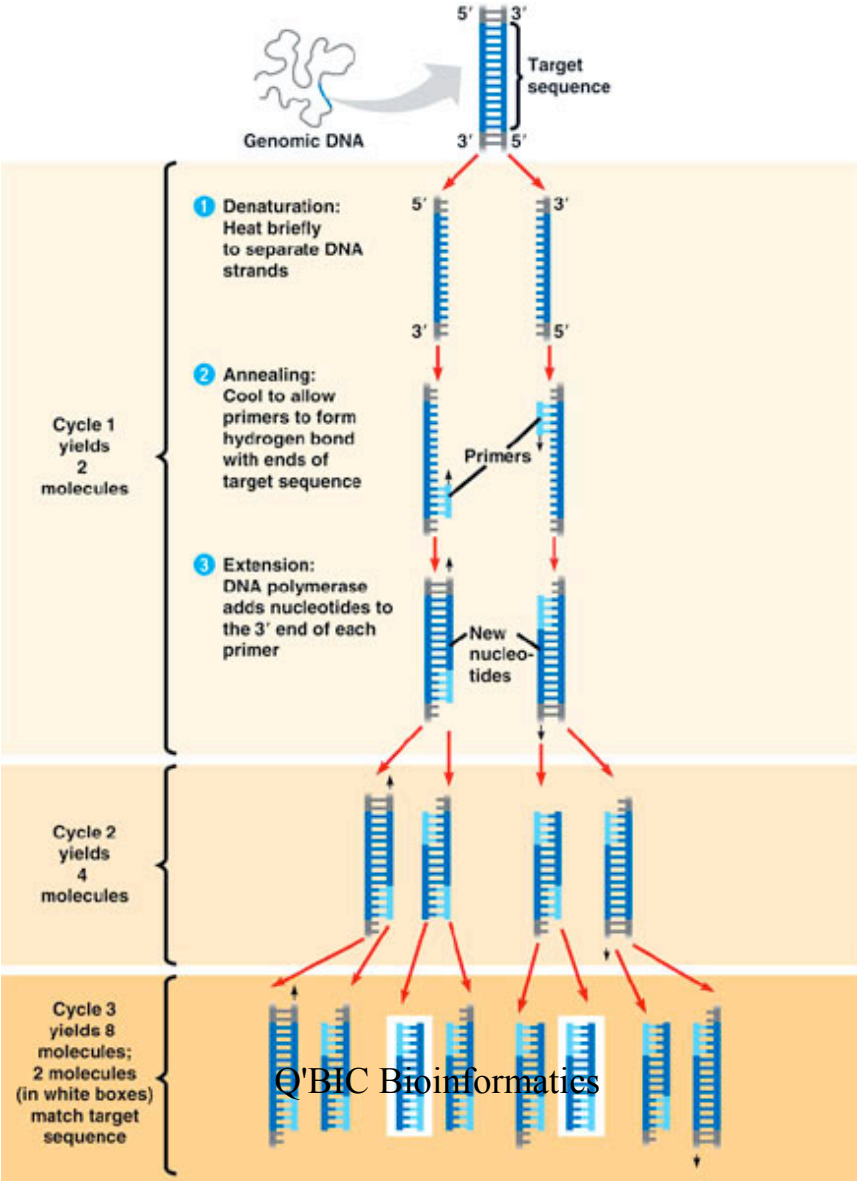
BSC 4934: Q'BIC Capstone Workshop

Giri Narasimhan
ECS 254A; Phone: x3748
giri@cs.fiu.edu
http://www.cs.fiu.edu/~giri/teach/BSC4934_Su10.html
July 2010

TP53 Exons

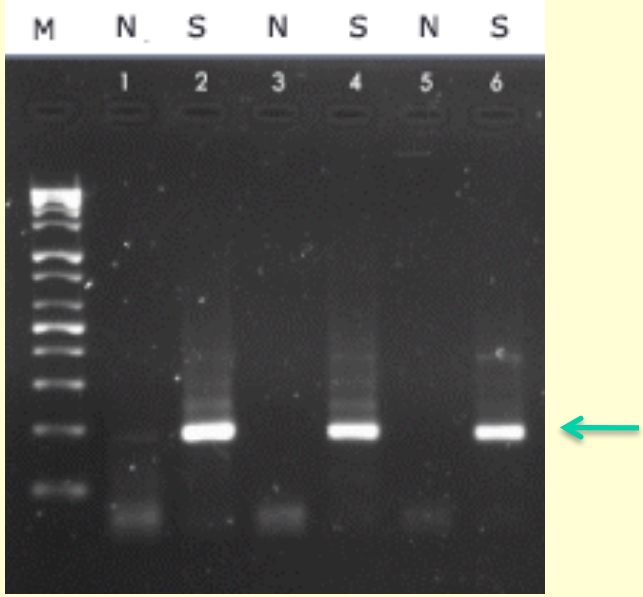
	mRNA				coding	
start	**end**	**length**		**start**	**end**	**length**
1	168	168		7572927	7573008	82
10926	11024	99		7573927	7574033	107
11142	11163	22		7576853	7576926	74
11273	11551	279		7577019	7577155	137
12309	12492	184		7577499	7577608	110
12574	12686	113		7578177	7578289	113
13255	13364	110		7578371	7578554	184
13708	13844	137		7579312	7579590	279
13937	14010	74		7579700	7579721	22
16830	16936	107		7579839	7579912	74
17855	19143	1289				

