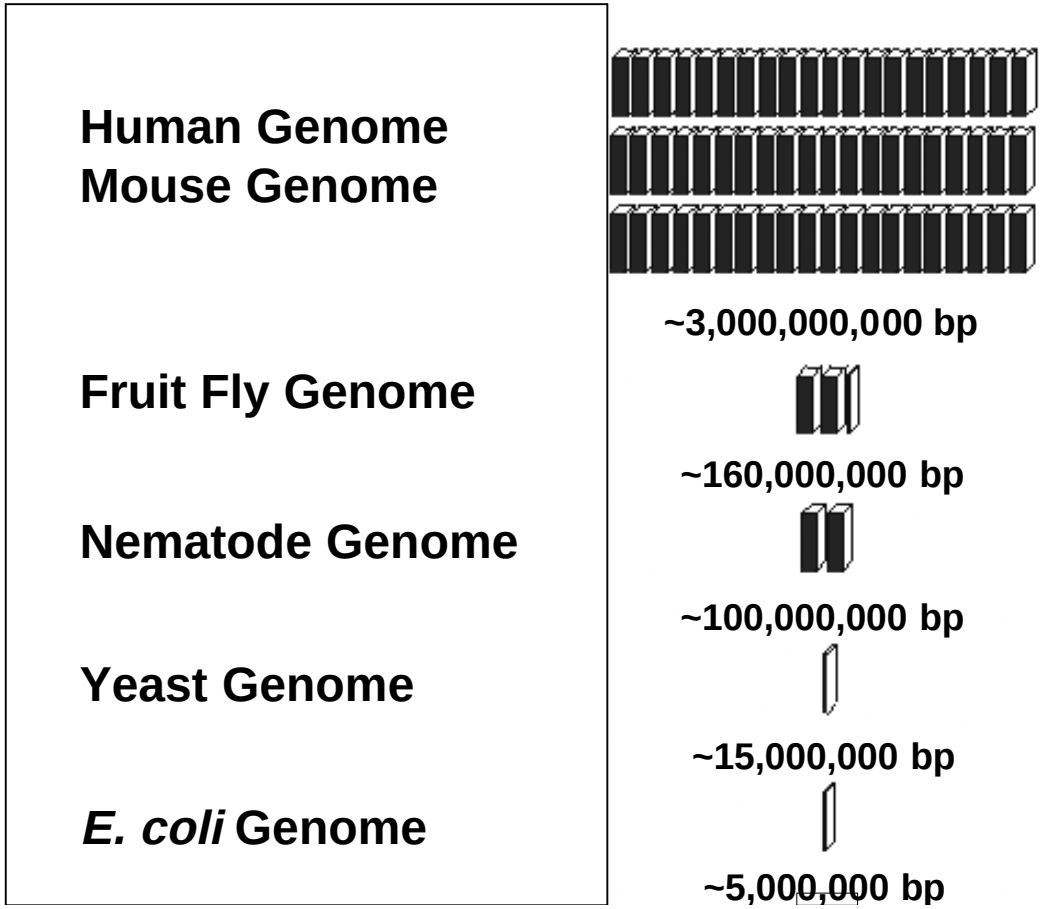
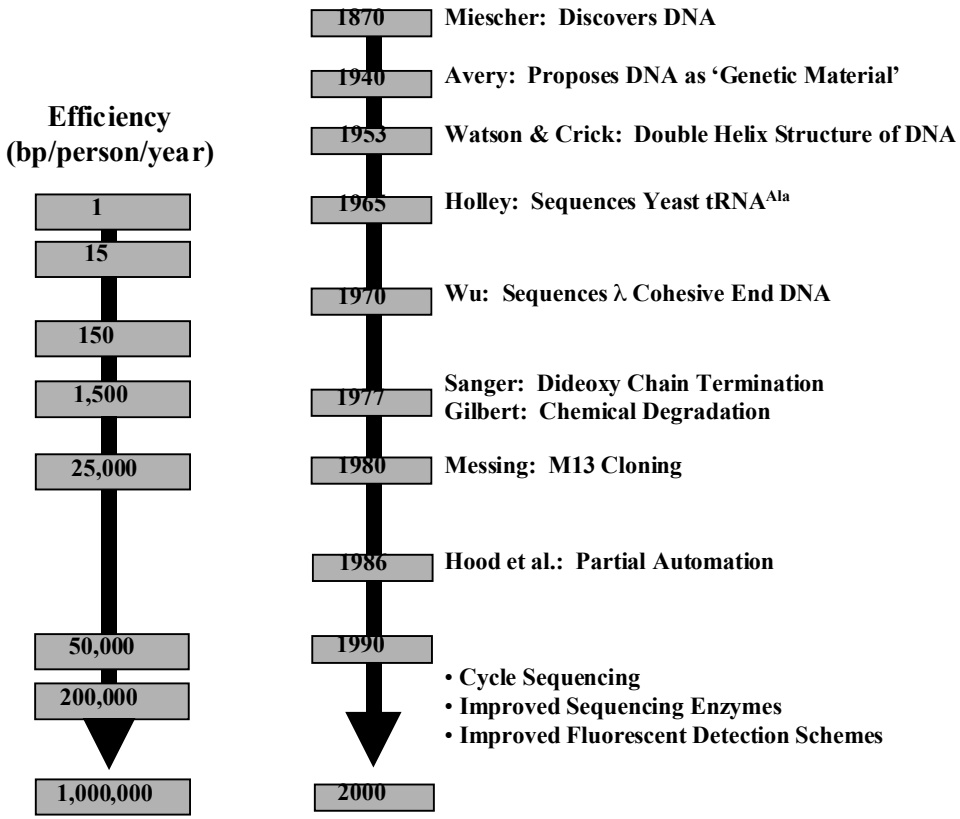
## Genome Sizes



