CAP 5510: Introduction to Bioinformatics
CGS 5166: Bioinformatics Tools

Giri Narasimhan

ECS 254; Phone: x3748

giri@cis.fiu.edu

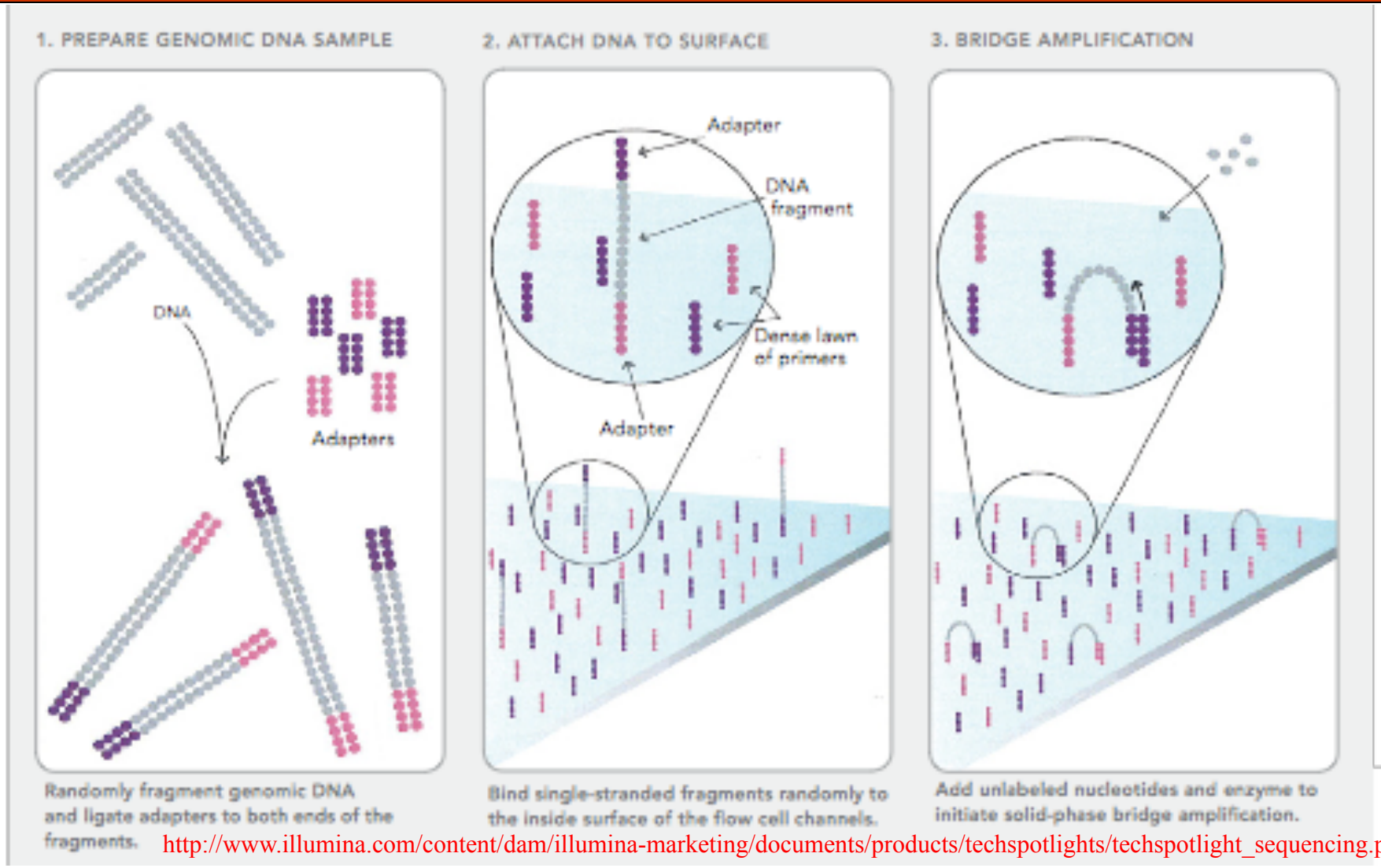
www.cis.fiu.edu/~giri/teach/BioinfF18.html

Next Generation Sequencing

History of NGS

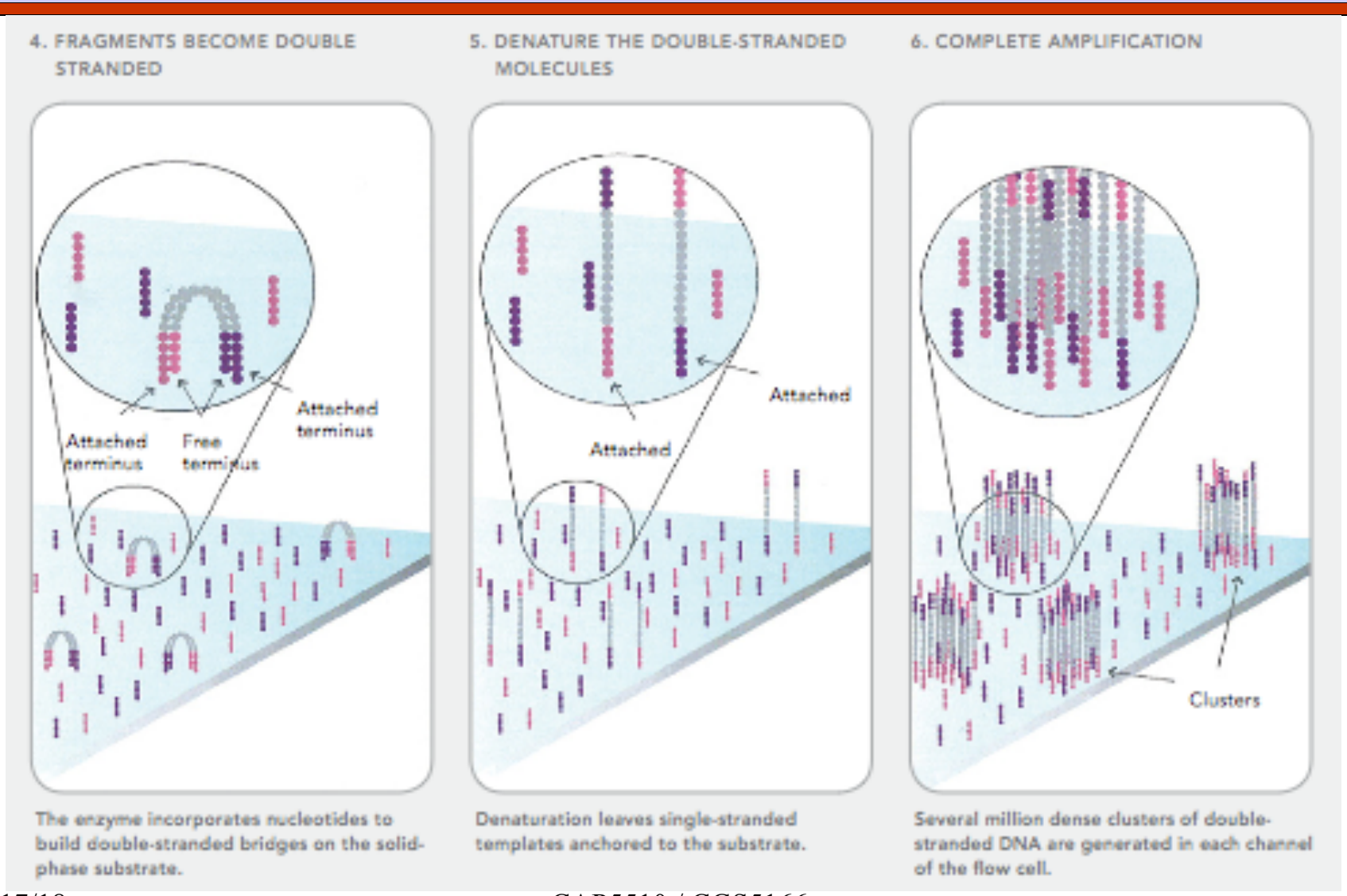
- 1977: Sanger Method (70Kbp/run)
- Sequencing by Hybridization (SBH); Dual end sequencing; Chromosome Walking (see page 5-6 of Pevzner's text);
- 1987: Automated Sequencer (AB Prism)
- 1996: Capillary Sequencer (ABI 310)
- 2005: 454 Sequencing (GS 20; 60Mbp/run)
- 2006: Solexa Sequencing (Illumina; 600Mbp/run)
- 2007 : SOLiD (AB)
- 2009 : Helicos single molecule sequencer
- 2011 : Ion Torrent (PGM)
- 2011 : Pacific Biosciences single molecule sequencer
- 2012 : Oxford Nanopore Tech. ultra long single mol. Reads; so small that they have put the whole technology into a USB drive.