PCR


Genomic DNA
5′
3′
Target sequence
3′
5′
Cycle 1 yields 2 molecules
1 Denaturation: Heat briefly to separate DNA strands
2 Annealing: Cool to allow primers to form hydrogen bond with ends of target sequence
Primers
3 Extension: DNA polymerase adds nucleotides to the 3′ end of each primer
New nucleo-tides
Cycle 2 yields 4 molecules
Cycle 3 yields 8 molecules; 2 molecules (in white boxes) match target sequence

Gel Electrophoresis

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

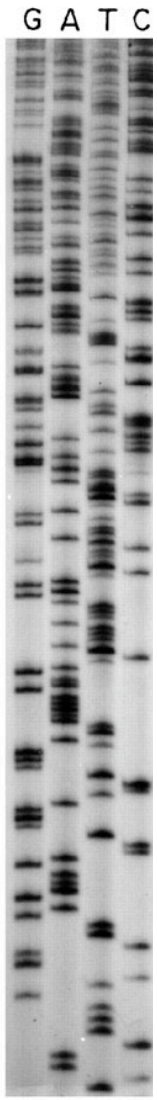


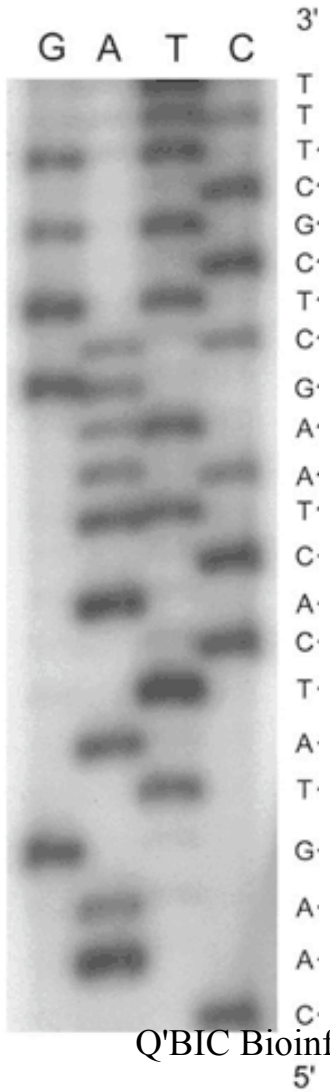
Sequencing

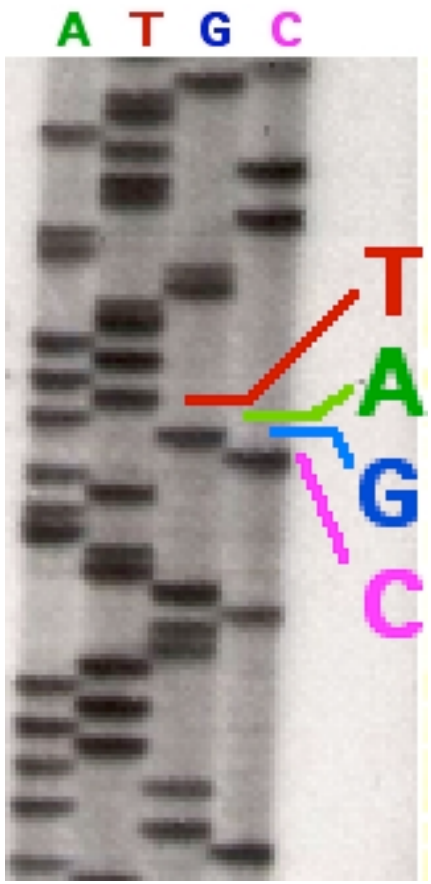
Original Sanger Method

- (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.

