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/BioinfS11.html

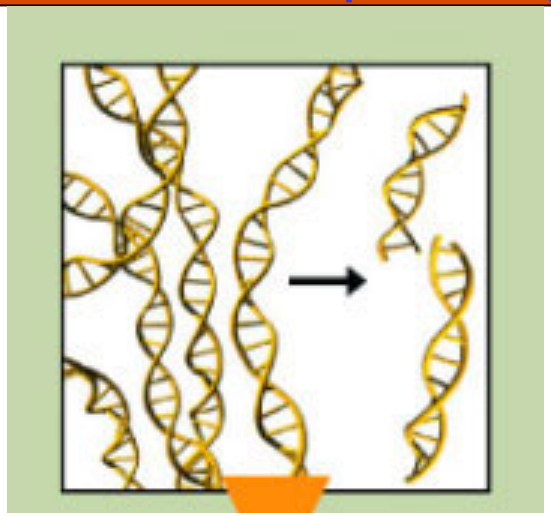
Other sequencing methods

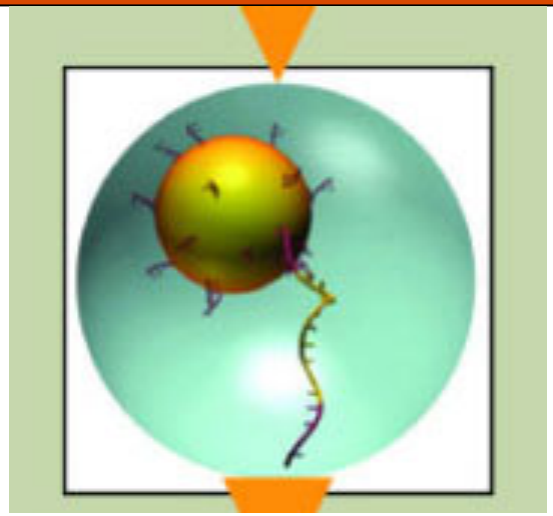
- Sanger Method (70Kbp/run)
- Sequencing by Hybridization (SBH)
- Dual end sequencing
- Chromosome Walking (see page 5-6 of Pevzner's text)
- 454 Sequencing (60Mbp/run)
- Solexa Sequencing (600Mbp/run) [Illumina]

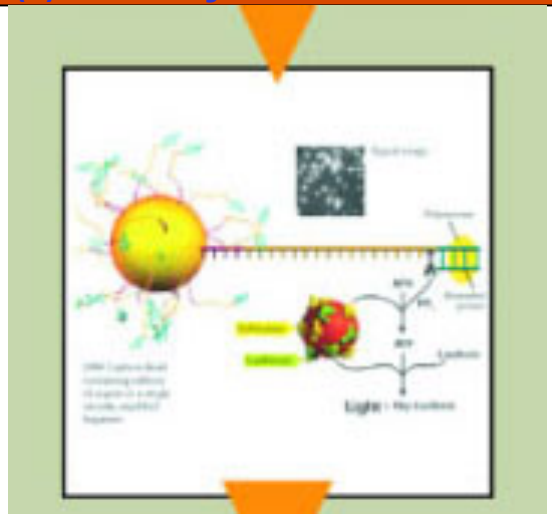
454 Sequencing: New Sequencing Technology

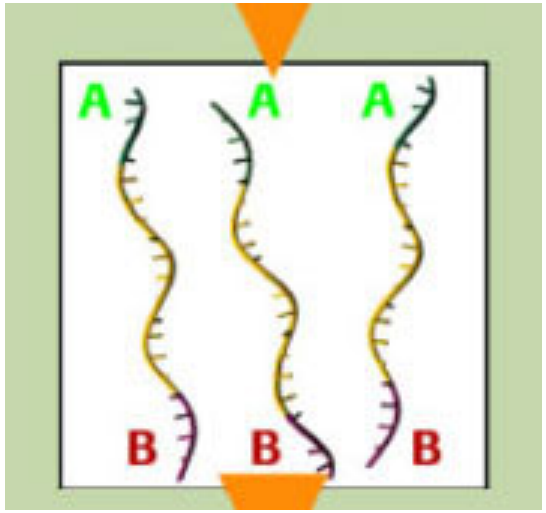
- 454 Life Sciences, Roche
- Fast (20 million bases per 4.5 hour run)
- Low cost (lower than Sanger sequencing)
- Simple (entire bacterial genome in days with one person -- without cloning and colony picking)
- Convenient (complete solution from sample prep to assembly)
- PicoTiterPlate Device
  - Fiber optic plate to transmit the signal from the sequencing reaction
- Process:
  - Library preparation: Generate library for hundreds of sequencing runs
  - Amplify: PCR single DNA fragment immobilized on bead
  - Sequencing: "Sequential" nucleotide incorporation converted to chemilluminscent signal to be detected by CCD camera.

