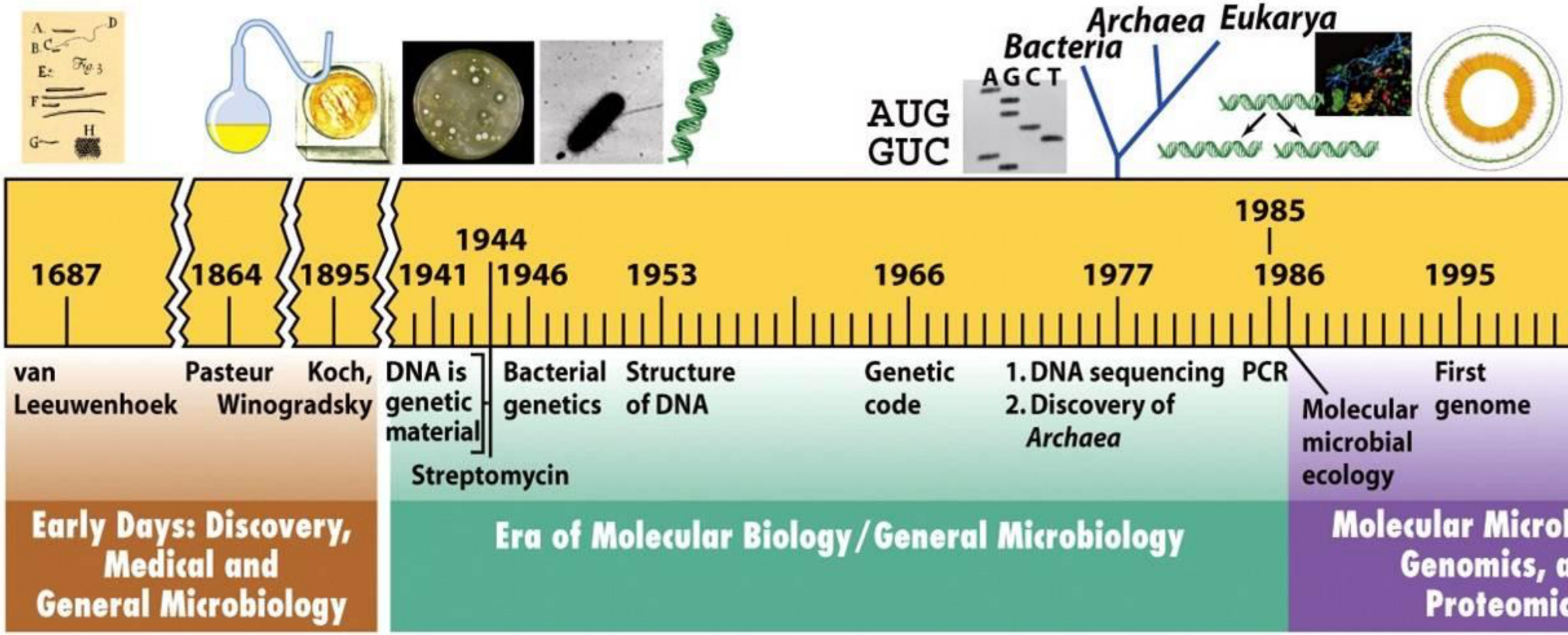


Figure 1-17 Brock Biology of Microorganisms 11/e

An approach for genome analysis based on sequencing and assembly of unselected pieces of DNA from the whole chromosome has been applied to obtain the complete nucleotide sequence (1,830,137 base pairs) of the genome from the bacterium *Haemophilus influenzae* Rd. This approach eliminates the need for initial mapping efforts and is therefore applicable to the vast array of microbial species for which genome maps are unavailable. The *H. influenzae* Rd genome sequence (Genome Sequence DataBase accession number L42023) represents the only complete genome sequence from a free-living organism.

What Genomes Done Before?

- What Genomes Had Been Sequenced?

A prerequisite to understanding the complete biology of an organism is the determination of its entire genome sequence. Several viral and organellar genomes have been completely sequenced. Bacteriophage ϕX174 [5386 base pairs (bp)] was the first to be sequenced, by Fred Sanger and colleagues in 1977 (*1*). Sanger *et al.*

How Were Genomes Being Sequenced?

- How Were Genomes Being Sequenced?

Homo sapiens (*11*). These projects, as well as viral genome sequencing, have been based primarily on the sequencing of clones usually derived from extensively mapped restriction fragments, or λ or cosmid clones. Despite advances in DNA sequencing technology

What is new here?

- What did they do that was new in terms of sequencing?

The computational methods developed to create assemblies from hundreds of thousands of 300- to 500-bp complementary DNA (cDNA) sequences (*13*) led us to test the hypothesis that segments of DNA several megabases in size, including entire microbial chromosomes, could be sequenced rapidly, accurately, and cost-effectively by applying a shotgun sequencing strategy to whole genomes. With this strategy, a single random DNA fragment library may be prepared, and the ends of a sufficient number of randomly selected fragments may be sequenced and assembled to produce the complete genome. We chose the free-living organism *Haemophilus influenzae* Rd as a pilot project because its genome size (1.8 Mb) is typical among bacteria, its G+C base composition (38 percent) is close to that of human, and a physical clone map did not exist.

Table 1. Whole-genome sequencing strategy.