Sequencing Gel





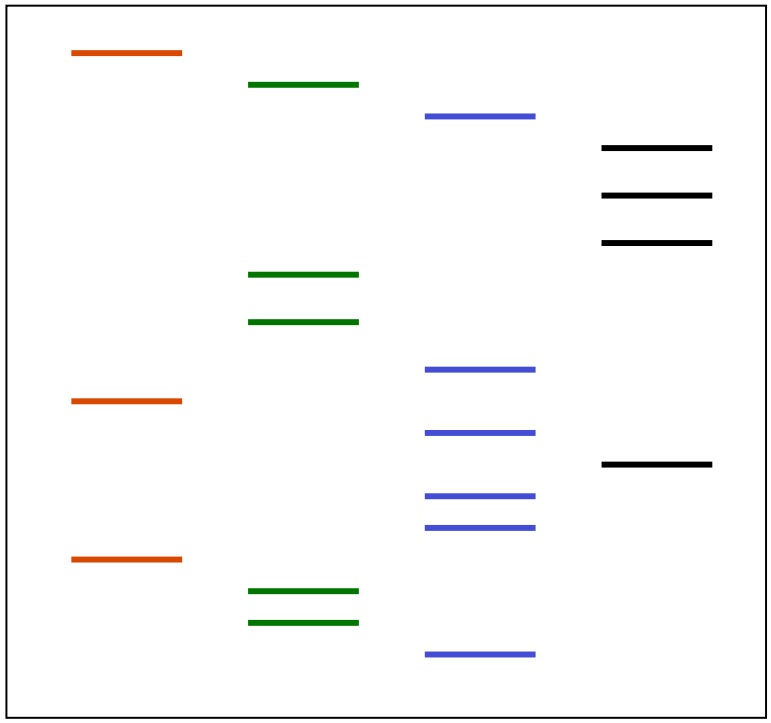


Modified Sanger

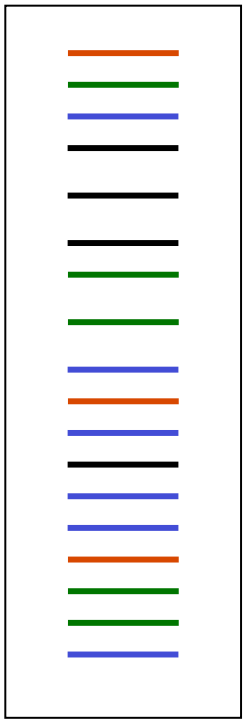
- Reactions performed in a single tube containing all four ddNTP's, each labeled with a different color fluorescent dye

Sequencing Gels: Separate vs Single Lanes

GCCAGGTGAGCCTTTGCA



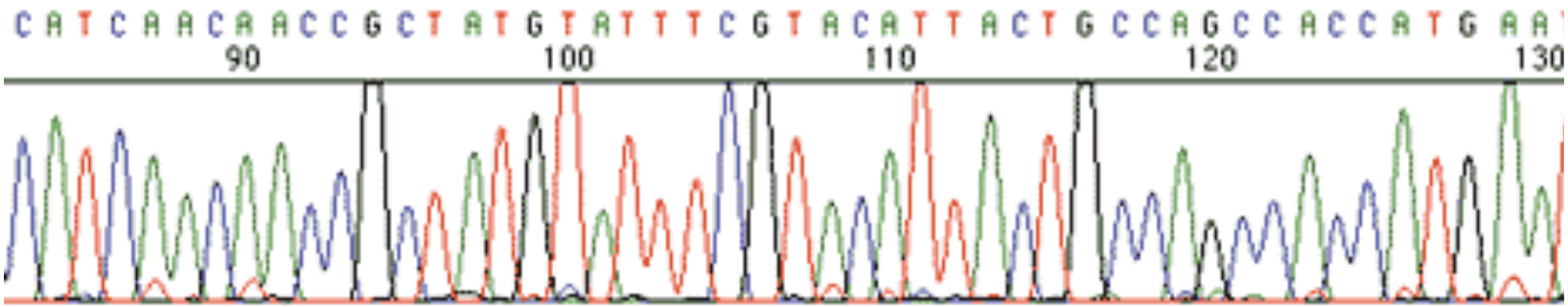
A C G T



Automated Sequencing Instruments

Sequencing

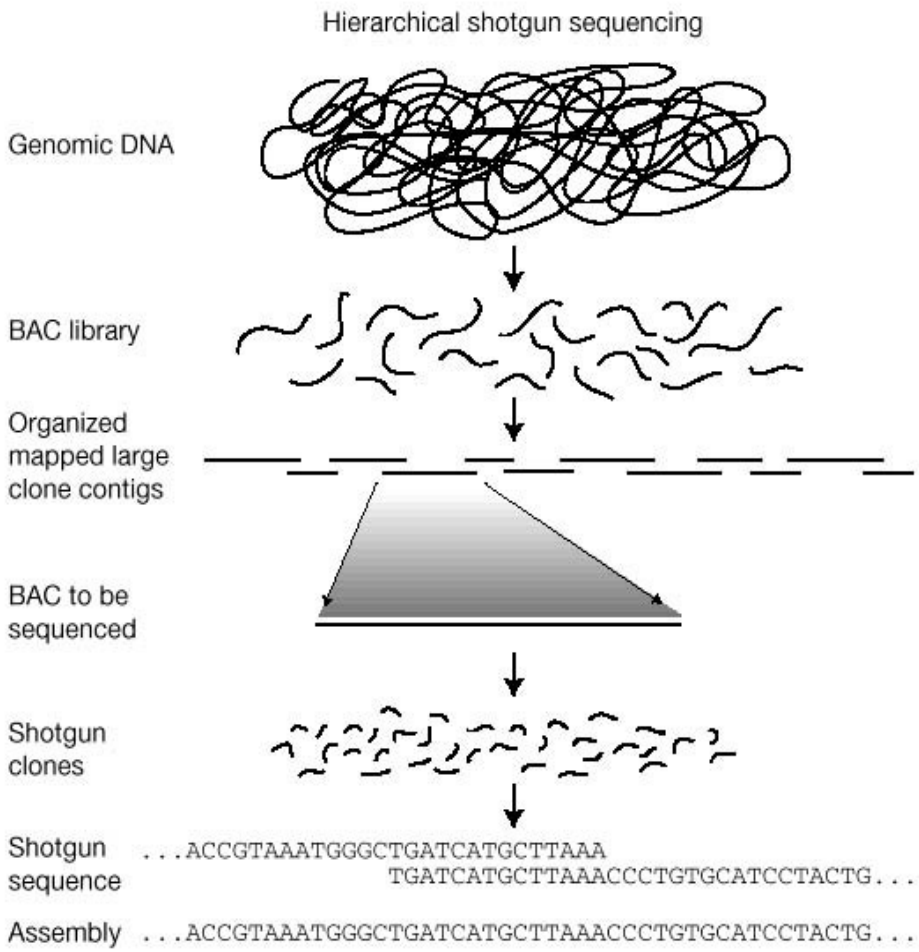
- Flourescence sequencer
- Computer detects specific dye
- Peak is formed
- Base is detected
- Computerized