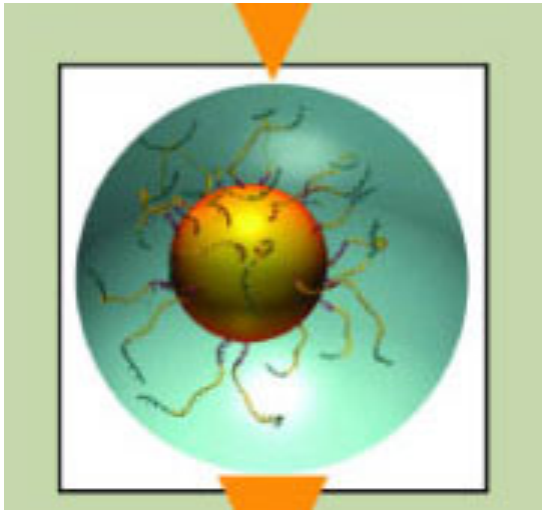
(a) Fragment, (b) add adaptors, (c) “1 fragment, 1 bead”, (d) emPCR on bead, (e) put beads in PicoTiterPlate and start sequencing: “1 bead, 1 read”, and (f) analyze

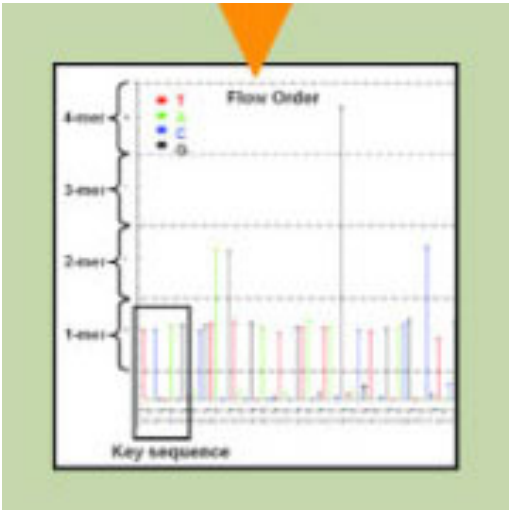






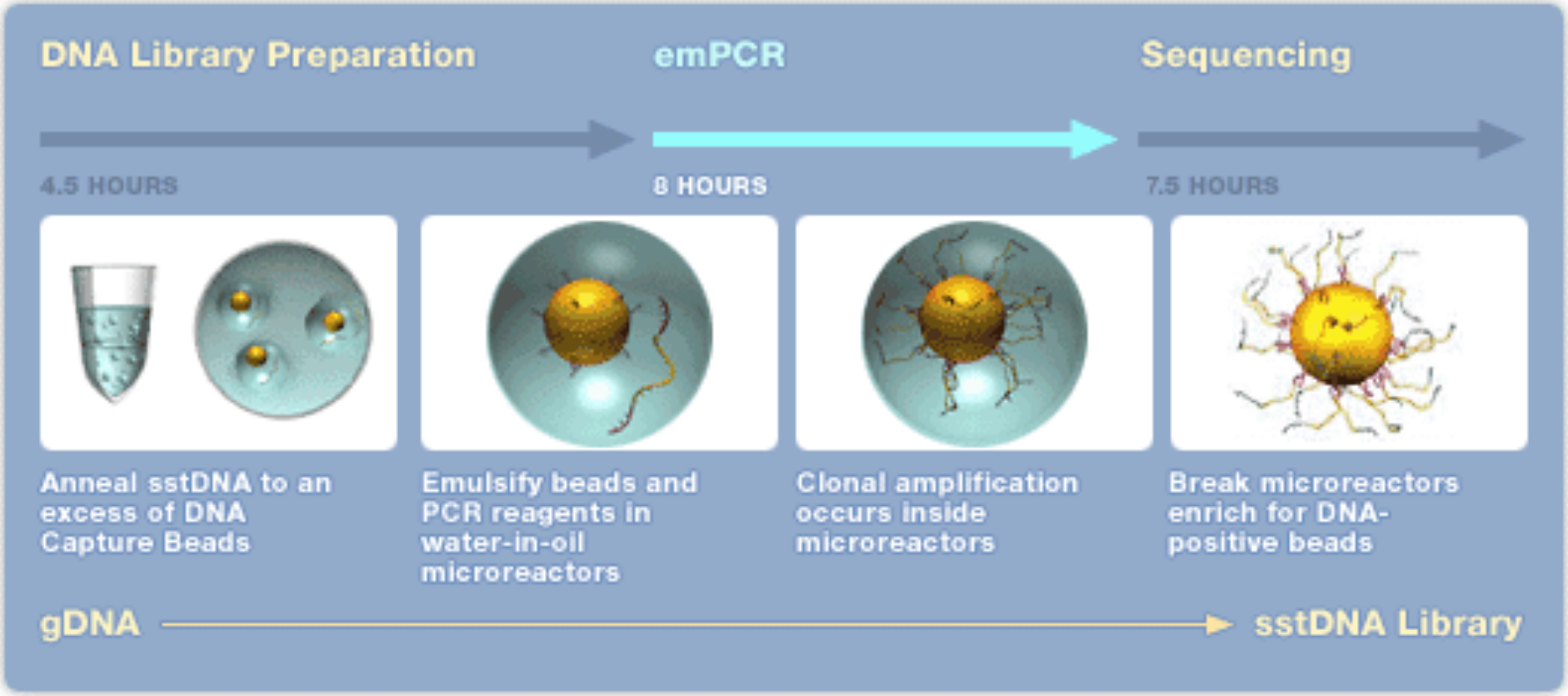






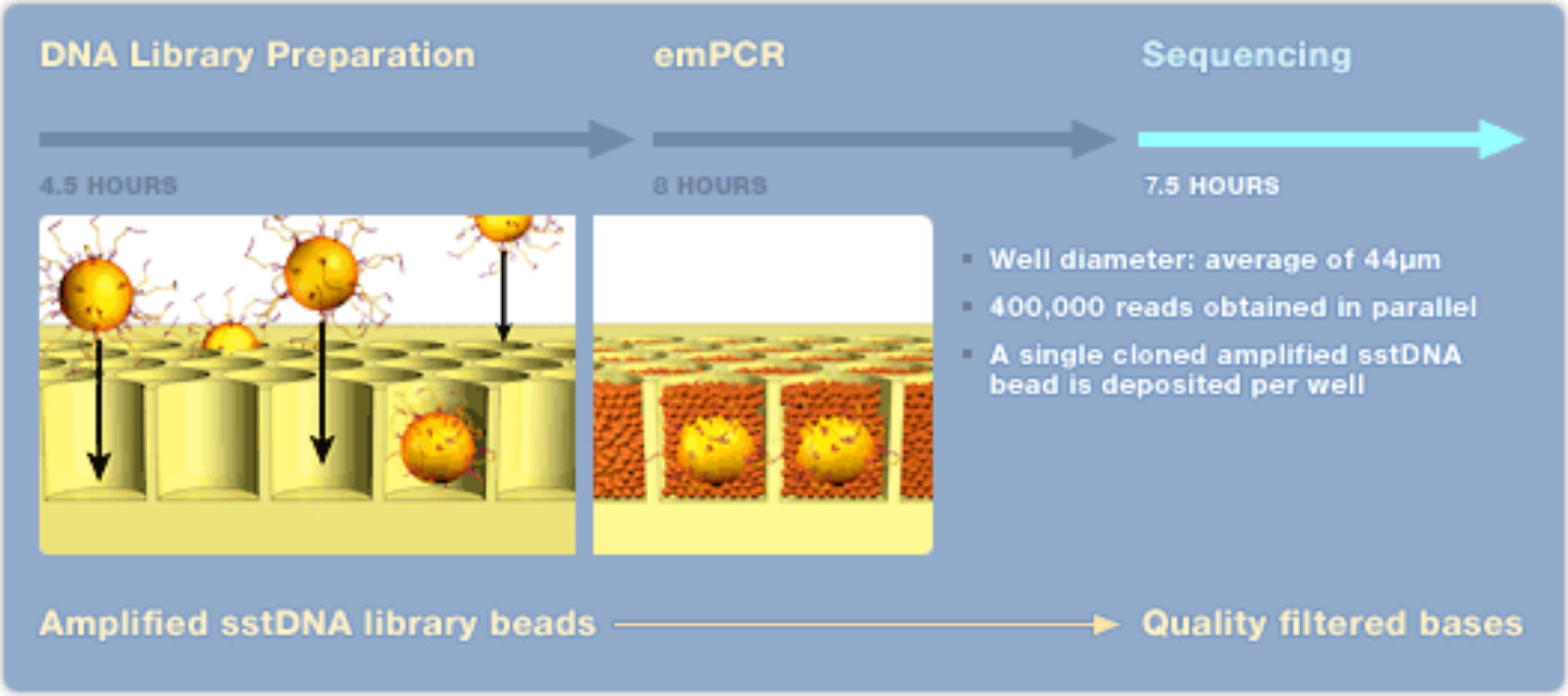
emPCR

FIGURE 8