Stage	Description
Random small insert and large insert library construction	Shear genomic DNA randomly to ~2 kb and 15 to 20 kb, respectively
Library plating	Verify random nature of library and maximize random selection of small insert and large insert clones for template production
High-throughput DNA sequencing	Sequence sufficient number of sequence fragments from both ends for 6× coverage
Assembly	Assemble random sequence fragments and identify repeat regions
Gap closure	
Physical gaps	Order all contigs (fingerprints, peptide links, λ clones, PCR) and provide templates for closure
Sequence gaps	Complete the genome sequence by primer walking
Editing	Inspect the sequence visually and resolve sequence ambiguities, including frameshifts
Annotation	Identify and describe all predicted coding regions (putative identifications, starts and stops, role assignments, operons, regulatory regions)

Organisms Chosen

- What is H. Influenzae?
- Why did they choose it?

complete genome. We chose the free-living organism *Haemophilus influenzae* Rd as a pilot project because its genome size (1.8 Mb) is typical among bacteria, its G+C base composition (38 percent) is close to that of human, and a physical clone map did not exist.

Sequencing Details?

- Outline of sequencing approach?

Shotgun Sequencing

- What is shotgun sequencing?
- How does it work?

Shotgun Sequencing

- What is shotgun sequencing?
- How does it work?
- How can you complete a genome this way?

Genome sequencing. The strategy for a shotgun approach to whole genome sequencing is outlined in Table 1. The theory follows from the Lander and Waterman (*14*) application of the equation for the Poisson distribution. The probability that a base is not sequenced is $P_o = e^{-m}$, where m is the sequence coverage. Thus after 1.83 Mb of sequence has been randomly generated for the *H. influenzae* genome ($m = 1$, 1 × coverage), $P_o = e^{-1} = 0.37$ and approximately 37 percent of the genome is unsequenced. Fivefold coverage (approximately 9500 clones sequenced from both insert ends and an average sequence read length of 460 bp) yields $P_o = e^{-5} = 0.0067$, or 0.67 percent unsequenced. If L is genome length and n is the number of random sequence segments done, the total gap length is Le^{-m}, and the average gap size is L/n. Fivefold coverage would leave about 128 gaps averaging about 100 bp in size.

$$P_0 = e^{-m}$$

P_0 = probability a base is not sequenced

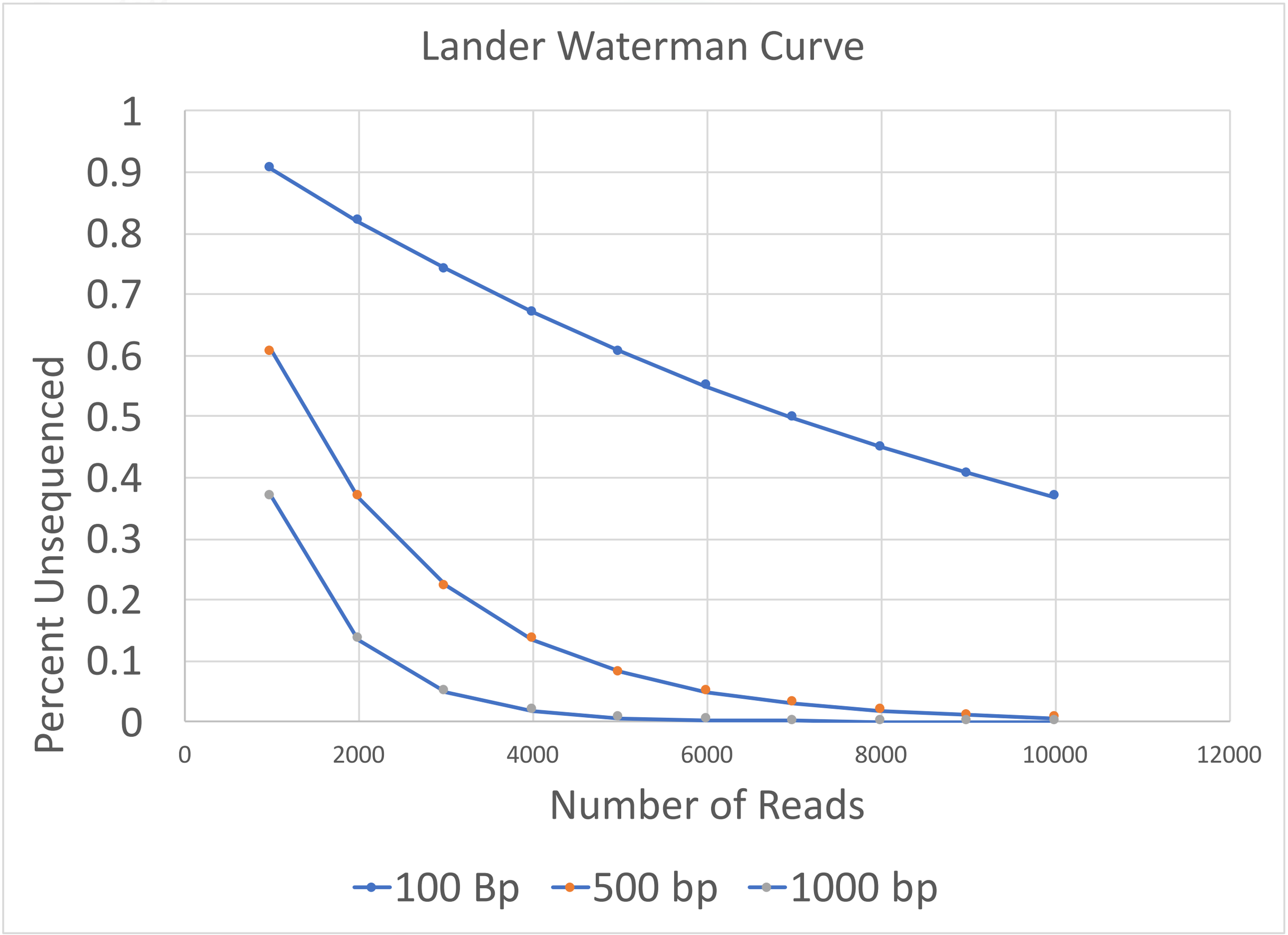
m = coverage

$P_0 = e^{-m}$

P_0 = probability a base is not sequenced

WHY DOES RANDOMNESS MATTER?

