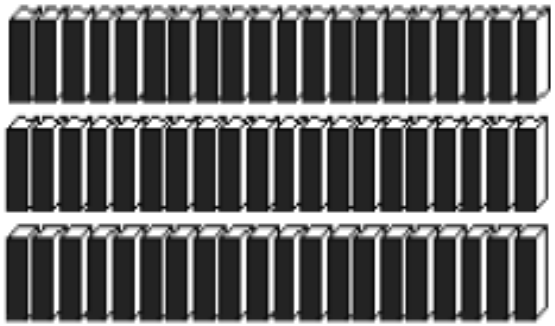
Large-Scale DNA Sequencing

Current Topics in Genome Analysis, 2000

Jeff Touchman, Ph.D.
Director, Sequence Production Group
NIH Intramural Sequencing Center
Tel: 301-435-6156
Fax: 301-435-6170
Email: jefft@nhgri.nih.gov

Genome Sizes

Human Genome
Mouse Genome



~3,000,000,000 bp

Fruit Fly Genome

~160,000,000 bp

Nematode Genome

~100,000,000 bp

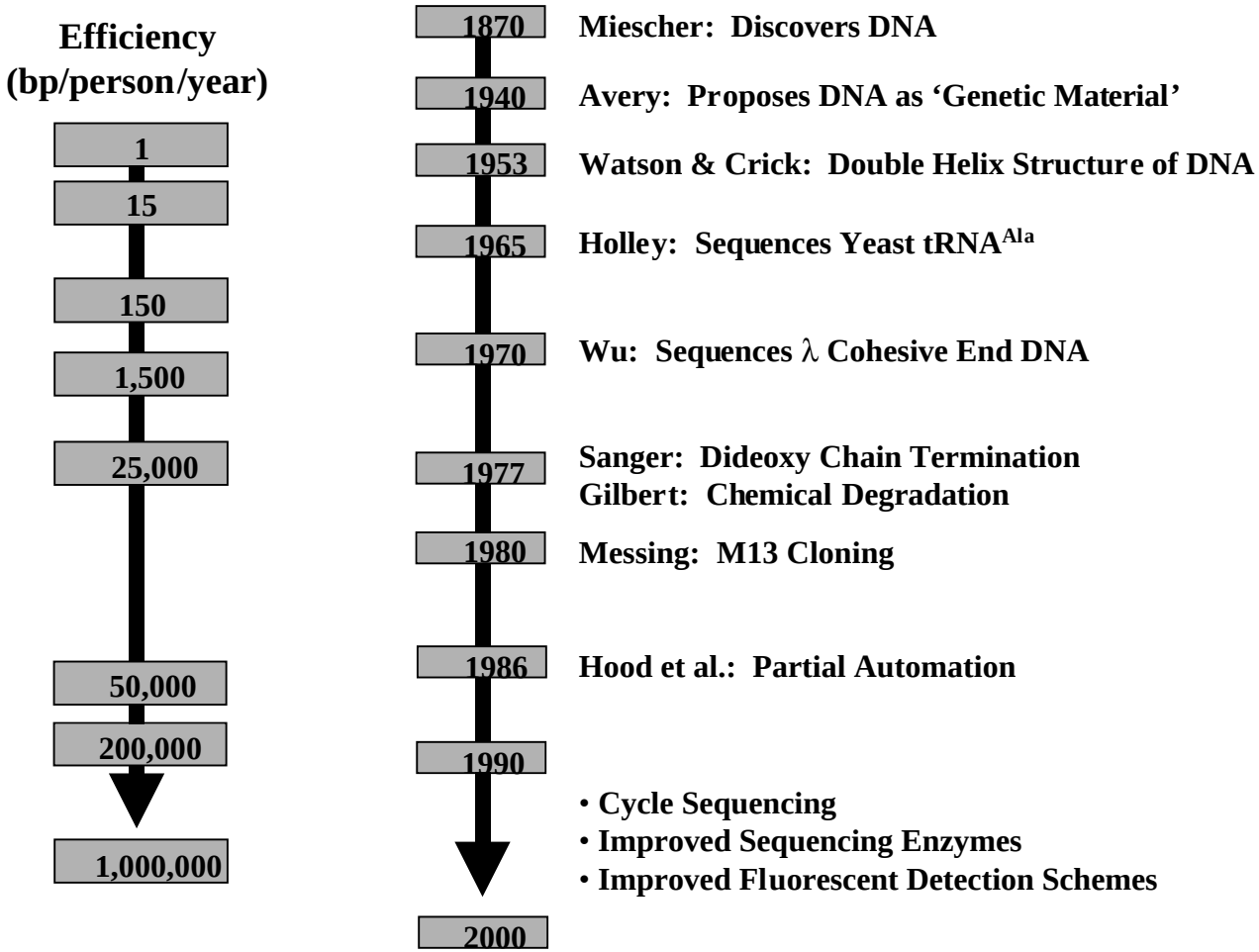
Yeast Genome

~15,000,000 bp

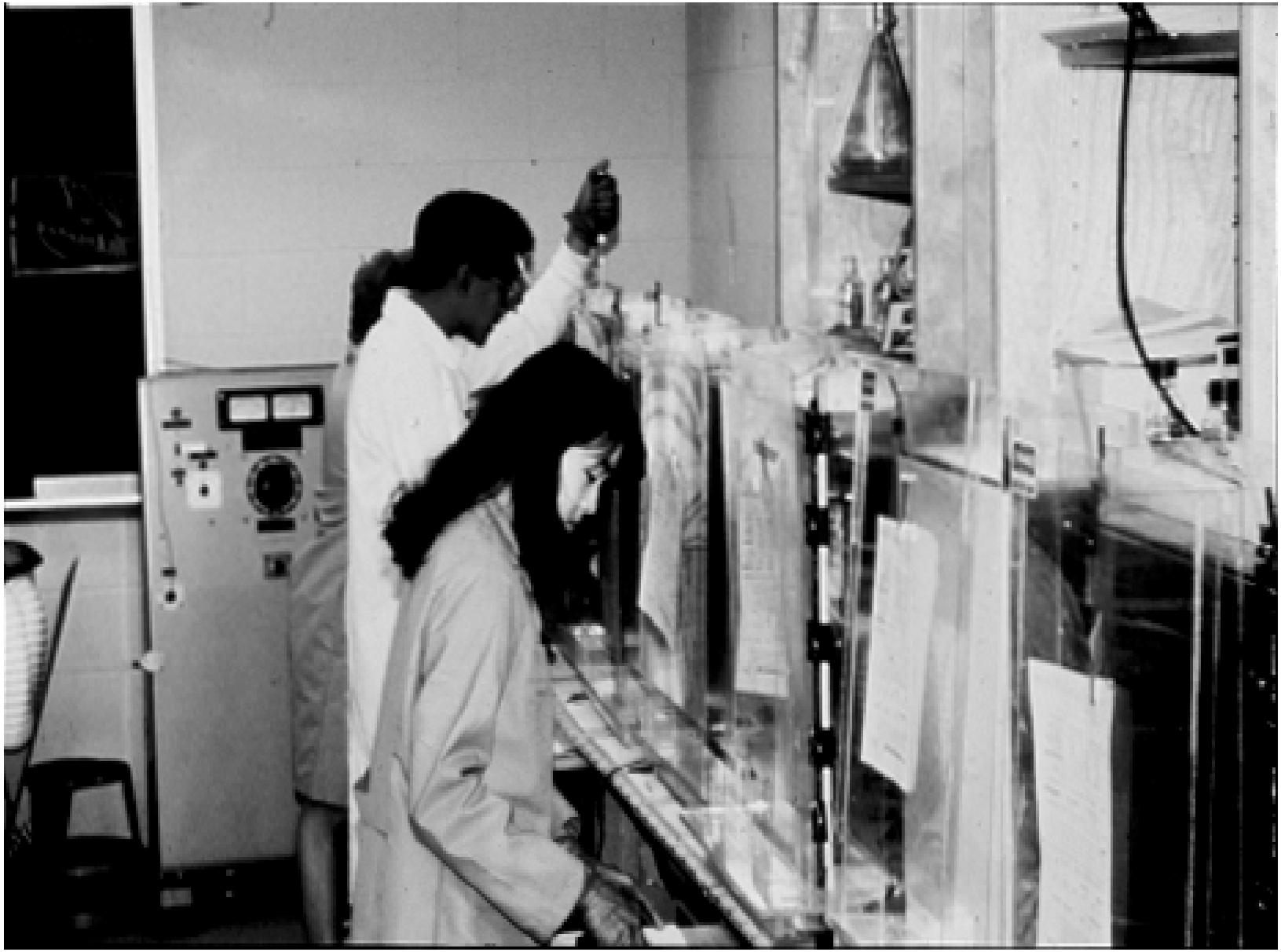
***E. coli* Genome**

~5,000,000 bp

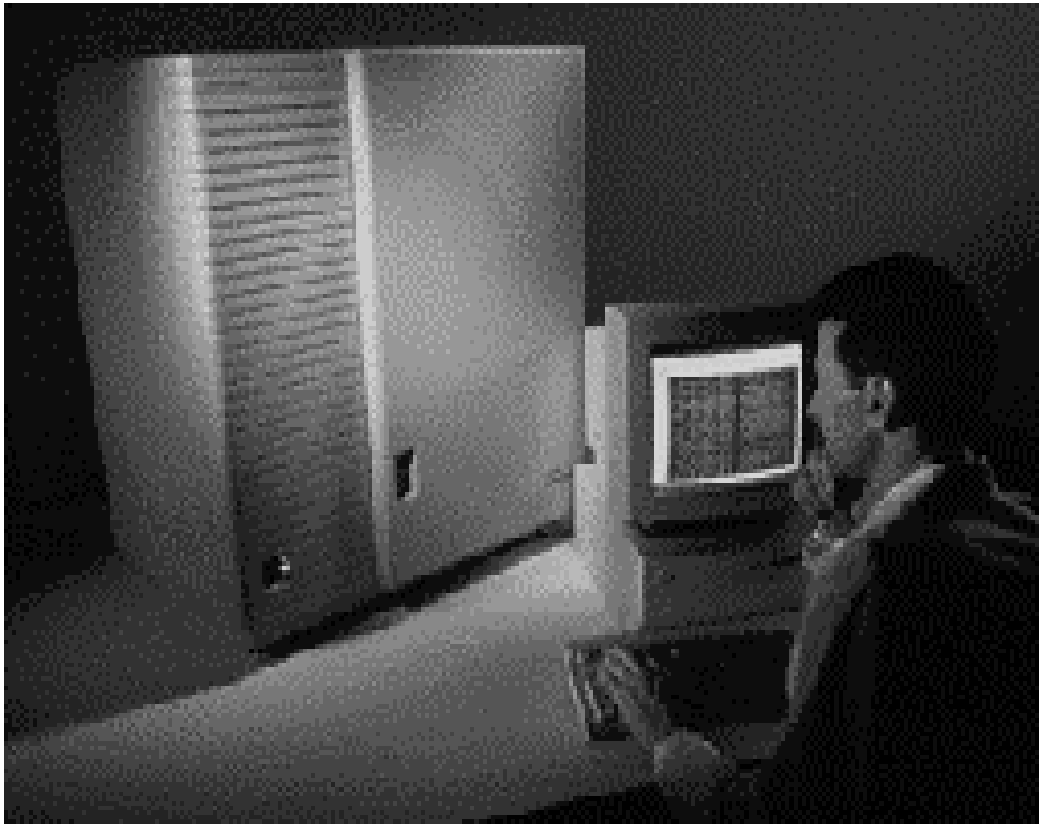
History of DNA Sequencing



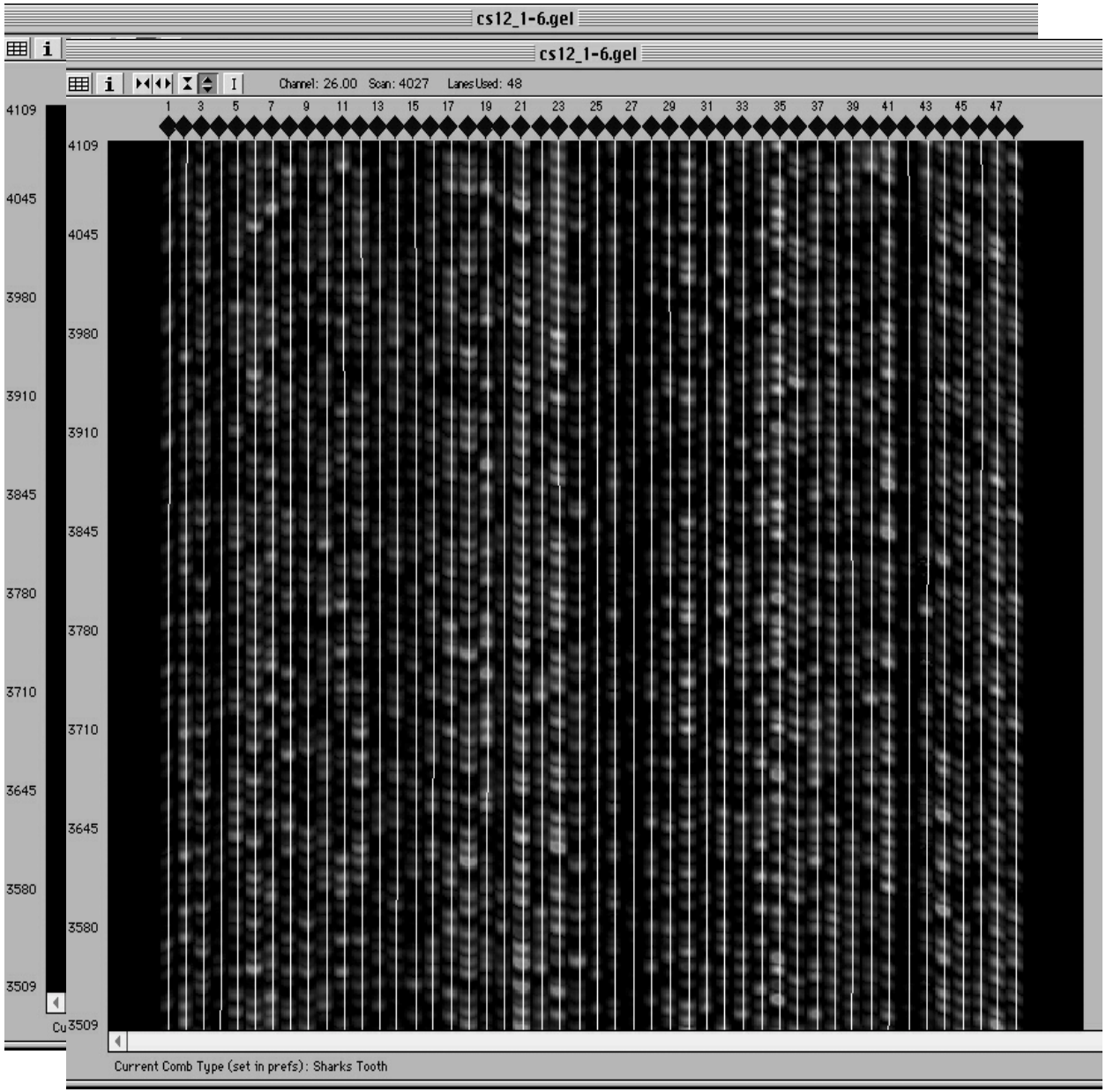
Radioactive Sequencing