Illumina's Sequencing-by-Synthesis

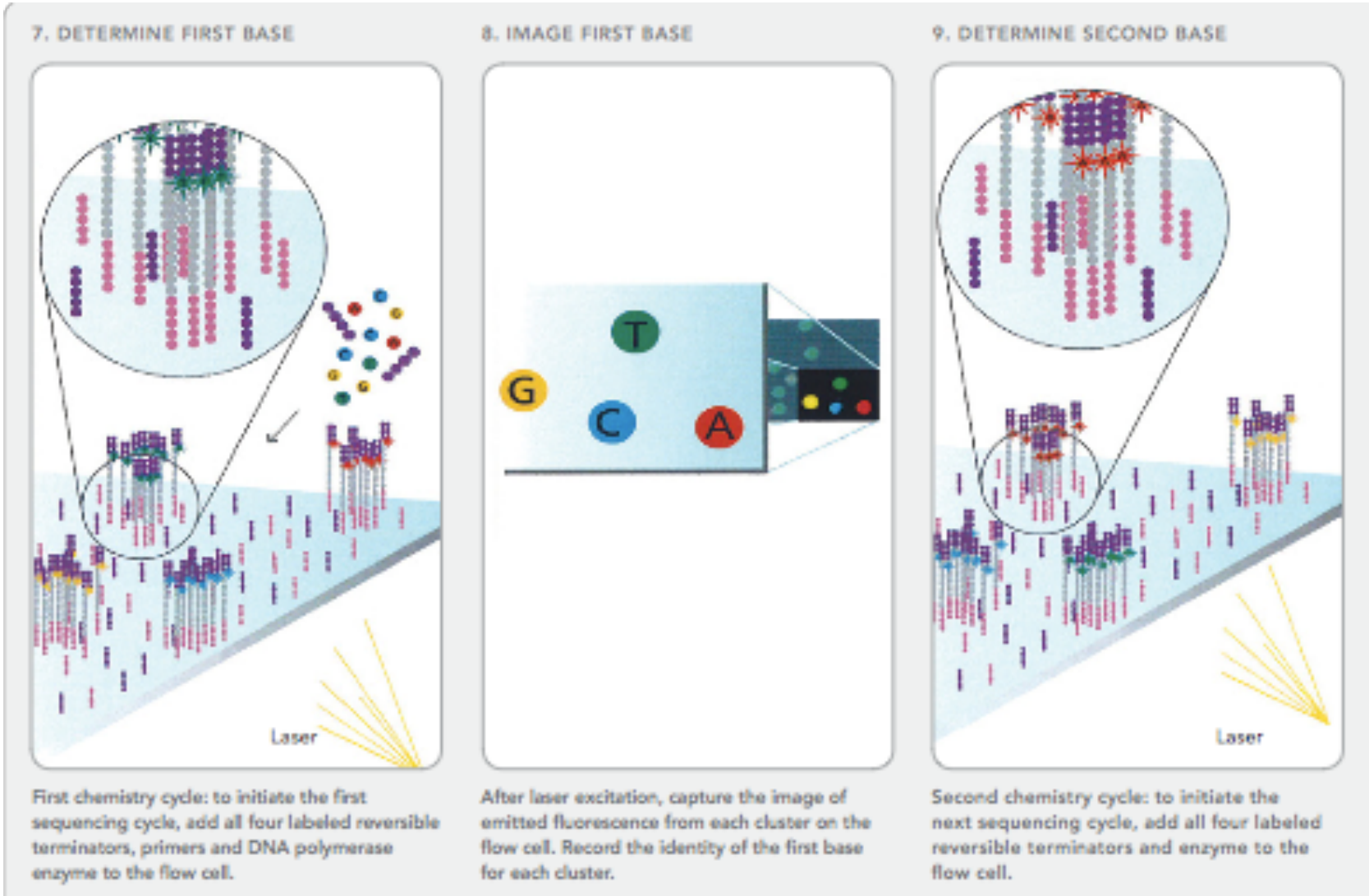


http://www.illumina.com/content/dam/illumina-marketing/documents/products/techspotlights/techspotlight_sequencing.pdf