# History of DNA Sequencing



# Radioactive Sequencing

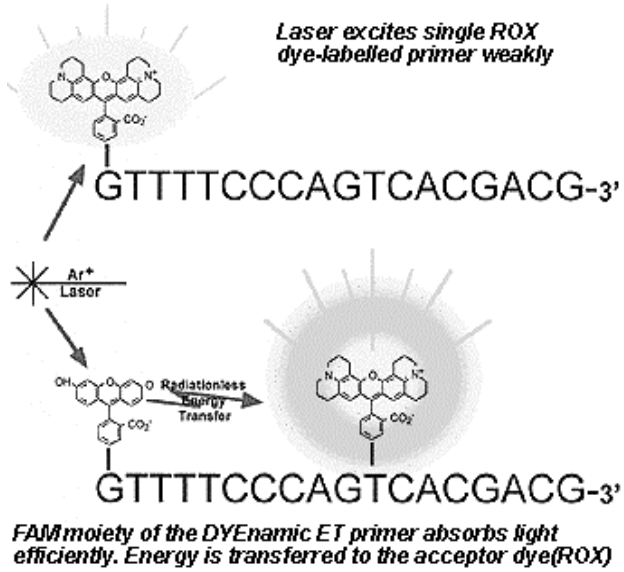
# Perkin Elmer/Applied Biosystems 377

# Cycle Sequencing

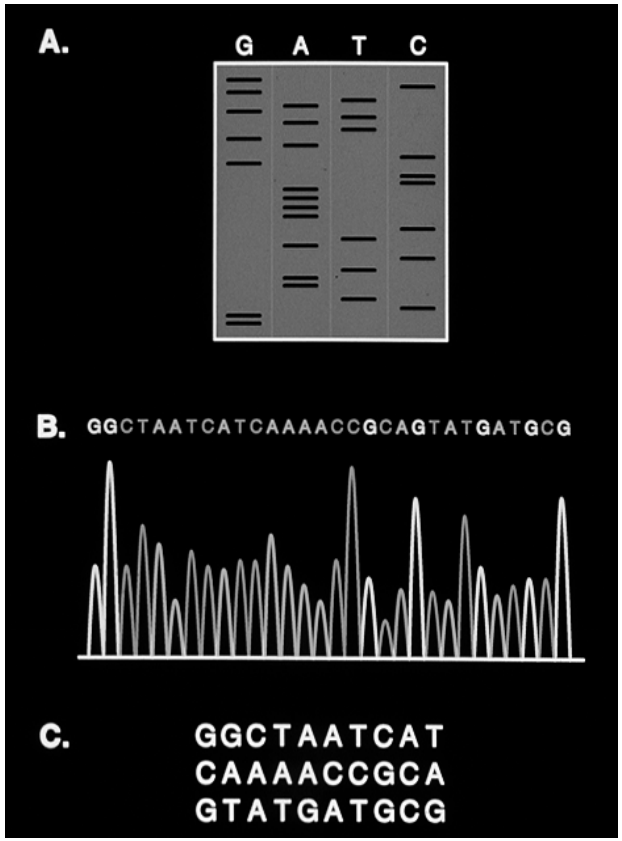
# Energy Transfer™ Technology

(Ju et al., 1996; Lee et al., 1997)