Maxam-Gilbert Sequencing

- Not popular
- Involves putting copies of the nucleic acid into separate test tubes
- Each of which contains a chemical that will cleave the molecule at a different base (either adenine, guanine, cytosine, or thymine)
- Each of the test tubes contains fragments of the nucleic acid that all end at the same base, but at different points on the molecule where the base occurs.
- The contents of the test tubes are then separated by size with gel electrophoresis (one gel well per test tube, four total wells), the smallest fragments will travel the farthest and the largest will travel the least far from the well.
- The sequence can then be determined from the picture of the finished gel by noting the sequence of the marks on the gel and from which well they came from.

Shotgun Sequencing



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

Sequencing

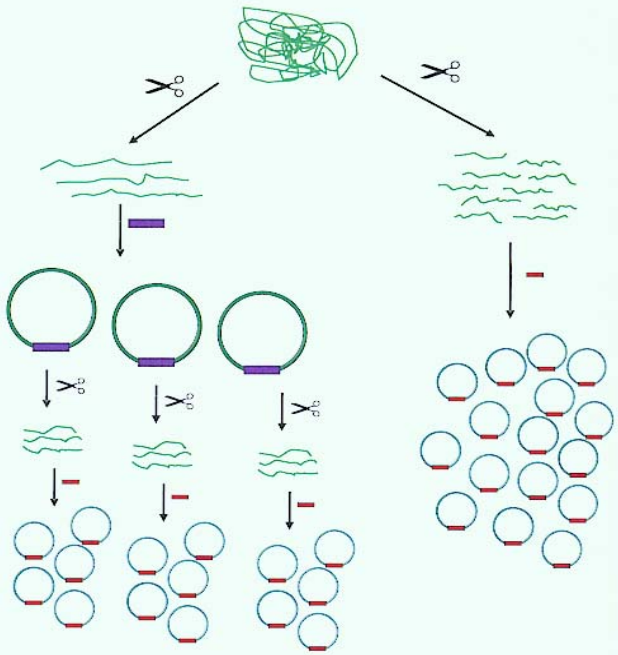
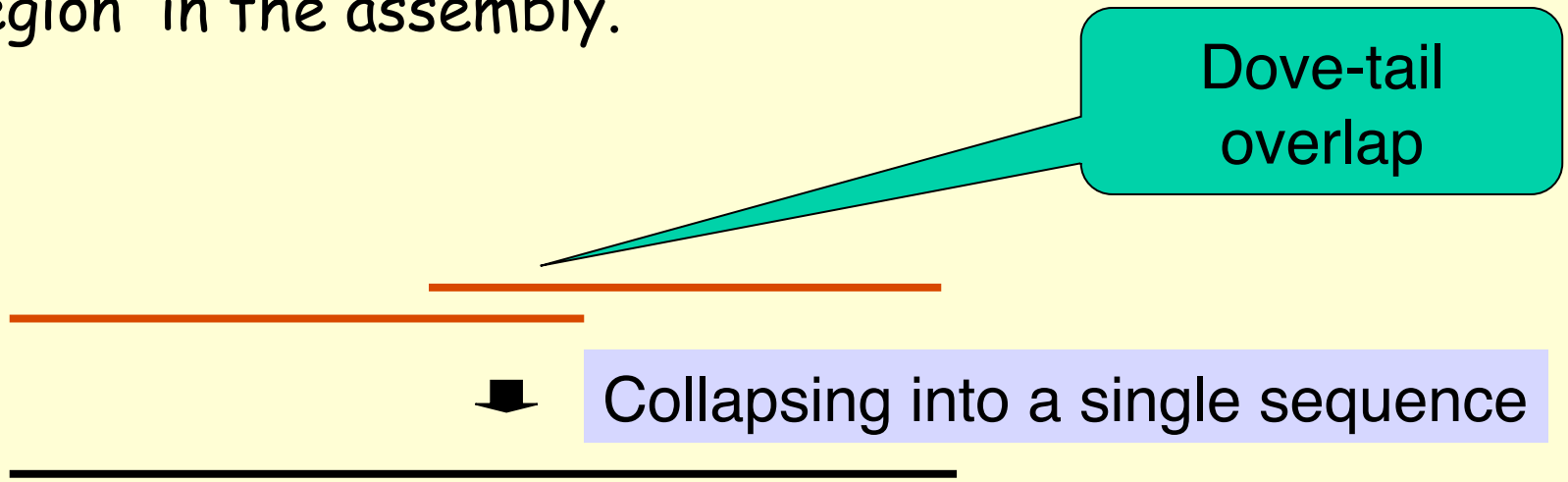