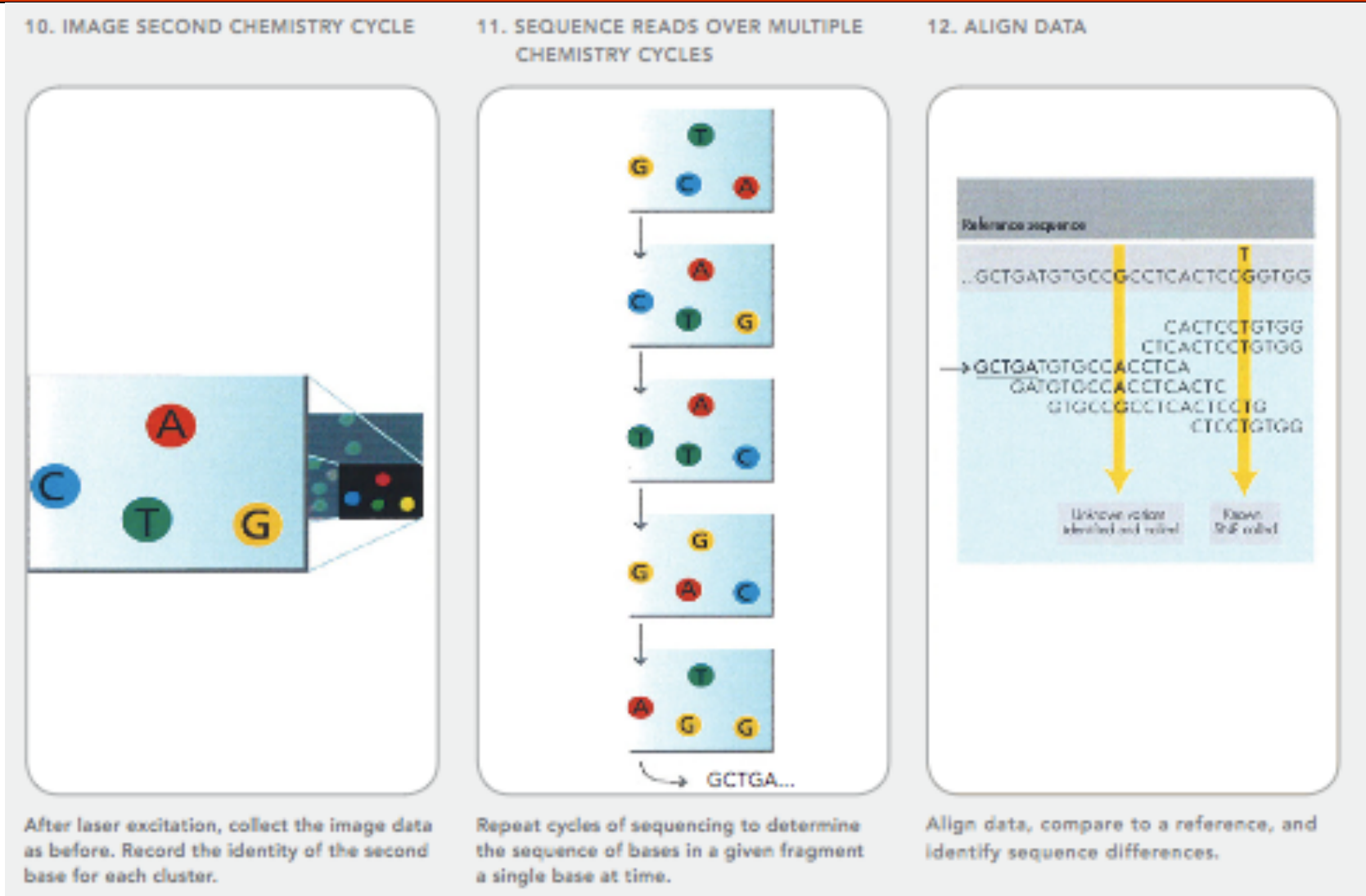
Solexa Sequencing



Solexa Sequencing



Solexa Sequencing



Ion Torrent Sequencer

- Harness power of semiconductor technology
- During nucleotide synthesis, a proton is released
- This can be detected by measuring pH, not fluorescence
- The dNTPs are flowed over the surface in a predetermined sequence & the ligations are detected

PacBio Sequencing

- Single molecule technology
- Extraordinarily long reads
- Non-trivial error, but unbiased

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)

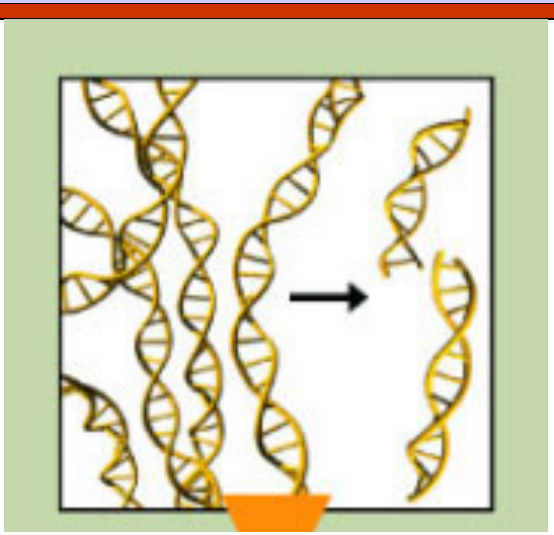
Applications of Sequencing

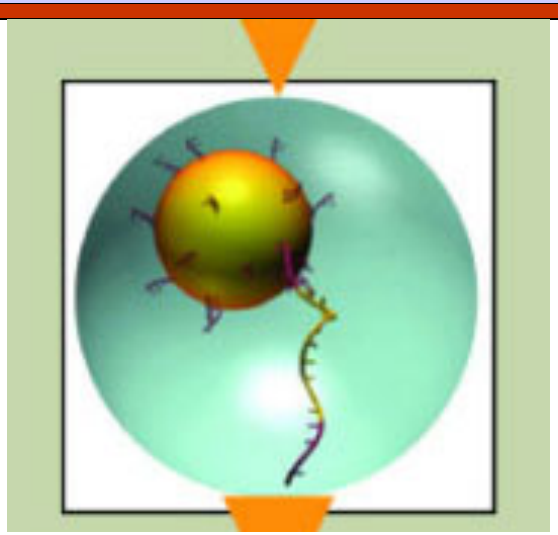
- Sequencing
- Resequencing
- SNP detection
- RNA-Seq
- CHiP-Seq
- Metagenomics

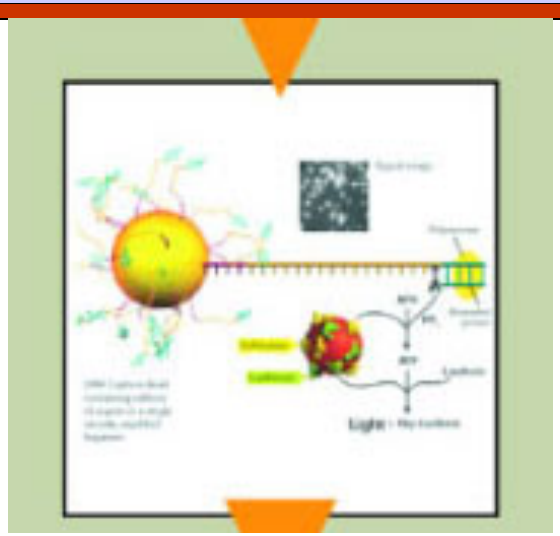
454 Sequencing: New Sequencing Technology

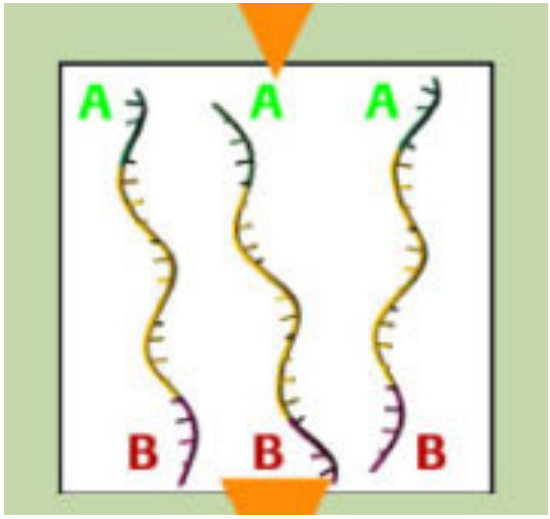
- This technology, started in 2005 and is now being phased out
- 454 Life Sciences, Roche
- Fast (20 million bases per 4.5 hour run)
- Low cost (lower than Sanger sequencing)
- Simple (entire bacterial genome in days with one person -- without cloning and colony picking)
- Convenient (complete solution 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.