- Less DNA in sequencing reactions
- Fewer PCR cycles allowable
- Longer read lengths
- Direct loading onto gels possible



# Dideoxy Chain Termination Sequencing



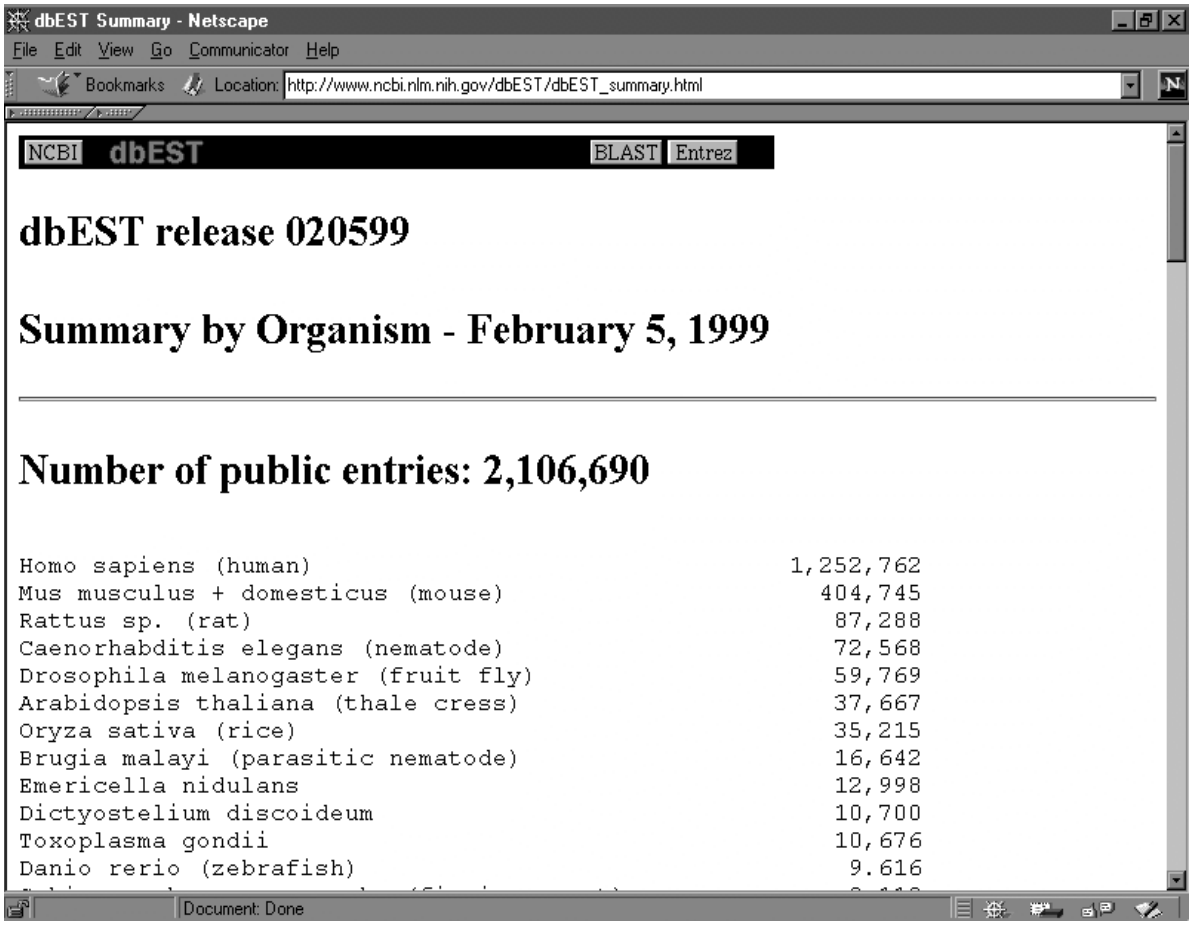
# Washington U. Genome Sequencing Center