FIGURE 13.1 Shotgun cloning. Genomic DNA sequencing begins with isolated genomic DNA in green at the top of the figure. In the hierarchical clone-based shotgun approach on the left, DNA is sheared and the size is selected for large fragments on the order of 200 Kb, then ligated to a suitable vector, such as a BAC vector shown in blue. Individually isolated clones in turn are sheared independently, generating fragments of approximately 4 Kb, which are then ligated to a small-scale vector, typically a plasmid (red bar) suitable for sequencing reactions. The whole genome shotgun approach bypasses the intermediate large-insert clone and generates large numbers of small fragments, typically 4 Kb and 10 Kb.

Sequencing: Generate Contigs

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



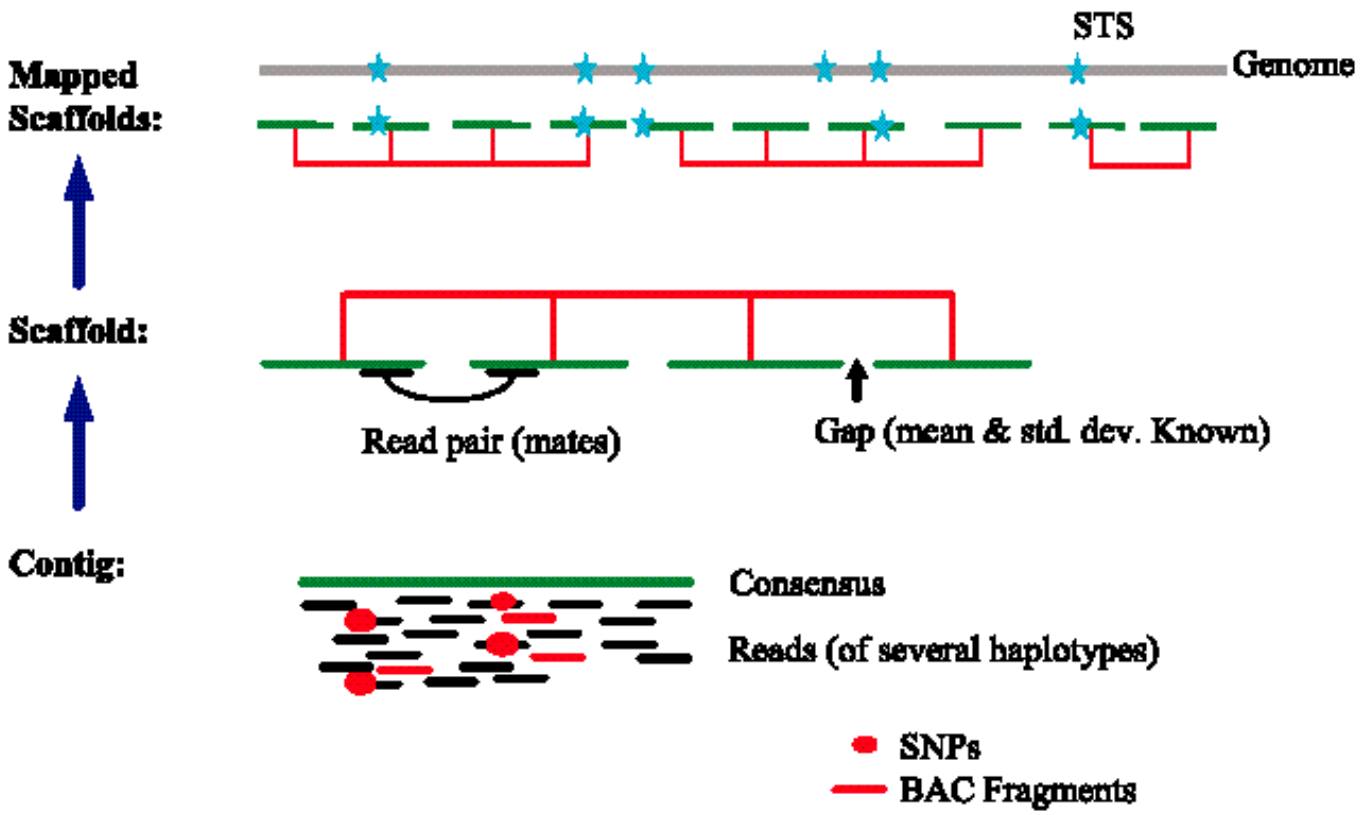
- Jang W et al (1999) Making effective use of human genomic sequence data. Trends Genet. 15(7): 284-6.
  Kent WJ and Haussler D (2001) Assembly of the working draft of the human genome with GigAssembler. Genome Res 11(9): 1541-8.