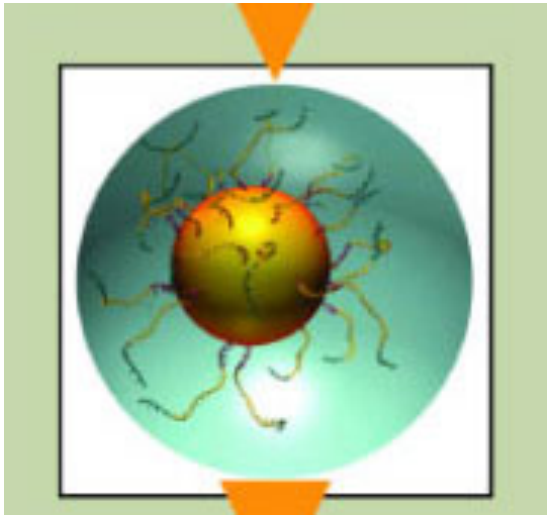
(a) Fragment, (b) add adaptors, (c) “1 fragment, 1 bead”, (d) emPCR on bead, (e) put beads in PicoTiterPlate and start sequencing: “1 bead, 1 read”, and (f) analyze

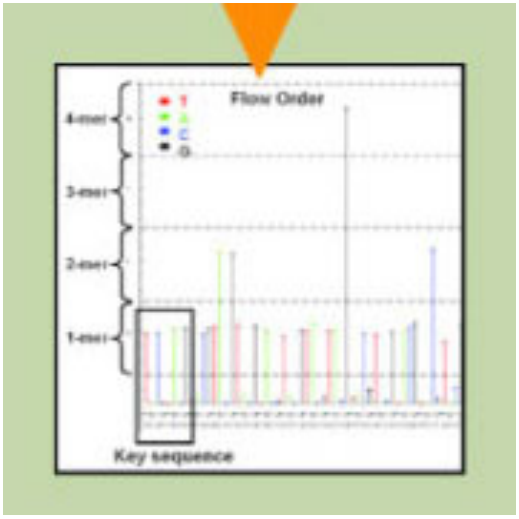












emPCR

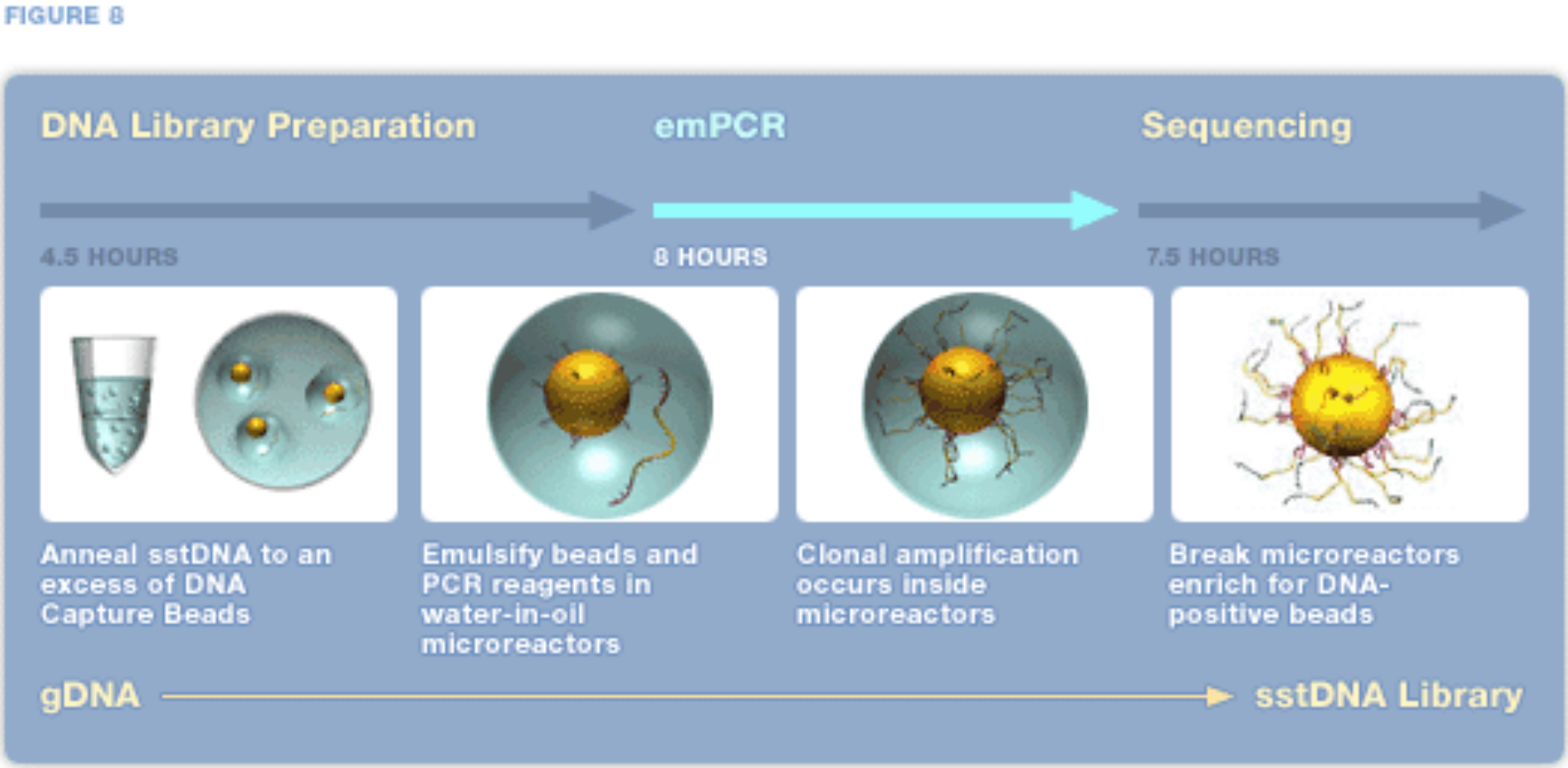
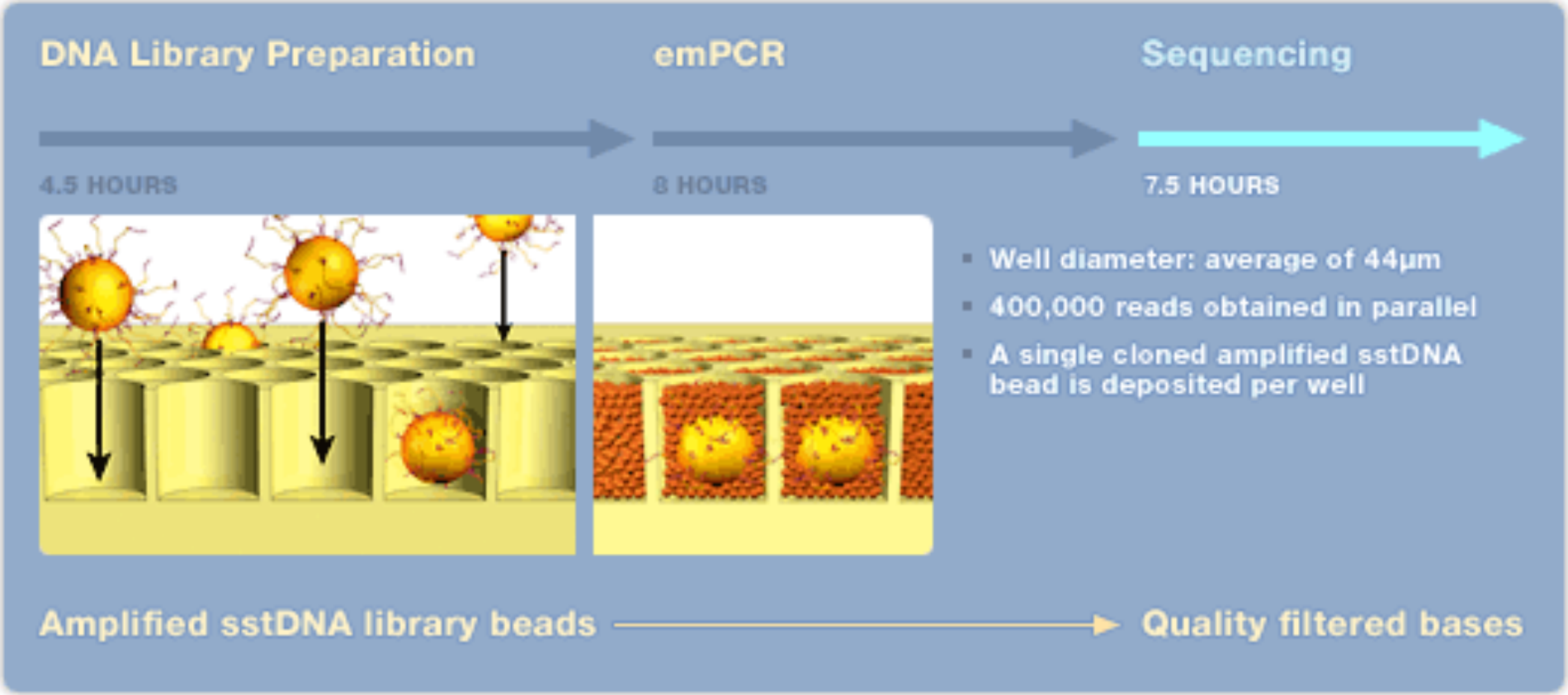