Show Curve


Lander Waterman Curve
Percent Unsequenced
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
0
2000
4000
6000
8000
10000
12000
Number of Reads
100 Bp
500 bp
1000 bp

Other issues

- Why do they talk about ESTS?
- Why do they need a database?

Assembly

TIGR ASSEMBLER is the software component that enabled us to assemble the *H. influenzae* genome. It simultaneously clusters and assembles fragments of the genome. In order to obtain the speed necessary to assemble more than 10^4 fragments, the algorithm builds a table of all 10-bp oligonucleotide subsequences to generate a list of potential sequence fragment overlaps. When TIGR ASSEMBLER is used, a single fragment begins the initial contig; to extend the contig, a candidate fragment is chosen with the best overlap based on oligonucleotide content. The current contig and candidate fragment are aligned by a modified version of the Smith-Waterman (*23*) algo-

rithm, which provides for optimal gapped alignments. The contig is extended by the fragment only if strict criteria for the quality of the match are met. The match criteria include the minimum length of overlap, the maximum length of an unmatched end, and the minimum percentage match. The algorithm automatically lowers these criteria in regions of minimal coverage and raises them in regions with a possible repetitive element. The number of potential overlaps for each fragment determines which fragments are likely to fall into repetitive elements. Fragments representing the boundaries of repetitive elements and potentially chimeric fragments are often rejected on the basis of partial mismatches at the needs of alignments and excluded from the contig.

