

CAP 5510: Introduction to Bioinformatics
CGS 5166: Bioinformatics Tools

Giri Narasimhan

ECS 254; Phone: x3748

giri@cis.fiu.edu

www.cis.fiu.edu/~giri/teach/BioinfF18.html

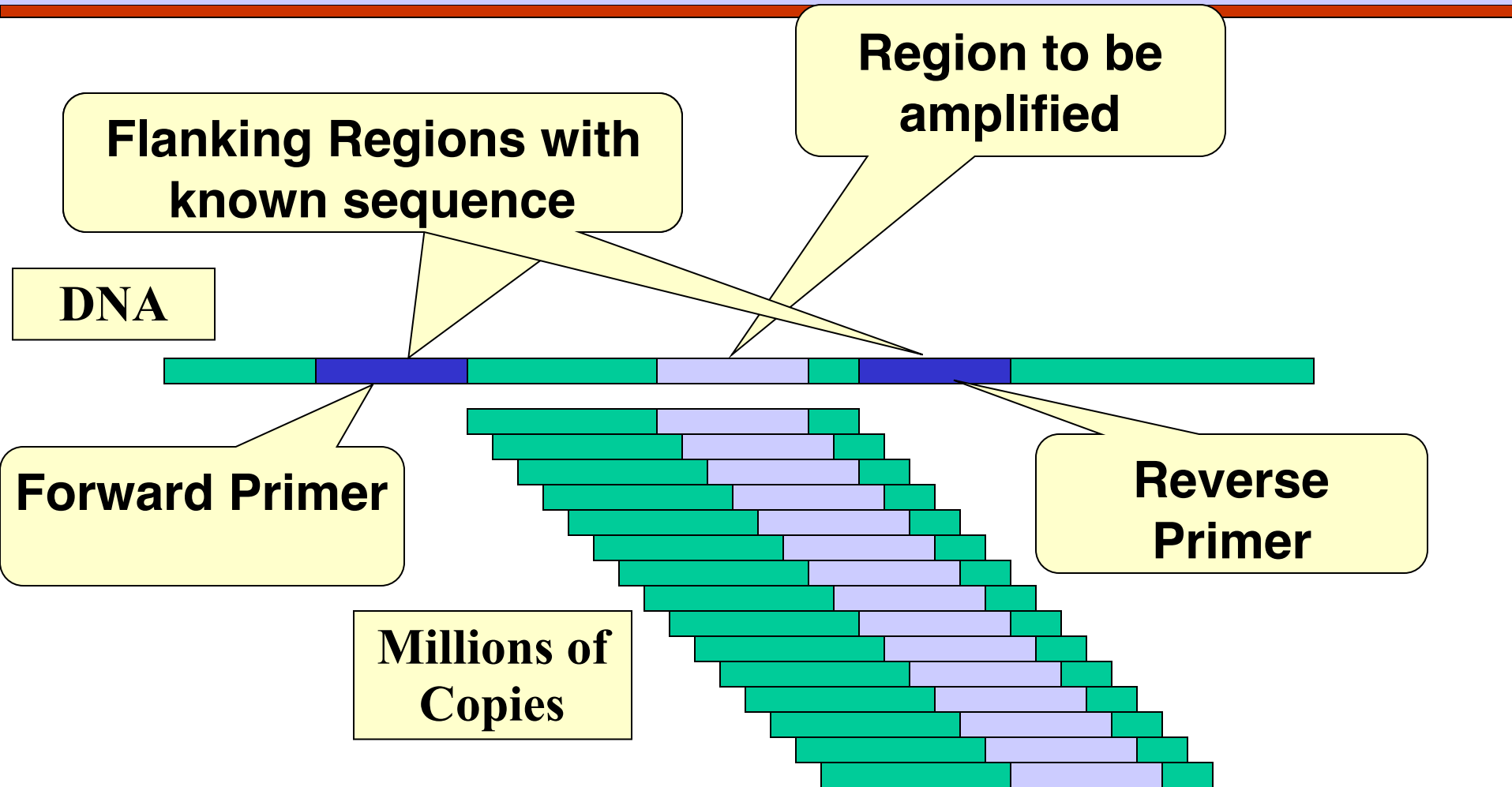
PCR and Sequencing



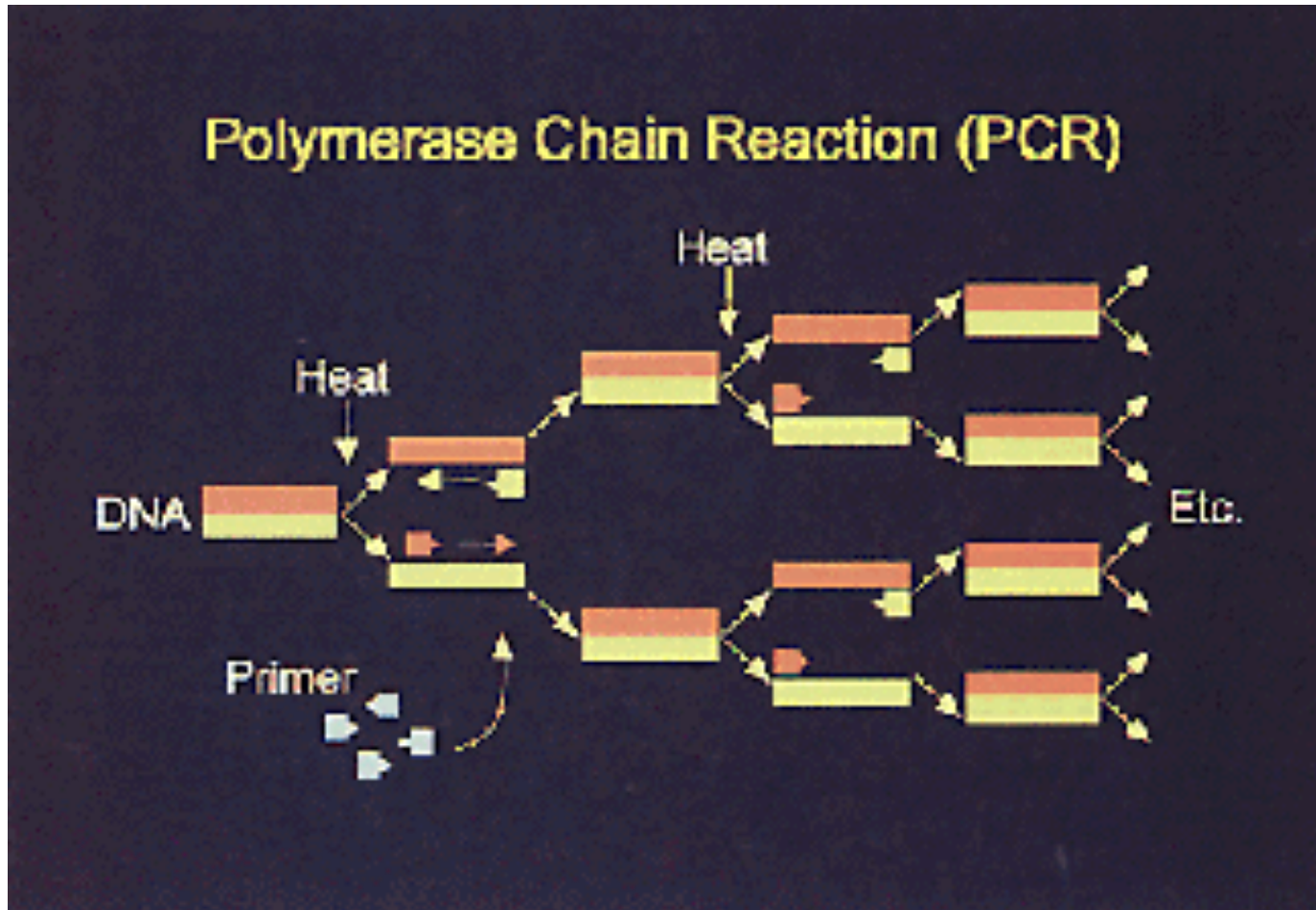
Polymerase Chain Reaction (PCR)

