# Large-scale DNA Sequencing

## Current Topics in Genome Analysis
## October 28, 1997

**Jeff Touchman, Ph.D.**
**NIH Intramural Sequencing Center, and**
**National Human Genome Research Institute**
**jefft@nhgri.nih.gov**
**http://www.nhgri.nih.gov/NISC/**

# Large-scale Sequencing

I Current technology

- Sanger method
- Fluorescent sequencing chemistry: dye-primer and dye-terminator
- Four-color fluorescent detection

II Classical sequencing strategies

- Directed
  - Primer "walking"
  - Controlled nested deletions
  - Transposon insertions
- Random

III Shotgun sequencing

- Laboratory techniques
- Poisson distribution
- Directed vs. random strategies

IV EST sequencing

V Bacterial genomes

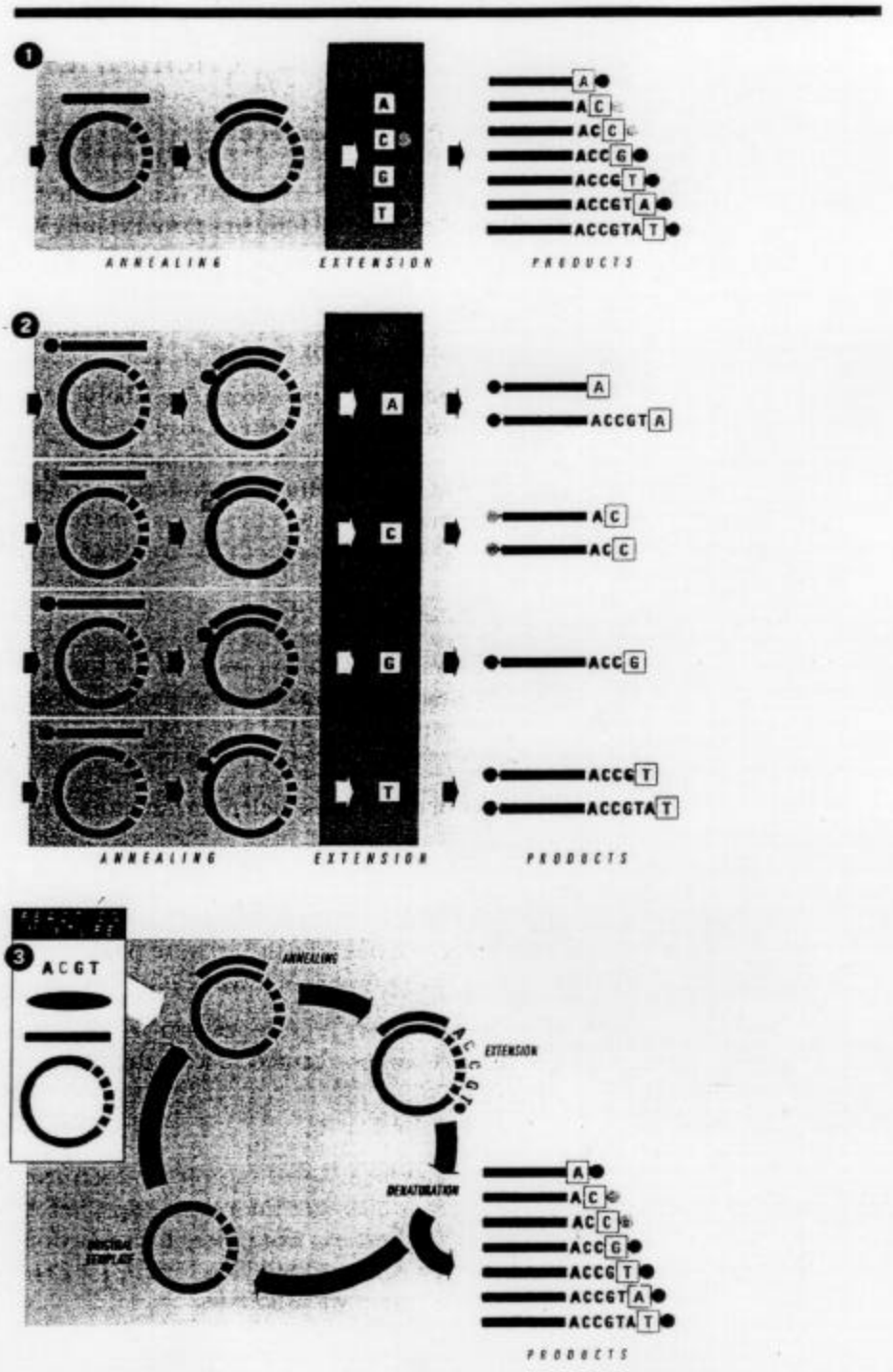
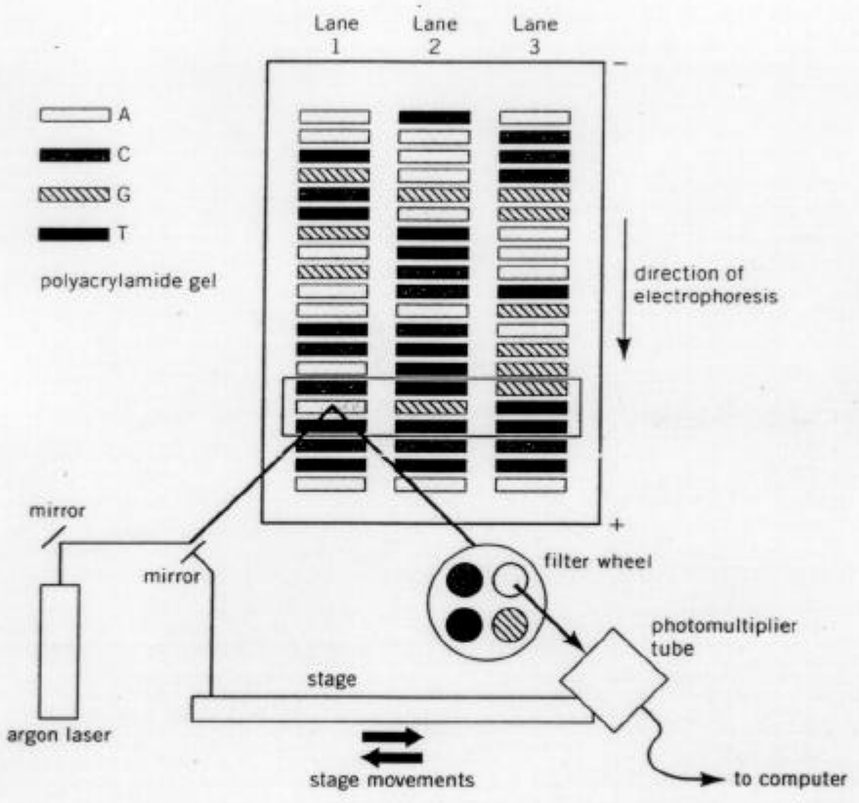