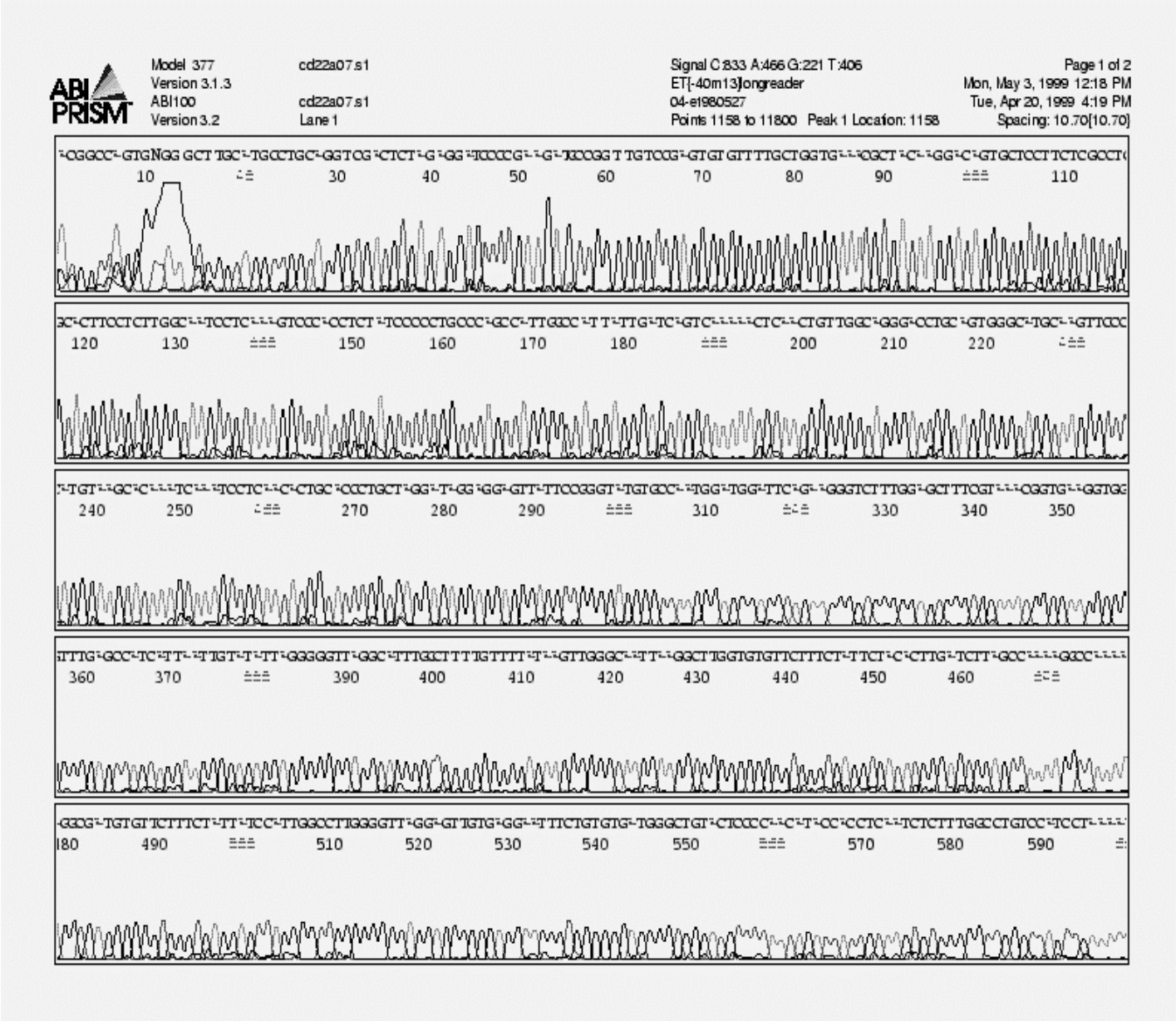
Perkin Elmer/Applied Biosystems 377



Fluorescent DNA Sequencing: Lane Tracking

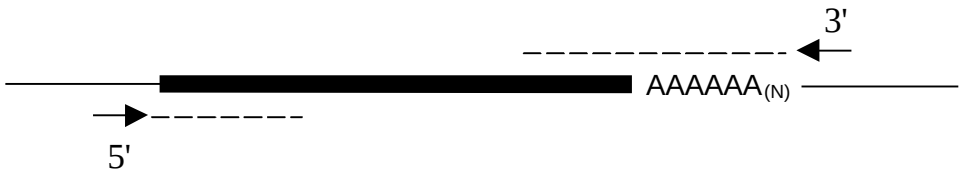


Fluorescent DNA Sequence Trace



Expressed-Sequence Tags (ESTs)

- **Single-Pass Sequence of Random cDNA Clone**
- **Often from Normalized cDNA Libraries**



- **3' ESTs More Likely to be Unique Among Gene Family Members**
- **5' ESTs More Likely to Yield Homology Information Indicative of Gene Function**

Publicly Available ESTs

dbEST Summary - Netscape

File Edit View Go Communicator Help

Bookmarks Location: http://www.ncbi.nlm.nih.gov/dbEST/dbEST_summary.html

NCBI dbEST BLAST Entrez

dbEST release 020599

Summary by Organism - February 5, 1999

Number of public entries: 2,106,690

Homo sapiens (human)	1,252,762
Mus musculus + domesticus (mouse)	404,745
Rattus sp. (rat)	87,288
Caenorhabditis elegans (nematode)	72,568
Drosophila melanogaster (fruit fly)	59,769
Arabidopsis thaliana (thale cress)	37,667
Oryza sativa (rice)	35,215
Brugia malayi (parasitic nematode)	16,642
Emericella nidulans	12,998
Dictyostelium discoideum	10,700
Toxoplasma gondii	10,676
Danio rerio (zebrafish)	9.616

Document: Done

Publicly Available ESTs

dbEST Summary - Netscape

File Edit View Go Communicator Help

NCBI

dbEST: database of "Expressed Sequence Tags"