FIGURE 8
DNA Library Preparation
emPCR
Sequencing
4.5 HOURS
8 HOURS
7.5 HOURS
Anneal sstDNA to an excess of DNA Capture Beads
Emulsify beads and PCR reagents in water-in-oil microreactors
Clonal amplification occurs inside microreactors
Break microreactors enrich for DNA-positive beads
gDNA
sstDNA Library

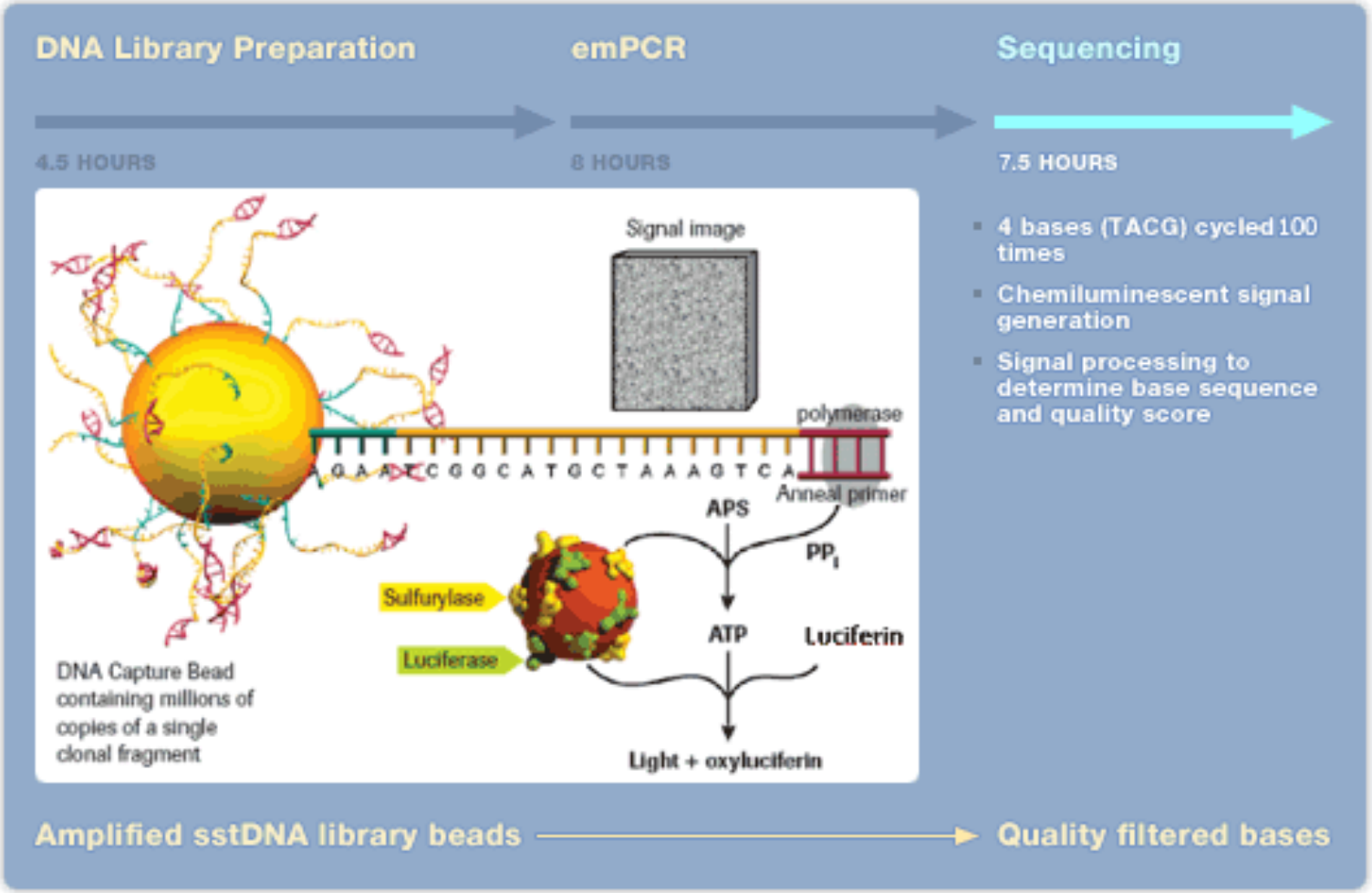
Sequencing

FIGURE 9



Sequencing

FIGURE 10



Assembly Challenges

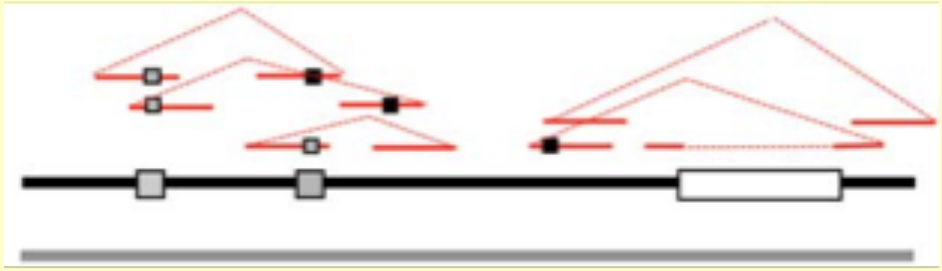


Figure 1: *Illustration of the HTS technologies. Dark thick line: genome being sequenced. Gray thick line: reference. Small box on genome: SNP. Long box on genome: deletion. The short, red lines are the short sequence reads mapped to the proper location. Light box on reads: variations. Dark box on reads: sequencing error. Paired end reads are connected by dashed lines. Split-read: split into two segments.*

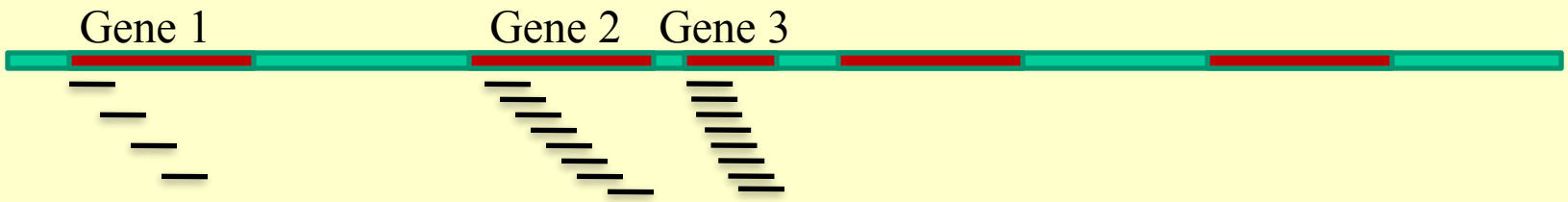
See note below

NGS Applications

Applications of NGS

- RNA-Seq
- ChIP-Seq
- SNP-Seq
- Metagenomics
- Alternative Splicing
- Copy Number Variations (CNV)
- …