**Figure 1-1. (1) Dye terminator labeling, (2) Dye primer labeling, (3) Cycle sequencing**

# Four Color Fluorescent Sequencing



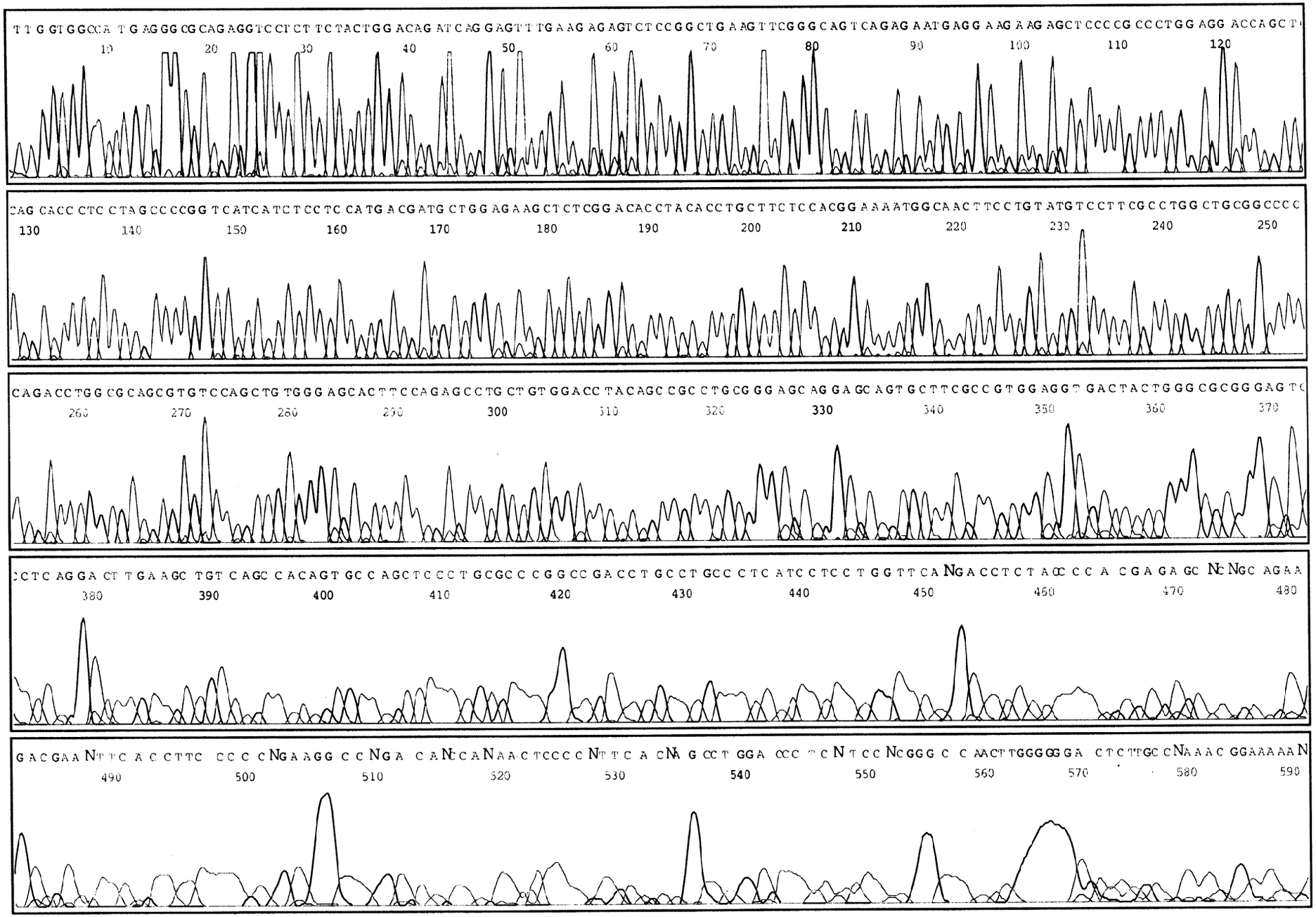
DYE TERMINATOR

ABI PRISM™

Model 373 31
Version 3.0
ABI50 31
Version 3.0 Lane 5

Signal G:212 A:580 T:219 C:225
DT6%Ac{A Set-AnyPrimer}
94061942S
Points 944 to 7164 Base 1: 944


Mon, Oct 20, 1997 10:32 PM
Sat, Feb 1, 1997 12:41 PM
Spacing: 10.79{10.79}



DYE PRIMER

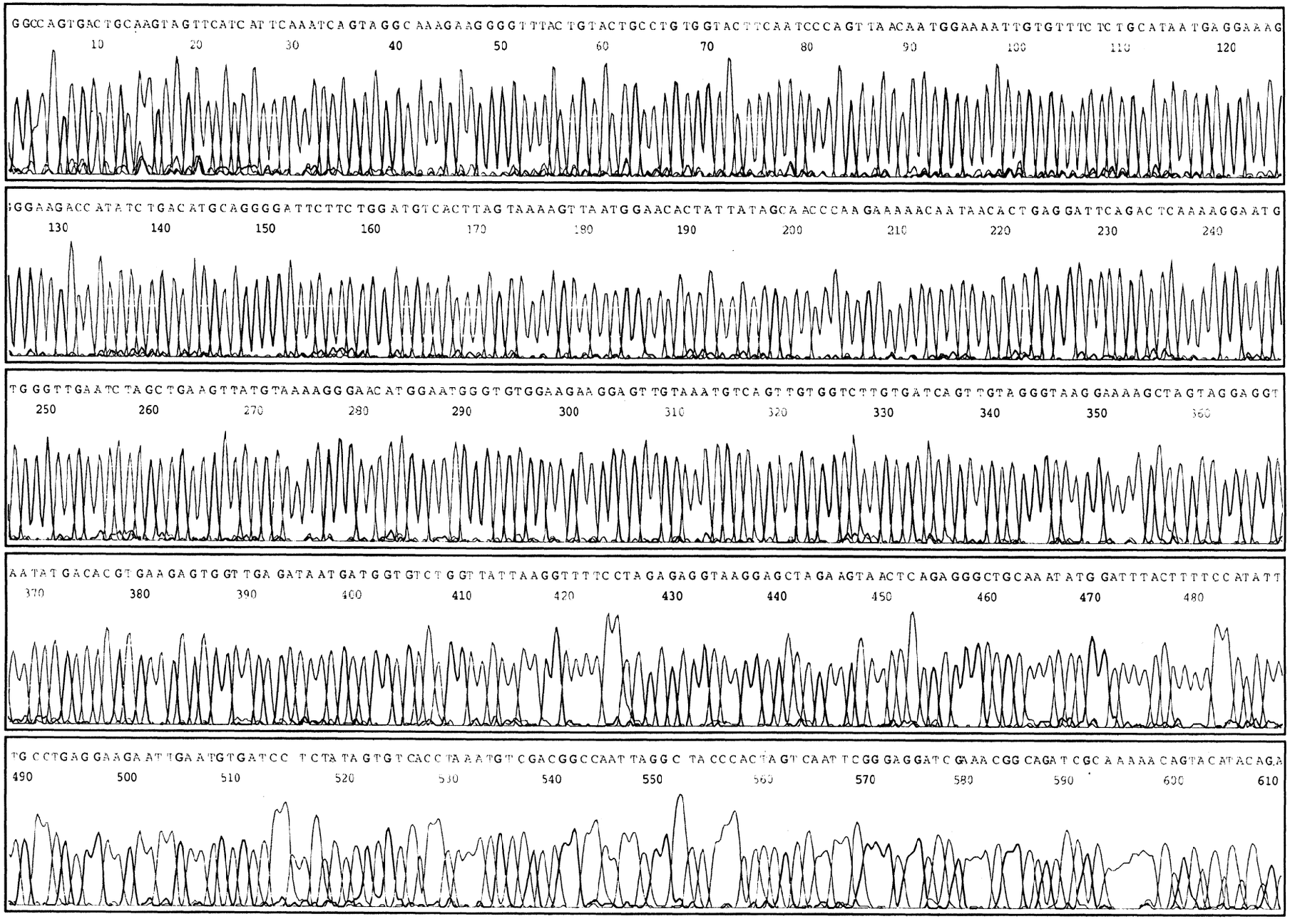
ABI PRISM™

Model 377
Version 3.0
ABI200
Version 2.1.1

c58t01d09.s1
c58t01d09.s1
Lane 21

Signal C:392 A:367 G:272 T:365
ET{-40m13}960208
ET-Matrix960720
Points 1121 to 8008 Base 1: 1121