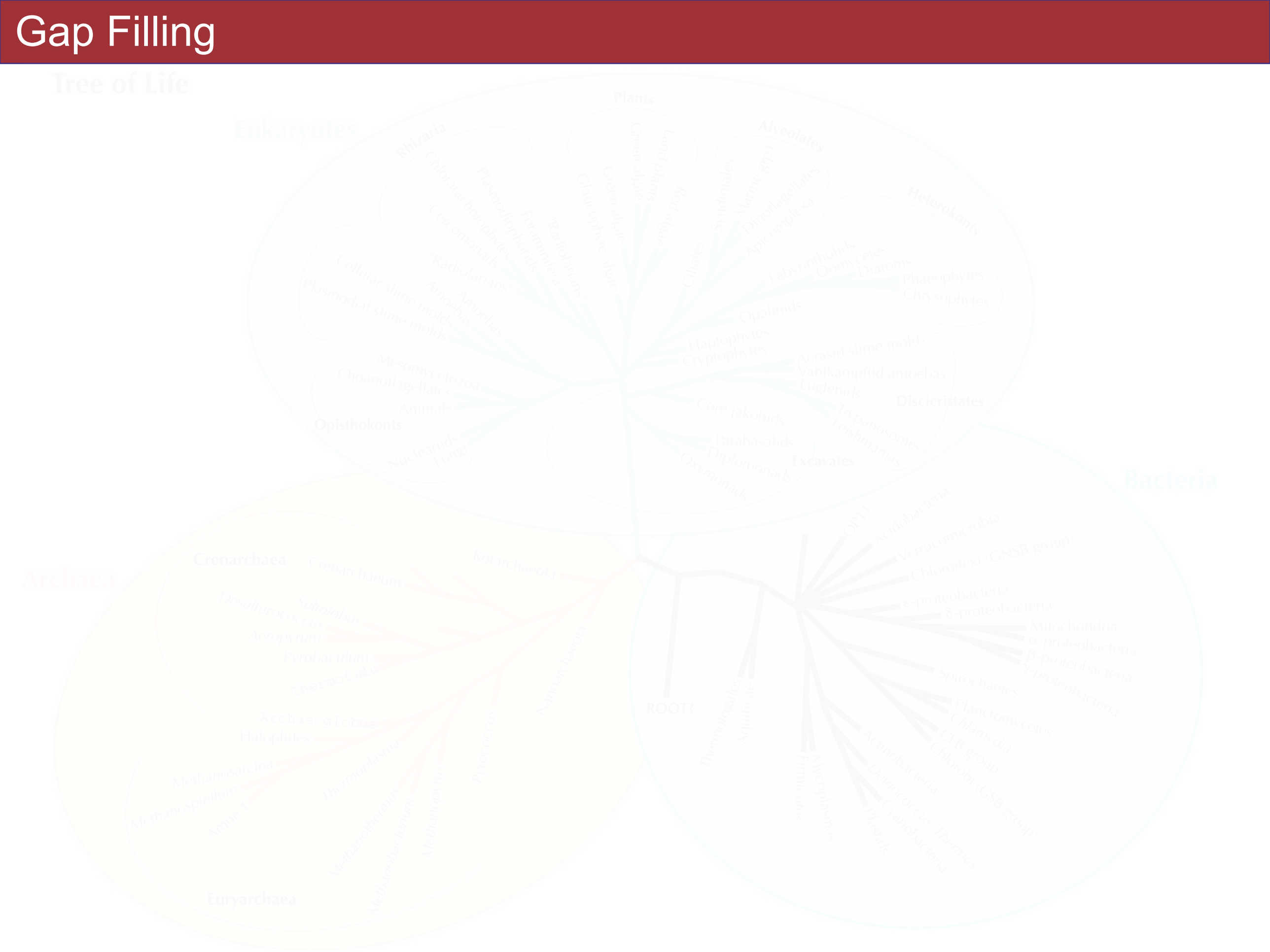
Assembly Method

- TIGR Assembler
- Smith Waterman
- Paired Ends

Assembly Results

- Contigs
- Sequencing Gaps
- Physical Gaps

Gap Filling


Tree of Life
Eukaryotes
Plants
Green algae
Land plants
Red algae
Rhizaria
Chlorarachniophytes
Cercomonads
Plasmodiophorids
Foraminifera
"Radiolarians"
Glaucophyte algae
Ciliates
Syndinians
Alveolates
Marine alveolates
Dinoflagellates
Apicomplexa
Heterokonts
Labyrinthulids
Oomycetes
Diatoms
Phaeophytes
Chrysophytes
Opalinids
Haptophytes
Cryptophytes
Acrasid slime molds
Vahlkampfiid amoebas
Euglenids
Discicristates
Trypanosomes
Leishmanias
Core Jakobids
Parabasalids
Diplomonads
Oxymonads
Excavates
Bacteria
Cellular slime molds
Plasmodial slime molds
Amoebas
Amoebas
Mesomycetozoa
Choanoflagellates
Animals
Opisthokonts
Nucleariids
Fungi
Archaea
Crenarchaea
Crenarchaeum
Korarchaeota
Desulfurococcus
Sulfolobus
Aeropyrum
Pyrobaculum
Thermofilum
Nanoarchaeota
Archaeoglobus
Halophiles
Methanosarcina
Methanospirillum
Thermoplasma
Methanothermus
Methanobacterium
Methanopyrus
Pyrococcus
Euryarchaea
ROOT?
Thermotogales
Aquifex
Cyanobacteria
Plastids
Mitochondria
α-proteobacteria
β-proteobacteria
γ-proteobacteria
ε-proteobacteria
δ-proteobacteria
Spirochaetes
Planctomycetes
Chlamydia
CFB group
Chlorobi (GSB group)
Actinobacteria
Deinococcus/Thermus
Mycoplasmas
Firmicutes
Chloroflexi (GNSB group)
Verrucomicrobia
Acidobacteria
OP11

Gap Filling

- DNA Hybridization
- Peptide Links
- Phage lambda libraries
- PCR

Accuracy Checks?

Accuracy Checks?

- Frameshift in protein sequences vs. Homologs
- Coverage > 1x
- Few ambiguities
- Comparison to H. Influenza sequence in DBs

Cost

Tree of Life
Eukaryotes
Plants
Green algae
Land plants
Red algae
Rhizaria
Chlorarachniophytes
Cercomonads
Plasmodiophorids
Foraminifera
"Radiolarians"
Glaucophyte algae
Alveolates
Ciliates
Syndinians
Marine alveolates
Dinoflagellates
Apicomplexa
Heterokonts
Labyrinthulids
Oomycetes
Diatoms
Phaeophytes
Chrysophytes
Opalinids
Haptophytes
Cryptophytes
Discicristates
Acrasid slime molds
Vahlkampfiid amoebas
Euglenids
Trypanosomes
Leishmanias
Core Jakobids
Parabasalids
Diplomonads
Oxymonads
Excavates
Bacteria
Cellular slime molds
Plasmodial slime molds
Amoebas
Amoebas
Mesomycetozoa
Choanoflagellates
Animals
Opisthokonts
Nucleariids
Fungi
Archaea
Crenarchaea
Crenarchaeum
Korarchaeota
Desulfurococcus
Sulfolobus
Aeropyrum
Pyrobaculum
Thermofilum
Nanoarchaeota
Archaeoglobus
Halophiles
Methanosarcina
Methanospirillum
Thermoplasma
Methanothermus
Methanobacterium
Methanopyrus
Pyrococcus
Euryarchaea
ROOT?
Thermotogales
Aquifex
Gram+
Mycoplasmas
Actinobacteria
Deinococcus/Thermus
Cyanobacteria
Plastids
Chlamydia
CFB group
Chlorobi (GSB group)
Spirochaetes
Planctomycetes
Mitochondria
α-proteobacteria
β-proteobacteria
γ-proteobacteria
ε-proteobacteria
δ-proteobacteria
Chloroflexi (GNSB group)
Verrucomicrobia
Acidobacteria
OP11