dbEST release 070700

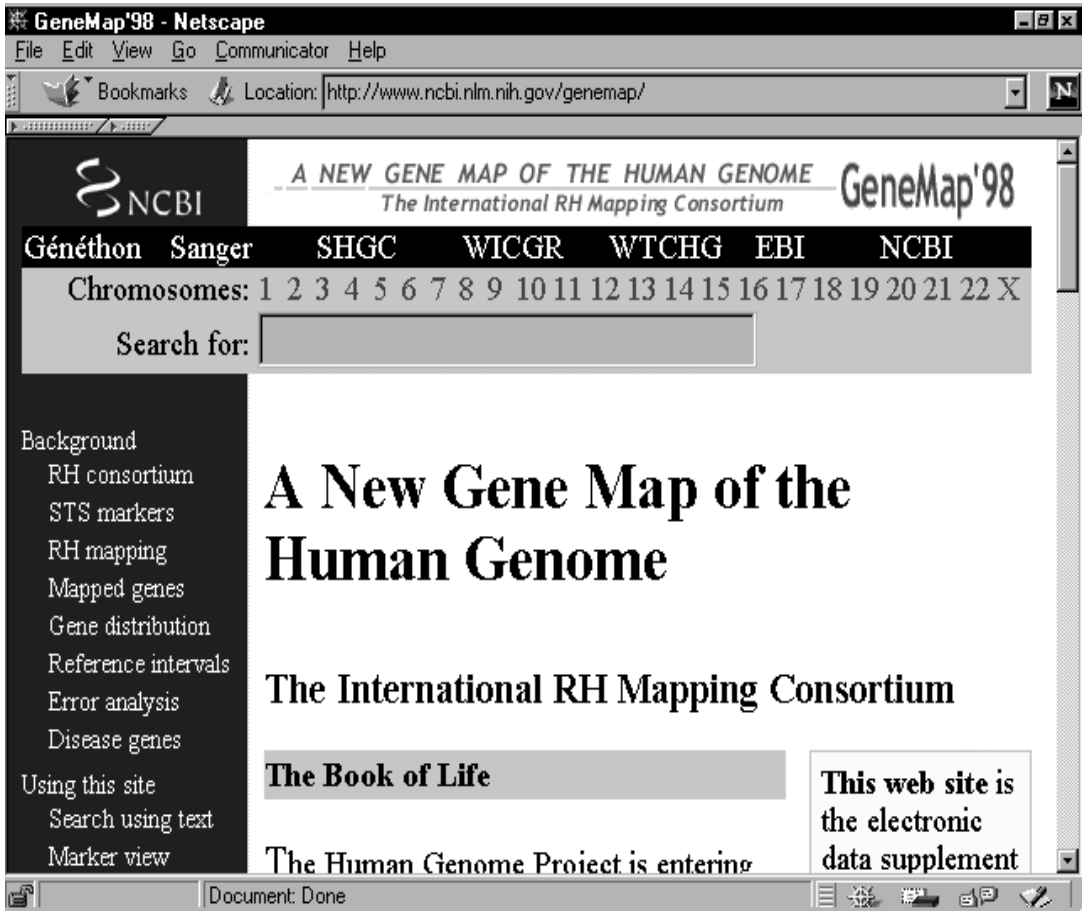
Summary by Organism - July 7, 2000

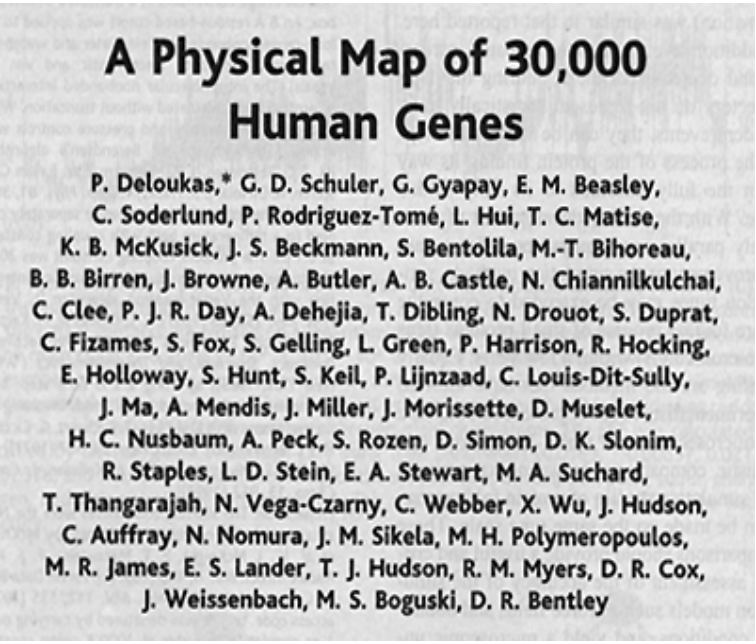
Number of public entries: 4,800,090

Homo sapiens (human)	2,121,173
Mus musculus + domesticus (mouse)	1,274,400
Rattus sp. (rat)	187,828
Caenorhabditis elegans (nematode)	101,252
Arabidopsis thaliana (thale cress)	100,998
Drosophila melanogaster (fruit fly)	90,777
Glycine max (soybean)	89,288
Lycopersicon esculentum (tomato)	72,648
Danio rerio (zebrafish)	71,186
Zea mays (maize)	63,626
Oryza sativa (rice)	59,960
Bos taurus (cattle)	48,003
Medicago truncatula (barrel medic)	34,352
Xenopus laevis (African clawed frog)	30,262
Sus scrofa (pig)	28,317
Lotus japonicus	25,946
Neurospora crassa	24,626
Brugia malayi (parasitic nematode)	22,121

Document: Done

RH Mapping-Based Gene Map





A Physical Map of 30,000 Human Genes

P. Deloukas,* G. D. Schuler, G. Gyapay, E. M. Beasley, C. Soderlund, P. Rodriguez-Tomé, L. Hui, T. C. Matise, K. B. McKusick, J. S. Beckmann, S. Bentolila, M.-T. Bihoreau, B. B. Birren, J. Browne, A. Butler, A. B. Castle, N. Chiannilkulchai, C. Clee, P. J. R. Day, A. Dehejia, T. Dibling, N. Drouot, S. Duprat, C. Fizames, S. Fox, S. Gelling, L. Green, P. Harrison, R. Hocking, E. Holloway, S. Hunt, S. Keil, P. Lijnzaad, C. Louis-Dit-Sully, J. Ma, A. Mendis, J. Miller, J. Morissette, D. Muselet, H. C. Nusbaum, A. Peck, S. Rozen, D. Simon, D. K. Slonim, R. Staples, L. D. Stein, E. A. Stewart, M. A. Suchard, T. Thangarajah, N. Vega-Czarny, C. Webber, X. Wu, J. Hudson, C. Auffray, N. Nomura, J. M. Sikela, M. H. Polymeropoulos, M. R. James, E. S. Lander, T. J. Hudson, R. M. Myers, D. R. Cox, J. Weissenbach, M. S. Boguski, D. R. Bentley

***Science* 282:744-746, 1998**

The Next Challenge with cDNAs

- **Construction of Full-Length cDNA Libraries**
- **Identification of Complete Sets of Full-Length cDNA Clones**
- **Sequencing of Complete Sets of Full-Length cDNA Clones**

Mammalian
Gene
Collection

VIEWPOINT

The Mammalian Gene Collection

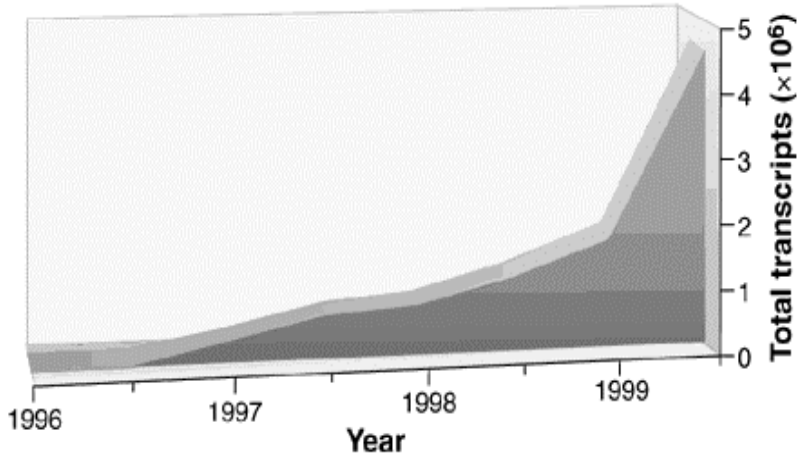
Robert L. Strausberg,[1] Elise A. Feingold,[2] Richard D. Klausner,[1*] Francis S. Collins,[2*]

Science 286:455-457, 1999

SAGE

Serial Analysis of Gene Expression

Designed to take advantage of high-throughput sequencing technology to obtain a quantitative profile of gene expression



Velculescu VE et al. *Science* 270: 484-487, 1995

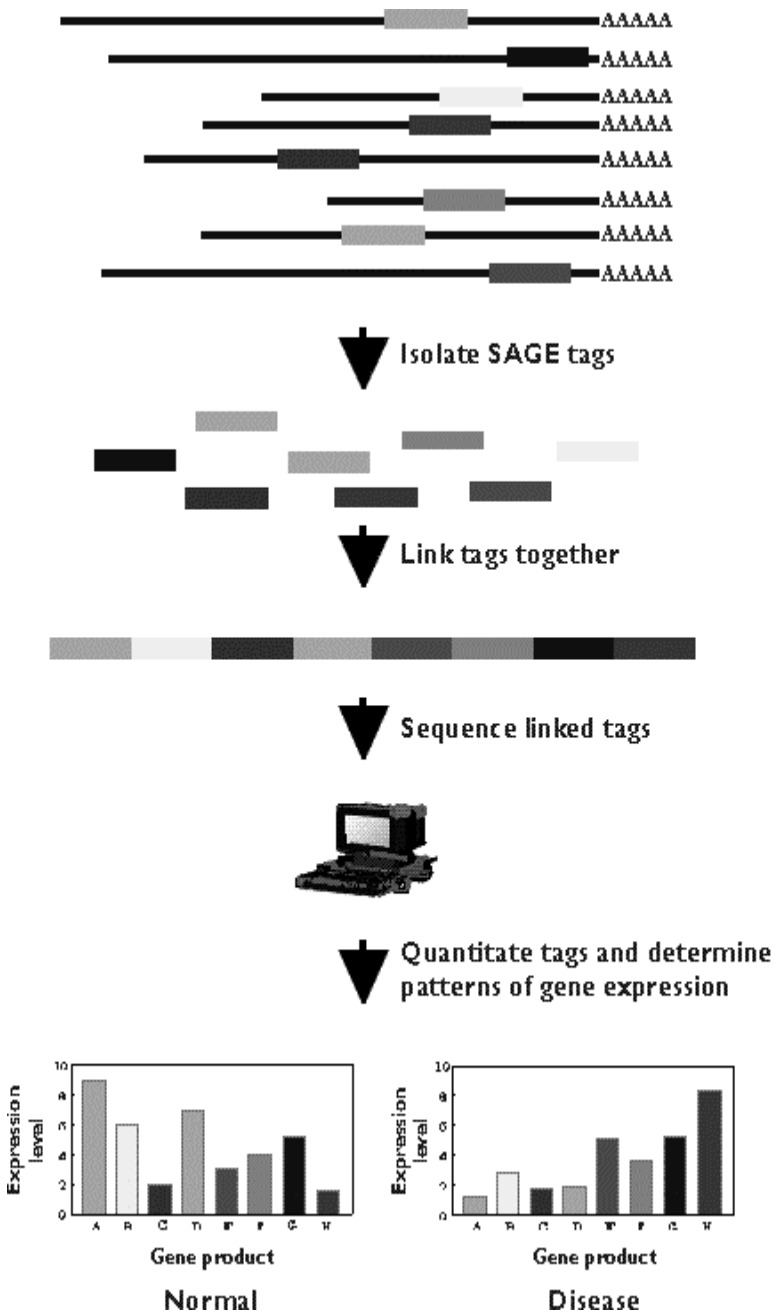
www.sagenet.org
www.ncbi.nlm.nih.gov/SAGE/