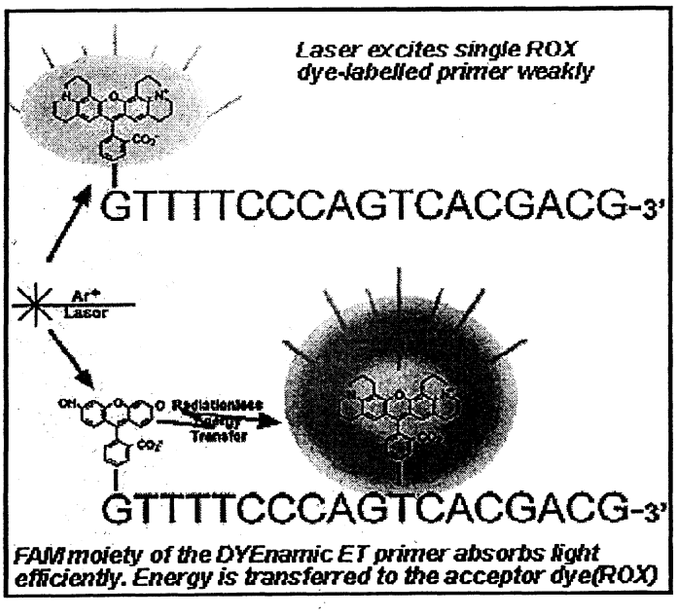

Tue, Jul 30, 1996 9:36 AM
Mon, Jul 29, 1996 12:51 PM
Spacing: 10.30{ 0}



# Energy Transfer Technology

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

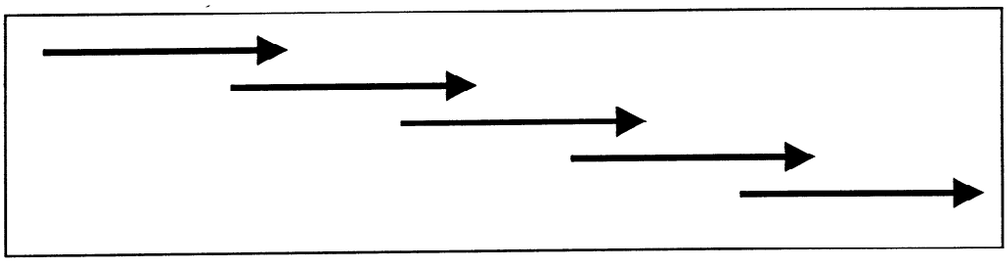
- Less DNA needed in sequencing reaction
- Fewer PCR cycles allowable
- Longer read lengths
- Direct loading onto sequencing gels possible (primers only)



# Classical large-scale sequencing strategies

## 1. Directed or Ordered Approaches

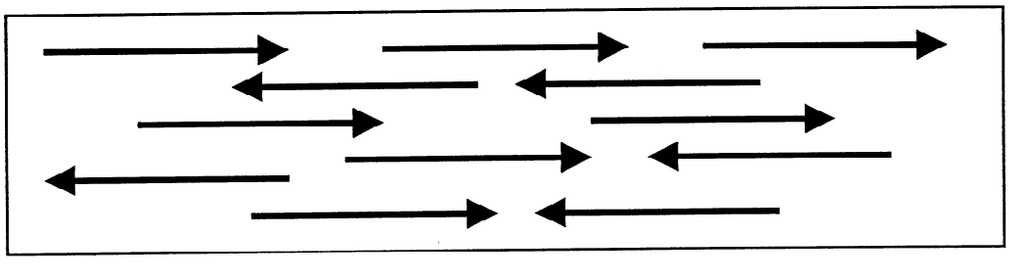
- Generate ordered clones
- Minimal redundancy sequencing



- Primer walking
- Controlled nested deletions
  (Bal31 digestion, ExoIII/S1 digestion, ssDNA deletions)
- Transposon insertion
- Restriction fragment subcloning

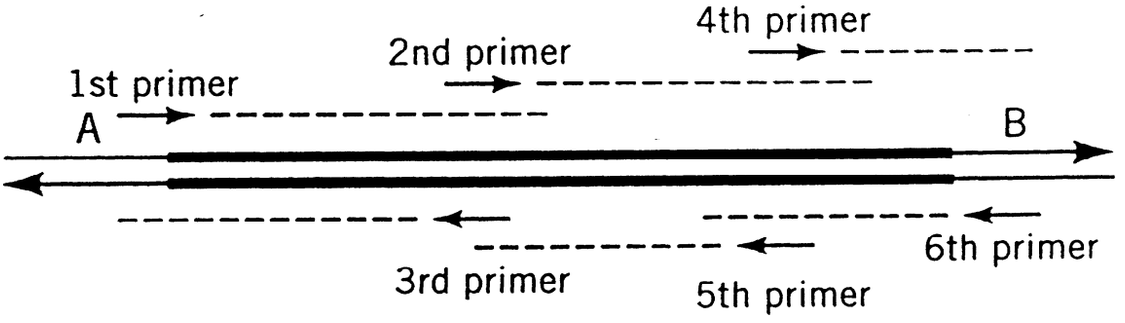
## 2. Random Approaches

- High redundancy sequencing
- Gap closure



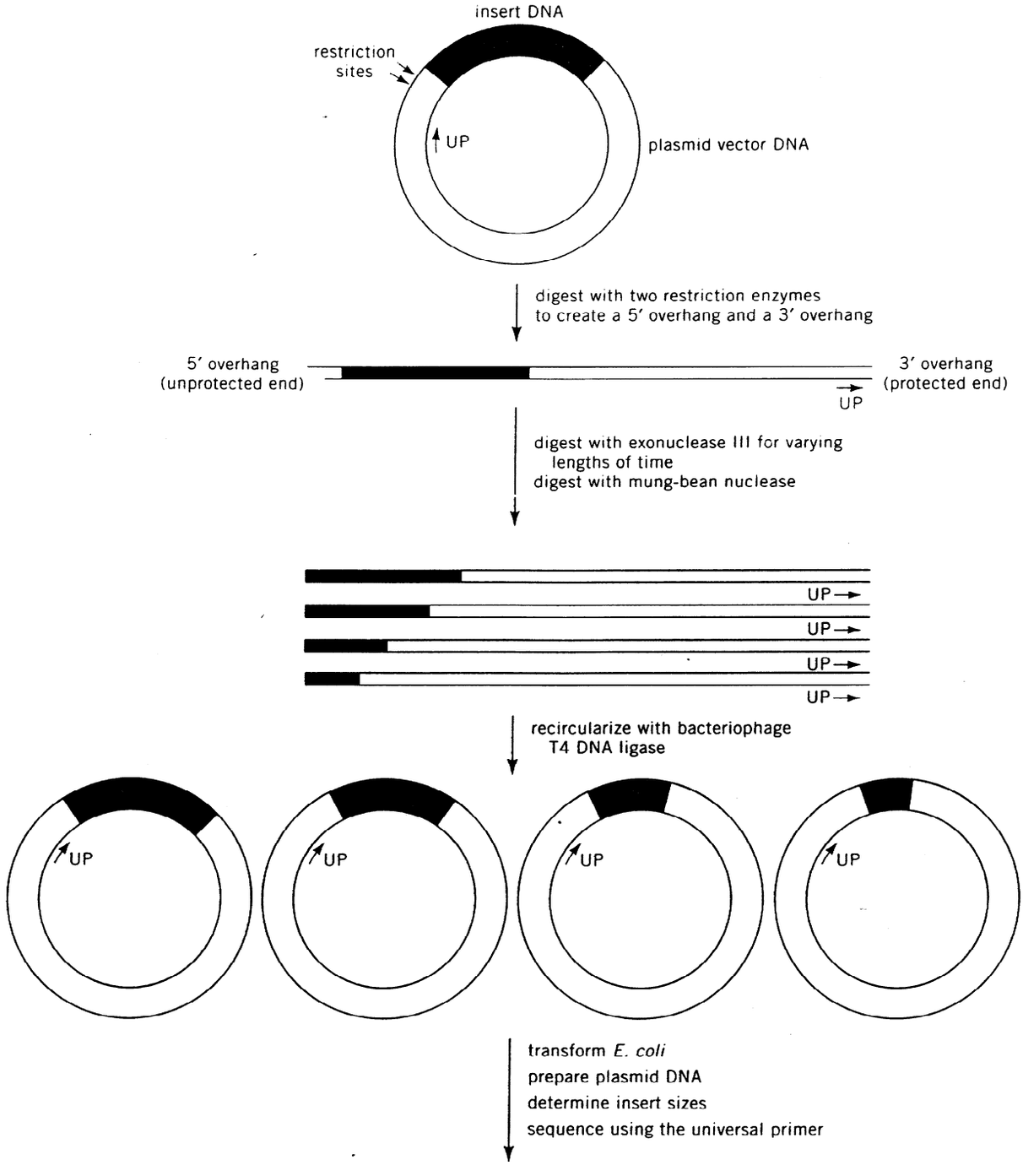
- "Shotgun"