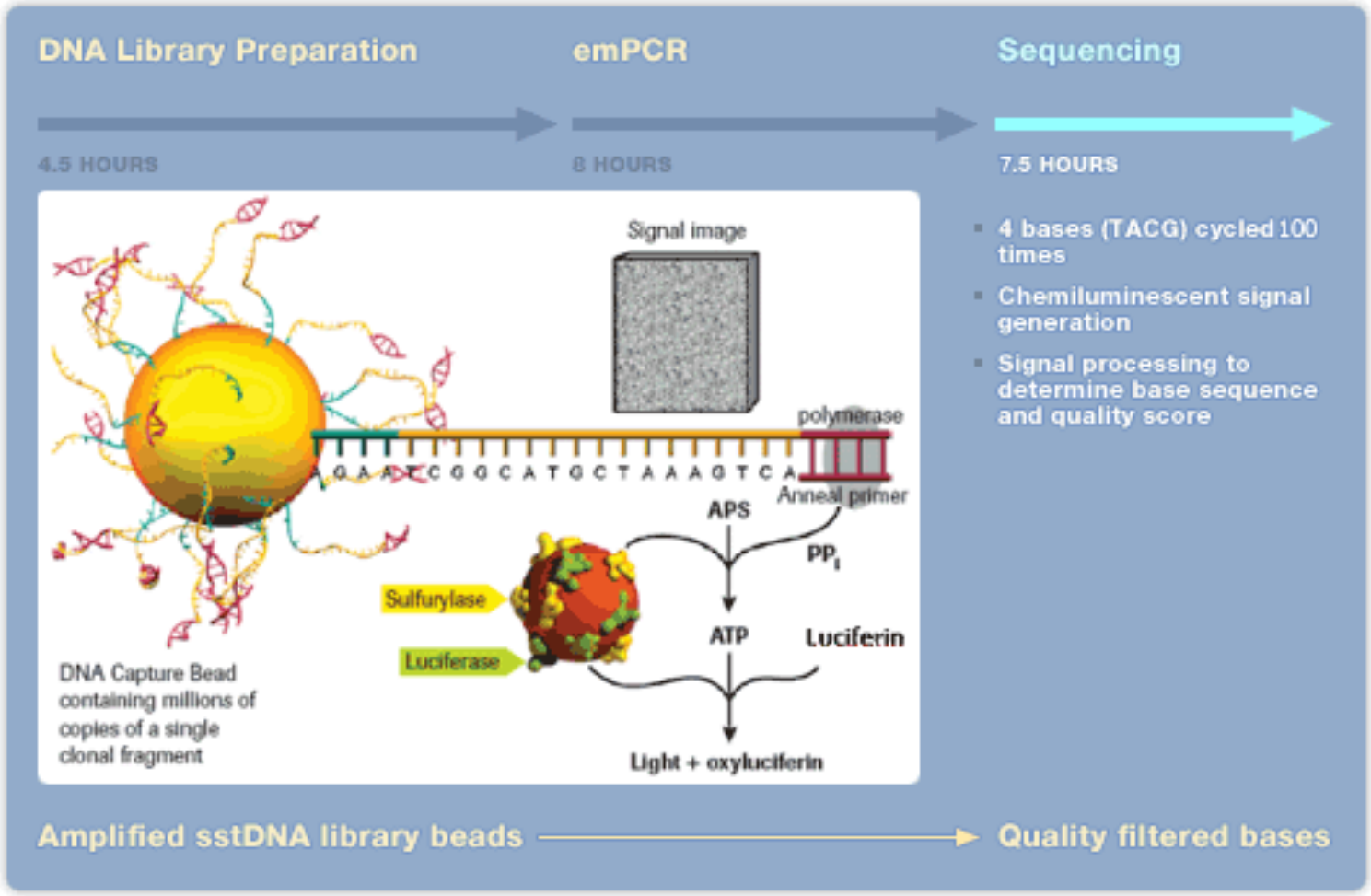
Sequencing

FIGURE 9

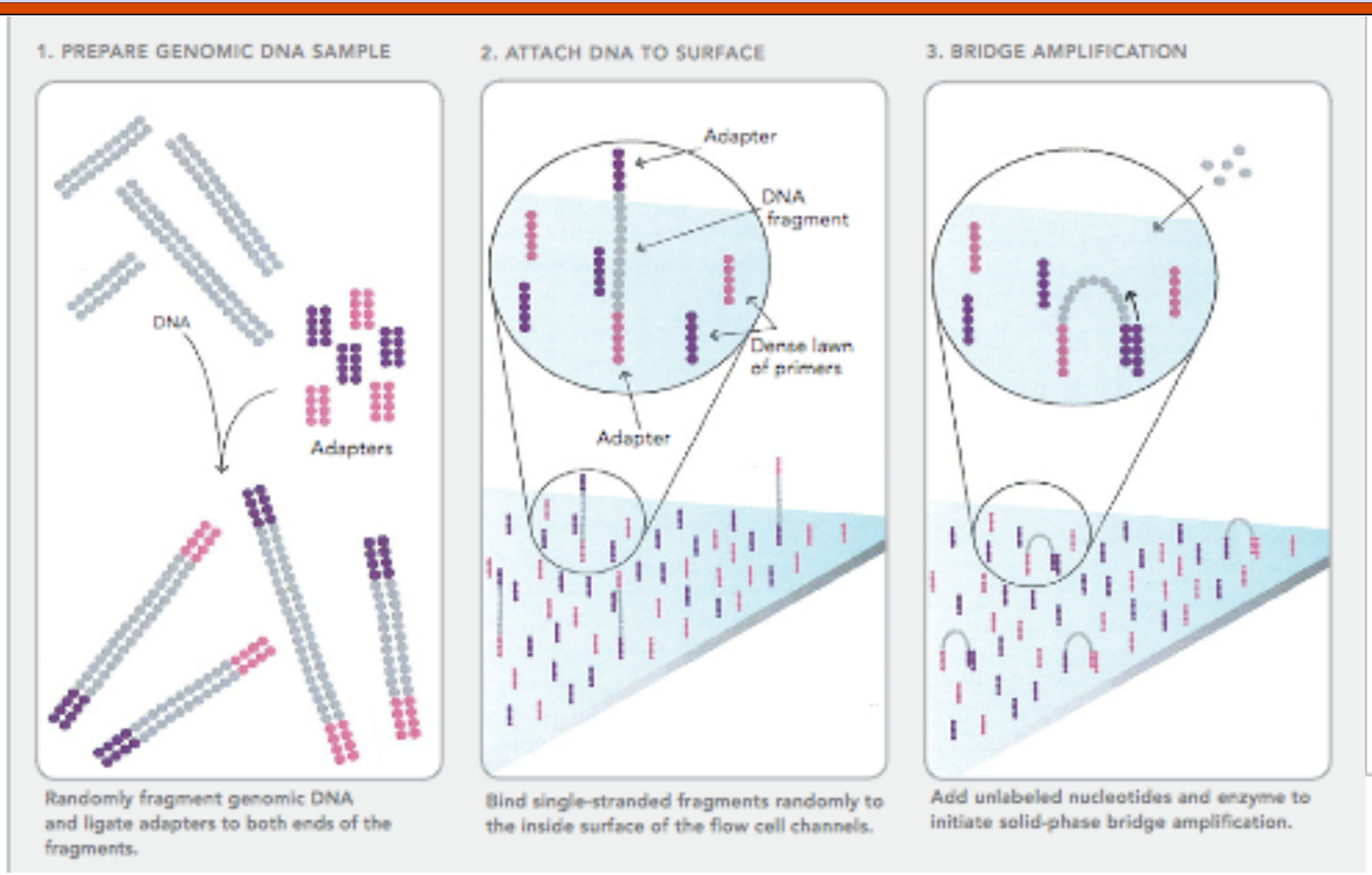


Sequencing

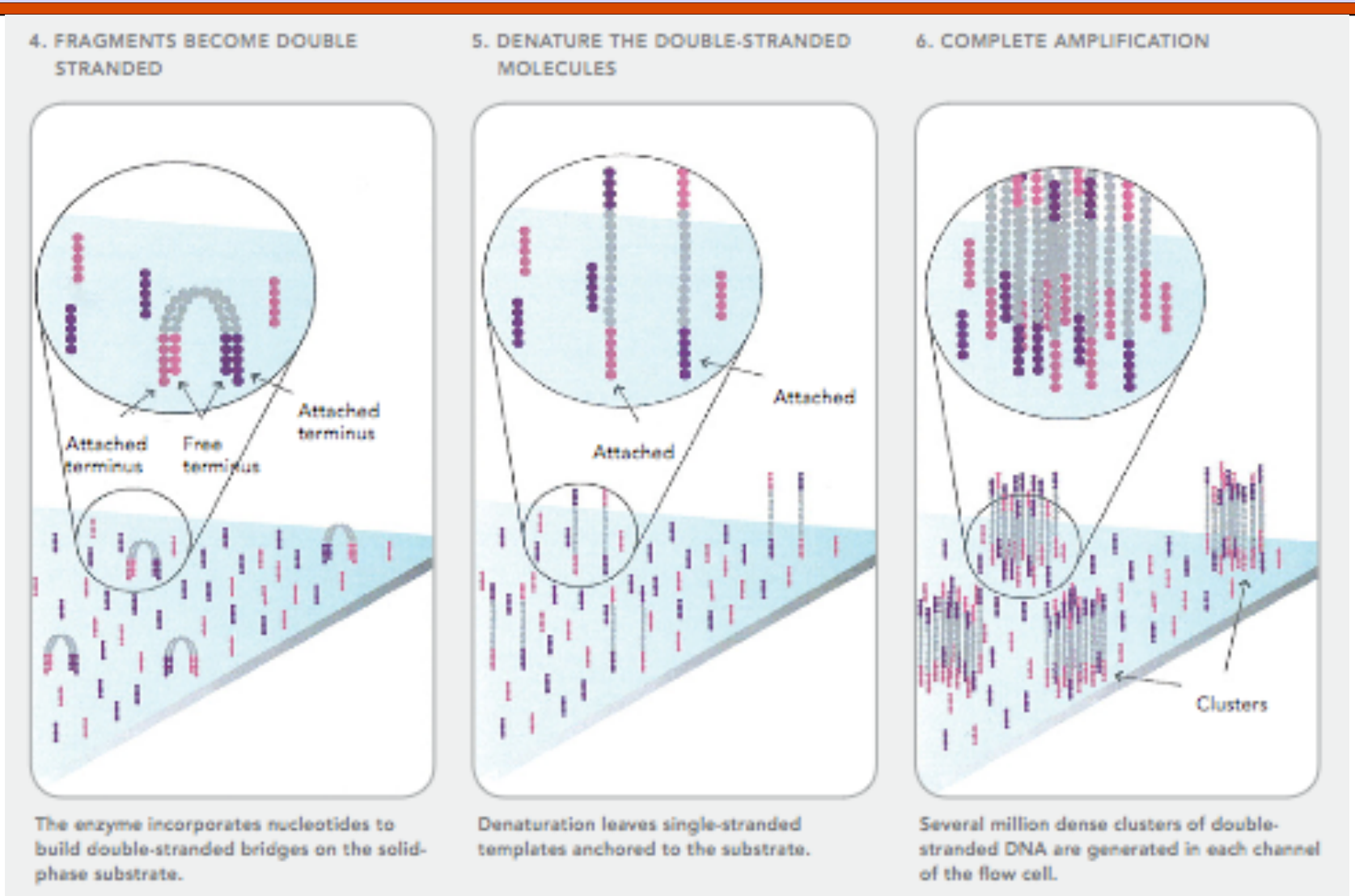
FIGURE 10



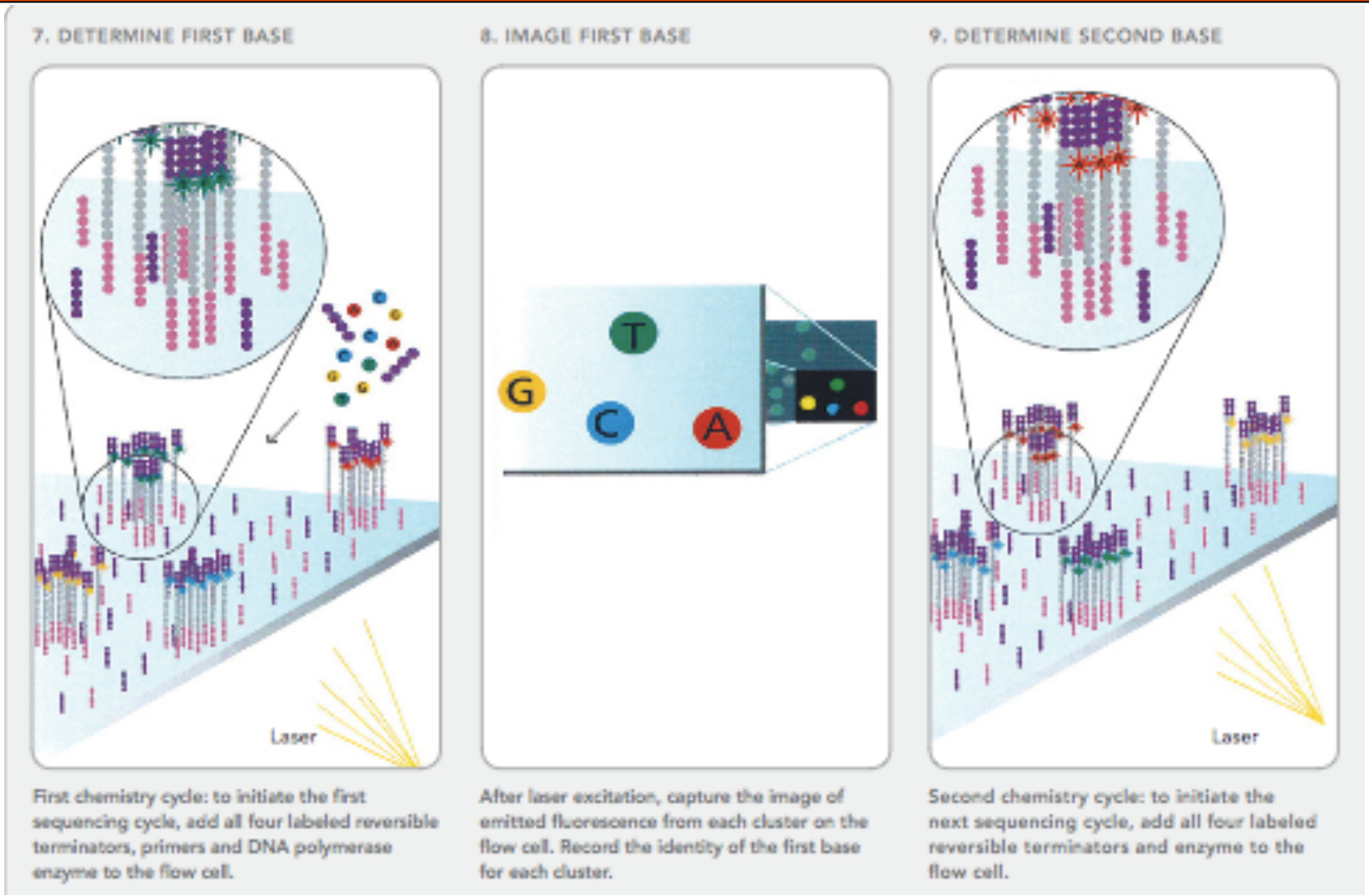