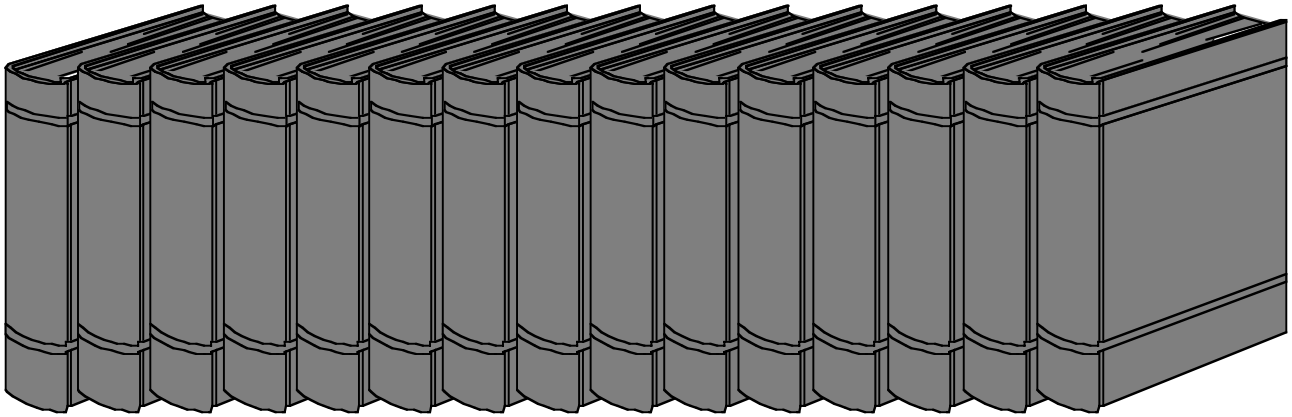
SAGE Principles

1) A short sequence tag (10-14bp) contains sufficient information to uniquely identify a transcript provided that the tag is obtained from a unique position within each transcript

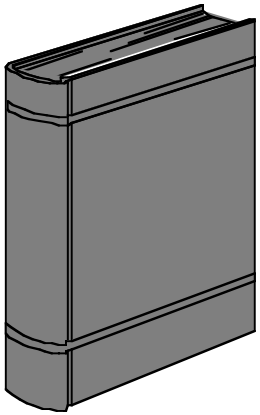
2) Sequence tags can be linked together to from long serial molecules that can be cloned and sequenced

3) Quantitation of the number of times a particular tag is observed provides the expression level of the corresponding transcript





Genome
(~3000 Mb)



Chromosome
(~130 Mb)

BAC
(~0.1-0.2 Mb)

“Sequence-Ready Maps”

Genomic Sequencing: Strategies

- **Transposon-Mediated Sequencing**

  Refined within Drosophila Sequencing Effort

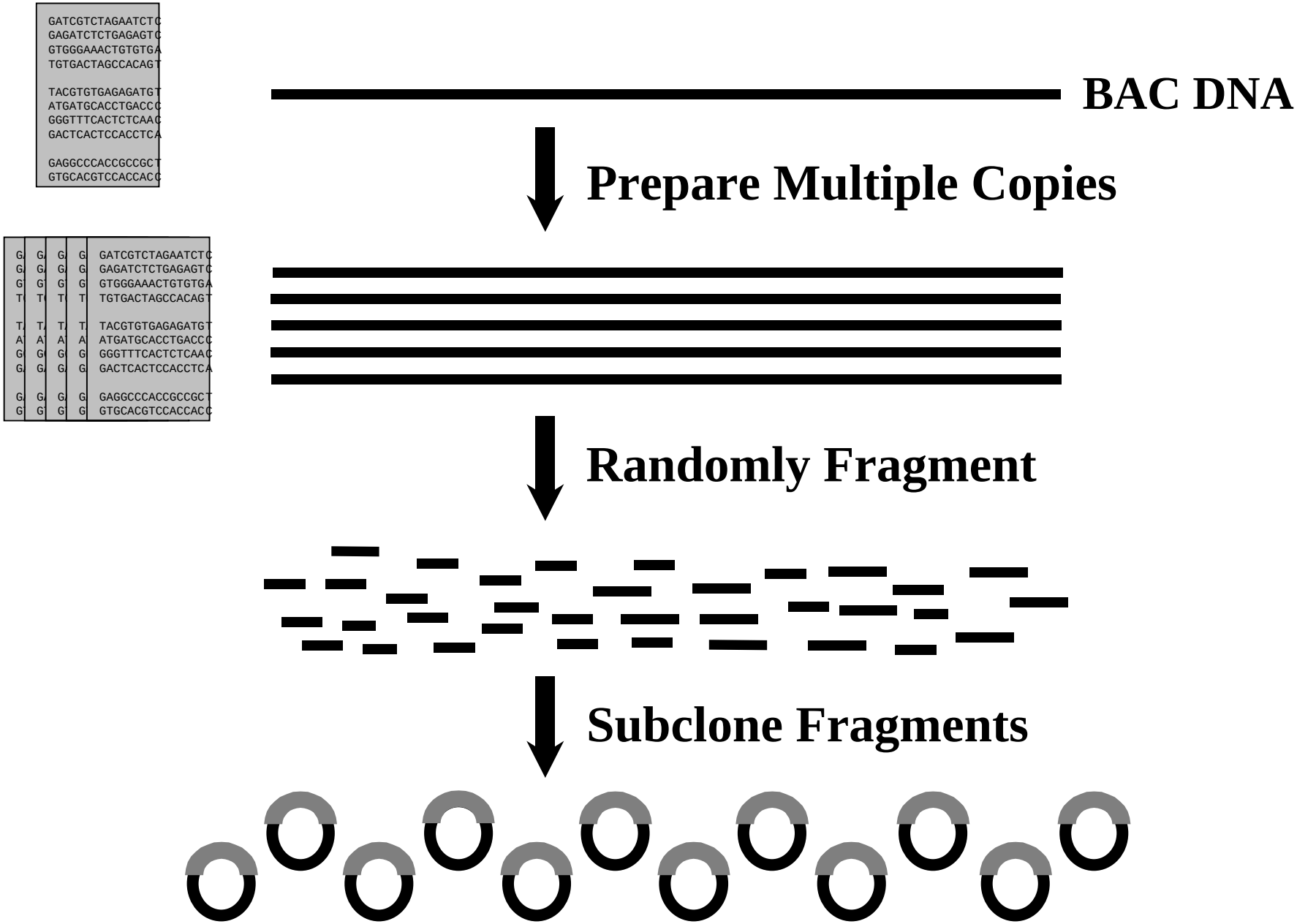
  Kimmel et al., *Genome Analysis*
  Vol. 1 (CSHL Press)