# Primer-directed Strategy

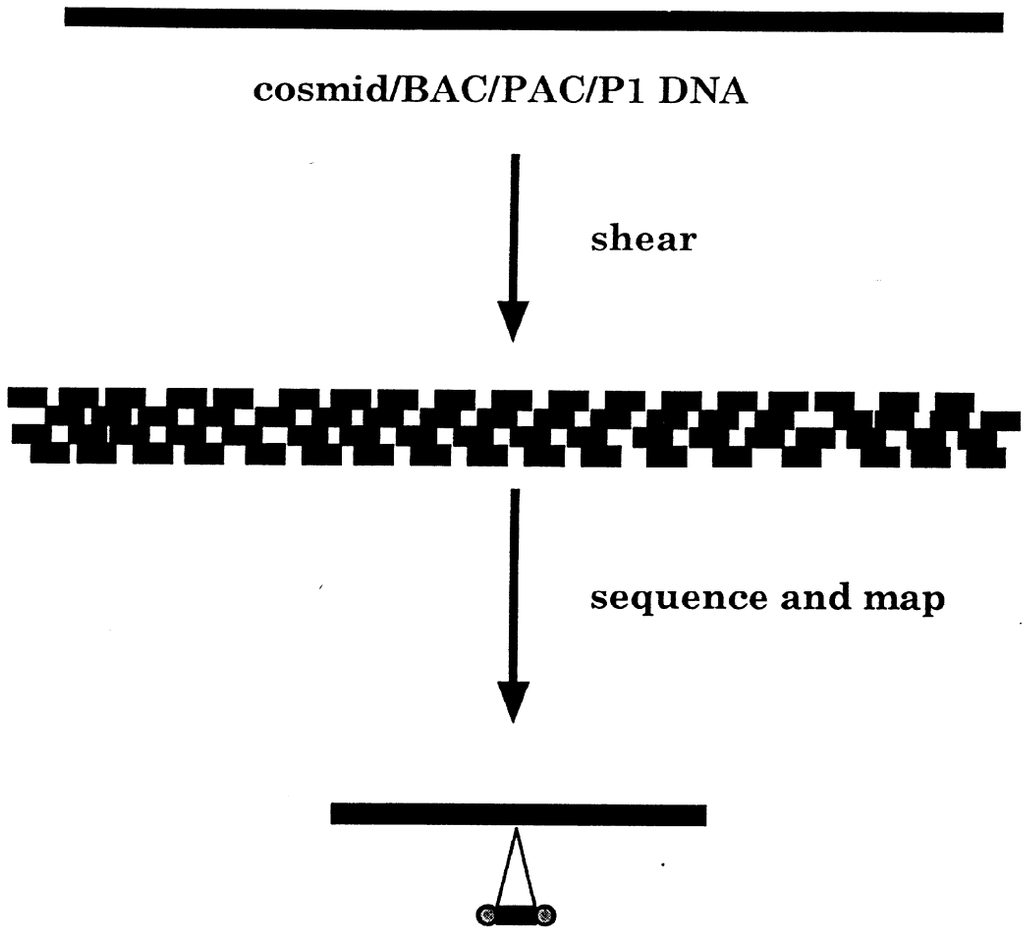



*Note*: 1st and 6th primers anneal to universal (i.e., vector) sequences

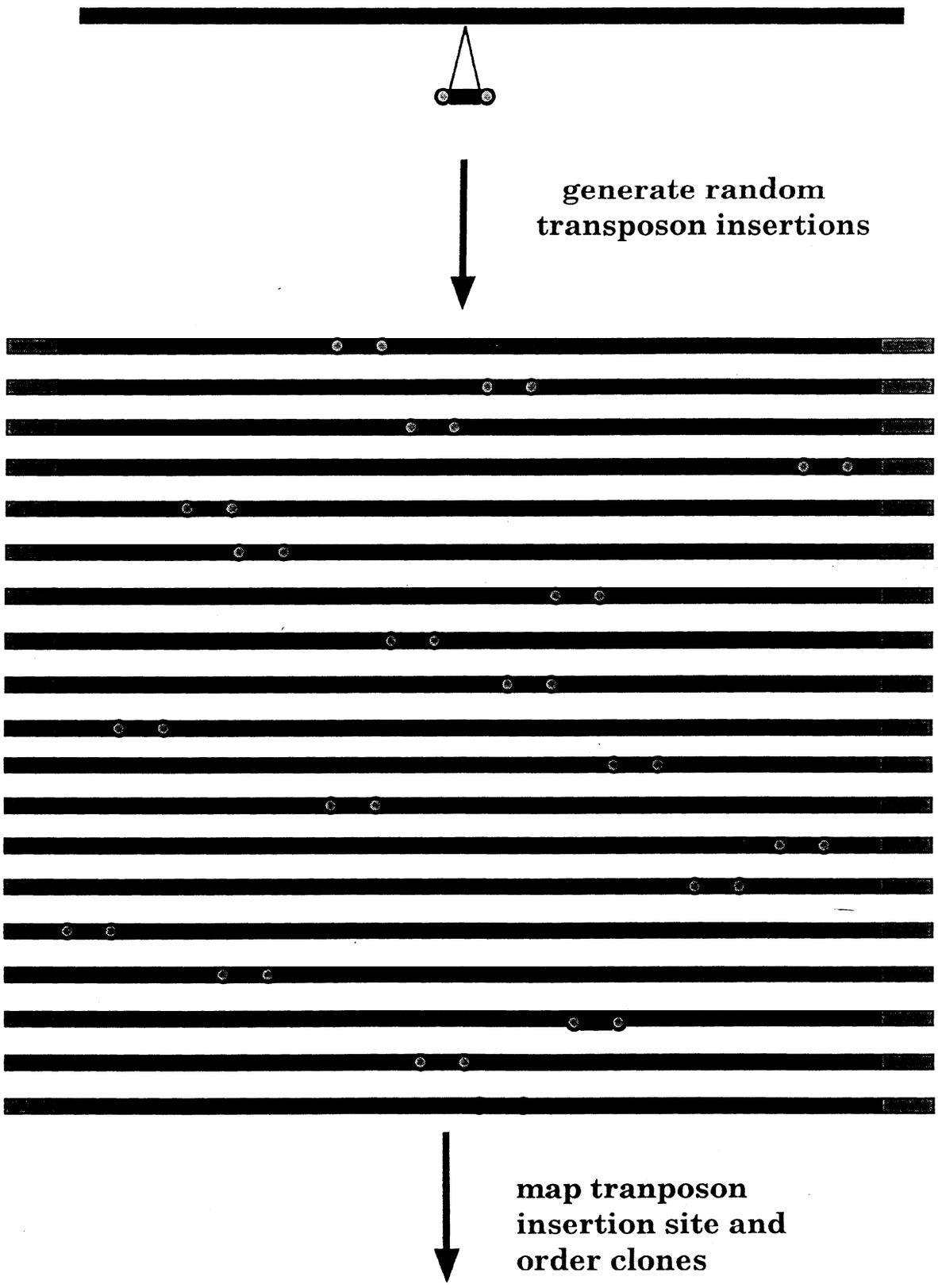
# Nested Deletions



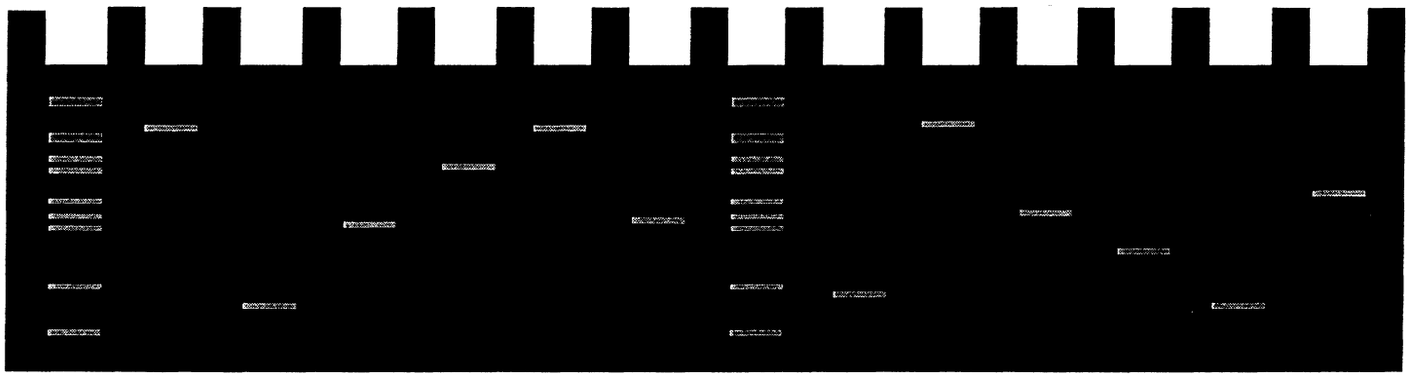
# Sequencing from ordered transposon insertions into mapped random subclones