- For testing, large amount of DNA is needed
 - Identifying individuals for forensic purposes
 - (0.1 microliter of saliva contains enough epithelial cells)
 - Identifying pathogens (viruses and/or bacteria)
- PCR is a technique to amplify the number of copies of a specific region of DNA.
- Useful when exact DNA sequence is unknown
- Need to know “flanking” sequences
- Primers designed from “flanking” sequences

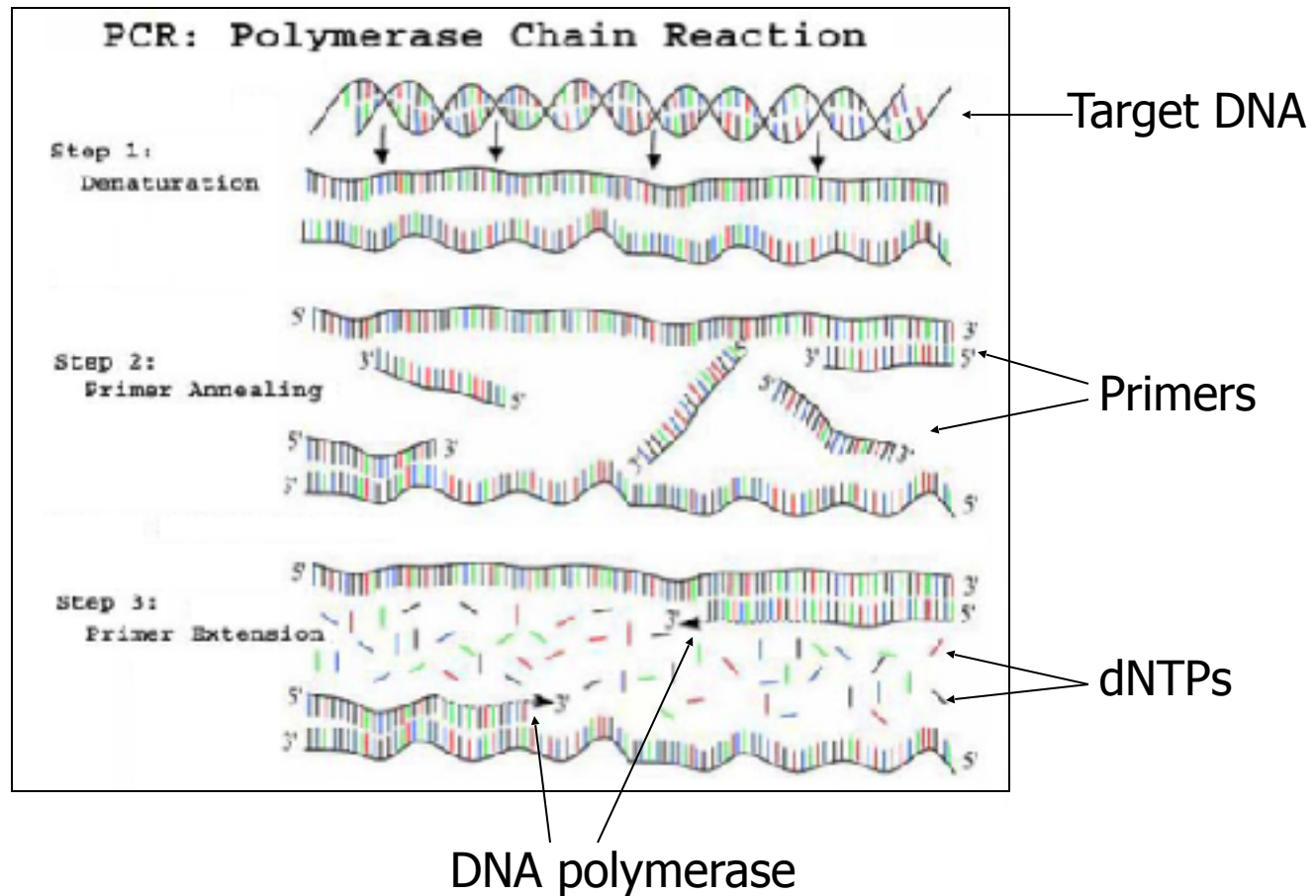
PCR



PCR

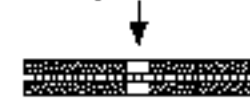


Schematic outline of a typical PCR cycle

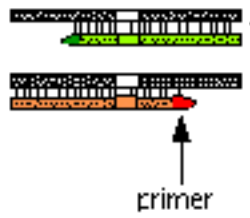


POLYMERASE CHAIN REACTION

DNA region of interest.