# Expressed-Sequence Tags (ESTs)

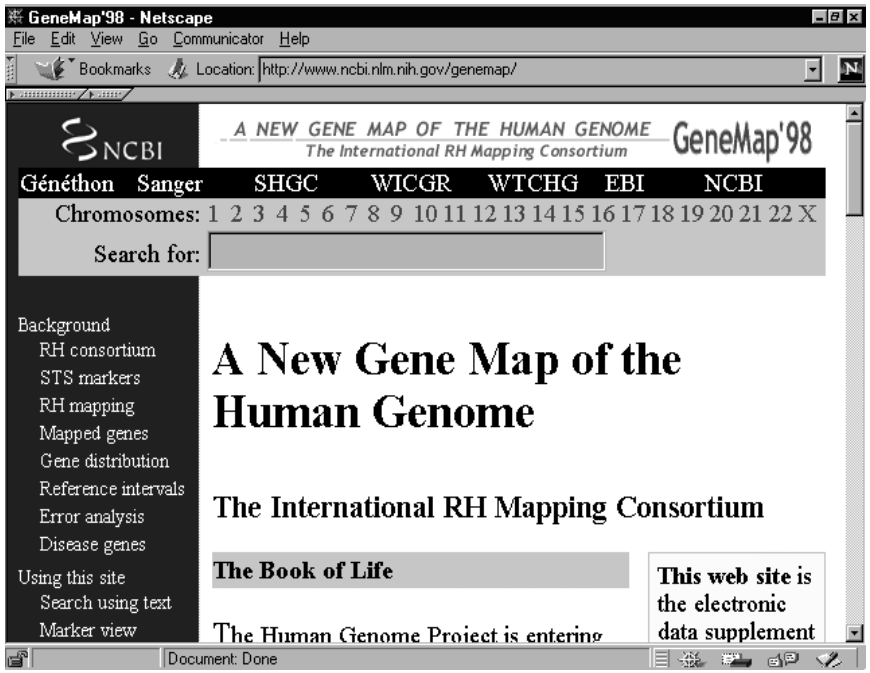
- **Single-Pass Sequence of Random cDNA Clone**
- **Often from Normalized cDNA Libraries**


- **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



# 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**

## Genomic Sequencing: Strategies

- **Transposon-Mediated Sequencing**

  **Refined within Drosophila Sequencing Effort**

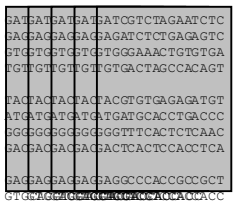
  **Kimmel et al., *Genome Analysis***
  **Vol. 1 (CSHL Press)**