- **Shotgun Sequencing**

  Refined within Nematode Sequencing Effort

  Wilson & Mardis, *Genome Analysis*
  Vol. 1 (CSHL Press)

Subclone Construction



Poisson calculations

The sequencing strategy for the shotgun approach follows the Lander and Waterman application of the Poisson distribution

The probability a base is not sequenced is given by:

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

Where:

- **c = fold sequence coverage (c=LN/G),**
- **LN = # bases sequenced, i.e. L = average sequencing read length and N = # reads**
- **G = target sequence length**
- **e = 2.718 (e=2.718281828459)**

Fold Coverage	$P_0=e^{-c}$	% not sequenced	% sequenced
1	0.37	37%	63%
2	0.135	13.5%	87.5%
3	0.05	5%	95%
4	0.018	1.8%	98.2%
5	0.0067	0.6%	99.4%
6	0.0025	0.25%	99.75%
7	0.0009	0.09%	99.91%
8	0.0003	0.03%	99.97
9	0.0001	0.01%	99.99%
10	0.000045	0.005%	99.995%

Total Gap Length

Total Gap Length (bp) = Ge^{-c}

Where:

- **c = fold coverage**
- **G = target sequence length**
- **$e^{-c} = P_0$**

Genome size =	50 kb	150 kb	300 kb	2 Mb	4 Mb
Fold coverage	Ge^{-c}	Ge^{-c}	Ge^{-c}	Ge^{-c}	Ge^{-c}
1	18,500	55,500	111,000	740,000	1,480,000
2	6,750	20,250	40,500	270,000	540,000
3	2,500	7,500	15,000	100,000	200,000
4	900	2,700	5,400	36,000	72,000
5	335	1,005	2,010	13,400	26,800
6	125	375	750	5,000	10,000
7	45	135	270	1,800	3,600
8	15	45	90	600	1,200
9	5	15	30	200	400
10	2	6	12	90	180

Total Number of Gaps

Total number of gaps = Ne^{-c}

Where:

- **N = Gc/L = number of reads for x-fold coverage**

  G = Target sequence length
  c = Fold Coverage
  L = Average sequencing read length

- **$e^{-c} = P_0$**

50 kb Target Clone:

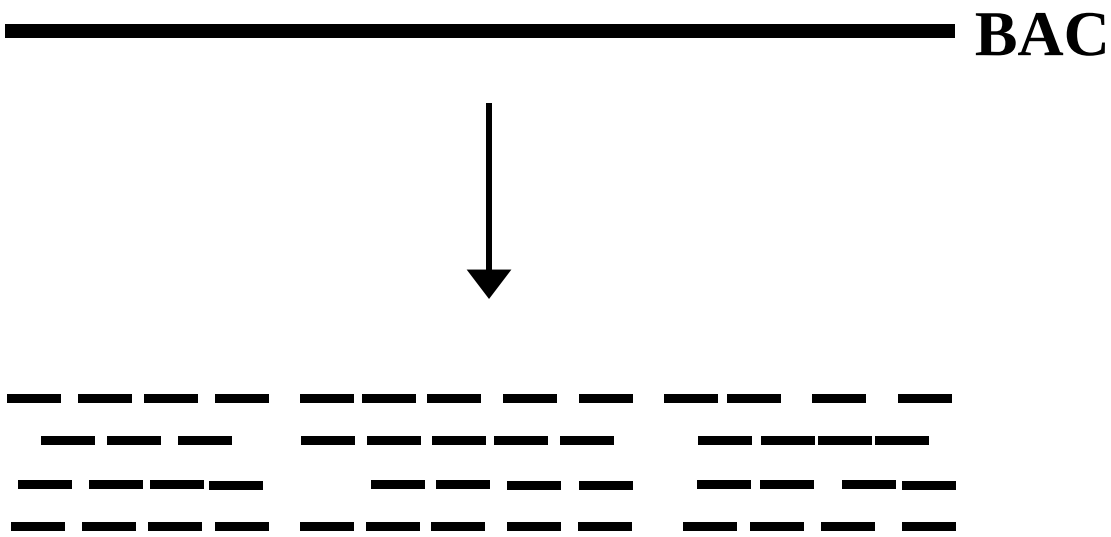
Read length	400			500			600		
Fold Cov.	N	e^{-c}	#Gaps $=Ne^{-c}$	N	e^{-c}	#Gaps $=Ne^{-c}$	N	e^{-c}	#Gaps $=Ne^{-c}$
1	125	0.37	46	100	0.37	37	84	0.37	31
2	250	0.135	34	200	0.135	27	168	0.135	23
3	375	0.05	19	300	0.05	15	242	0.05	12
4	500	0.018	9	400	0.018	7	326	0.018	6
5	625	0.0067	4	500	0.0067	3	410	0.0067	3
6	750	0.0025	2	600	0.0025	2	500	0.0025	1
7	875	0.0009	1	700	0.0009	1	583	0.0009	1
8	1000	0.0003	0	800	0.0003	0	667	0.0003	0
9	1125	0.0001	0	900	0.0001	0	750	0.0001	0
10	1250	0.000045	0	1000	0.000045	0	833	0.000045	0

The values for each fold coverage for a 150kb BAC (G=150,000) with average read length of 500 bases are:

Fold coverage	Total bases sequenced	e^{-c}	Total gap length in bases $=Ge^{-c}$	Number of Gaps $= Ne^{-c}$	Gap Length/# gaps = # bases per gap	% complete
1	150000	0.37	55,500	111	500	63
2	300000	0.135	20,250	81	250	87.5
3	450000	0.05	7,500	45	167	95
4	600000	0.018	2,700	22	123	98.2
5	750000	0.0067	1,005	10	101	99.4
6	900000	0.0025	375	5	75	99.75
7	1050000	0.0009	135	2	68	99.91
8	1200000	0.0003	45	1	45	99.97
9	1350000	0.0001	15	1	15	99.99
10	1500000	0.000045	6	1	6	99.995