Cost

- 0.48$ / finished bp

Annotation

- Structural
- Functional

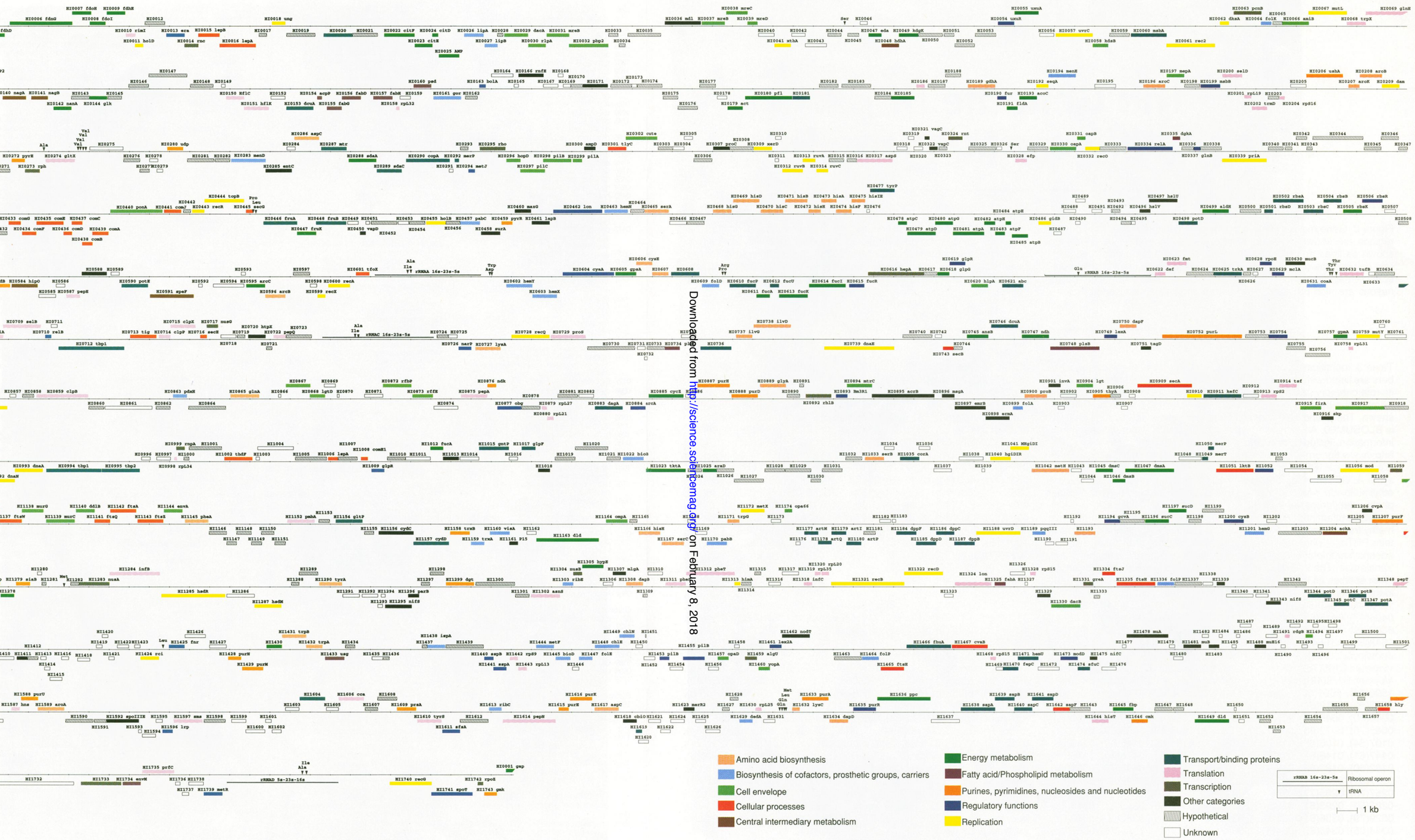
Annotation

- Structural
- Functional

Table 2. Summary of features of whole-genome sequencing of *H. influenzae* Rd.

Description	Number
Double-stranded templates	19,687
Forward-sequencing reactions (M13-21 primer)	19,346
Successful (%)	16,240 (84)
Average edited read length (bp)	485
Reverse sequencing reactions (M13RP1 primer)	9,297
Successful (%)	7,744 (83)
Average edited read length (bp)	444
Sequence fragments in random assembly	24,304
Total base pairs	11,631,485
Contigs	140
Physical gap closure	42
PCR	37
Southern analysis	15
λ clones	23
Peptide links	2
Terminator sequencing reactions*	3,530
Successful (%)	2,404 (68)
Average edited read length (bp)	375
Genome size (bp)	1,830,137
G+C content (%)	38
rRNA operons	6
rrnA, rrnC, rrnD (spacer region) (bp)	723
rrnB, rrnE, rrnF (spacer region) (bp)	478
tRNA genes identified	54
Number of predicted coding regions	1,743
Unassigned role (%)	736 (42)
No database match	389
Match hypothetical proteins	347
Assigned role (%)	1,007 (58)
Amino acid metabolism	68 (6.8)
Biosynthesis of cofactors, prosthetic groups, and carriers	54 (5.4)
Cell envelope	84 (8.3)
Cellular processes	53 (5.3)
Central intermediary metabolism	30 (3.0)
Energy metabolism	105 (10.4)
Fatty acid and phospholipid metabolism	25 (2.5)
Purines, pyrimidines, nucleosides and nucleotides	53 (5.3)
Regulatory functions	64 (6.3)
Replication	87 (8.6)
Transcription	27 (2.7)
Translation	141 (14.0)
Transport and binding proteins	123 (12.2)
Other	93 (9.2)

*Includes gap closure, walks on rRNA repeats, random end-sequencing of λ clones for assembly confirmation, and alternative reactions for ambiguity resolution.