1. DNA is denatured. Primers attach to each strand. A new DNA strand is synthesized behind primers on each template strand.



primer



2. Another round: DNA is denatured, primers are attached, and the number of DNA strands are doubled.

3. Another round: DNA is denatured, primers are attached, and the number of DNA strands are doubled.

4. Another round: DNA is denatured, primers are attached, and the number of DNA strands are doubled.

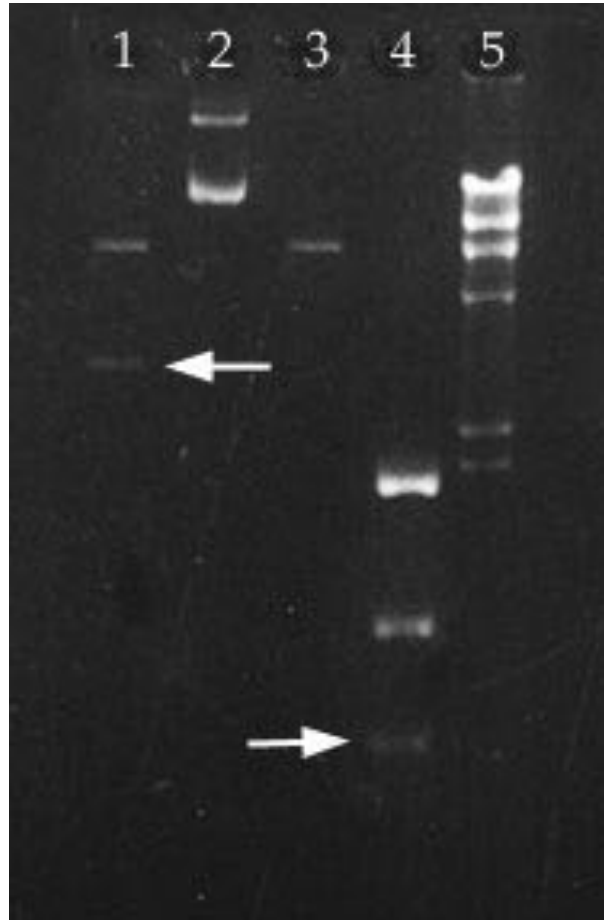
5. Continued rounds of amplification swiftly produce large numbers of identical fragments. Each fragment contains the DNA region of interest.



Gel Electrophoresis

- ❑ Used to measure the lengths of DNA fragments.
- ❑ When voltage is applied to DNA, different size fragments migrate to different distances (smaller ones travel farther).

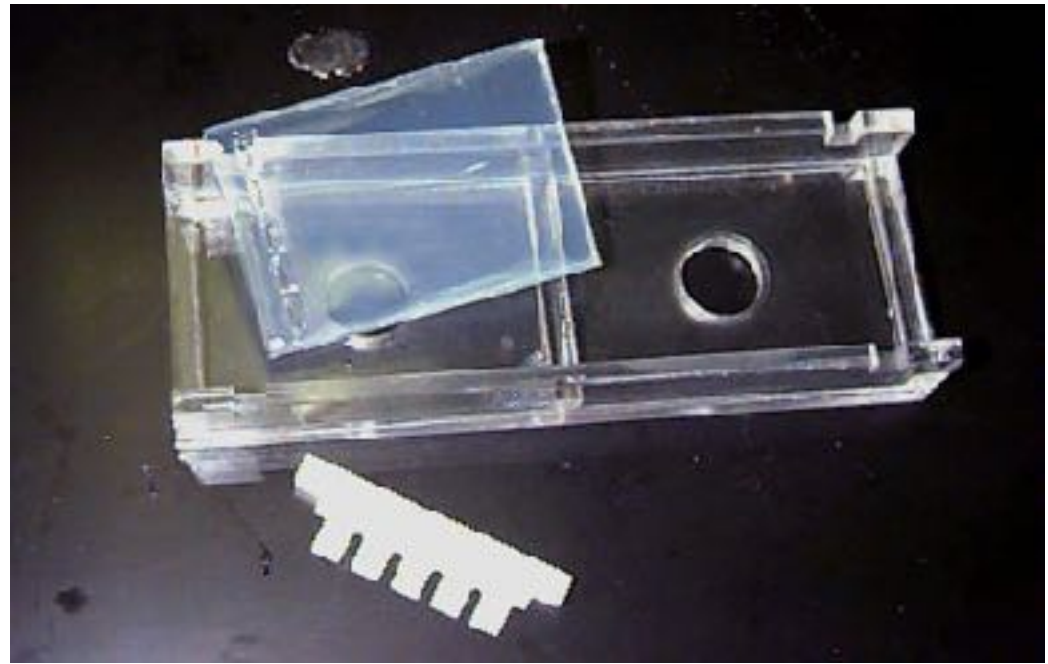
Gel Pictures



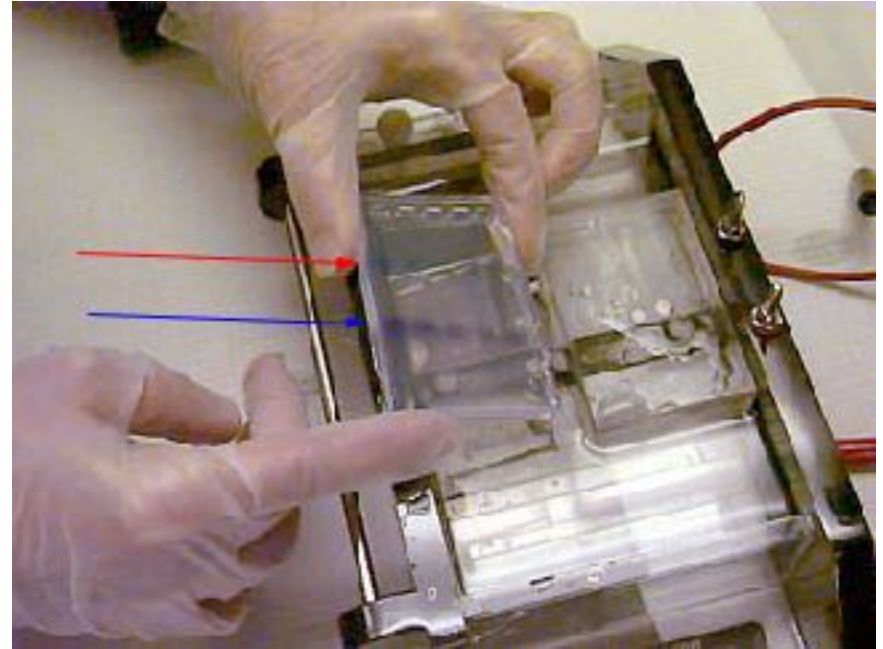
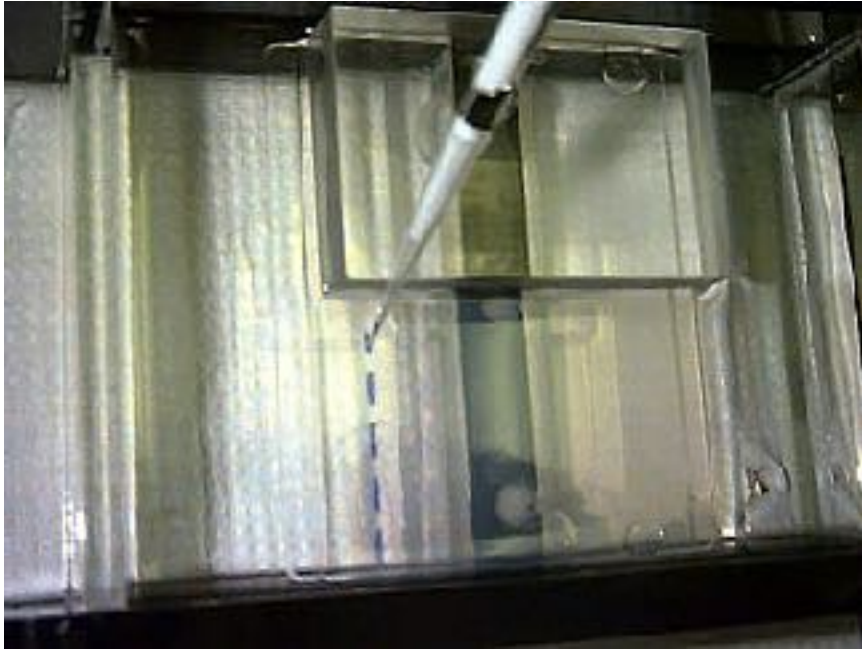
Gel Electrophoresis: Measure sizes of fragments