Supercontigs/Scaffolds

- A supercontig is formed when an association can be made between two contigs that have no sequence overlap.
  - This commonly occurs using information obtained from paired plasmid ends. For example, if both ends of a BAC clone are sequenced, then it can be inferred that these two sequences are approximately 150-200 Kb apart (based on the average size of a BAC). If the sequence from one end is found in a particular sequence contig, and the sequence from the other end is found in a different sequence contig, the two sequence contigs are said to be linked. In general, it is useful to have end sequences from more than one clone to provide evidence for linkage.

[NCBI Genome Glossary]

Shotgun Sequencing



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

Human Genome Project

Play the Sequencing Video:

- Download Windows file from http://www.cs.fiu.edu/~giri/teach/Bioinf/Papers/Sequence.exe
- Then run it on your PC.

Human Genome Project

1980 The sequencing methods were sufficiently developed

International collaboration: International Human Genome Consortium of 20 groups - a Public Effort (James Watson as chair!)

Estimated expense: $3 billion and 15 years

Part of this project was to sequence (started Oct '90): *E. coli, S. cerevisiae, D. melanogaster, A. thaliana, C. elegans*

Automated sequencing and computerized analysis

Public effort: 150,000 bp fragments into artificial chromosomes

In three years large scale physical maps were available

Venter vs Collins

National Human Genome Research Institute

Venter's lab in NIH (joined NIH in 1984) is the first test site for ABI automated sequences; he developed strategies (Expressed Sequence Tags - ESTs)

1992 - decided to patent the genes expressed in brain - "Outcry"

Resistance to his idea

Watson publicly made the comment that Venter's technique during senate hearing - "wasn't science - it could be run by monkeys"

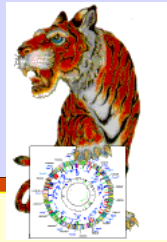
In April 1992 Watson resigned from the HGP

Craig Venter and his wife Claire Fraser left the NIH to set up two companies

- the not-for-profit TIGR The Institute for Genomic Research, Rockville, Md
- A sister company FOR-profit with William Hazeltine - HGSI - Human Genome Sciences Inc., which would commercialize the work of TIGR
- Financed by Smith-Kline Beecham ($125 million) and venture capitalist Wallace Steinberg.

Francis Collins of the University of Michigan replaced Watson as head of NHGRI.

Venter vs Collins



HGSI promised to fund TIGR with $70 million over ten years in exchange for marketing rights to TIGR's discoveries

PE Biosystems (aka Perkin Elmer / Applied Biosystems / Applera) developed the automated sequencer & Venter - Whole-genome shotgun approach

In May 1998, Venter, in collaboration with Michael Hunkapiller at PE Biosystems, formed Celera Genomics

Goal: sequence the entire human genome by Dec 31, 2001 - 2 years before the completion by the HGP, and for a mere $300 million

April 6, 2000 - Celera announces completion "Cracks human code"

Agrees to wait for HGP

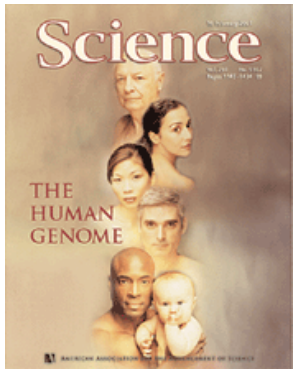
Summer 2000 - both groups announced the rough draft is ready

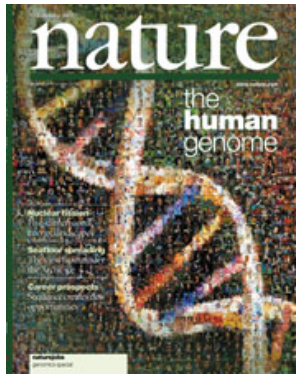
Human Genome Sequence

6 months later it was published - 5 years ahead of schedule with $3B

50 years after the discovery of DNA structure

Human Genome Project was completed - 3.1 billion basepairs





Pros: No guessing of where the genes are

Study individual genes and their contribution

Understand molecular evolution

Risk prediction and diagnosis

Con: Future Health Diary --> physical and mental

Who should be entrusted? Future Partners, Agencies, Government

Right to "Genetic Privacy"