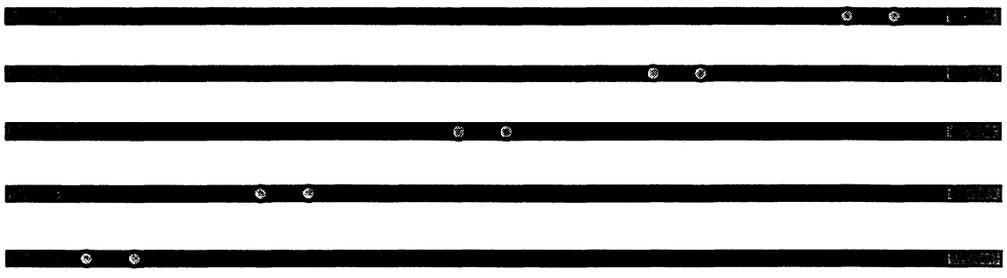


# Ordering transposon insertions

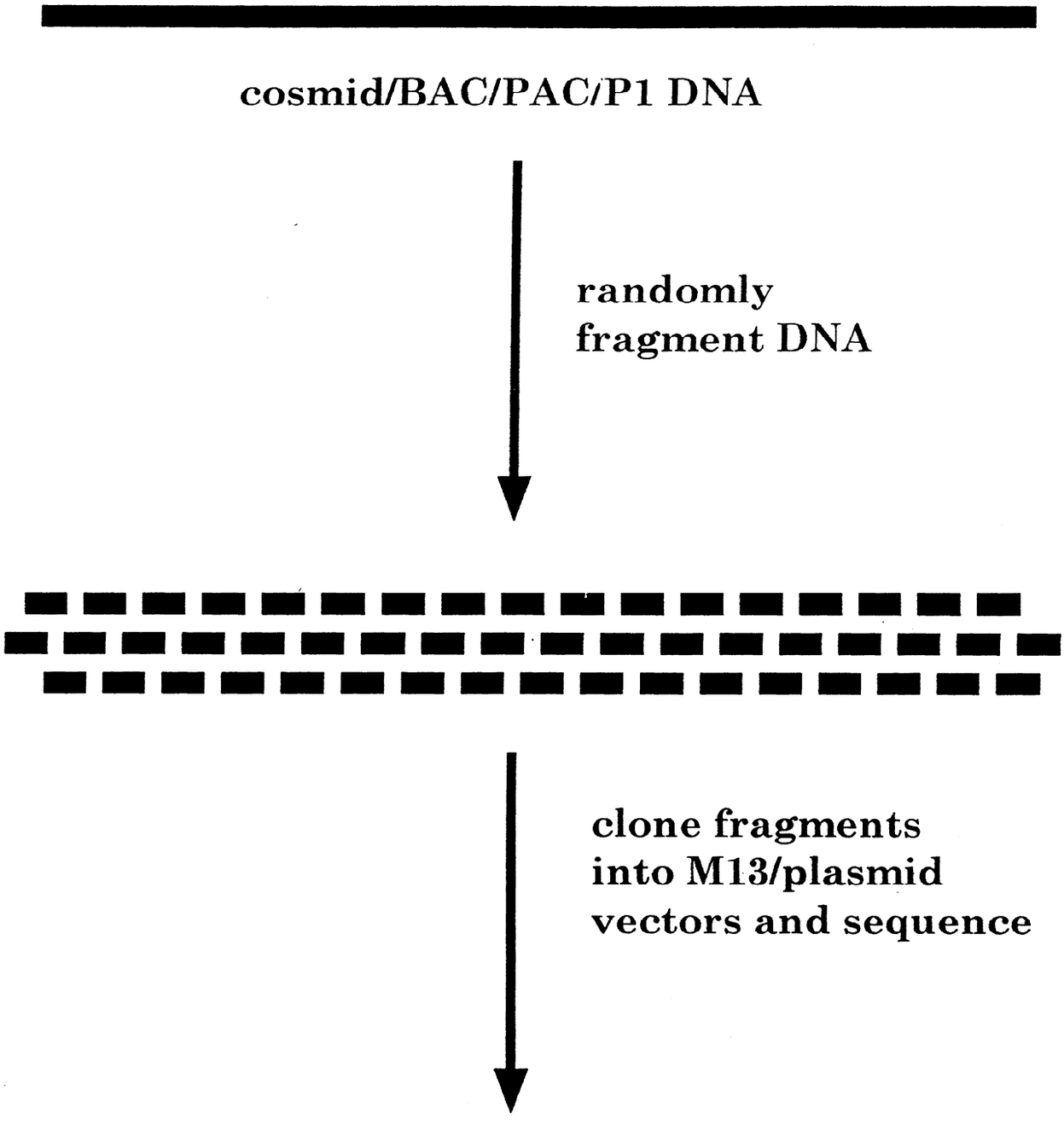


# Ordering transposon insertions



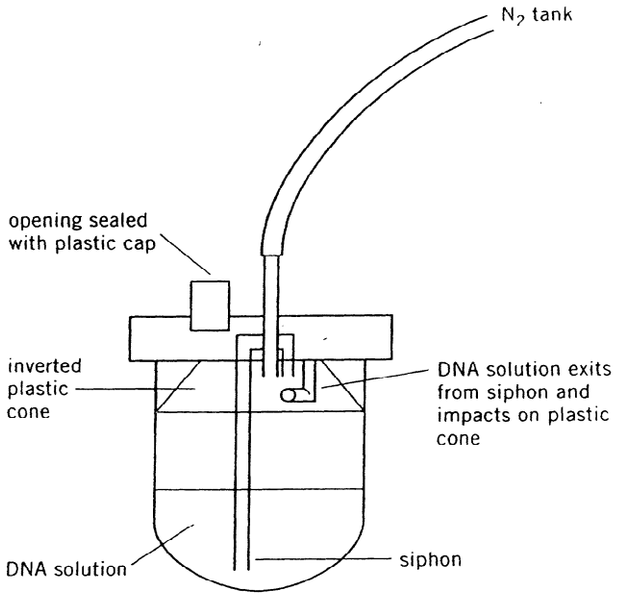


# Random Shotgun Sequencing with Directed Gap Closure

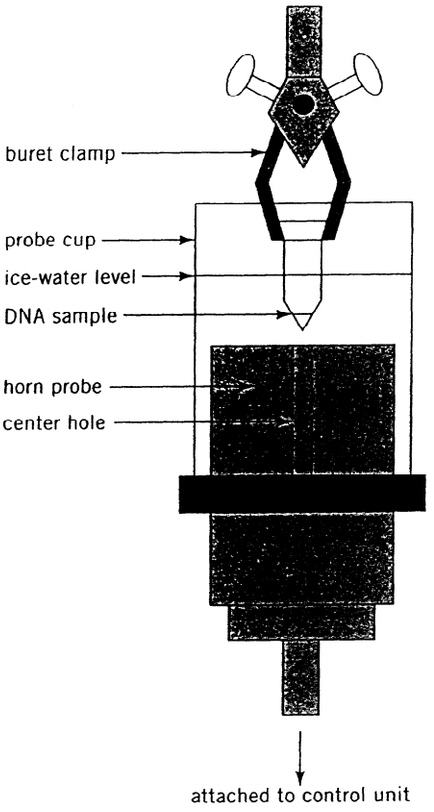


# Popular DNA Shearing Strategies

- **Sonication**
  Physical shearing with high frequency sound waves
- **Nebulization**
  Physical shearing with high pressure
- **French Press Shearing**
  Physical shearing with high pressure
- **Partial Digestion**
  Enzymatic, e.g. Sau3AI, DnaseI/$Mn^{++}$

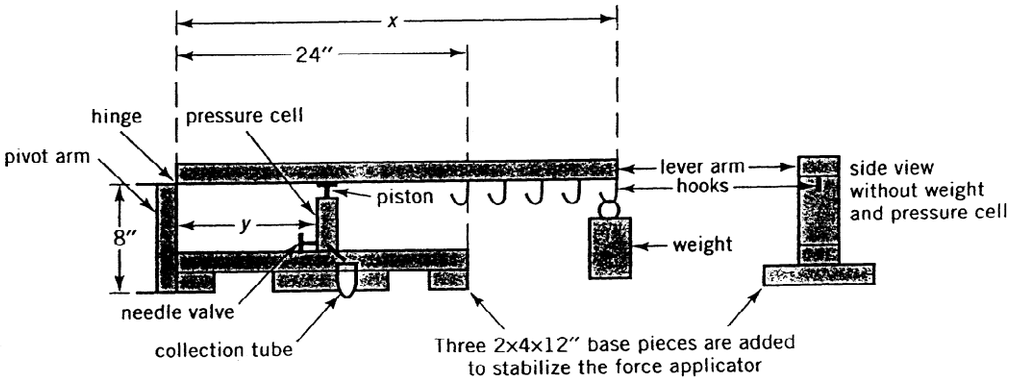



## Nebulizer

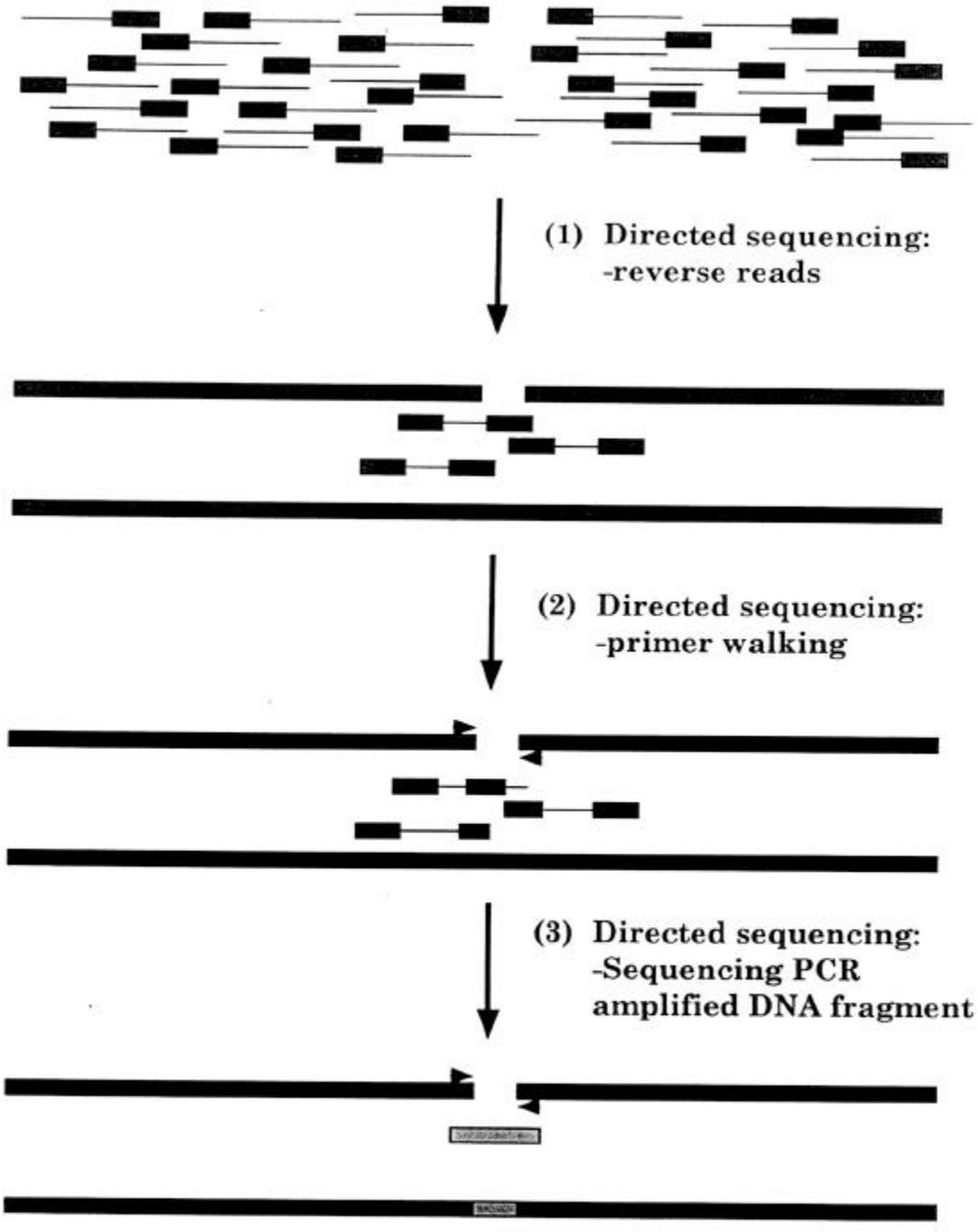



## Cup Horn Sonicator

## French Press



# Random Shotgun Sequencing with Directed Gap Closure



# Sequence Assembly Software

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

# Poisson calculations

**The sequencing strategy for the shotgun approach follows the Lander and Waterman application of the Poisson distribution (Lander and Waterman, 1988)**

**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% |

% sequenced is independent of read length when fold coverage is considered

# 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 50kb cosmid (G=50,000) 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 | 50000 | 0.37 | 18,500 | 37 | 500 | 63 |
| 2 | 100000 | 0.135 | 6,750 | 27 | 250 | 87.5 |
| 3 | 150000 | 0.05 | 2,500 | 15 | 167 | 95 |
| 4 | 200000 | 0.018 | 900 | 7 | 129 | 98.2 |
| 5 | 250000 | 0.0067 | 335 | 3 | 112 | 99.4 |
| 6 | 300000 | 0.0025 | 125 | 2 | 63 | 99.75 |
| 7 | 350000 | 0.0009 | 45 | 1 | 45 | 99.91 |
| 8 | 400000 | 0.0003 | 15 | 1 | 15 | 99.97 |
| 9 | 450000 | 0.0001 | 5 | 1 | 5 | 99.99 |
| 10 | 500000 | 0.000045 | 2 | 1 | 2 | 99.995 |