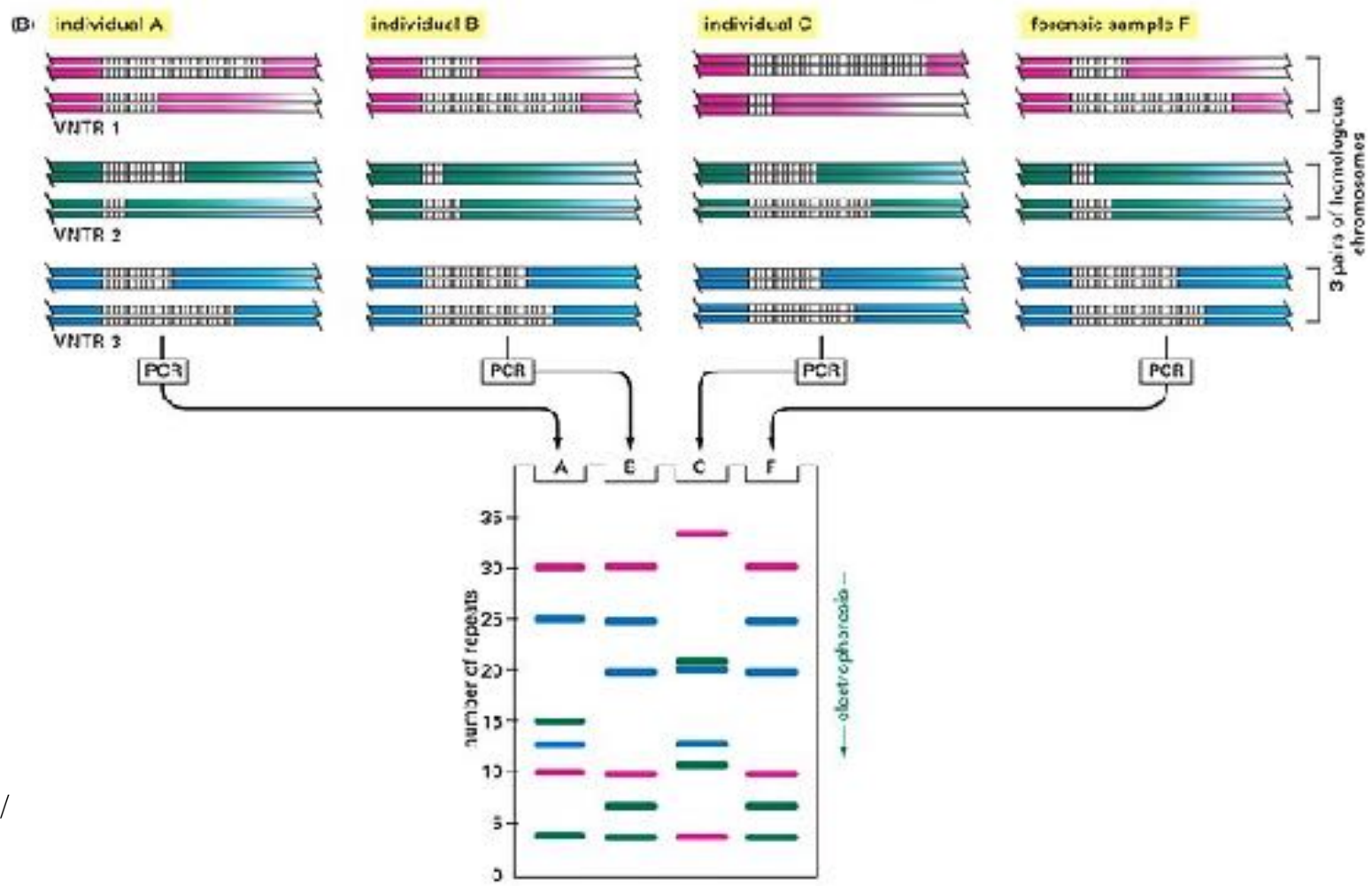
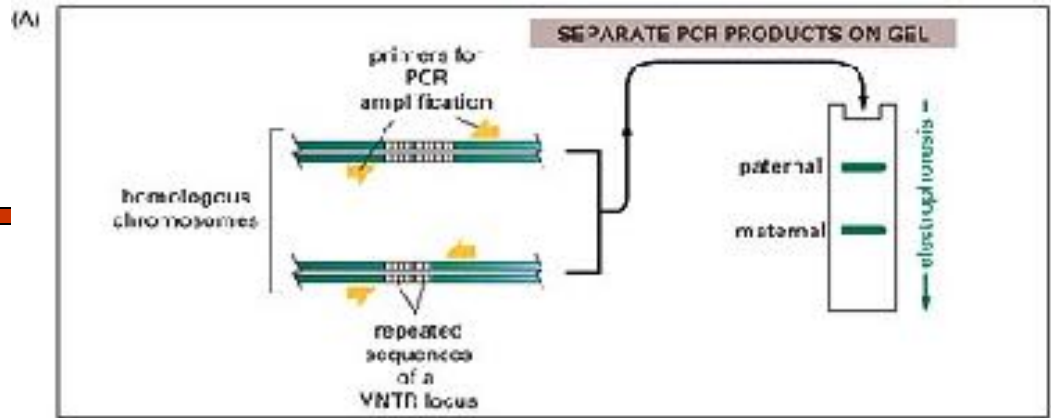
- ❑ The phosphate backbone makes DNA a highly negatively charged molecule.
- ❑ DNA can be separated according to its size.
- ❑ **Gel**: allow hot 1% solution of purified agarose to cool and solidify/polymerize.
- ❑ DNA sample added to wells at the top of a gel and voltage is applied. Larger fragments migrate through the pores slower.
- ❑ Varying concentration of agarose makes different pore sizes & results.
- ❑ Proteins can be separated in much the same way, only acrylamide is used as the crosslinking agent.