Modern Sequencing methods

- 454 Sequencing (60Mbp/run) [Roche]
- Solexa Sequencing (600Mbp/run) [Illumina]

Compare to

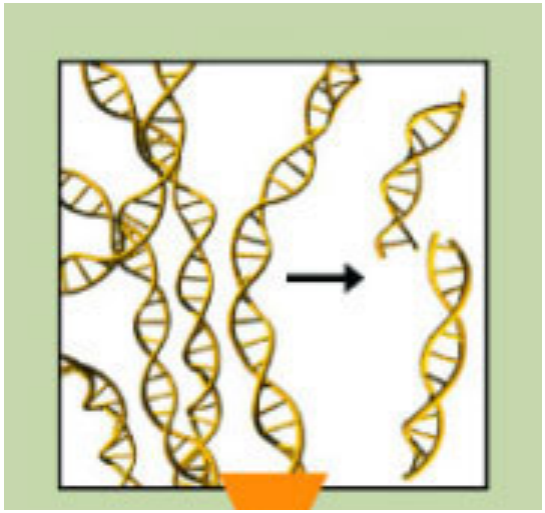
- Sanger Method (70Kbp/run)
- Shotgun Sequencing (??)

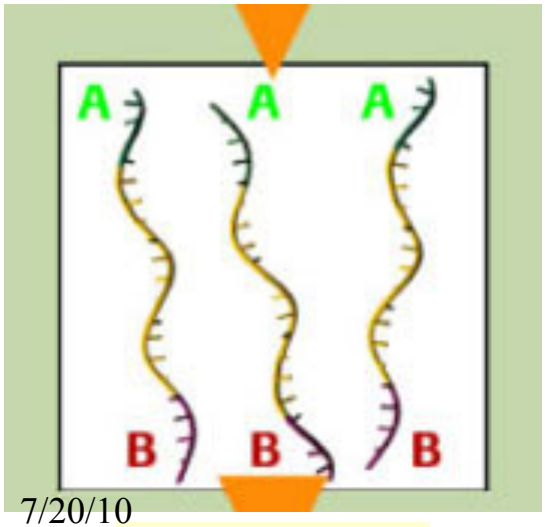
454 Sequencing: New Sequencing Technology

- 454 Life Sciences, Roche
- Sequencing by synthesis - pyrosequencing
- Parallel pyrosequenicng
- Fast (20 million bases per 4.5 hour run)
- Low cost (lower than Sanger sequencing)
- Simple (entire bacterial genome in on day 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.

454 Sequening

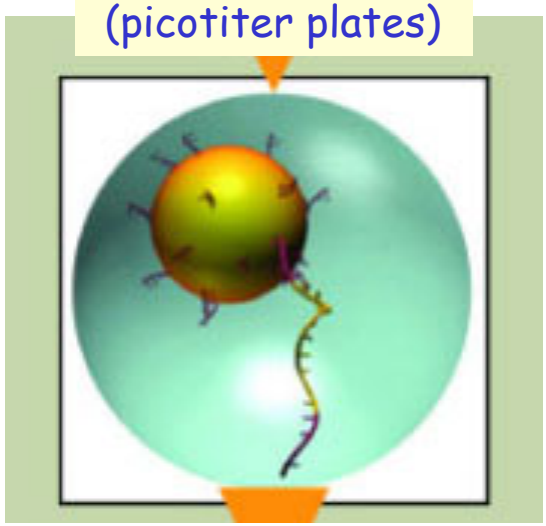
Fragment

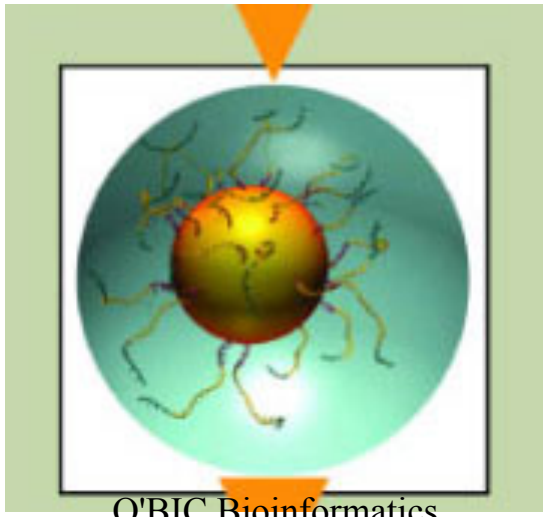




Add Adaptors

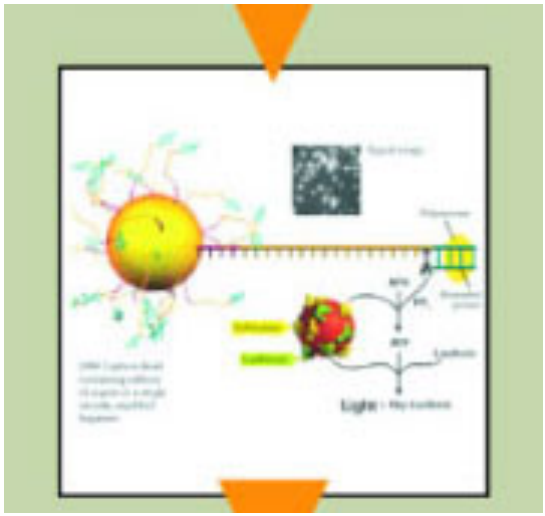
1 fragment-1 bead
(picotiter plates)