Solexa Sequencing



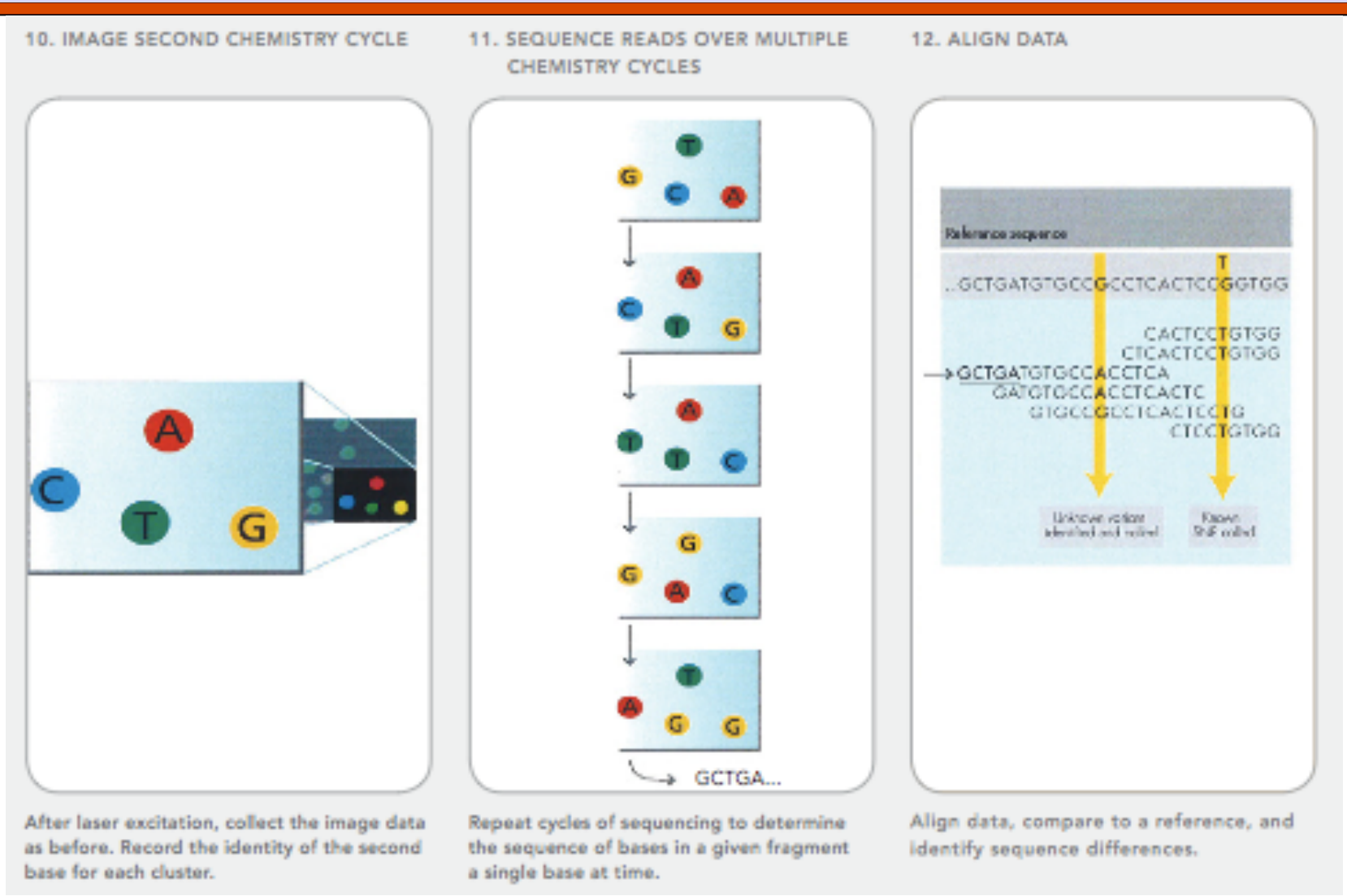
Solexa Sequencing



Solexa Sequencing



Solexa Sequencing



Assemblers

- TIGR Assembler (TIGR)
- Phrap (U Washington)
- Celera Assembler (Celera Genomics)
- Arachne (Broad Institute of MIT & Harvard)
- Phusion (Sanger Center)
- Atlas (Baylor College of Medicine)

Applications of Sequencing

- Sequencing
- Resequencing
- SNP detection
- RNA-Seq
- CHiP-Seq
- Metagenomics

Basic Assembler

- **Read**: sequenced fragment; **Contig**: contiguous segment. How to assemble a contig?