Gel Electrophoresis

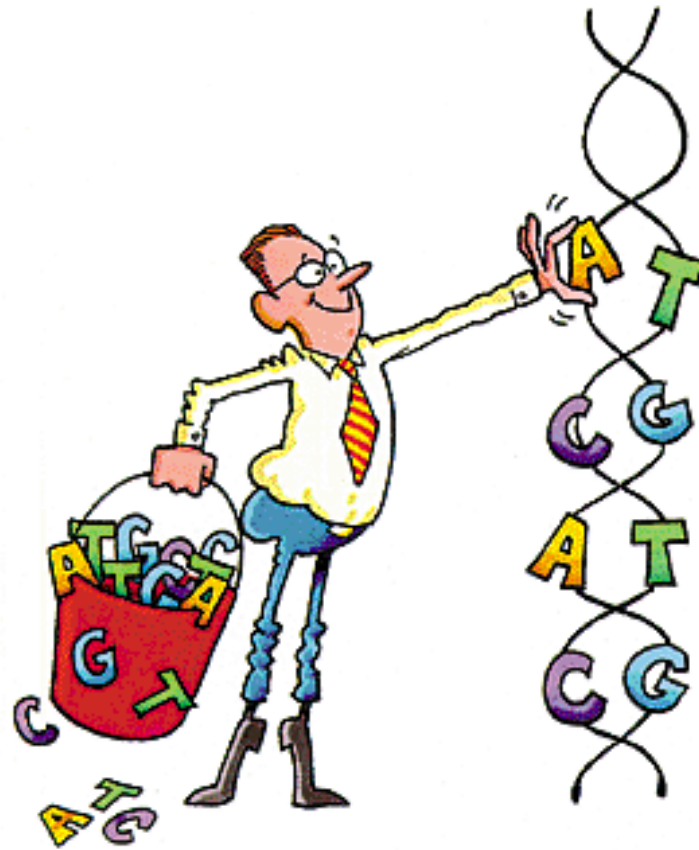


Gel Electrophoresis





Sequencing



Why sequencing?

□ Useful for further study:

- Locate gene sequences, regulatory elements
- Compare sequences to find similarities
- Identify mutations
- Use it as a basis for further experiments

Next 4 slides contains material prepared by Dr. Stan Metzenberg. Also see:
<http://stat-www.berkeley.edu/users/terry/Courses/s260.1998/Week8b/week8b/node9.html>

History

- Two methods independently developed in 1974
 - Maxam & Gilbert method
 - Sanger method: became the standard
- Nobel Prize in 1980

Original Sanger Method

- (Labeled) Primer is annealed to template strand of denatured DNA. This primer is specifically constructed so that its 3' end is located next to the DNA sequence of interest. Once the primer is attached to the DNA, the solution is divided into four tubes labeled "G", "A", "T" and "C". Then reagents are added to these samples as follows:
 - "G" tube: ddGTP, DNA polymerase, and all 4 dNTPs
 - "A" tube: ddATP, DNA polymerase, and all 4 dNTPs
 - "T" tube: ddTTP, DNA polymerase, and all 4 dNTPs
 - "C" tube: ddCTP, DNA polymerase, and all 4 dNTPs
- DNA is synthesized, & nucleotides are added to growing chain by the DNA polymerase. Occasionally, a ddNTP is incorporated in place of a dNTP, and the chain is terminated. Then run a gel.
- All sequences in a tube have same prefix and same last nucleotide.
- <http://www.wellcome.ac.uk/Education-resources/Teaching-and-education/Animations/DNA/WTDV026689.htm>

Sanger Method

- Example of sequences seen in gel from "G" tube:

```
5'-GAATGTCCTTTCTCTAAGTCCTAAG
3'-GGAGACTTACAGGAAAGAGATTCAGGATTCAGGAGGCCTACCATGAAGATCAAG-5'

5'-GAATGTCCTTTCTCTAAGTCCTAAGTCCTCCG
3'-GGAGACTTACAGGAAAGAGATTCAGGATTCAGGAGGCCTACCATGAAGATCAAG-5'

5'-GAATGTCCTTTCTCTAAGTCCTAAGTCCTCCG
3'-GGAGACTTACAGGAAAGAGATTCAGGATTCAGGAGGCCTACCATGAAGATCAAG-5'

5'-GAATGTCCTTTCTCTAAGTCCTAAGTCCTCCGGATG
3'-GGAGACTTACAGGAAAGAGATTCAGGATTCAGGAGGCCTACCATGAAGATCAAG-5'

5'-GAATGTCCTTTCTCTAAGTCCTAAGTCCTCCGGATG
3'-GGAGACTTACAGGAAAGAGATTCAGGATTCAGGAGGCCTACCATGAAGATCAAG-5'

5'-GAATGTCCTTTCTCTAAGTCCTAAGTCCTCCGGATGGTACTTCTAG
3'-GGAGACTTACAGGAAAGAGATTCAGGATTCAGGAGGCCTACCATGAAGATCAAG-5'
```