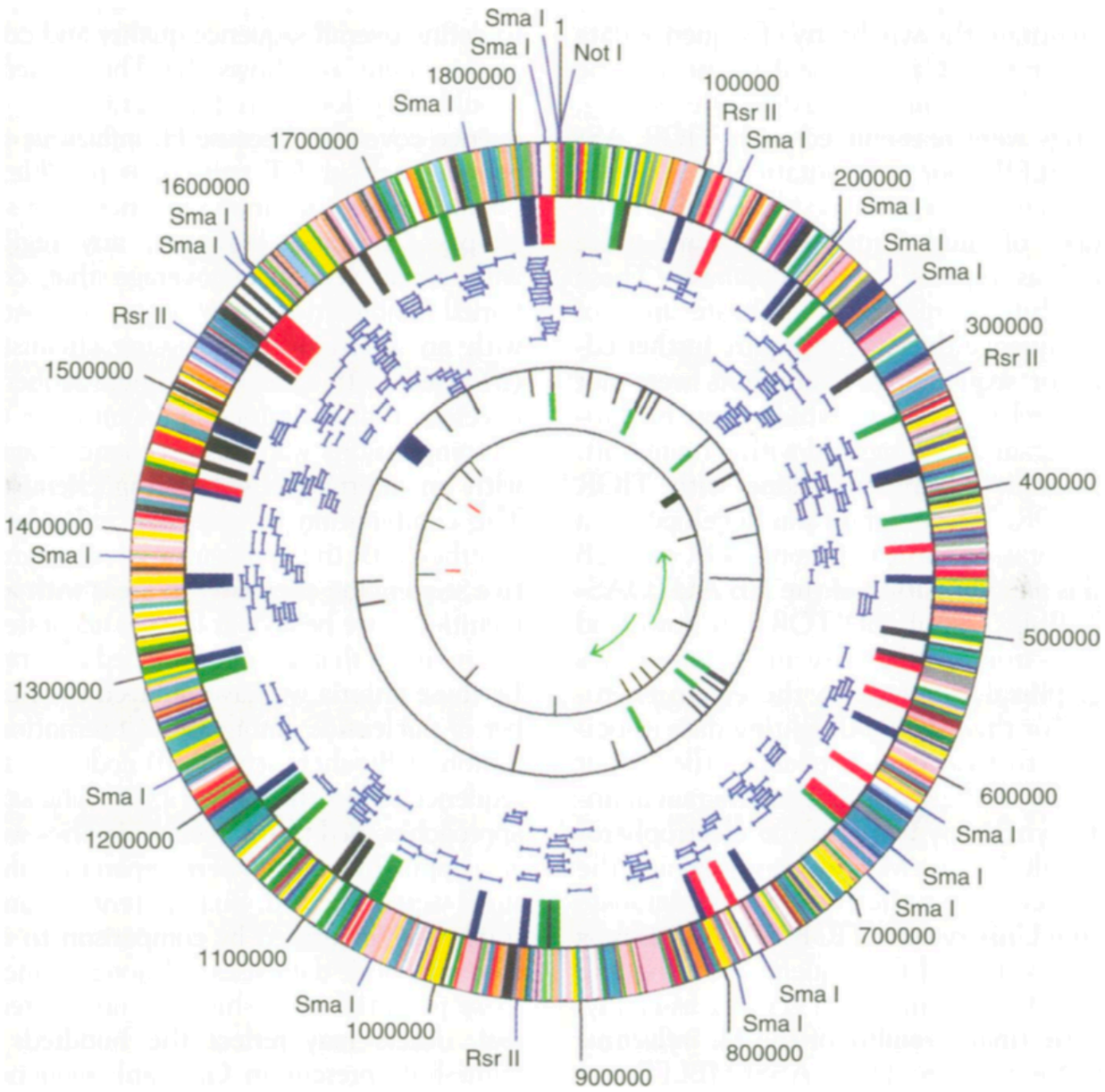


Fig. 1. A circular representation of the *H. influenzae* Rd chromosome illustrating the location of each predicted coding region containing a database match as well as selected global features of the genome. Outer perimeter: The location of the unique Not I restriction site (designated as nucleotide 1), the Rsr II sites, and the Sma I sites. Outer concentric circle: Coding regions for which a gene identification was made. Each coding region location is classified as to role according to the color code in Fig. 2. Second concentric circle: Regions of high G+C content (>42 percent, red; >40 percent, blue) and high A+T content (>66 percent, black; >64 percent, green). Third concentric circle: Coverage by λ clones (blue). More than 300 λ clones were sequenced from each end to confirm the overall structure of the genome and identify the six ribosomal operons. Fourth concentric circle: The locations of the six ribosomal operons (green), the tRNAs (black) and the cryptic mu-like prophage (blue). Fifth concentric circle: Simple tandem repeats. The locations of the following repeats are shown: CTGGCT, GTCT, ATT, AATGGC, TTGA, TTGG, TTTA, TTATC ,TGAC, TCGTC, AACC, TTGC, CAAT, CCAA. The putative origin of replication is illustrated by the outward pointing arrows (green) originating near base 603,000. Two potential termination sequences are shown near the opposite midpoint of the circle (red).

Table 4. Two-component systems in *H. influenzae* Rd. ID, identity; Sim, similarity.

Identification number	Location	Best match*	Id (%)	Sim (%)	Length (bp)
		Sensors			
HI0220	239,378	*arcB*	39.5	63.9	200
HI0267	299,541	*narQ*	38.1	68.0	562
HI1707	1,781,143	*basS*	27.7	51.5	250
HI1378	1,475,017	*phoR*	38.1	61.6	280
		Regulators			
HI0726	777,934	*narP*	59.3	77.0	209
HI0837	887,011	*cpxR*	51.9	73.0	229
HI0884	936,624	*arcA*	77.2	87.8	236
HI1379	1,475,502	*phoB*	52.9	71.4	228
HI1708	1,781,799	*basR*	43.5	59.3	219

*In all cases, the best match was to a gene of *E. coli*.