**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

## Directed Approaches

+ Reduction of sequencing redundancy (to 2-3X)
+ Better organization of raw sequence data
+ Ability to sequence highly repetitive regions

- Manual work requires highly skilled personnel
- Labor-intensive procedures difficult to automate
- Works only from 5-10 kb each time
- Primer walking expensive and slow

## Random Approaches

+ Rapid accumulation of data
+ Over-determination of each nucleotide minimizes errors in the sequence
+ More convenient to automate
+ Currently works up to 50-250 kb

- Large redundancy (5-10X) increases cost
- Gap-filling ("finishing") is tedious and slow
- Progress of a project is more difficult to control
- Highly repetitive regions difficult to handle

# Goals

**Maximize efficiency and accuracy**

**Minimize cost**

# Method of choice:

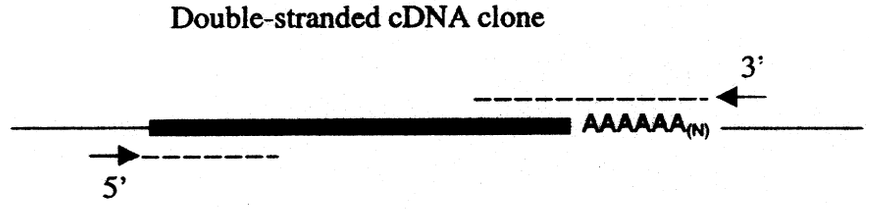
**Random sequencing** **98-99%**

**followed by**

**Directed sequencing** **1-2%**

# Expressed Sequence Tags (ESTs)

- **Single-pass partial sequencing of random cDNA clones**




- Sample or "tag" sequence from both 5' and 3' ends of normalized, tissue-specific 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

# Usefulness of ESTs

- Rapid means of gene discovery for those searching for human disease genes
- The study of temporal and spatial expression patterns
- Construction of a genome-wide STS based transcript map

NCBI dbEST BLAST Entrez

# dbEST release 101097

## Summary by Organism - October 10, 1997

## Number of public entries: 1,266,975

| | | |
|---|---|---|
| Homo sapiens (human) | 826,089 | |
| Mus musculus + domesticus (mouse) | 234,814 | |
| Caenorhabditis elegans (nematode) | 72,521 | |
| Arabidopsis thaliana (thale cress) | 32,964 | |
| Oryza sativa (rice) | 20,708 | |
| Drosophila melanogaster (fruit fly) | 16,859 | |
| Brugia malayi (parasitic nematode) | 10,721 | |
| Toxoplasma gondii | | 10,671 |
| Rattus sp. (rat) | 6,703 | |
| Schistosoma mansoni (blood fluke) | 3,394 | |
| Danio rerio (zebrafish) | 3,145 | |
| Saccharomyces cerevisiae (baker's yeast) | 3,042 | |
| Plasmodium falciparum (malaria parasite) | 2,872 | |
| Trypanosoma brucei rhodesiense | 2,626 | |
| Caenorhabditis briggsae | 2,424 | |
| Zea mays (maize) | 1,757 | |
| Onchocerca volvulus | 1,616 | |
| Brassica napus (oilseed rape) | 1,427 | |
| Sus scrofa (pig) | 1,323 | |
| Brassica campestris (field mustard) | 965 | |
| Dictyostelium discoideum | 795 | |
| Ricinus communis (castor bean) | 750 | |
| Pristionchus pacificus | 703 | |
| Pyrococcus furiosus (hyperthermophilic archaeon) | 577 | |
| Trypanosoma cruzi | 575 | |
| Cryptosporidium parvum | 567 | |
| Paralichthys olivaceus | | 550 |
| Leishmania major | | 535 |
| Schistosoma japonicum (blood fluke) | 455 | |
| Citrus unshiu | 426 | |
| Emericella nidulans | 272 | |
| Pinus taeda (loblolly pine) | 256 | |
| Capra hircus (goat) | 245 | |
| Saccharum sp. | 245 | |
| Strongylocentrotus purpuratus (purple urchin) | 242 | |
| Gallus gallus (chicken) | 216 | |
| Oryctolagus cuniculus (rabbit) | 207 | |
| Gracilaria gracilis | 200 | |
| Boophilus microplus (southern cattle tick) | 143 | |
| Trypanosoma brucei brucei | 143 | |
| Avena sativa (oat) | 125 | |
| Wuchereria bancrofti | 120 | |
| Biomphalaria glabrata | | 111 |
| Canis familiaris (dog) | 107 | |
| Entamoeba histolytica | 106 | |
| Glycine max (soybean) | 97 | |
| Sorghum bicolor (sorghum) | 91 | |
| Antheraea yamamai | 90 | |

# Other Strategies for EST/cDNA Sequencing

## Directed Methods

- Nested deletions
- End sequencing and primer walking
- Sequencing via ordered transposon insertions

## Random Methods

- Random shearing of clone, or PCR product of clone
- Random shearing of multiple cDNA clones ligated together

REVIEW

# Big Time for Small Genomes

Eugene V. Koonin[1]

National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, Maryland 20894

## Completed microbial genomes

| Link | Genome | Strain | Domain | Size (Mb) | Institution | Funding | Publication |
|---|---|---|---|---|---|---|---|
| | ***Haemophilus influenzae* Rd** | KW20 | **B** | 1.83 | **TIGR** | **TIGR** | Fleischmann *et. al.*, *Science* **269**:496-512 (1995) |
| | ***Mycoplasma genitalium*** | G-37 | **B** | 0.58 | **TIGR** | **DOE** | Fraser *et. al.*, *Science* **270**:397-403 (1995) |
| | ***Methanococcus jannaschii*** | | **A** | 1.66 | **TIGR** | **DOE** | Bult *et. al.*, *Science* **273**:1058-1073 (1996) |
| | ***Synechocystis* sp.** | PCC 6803 | **B** | 3.57 | **Kazusa DNA Research Inst.** | | Kaneko *et. al.*, *DNA Res.* **3**: 109-136 (1996) |
| | ***Mycoplasma pneumoniae*** | | **B** | 0.81 | **Univ. of Heidelberg** | | Himmelreich et. al., *Nuc. Acid Res.* **24**:4420-4449 (1996) |
| | ***Saccharomyces cerevisiae*** | | **E** | 13 | **European & North American consortium** | | Goffeau et. al., *Nature* **387** (Suppl.) 5-105 (1997) |
| | ***Helicobacter pylori*** | 26695 | **B** | 1.66 | **TIGR** | **TIGR** | Tomb *et. al.*, *Nature* **388**:539-547 (1997) |
| | ***Escherichia coli*** | K-12 | **B** | 4.60 | **University of Wisconsin** | **NHGRI** | Blattner *et. al.*, *Science* **277**: (1997) |
| | ***Bacillus subtilis*** | | **B** | 4.20 | **European & Japanese Consortium** | | manuscript submitted |
| | ***Archaeoglobus fulgidus*** | VC-16, DSM4304 | **A** | 2.20 | **TIGR** | **DOE** | manuscript submitted |
| | ***Borrelia burgdorferi*** | | **B** | 1.30 | **TIGR** | **Mathers Foundation** | manuscript in preparation |