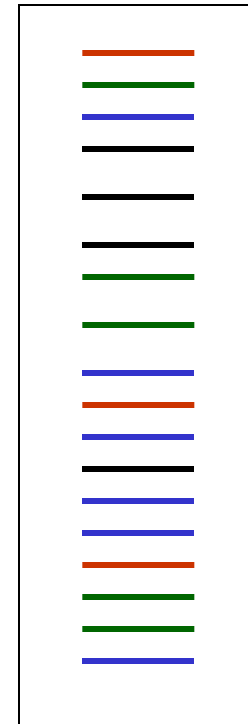
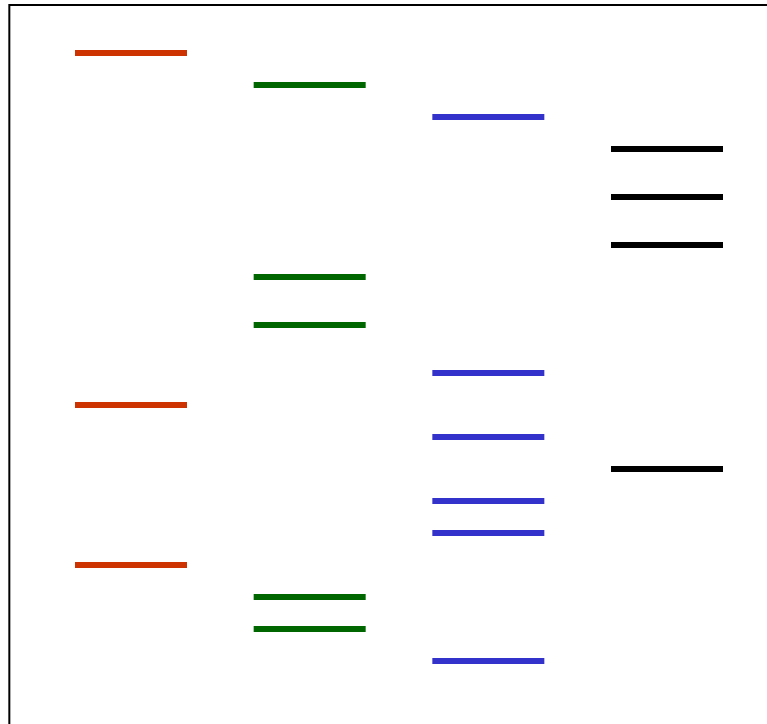
Modified Sanger

- Reactions performed in a single tube containing all four ddNTP's, each labeled with a different color dye



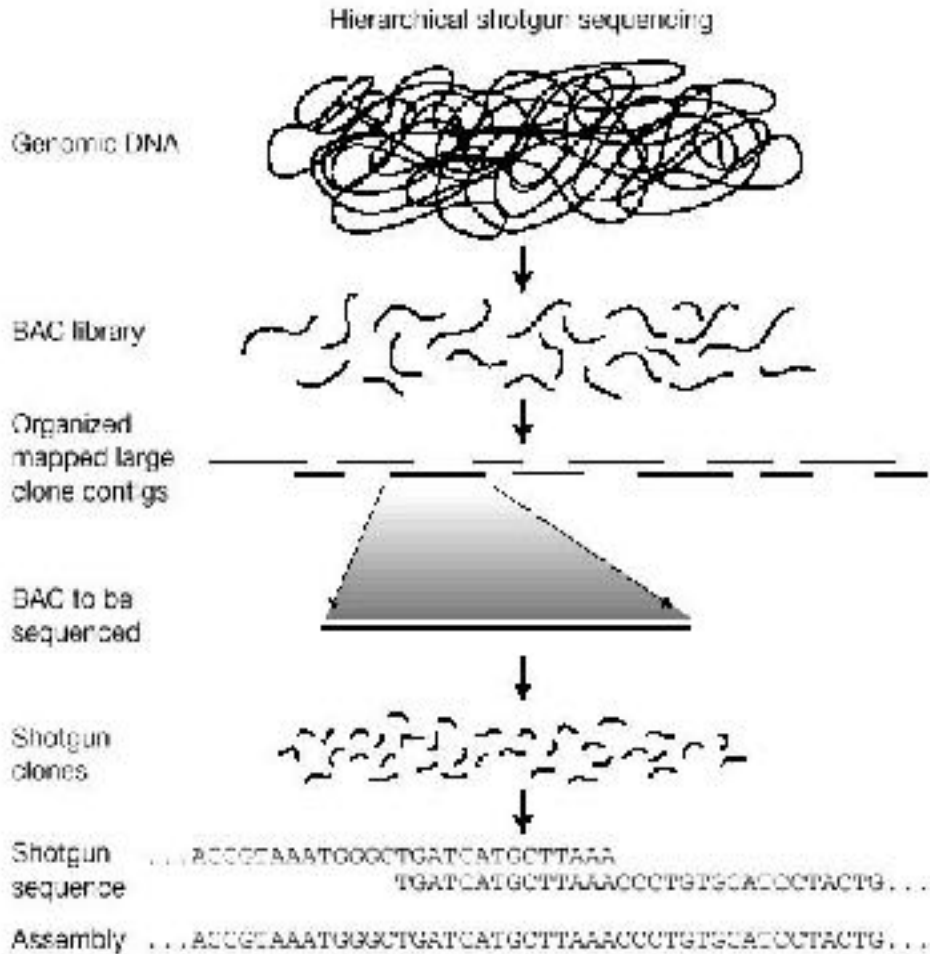
Sequencing Gels: Separate vs Single Lanes