RNA-Seq



- Align reads to genes and count
- Assume uniform sampling
  - Count of number of reads mapped per gene is a measure of its expression level
  - Expression of Gene 2 is twice that of Gene 1
  - Expression of Gene 3 is twice that of Gene 2

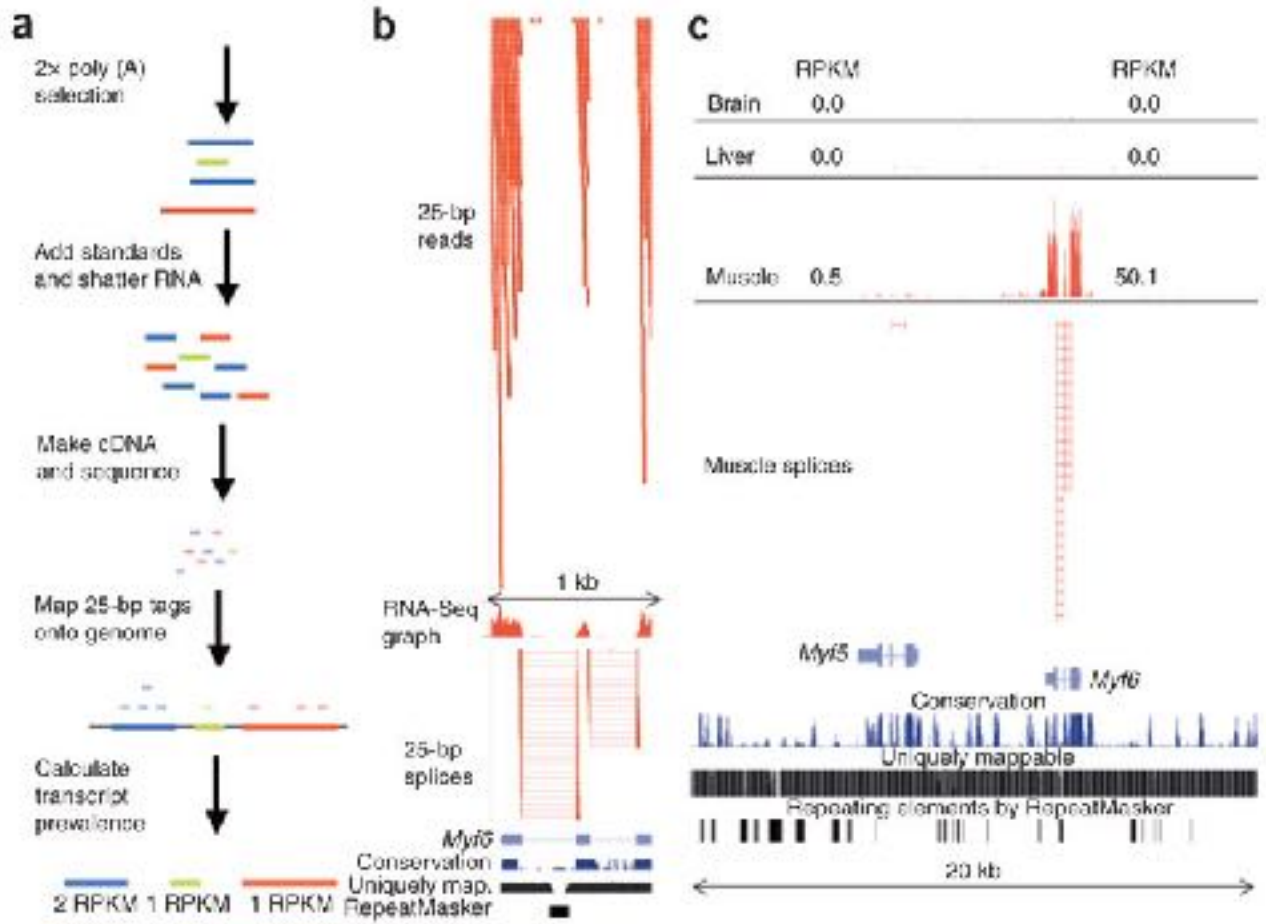
Expression Level of Gene

- RPKM = Ng / (N X L)
  - Ng = Number of reads mapped to gene
  - N = Total number of mapped reads (in millions)
  - L = Length of gene in KB
  - [Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B., Nat Methods. 2008 Jul;5(7):621-8. **Mapping and quantifying mammalian transcriptomes by RNA-Seq.**]

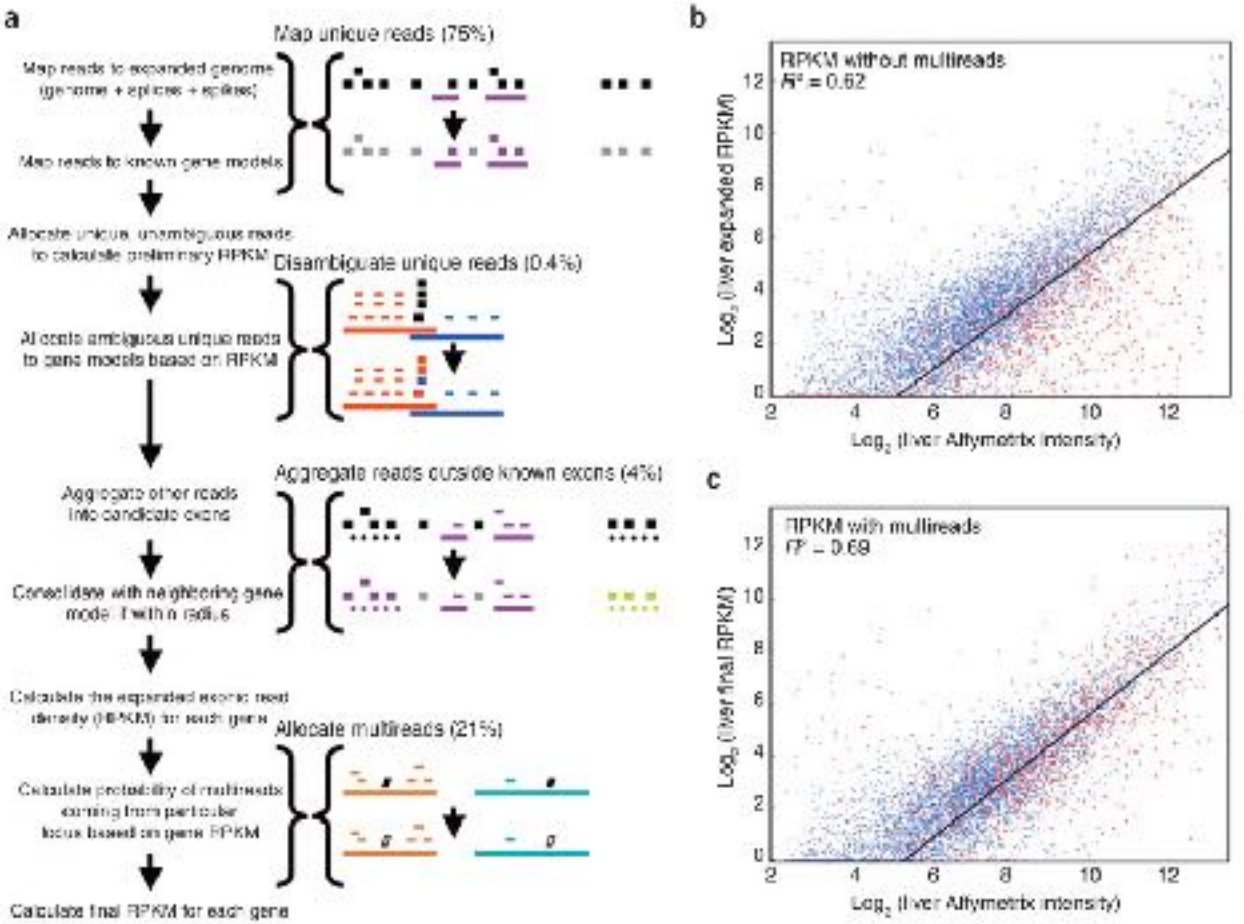
Complications

- Repeat regions
  - Paralogs and other homologous regions in genes
  - Ambiguities in maps
- Introns and Exons
  - Aligning reads to genome is more complex
- Alternative Splicing
- Transcription start site is upstream of ORFs
- Unknown ORFs and Small RNAs
- Other transcripts