# Microbial genomes in progress

| Genome | Strain | Domain | Size (Mb) | Institution | Funding | Anticipated Publication |
|---|---|---|---|---|---|---|
| *Actinobacillus actinomycetemcomitans* | | B | 2.2 | University of Oklahoma | NIH / NIDR | |
| *Aquifex aeolicus* | | B | 1.50 | Recombinant BioCatalysis Inc. | | |
| *Caulobacter crescentus* | | B | 3.80 | TIGR | | |
| *Chlamydia trachomatis* | serovar D (D/UW-3/Cx) | B | 1.05 | UC Berkeley & Stanford | NIAID | |
| *Clostridium acetobutylicum* | ATCC 824 | B | 4.1 | Genome Therapeutics | DOE | |
| *Deinococcus radiodurans* | R1 | B | 3.00 | TIGR | DOE | 1997 |
| *Enterococcus faecalis* | | B | 3.00 | TIGR | NIAID | 1999 |
| *Halobacterium salinarium* | | A | 4.0 | Max-Planck-Institute for Biochemistry | | |
| *Legionella pneumophila* | | B | 4.10 | TIGR | | |
| *Methanobacterium thermoautotrophicum* | delta H | A | 1.75 | Genome Therapeutics & Ohio State Univ. | DOE | |
| *Mycobacterium avium* | | B | 4.70 | TIGR | NIH | 2000 |
| *Mycobacterium tuberculosis* | CSU#93 (clinical isolate) | B | 4.40 | TIGR | NIAID | 1998 |
| *Mycobacterium tuberculosis* | H37Rv (lab strain) | B | 4.40 | Sanger Centre | Wellcome Trust | |
| *Neisseria gonorrhoeae* | | B | 2.20 | University of Oklahoma | NIAID | |
| *Neisseria meningitidis* | | B | 2.30 | TIGR / Oxford University | TIGR | |
| *Plasmodium falciparum* **Chr1** (clone 3D7) | | E | 0.8 | Sanger Centre | Wellcome Trust | |
| *Plasmodium falciparum* **Chr2** (clone 3D7) | | E | 1.00 | TIGR / NMRI | NIAID | 1997 |
| *Plasmodium falciparum* **Chr3** (clone 3D7) | | E | 1.20 | Sanger Centre | Wellcome Trust | |
| *Plasmodium falciparum* **Chr4** (clone 3D7) | | E | 1.5 | Sanger Centre | Wellcome Trust | |
| *Plasmodium falciparum* **Chr9** (clone 3D7) | | E | 1.80 | TIGR / NMRI | | |
| *Plasmodium falciparum* **Chr10** (clone 3D7) | | E | 2.10 | TIGR / NMRI | | |

| | | | | | | |
|---|---|---|---|---|---|---|
| *Plasmodium falciparum* **Chr12** (clone 3D7) | | **E** | 2.4 | Stanford University | Burroughs Wellcome Fund | |
| *Plasmodium falciparum* **Chr14** (clone 3D7) | | **E** | 3.4 | **TIGR / NMRI** | Burroughs Wellcome Fund | |
| ***Porphyromonas gingivalis*** | | **B** | 2.20 | **TIGR / Forsyth Dental Center** | **NIDR** | |
| *Pseudomonas aeruginosa* | PAO1 | **B** | 5.90 | **International collaboration** | | |
| ***Pyrobaculum aerophilum*** | | **A** | 2.3 | **Caltech / UCLA / Recombinant BioCatalysis Inc.** | **ONR / DOE** | |
| ***Pyrococcus furiosus*** | | **A** | 2.10 | **Center of Marine Biotechnology / Univ. Utah** | **DOE** | |
| ***Pyrococcus shinkaj*** | | **A** | 2.00 | **MITI / University of Tokyo** | | |
| ***Rickettsia prowazekii*** | | **B** | 1.10 | **University of Uppsala** | **SSF / NFR** | |
| ***Salmonella typhimurium*** | | **B** | 4.50 | **TIGR** | | |
| ***Shewanella putrefaciens*** | MR-1 | **B** | 4.50 | **TIGR** | **DOE** | |
| ***Streptococcus pneumoniae*** | | **B** | 2.20 | **TIGR** | **TIGR** | 1997 |
| *Streptococcus pyogenes* | | **B** | 1.98 | **University of Oklahoma** | **NIAID** | |
| *Sulfolobus solfataricus* | | **A** | 3.05 | **Canadian & European Consortium** | | |
| *Thermotoga maritima* | MSB8 | **B** | 1.80 | **TIGR** | **DOE** | 1998 |
| ***Treponema denticola*** | | **B** | 3.00 | **TIGR / Univ. Texas** | | |
| *Treponema pallidum* | | **B** | 1.05 | **TIGR / Univ. Texas** | **NIAID** | 1997 |
| ***Thermoplasma acidophilum*** | | **A** | 1.7 | **Max-Planck-Institute for Biochemistry** | | |
| ***Ureaplasma urealyticum*** | serovar 3 | **B** | 0.75 | **U. Alabama / PE-ABI** | **PE-ABI / NIH / UAB** | 1997 |
| *Vibrio cholerae* | serotype O1, Biotype El Tor, strain N16961 | **B** | 2.50 | **TIGR** | **NIAID** | 1998 |

**KEY**

| **Domain** | **A: Archaea** | **B: Eubacteria** | **E: Eucaryote** |
|---|---|---|---|

Comments and questions to: Anthony R. Kerlavage

# Reference List


Wu, R. and Taylor, E. (1971). Nucleotide sequence analysis of DNA. II. Complete nucleotide sequence of the cohesive ends of bacteriophage lambda DNA. J.Mol.Biol. *57*, 491-511.

Sanger, F., Nicklen, S., and Coulson, A.R. (1977). DNA sequencing with chain-terminating inhibitors. Proc.Natl.Acad.Sci.USA *74*, 5463-5467.

Wilson, R.K., Chen, C., Avdalovic, N., Burns, J., and Hood, L. (1990). Development of an automated procedure for fluorescent DNA sequencing. Genomics *6*, 626-634.

Lander, E.S. and Waterman, M.S. (1988). Genomic mapping by fingerprinting random clones: a mathematical analysis. Genomics *2*, 231-239.

Lee, L.G., Spurgeon, S.L., Heiner, C.R., Benson, S.C., Rosenblum, B.B., Menchen, S.M., Graham, R.J., Constantinescu, A., Upadhya, K.G., and Cassel, J.M. (1997). New energy transfer dyes for DNA sequencing. Nucleic.Acids.Res. *25*, 2816-2822.

Ju, J., Glazer, A.N., and Mathies, R.A. (1996). Energy transfer primers: a new fluorescence labeling paradigm for DNA sequencing and analysis. Nat.Med. *2*, 246-249.

Genome Analysis, A Laboratory Manual (1997). Analyzing DNA (Cold Spring Harbor: Cold Spring Harbor Laboratory Press).

Hillier, L., Lennon, G., Becker, M., Bonaldo, M., Chiapelli, B., Chissoe, S., Dietrich, N., DuBuque, T., Favello, A., Gish, W., et al. (1996). Generation and analysis of 280,000 human expressed sequence tags. Genome Res. *6*, 807-828.

Adams, M.D., Kerlavage, A.R., Fleischmann, R.D., Fuldner, R.A., Bult, C.J., Lee, N.H., Kirkness, E.F., Weinstock, K.J., Gocayne, J.D., White, O., Sutton, G., et al. (1995). Initial assessment of human gene diversity and expression patterns based upon 83 million nucleotides of cDNA sequence. Nature *377*, 3-174.

Boguski, M.S. and Schuler, G.D. (1995). ESTablishing a human transcript map. Nature Genet. *10*, 369-371.

Yu, W., Andersson, B., Worley, K.C., Muzny, D.M., Ding, Y., Liu, W., Ricafrente, J.Y., Wentland, M.A., Lennon, G., and Gibbs, R.A. (1997). Large-scale concatenation cDNA sequencing [letter]. Genome Res. *7*, 353-358.

Koonin, E.V. (1997). Big time for small genomes. Genome Res. *7*, 418-421.

# URLs

Links to almost ALL Genome related sites:

http://www-ls.lanl.gov/HGhotlist.html

Bacterial genomics database:

http://www.tigr.org/