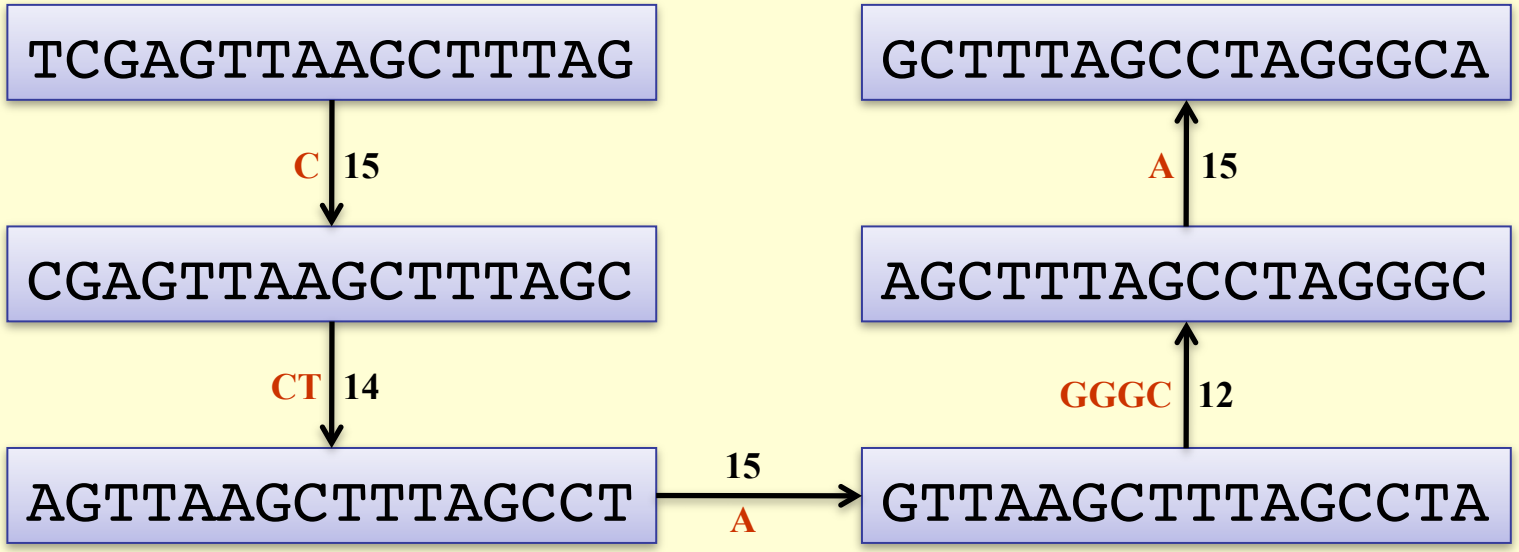
```
TCGAGTTAAGCTTTAG
 CGAGTTAAGCTTTAGC
   AGTTAAGCTTTAGCCT
    GTTAAGCTTTAGCCTA
       AGCTTTAGCCTAGGGC
          GCTTTAGCCTAGGCAG
              …
```

```
AGCTTTAGCCTAGGGC
AGTTAAGCTTTAGCCT
CGAGTTAAGCTTTAGC
GCTTTAGCCTAGGCAG
GTTAAGCTTTAGCCTA
TAAGCTTTAGCCTAGG
TCGAGTTAAGCTTTAG
```

Problem: Need to try every pair of reads!

Reduce to Graph Problem

- How to assemble a contig?
  - Node ⟷ Read
  - Edge between Nodes ⟷ Overlapping Reads
  - **Problem**: Find a path through each node in graph.



Issues: Problem is NP-Complete
nodes = # reads
of edges ≤ k(# nodes)

String graph

- Combine nodes that form paths into strings

A better solution

- Take each read and chop it into k-mers.
- Represent k-mers by nodes in a graph and edges between k-mers that overlap in k-1 bases.
- **Consequence**:
  - Number of nodes = 4^k ;
  - Number of edges = $k4^k$;
- **Issues**:
  - Problem (i.e., find path through all vertices) remains NP-Complete

A more efficient solution

- Represent every possible (k-1)-mer by a node.
- Edges connect 2 nodes if they share k-2 bases.
- Label each edge by k-mer.

- Problem:
  - Find a path through each edge in the graph
- The Eulerian path problem is NOT NP-Complete. It can be solved in linear time!

Sources of Assembly Errors

- Errors in reads – caused by technology
  - Error in base calls, color calls (SOLID Technology), or repeated base calls (454 Technology)
- Missing reads – sequencing bias
- Read orientation error
  - One or both orientations may occur
  - Not told which ones are present
- Sequence Variations – mixed sample study
  - SNP, cancer, metagenomics studies
- **REPEATS**
- Combinations of the above

How to deal with REPEAT Regions

- If no errors or repeat regions, then the graph has a unique path through all the edges.
- **Problem**: REPEAT regions cause branching in graph. If no errors in reads, then the graph has a unique path through all edges, but with some edges traversed more than once.
- How to identify REPEAT regions:
  - Higher coverage of repeat regions
  - Branching of nodes