GCCAGGTGAGCCTTTGCA



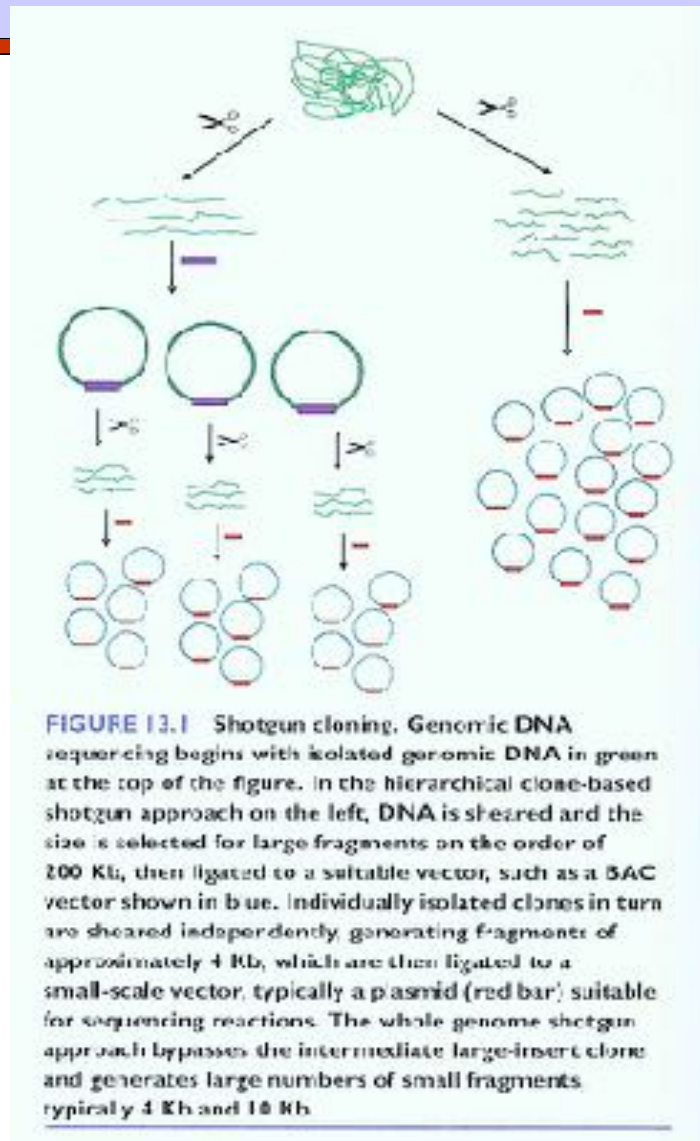
A C G T

Shotgun Sequencing



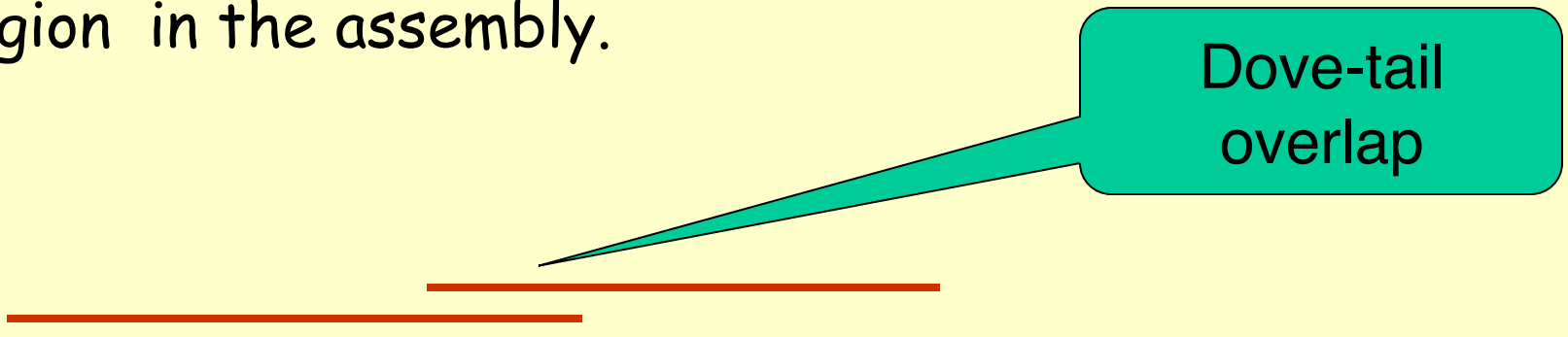
From <http://www.tulane.edu/~biochem/lecture/723/humgen.html>

Sequencing



Sequencing: Generate Contigs

- Short for “contiguous sequence”. A continuously covered region in the assembly.



Dove-tail overlap



Collapsing into a single sequence

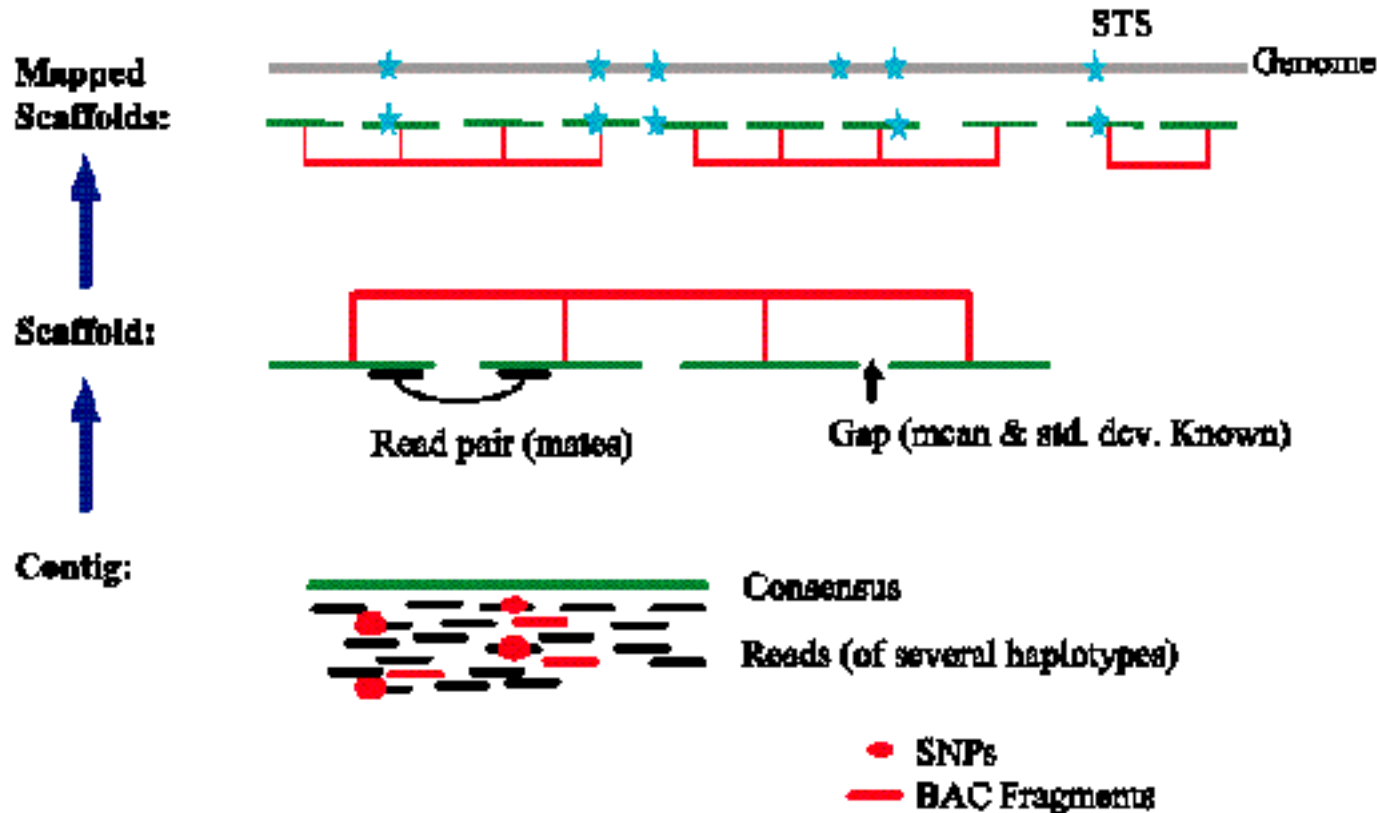
- Jang W et al (1999) Making effective use of human genomic sequence data. *Trends Genet.* 15(7): 284-6.
Kent WJ and Haussler D (2001) Assembly of the working draft of the human genome with GigAssembler. *Genome Res* 11(9): 1541-8.