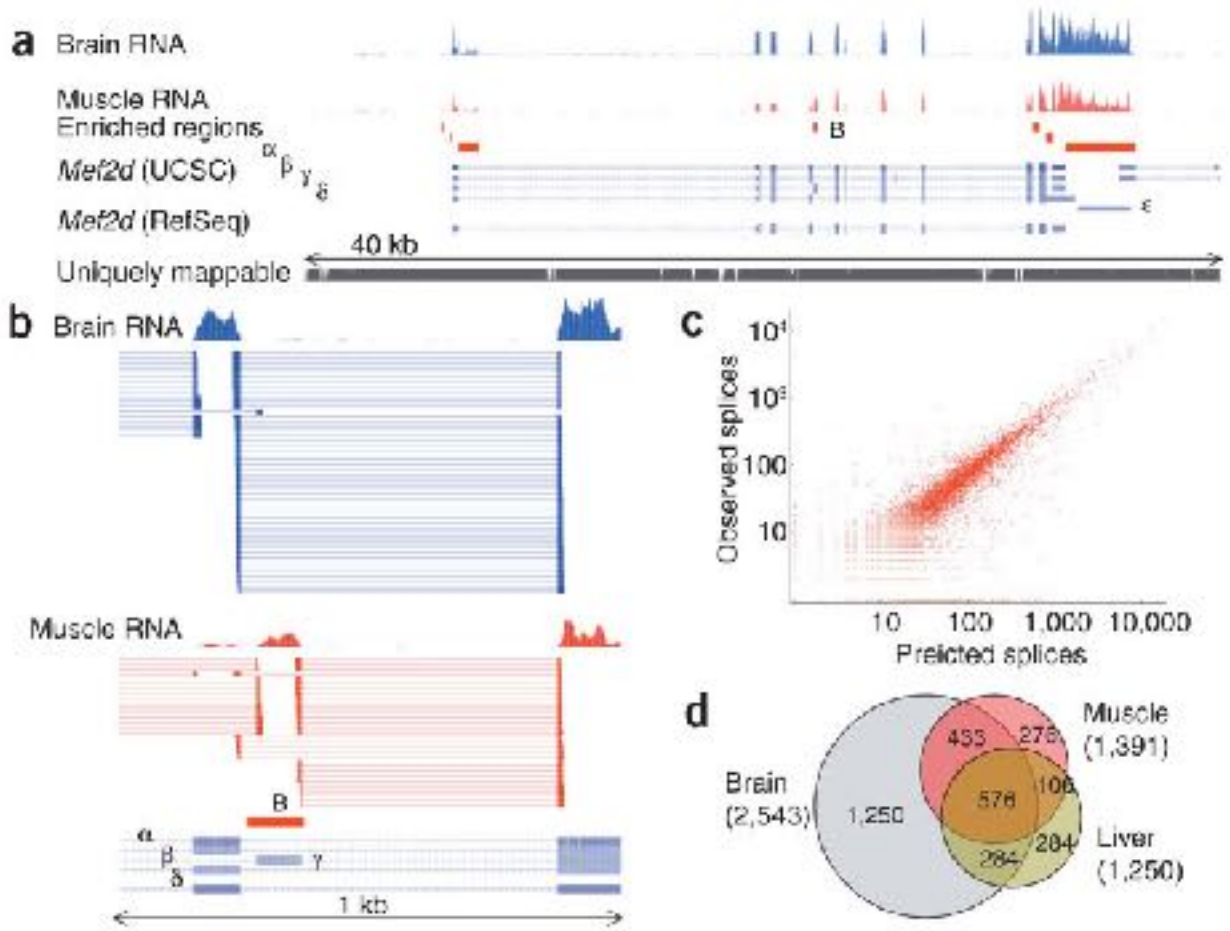
RNA-Seq Procedure



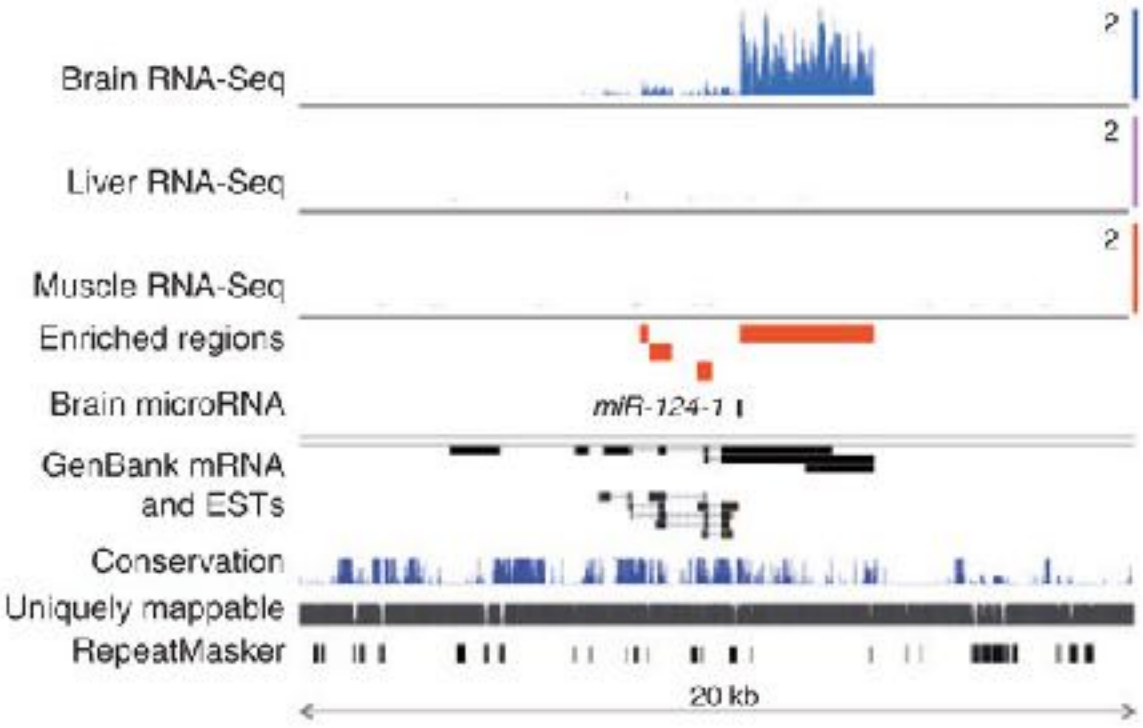
Mapping Reads to Reference



Alternative Splicing



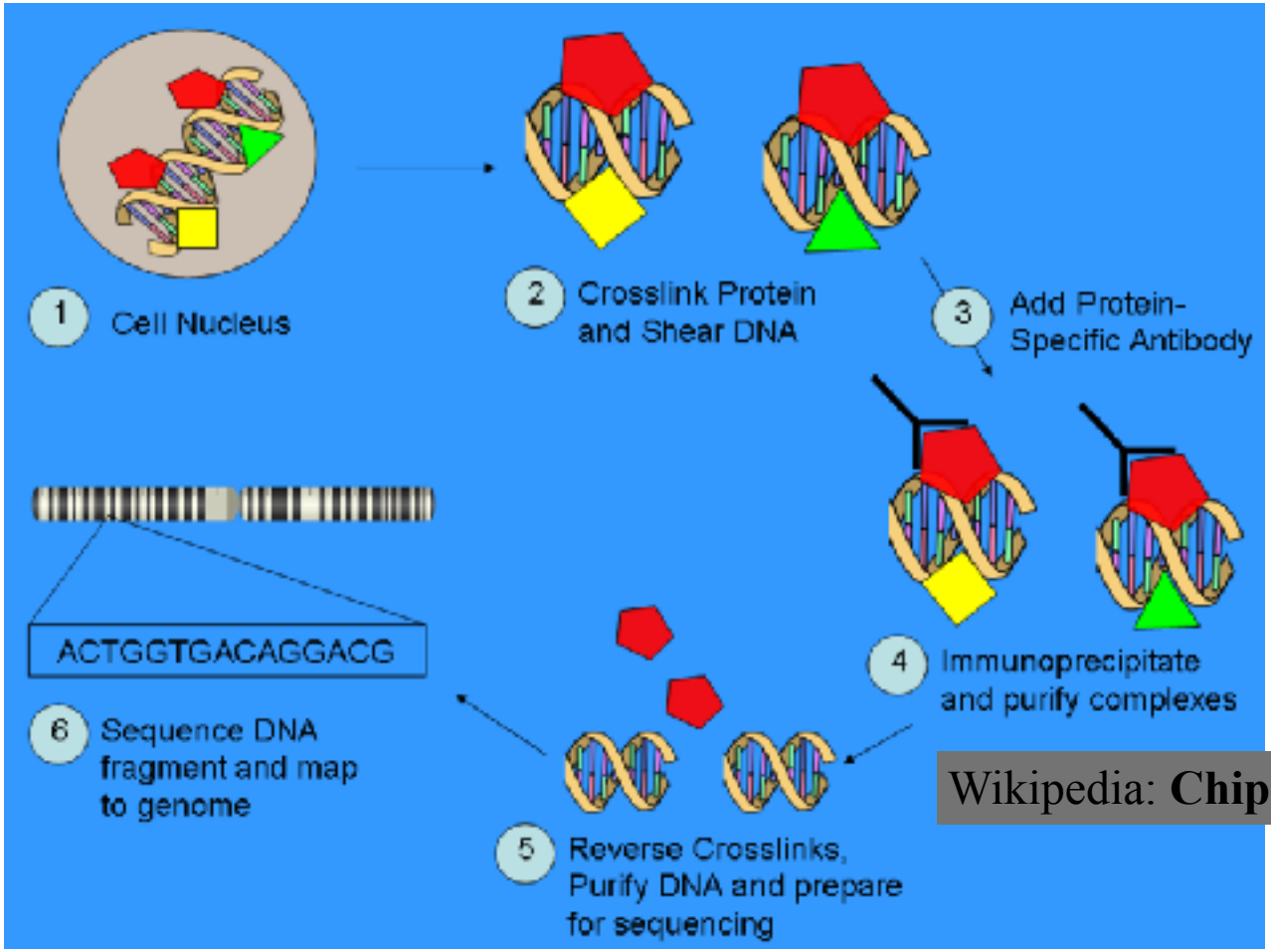
microRNA


Brain RNA-Seq
2
Liver RNA-Seq
2
Muscle RNA-Seq
2
Enriched regions
Brain microRNA
miR-124-1
GenBank mRNA
and ESTs
Conservation
Uniquely mappable
RepeatMasker
20 kb

Chromatin Immunoprecipitation

- Useful for pinpointing location of TFBS for TF
- High-throughput method to find all binding sites for a specific TF under specific conditions
- Identify sites using
  - ChIP-on-chip (Microarray technique)
  - ChIP-Seq (Sequencing technique)
- Problems: TFs bind to specific TFBS only under specific conditions – hard to predict

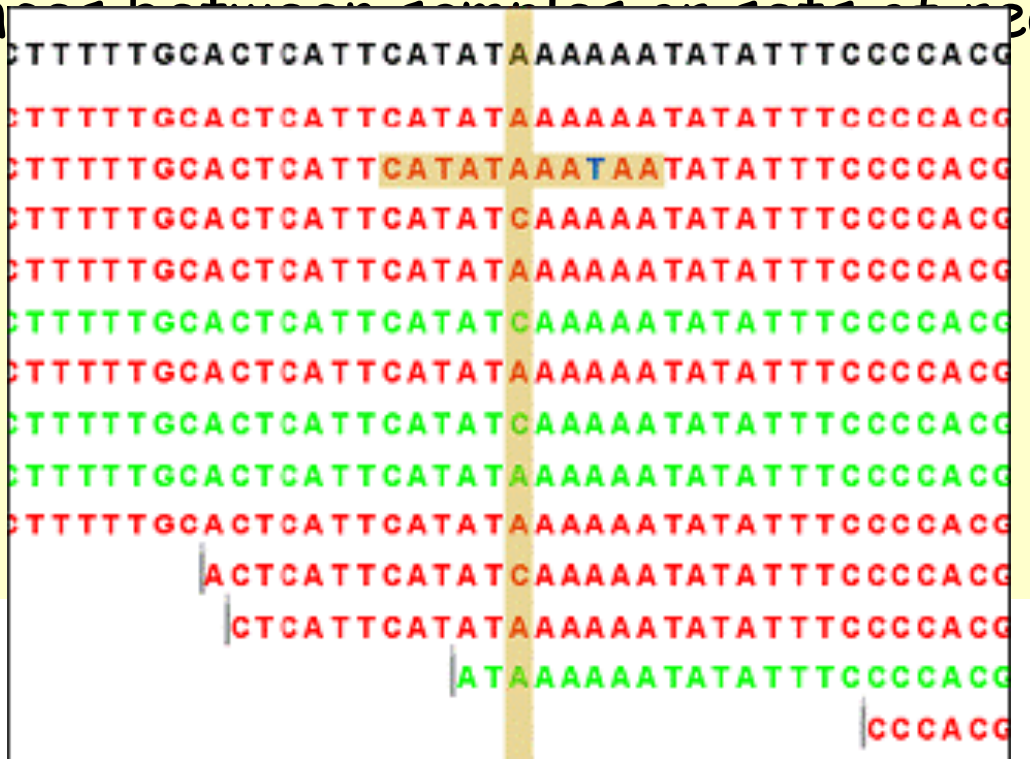
ChIP-Seq



Wikipedia: **Chip-Sequencing**

SNP-Seq

- Align reads and look for differences
  - Differences to reference
    - Align reads to reference sequence first
  - Differences within reads
  - Differences between samples or sets of reads



Environmental Microbiology

- Conventional methods
  - Culture, then identify
    - Slow, expensive, labor intensive, unculturable microbes
  - PCR-based length heterogeneity studies
- Microarray-based methods
  - Unique probes for organisms (e.g., Virochip)
    - Only works for sequenced regions of known organisms
- NGS-based methods

Metagenomics

- Detect known pathogens
- Diversity
  - Identity of individual species not needed
- Functional profile of community

NGS-based method

- Map reads against appropriate database
- Identify closest hits for each read
- Generate contigs
- Generate abundance information
- Clustering of reads can be beneficial to estimate abundance