For more calculations, see http://www.genome.ou.edu/poisson_calc.html

Shotgun Sequencing Strategy



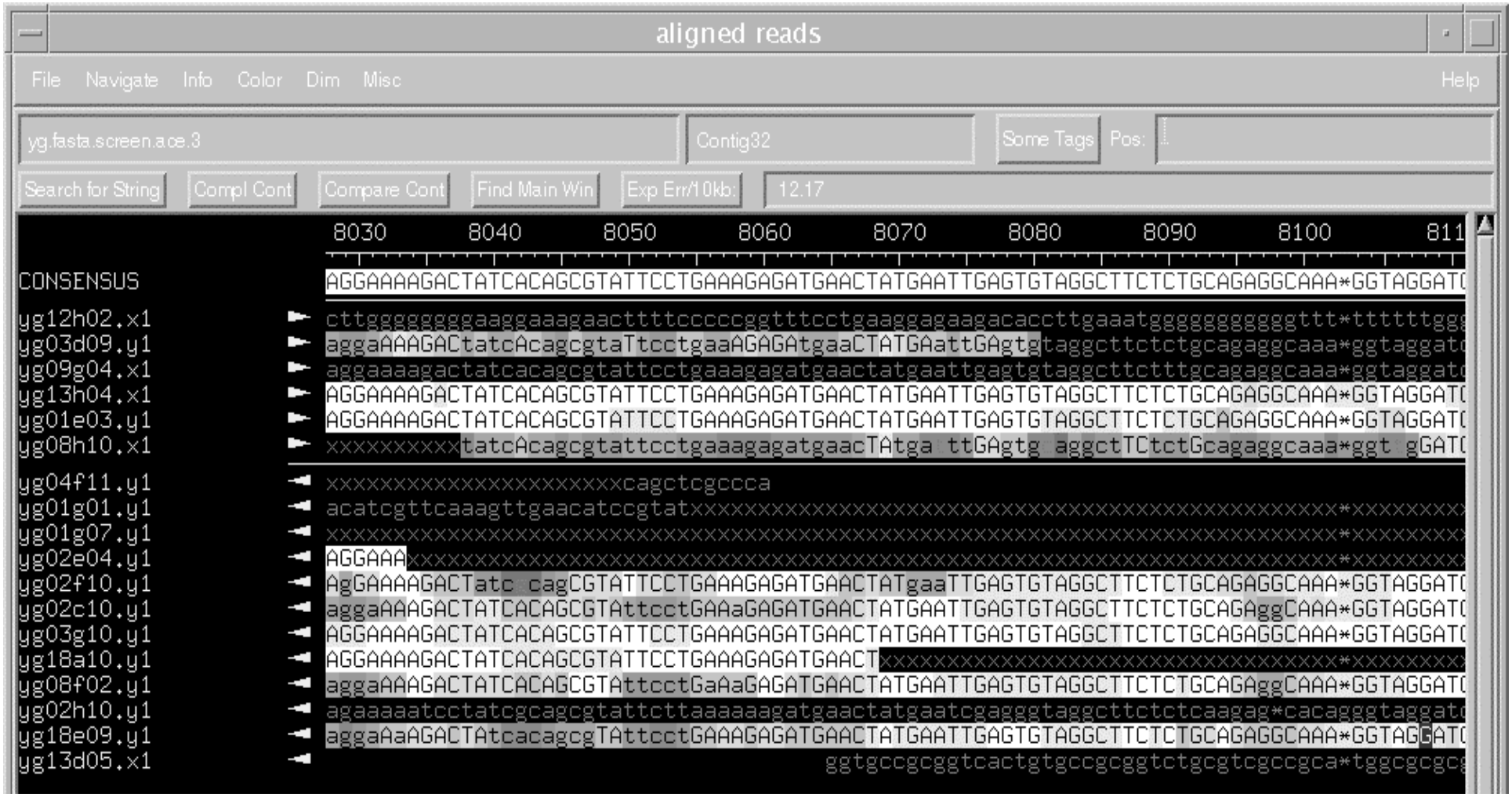
Sequence Assembly Software

DNA Star
Sequencher (Gene Codes)
Assembler (PE/ABI)
Gelassemble (GCG)
XBAP/XGAP (Staden)
Phrap (Green)

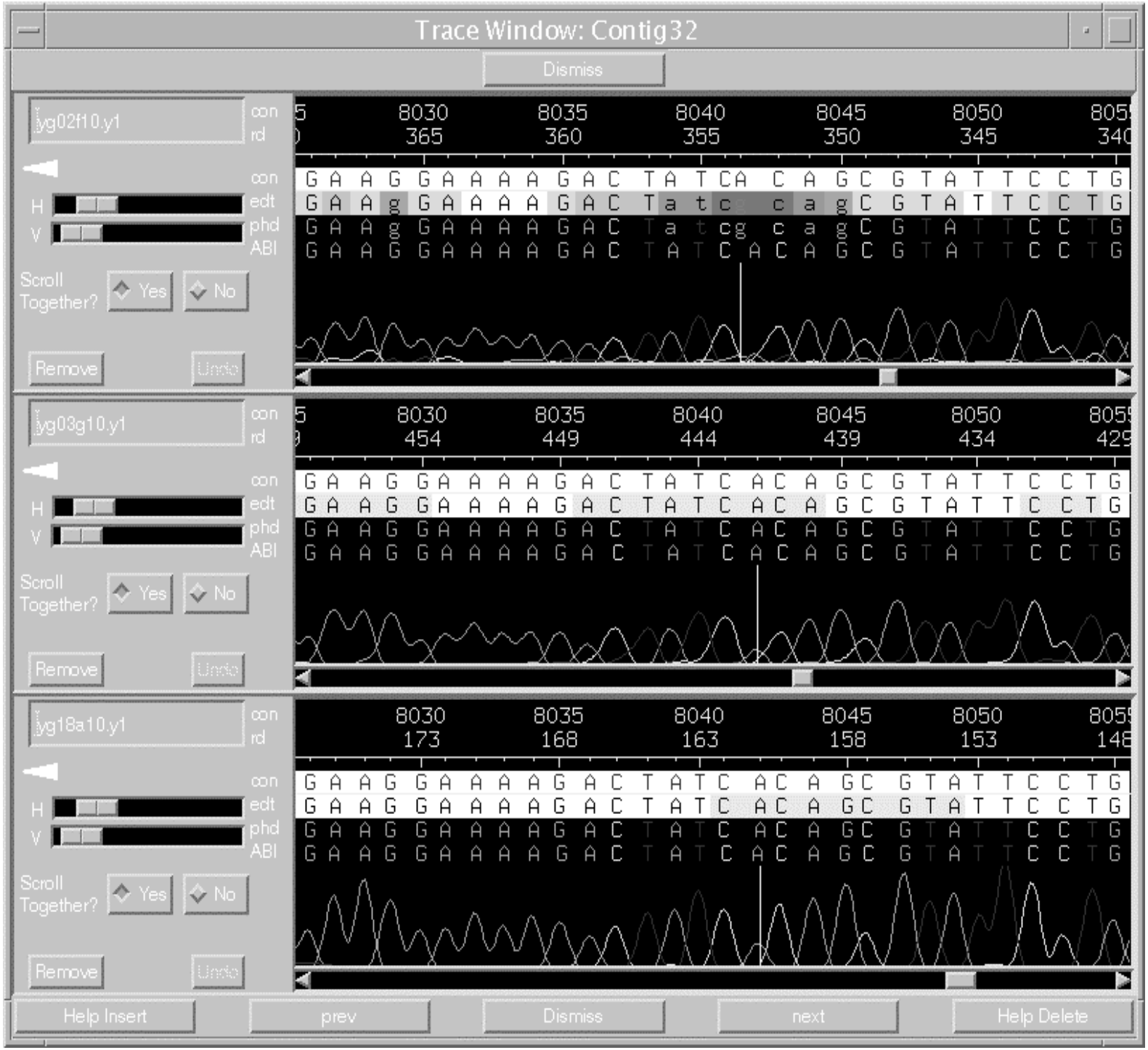
Shotgun Sequence Assembly

"Consed" (Gordon et al., *Genome Research* 8:195-202, 1998)