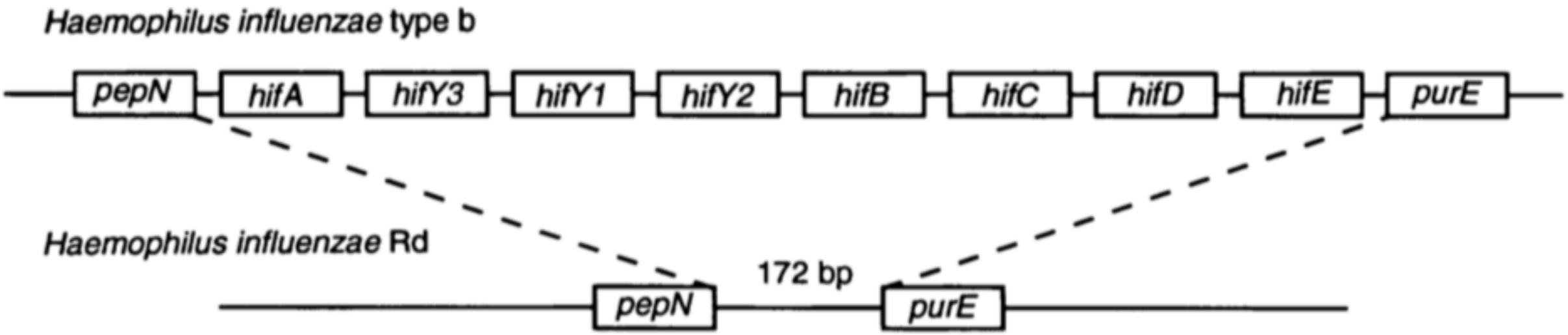


Fig. 3. A comparison of the region of the *H. influenzae* chromosome containing the eight genes of the fimbrial gene cluster present in *H. influenzae* type b and the same region in *H. influenzae* Rd. The region is flanked by *pepN* and *purE* in both organisms. However, in the noninfectious Rd strain the eight genes of the fimbrial gene cluster have been excised. A 172-bp spacer region is located in this region in the Rd strain and continues to be flanked by the *pepN* and *purE* genes.