Paired Reads

- **Scaffold (supercontig)**: formed when two **contigs** with no sequence overlap can be linked
 - Data from paired end reads help create scaffolds with known gaps
 - If two reads end up in two different contigs, then we can link contigs to form scaffold.



Shotgun Sequencing



From <http://www.tulane.edu/~biochem/lecture/723/humgen.html>

Human Genome Project

- ❑ Many videos available on youtube.com, dnatube.com, and elsewhere.
- ❑ Find some and watch them.

Assembly: Simple Example

□ ACCGT, CGTGC, TTAC, TACCGT

□ Total length = ~10

□

- **---ACCGT---**
- **-----CGTGC**
- **TTAC-----**
- **-TACCGT-**
- **TTACCGTGC**

Assembly: Complications

- ❑ Errors in input sequence fragments (~3%)
 - Indels or substitutions
- ❑ Contamination by host DNA
- ❑ Chimeric fragments (joining of non-contiguous fragments)
- ❑ Unknown orientation
- ❑ Repeats (long repeats)
 - Fragment contained in a repeat
 - Repeat copies not exact copies
 - Inherently ambiguous assemblies possible
 - Inverted repeats
- ❑ Inadequate Coverage

Assembly: Complications

```

w = AGTATTGGCAATC
z = AATCGATG
u = ATGCAAACCT
x = CCTTTTGG
y = TTGGCAATCACT

AGTATTGGCAATC---AATCGATG-----
-----ATGCAAACCT-----
---TTGGCAATCACT-----CCTTTTGG
-----
AGTATTGGCAATCACTAATCGATGCAAACCTTTTGG

```

FIGURE 4.20

A bad solution for an assembly problem, with a multiple alignment whose consensus is a shortest common superstring. This solution has length 36 and is generated by the Greedy algorithm. However, its weakest link is zero.

```

AGTATTGGCAATC-----CCTTTTGG-----
-----AATCGATG-----TTGGCAATCACT
-----ATGCAAACCT-----
-----
AGTATTGGCAATCGATGCAAACCTTTTGGCAATCACT

```

FIGURE 4.21

Solution according to the unique Hamiltonian path. This solution has length 37, but exhibits better linkage. Its weakest link is 3.

Assembly: Complications

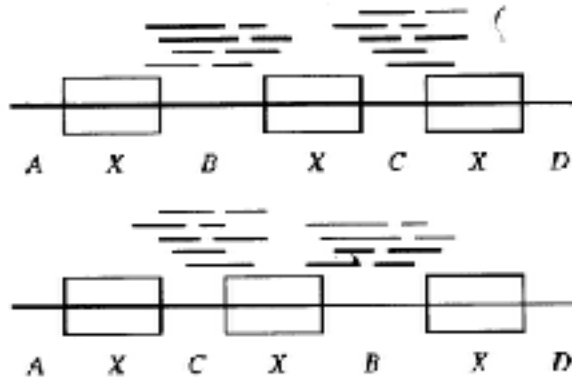


FIGURE 4.8

Target sequence leading to ambiguous assembly because of repeats of the form XXX.

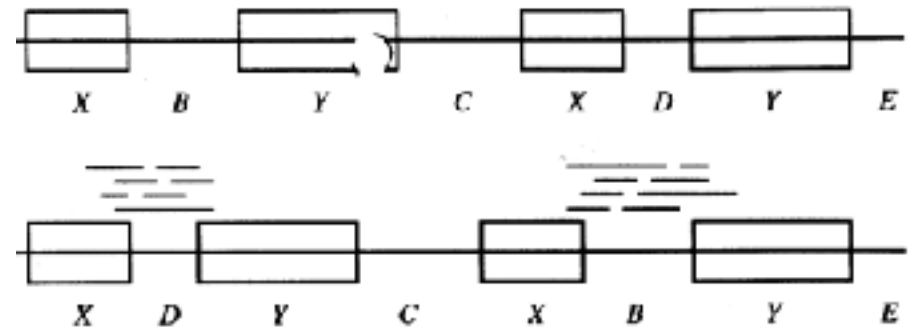


FIGURE 4.9

Target sequence leading to ambiguous assembly because of repeats of the form XYXY.

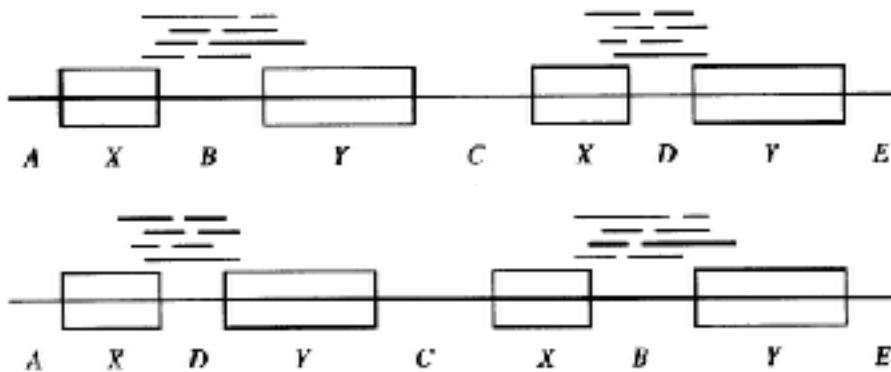


FIGURE 4.9

Target sequence leading to ambiguous assembly because of repeats of the form XYXY.

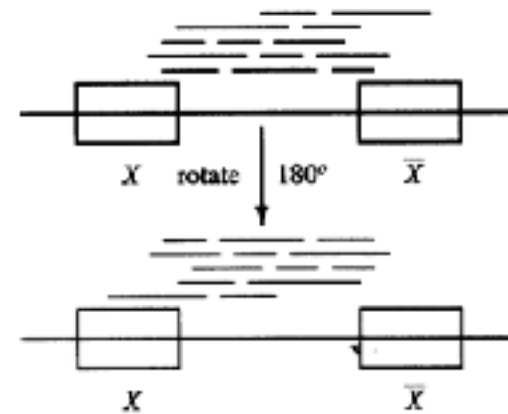


FIGURE 4.10

Target sequence with inverted repeat. The region marked \bar{X} is the reverse complement of the region marked X.