- **Shotgun Sequencing**

  **Refined within Nematode Sequencing Effort**

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

## Subclone Construction



BAC DNA

Prepare Multiple Copies

Randomly Fragment

Subclone Fragments

# 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:**

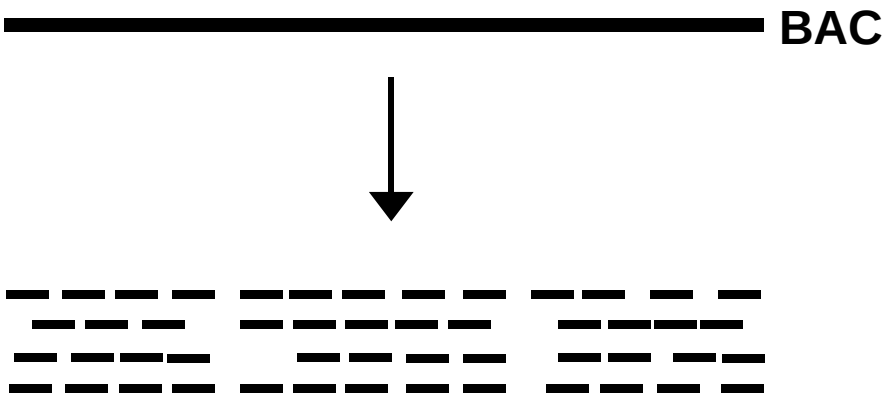
| Read length | 400 | | | 500 | | | 600 | | |
|---|---|---|---|---|---|---|---|---|---|
| Fold Cov. | N | $e^{-c}$ | #Gaps $=Ne^{-c}$ | N | $e^{-c}$ | #Gaps $=Ne^{-c}$ | N | $e^{-c}$ | #Gaps $=Ne^{-c}$ |
| 1 | 125 | 0.37 | 46 | 100 | 0.37 | 37 | 84 | 0.37 | 31 |
| 2 | 250 | 0.135 | 34 | 200 | 0.135 | 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

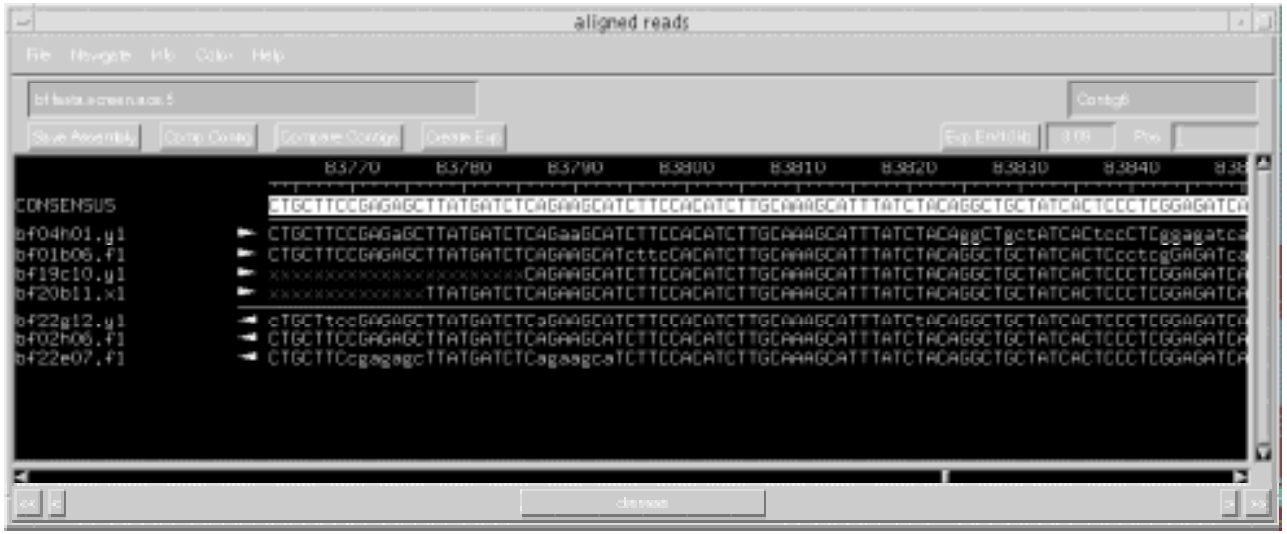
## Shotgun Sequencing Strategy



## 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)**