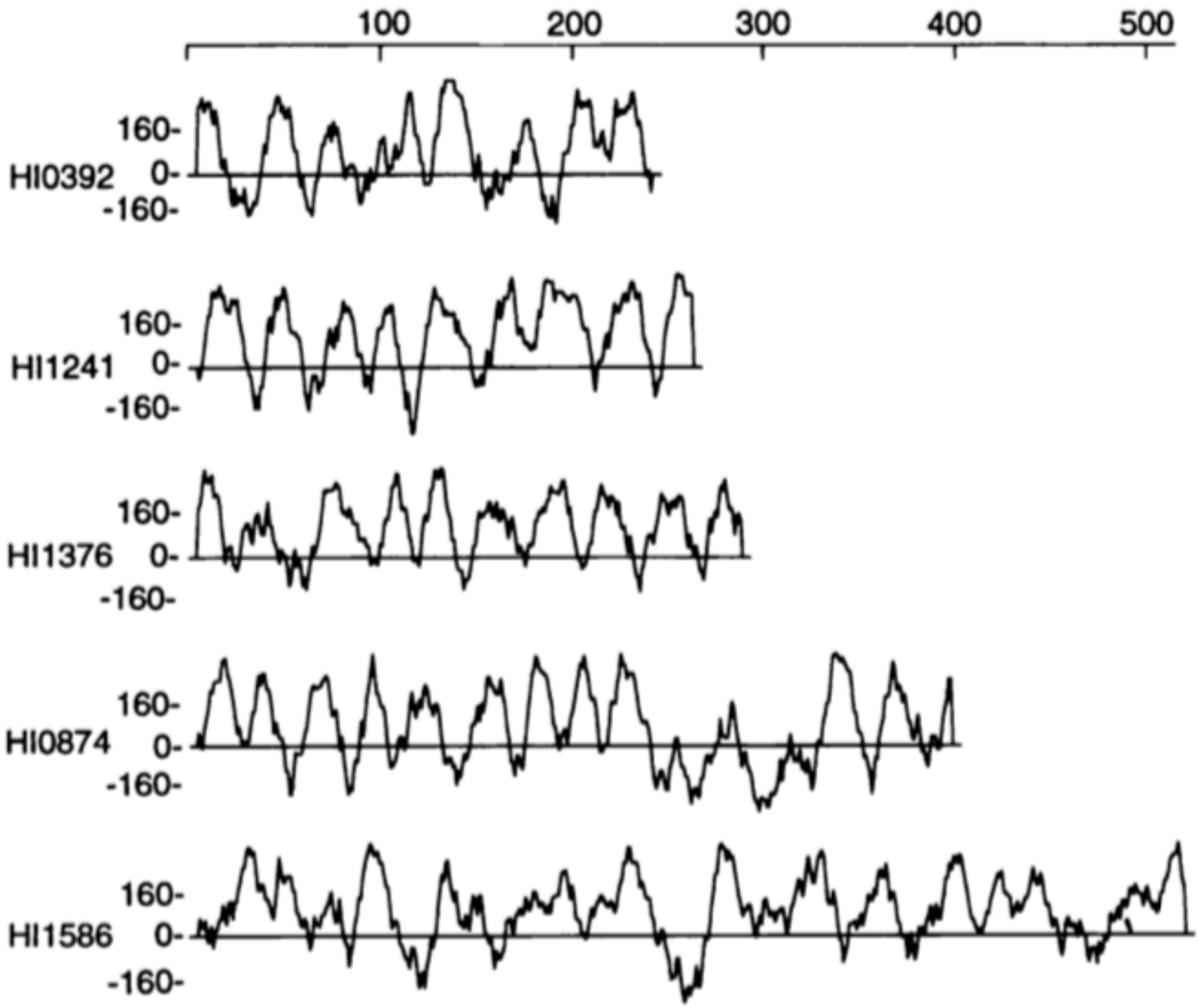


Fig. 4. Hydrophobicity analysis of five potential channel proteins. The amino acid sequences of five predicted coding regions that do not display similarity with known peptide sequences (GenBank release 87), each exhibit multiple hydrophobic domains that are characteristic of channel-forming proteins. The predicted coding region sequences were analyzed by the Kyte-Doolittle algorithm (*46*) (with a range of 11 residues) with the GENEWORKS software package (Intelligenetics).

Table 1. Whole-genome sequencing strategy.

Stage	Description
Random small insert and large insert library construction	Shear genomic DNA randomly to ~2 kb and 15 to 20 kb, respectively
Library plating	Verify random nature of library and maximize random selection of small insert and large insert clones for template production
High-throughput DNA sequencing	Sequence sufficient number of sequence fragments from both ends for 6× coverage
Assembly	Assemble random sequence fragments and identify repeat regions
Gap closure	
Physical gaps	Order all contigs (fingerprints, peptide links, λ clones, PCR) and provide templates for closure
Sequence gaps	Complete the genome sequence by primer walking
Editing	Inspect the sequence visually and resolve sequence ambiguities, including frameshifts
Annotation	Identify and describe all predicted coding regions (putative identifications, starts and stops, role assignments, operons, regulatory regions)

Tree of Life

Eukaryotes
Plants
Land plants
Green algae
Red algae
Glaucophyte algae
Rhizaria
Chlorarachniophytes
Cercomonads
Plasmodiophorids
Foraminifera
"Radiolarians"
Cellular slime molds
Plasmodial slime molds
Amoebas
Amoebas
"Radiolarians"
Alveolates
Ciliates
Syndinians
Marine alveolates
Dinoflagellates
Apicomplexa
Heterokonts
Labyrinthulids
Oomycetes
Diatoms
Phaeophytes
Chrysophytes
Opalinids
Haptophytes
Cryptophytes
Acrasid slime molds
Vahlkampfiid amoebas
Euglenids
Discicristates
Core Jakobids
Trypanosomes
Leishmanias
Parabasalids
Diplomonads
Excavates
Mesomycetozoa
Choanoflagellates
Animals
Opisthokonts
Nucleariids
Fungi
Bacteria
Archaea
Crenarchaea
Crenarchaeum
Korarchaeota
Desulfurococcus
Sulfolobus
Aeropyrum
Pyrobaculum
Thermofilum
Nanoarchaeota
Archaeoglobus
Halophiles
Methanosarcina
Methanospirillum
Thermoplasma
Methanothermus
Methanobacterium
Methanopyrus
Pyrococcus
Euryarchaea
ROOT?
Thermotogales
Aquifex
Gram+
Mycoplasmas
Thermus
Acidobacteria
Verrucomicrobia
Chloroflexi (GNSB group)
ε-proteobacteria
δ-proteobacteria
Mitochondria
α-proteobacteria
β-proteobacteria
γ-proteobacteria
Spirochaetes
Planctomycetes
Chlamydia
CFB group
Chlorobi (GSB group)
Actinobacteria
Deinococcus/Thermus
Cyanobacteria
Plastids

Tree of Life

Eukaryotes
Plants
Land plants
Green algae
Red algae
Glaucophyte algae
Rhizaria
Chlorarachniophytes
Cercomonads
Plasmodiophorids
Foraminifera
"Radiolarians"
Cellular slime molds
Plasmodial slime molds
Amoebas
Amoebas
"Radiolarians"
Alveolates
Ciliates
Syndinians
Marine alveolates
Dinoflagellates
Apicomplexa
Heterokonts
Labyrinthulids
Oomycetes
Diatoms
Phaeophytes
Chrysophytes
Opalinids
Haptophytes
Cryptophytes
Acrasid slime molds
Vahlkampfiid amoebas
Euglenids
Discicristates
Core Jakobids
Trypanosomes
Leishmanias
Parabasalids
Diplomonads
Excavates
Mesomycetozoa
Choanoflagellates
Animals
Opisthokonts
Nucleariids
Fungi
Bacteria
Archaea
Crenarchaea
Crenarchaeum
Korarchaeota
Desulfurococcus
Sulfolobus
Aeropyrum
Pyrobaculum
Thermofilum
Nanoarchaeota
Archaeoglobus
Halophiles
Methanosarcina
Methanospirillum
Thermoplasma
Methanothermus
Methanobacterium
Methanopyrus
Pyrococcus
Euryarchaea
ROOT?
Thermotogales
Aquifex
Gram+
Mycoplasmas
Thermus
Acidobacteria
Verrucomicrobia
Chloroflexi (GNSB group)
ε-proteobacteria
δ-proteobacteria
Mitochondria
α-proteobacteria
β-proteobacteria
γ-proteobacteria
Spirochaetes
Planctomycetes
Chlamydia
CFB group
Chlorobi (GSB group)
Actinobacteria
Deinococcus/Thermus
Cyanobacteria
Plastids