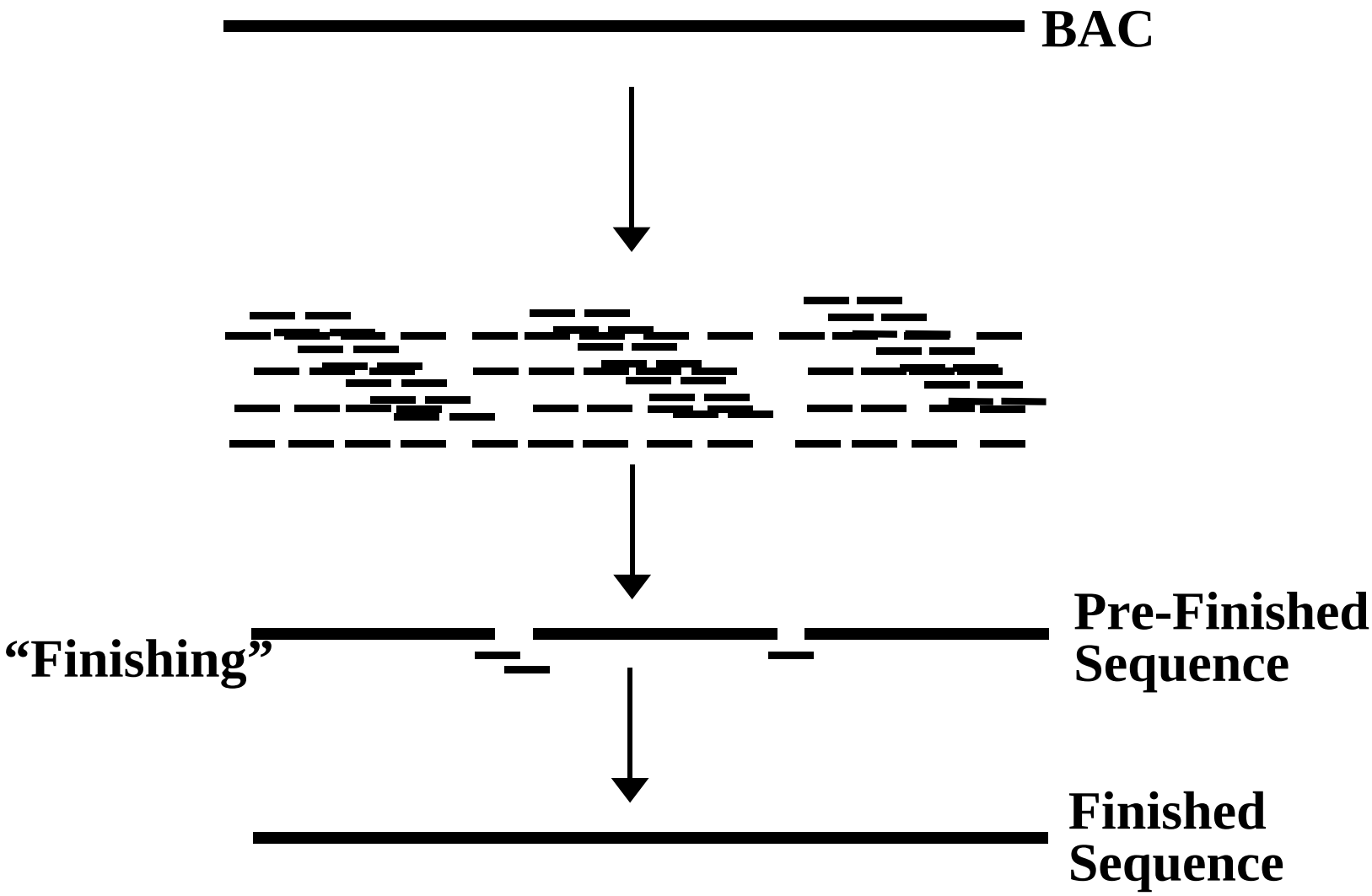
Shotgun Sequence Assembly



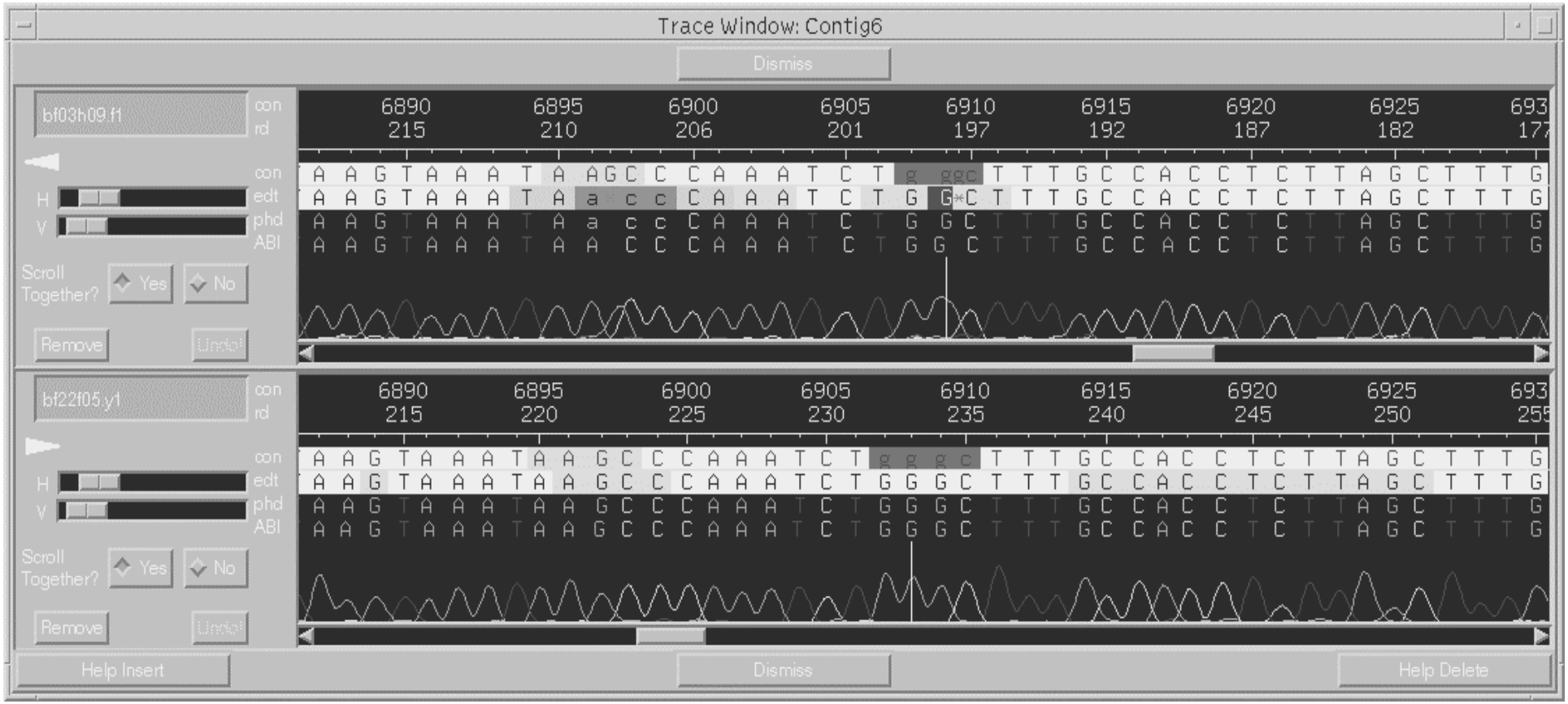
“Consed” (Gordon et al., *Genome Research* 8:195-202, 1998)


Trace Window: Contig32
Dismiss
yg02f10.y1
con
rd
8030 8035 8040 8045 8050
365 360 355 350 345
con
edt
phd
ABI
H
V
Scroll Together?
Yes
No
Remove
Undo
yg03g10.y1
454 449 444 439 434
yg18a10.y1
173 168 163 158 153
Help Insert
prev
Dismiss
next
Help Delete

Shotgun Sequencing Strategy



Resolve Ambiguities...



DNA Sequencing in the Human Genome Project

Complete Sequences of Microbial Genomes

TIGR Microbial Database - Netscape

File Edit View Go Communicator Help

TIGR Microbial Database:

a listing of microbial genomes and chromosomes completed and in progress

Published microbial genomes and chromosomes (scroll down for genomes in progress)

	Link	Genome	Strain	Domain	Size (Mb)	Institution	Funding	Publication
1		***Haemophilus influenzae* Rd**	KW20	B	1.83	TIGR	TIGR	Fleischmann *et. al.*, *Science* **269**:496-512 (1995)
2		***Mycoplasma genitalium***	G-37	B	0.58	TIGR	DOE	Fraser *et. al.*, *Science* **270**:397-403 (1995)
3		***Methanococcus jannaschii***	DSM 2661	A	1.66	TIGR	DOE	Bult *et. al.*, *Science* **273**:1058-1073 (1996)
4		***Synechocystis* sp.**	PCC 6803	B	3.57	Kazusa DNA Research Inst.		Kaneko *et. al.*, *DNA Res.* **3**: 109-136 (1996)
5		***Mycoplasma pneumoniae***	M129	B	0.81	Univ. of Heidelberg	DFG	Himmelreich et. al., *Nuc. Acid Res.* **24**:4420-4449 (1996)
6	SGD mips	***Saccharomyces cerevisiae***	S288C	E	13	International Consortium	EC, NHGRI, Welcome Trust, McGill U., RIKEN	Goffeau et. al., *Nature* **387** (Suppl.) 5-105 (1997)
7		***Helicobacter pylori***	26695	B	1.66	TIGR	TIGR	Tomb *et. al.*, *Nature***388**:539-547 (1997)
8		***Escherichia coli***	K-12	B	4.60	University of Wisconsin	NHGRI	Blattner *et. al.*, *Science* **277**:1453-1474 (1997)
9		***Methanobacterium thermoautotrophicum***	delta H	A	1.75	Genome Therapeutics & Ohio State Univ.	DOE	Smith *et.al.*, *J. Bacteriology*, **179**:7135-7155 (1997)
10		***Bacillus subtilis***	168	B	4.20	International Consortium	EC	Kunst *et.al.*, *Nature***390**: 249-256 (1997)

Document: Done

http://www.tigr.org/tdb/mdb