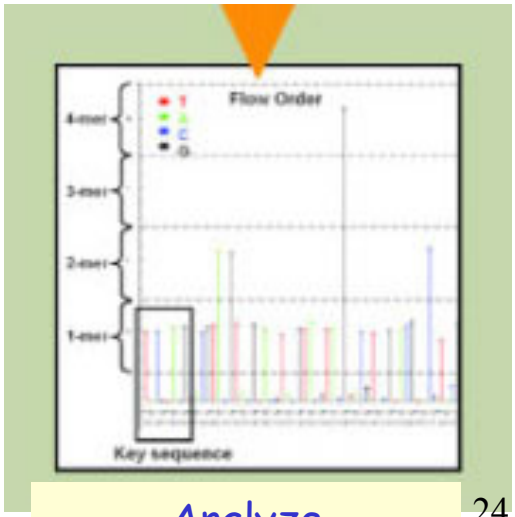




emPCR on bead

Sequence

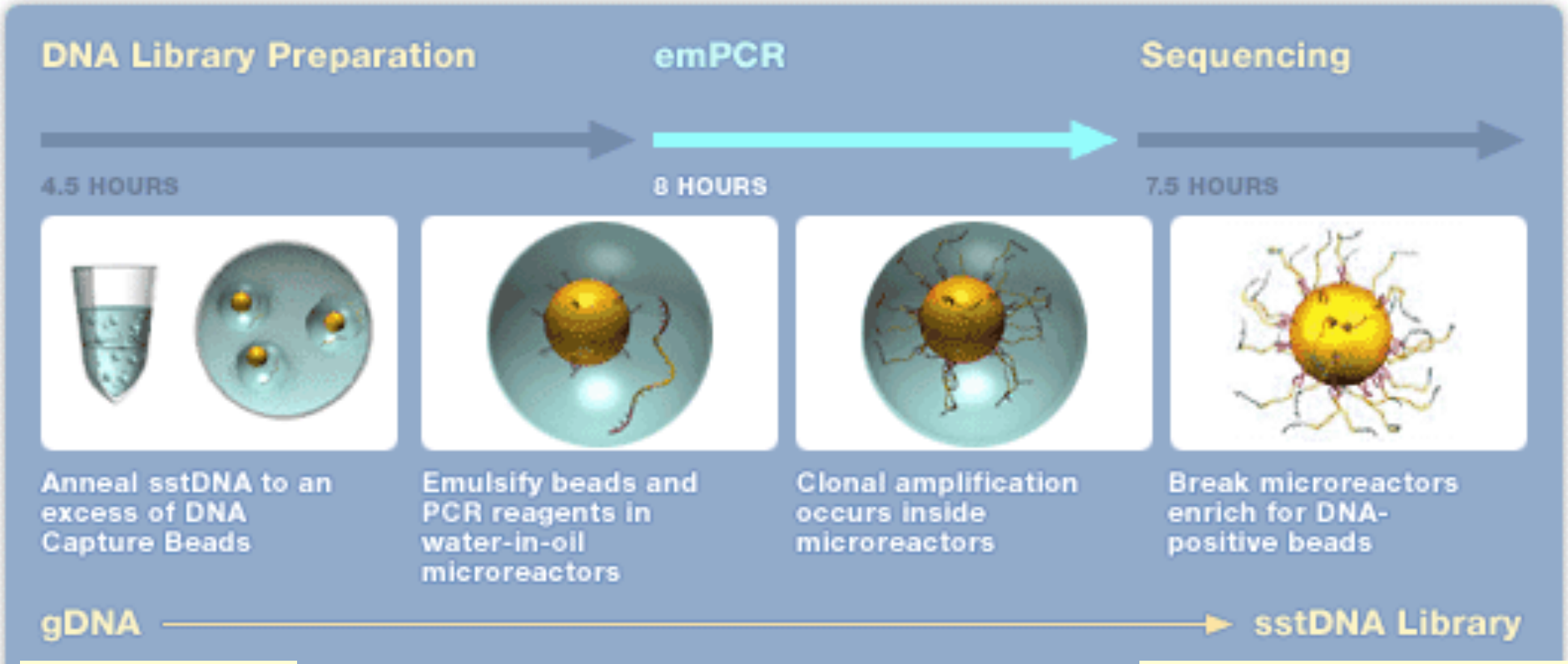




Analyze
one bead - one read

emPCR

FIGURE 8

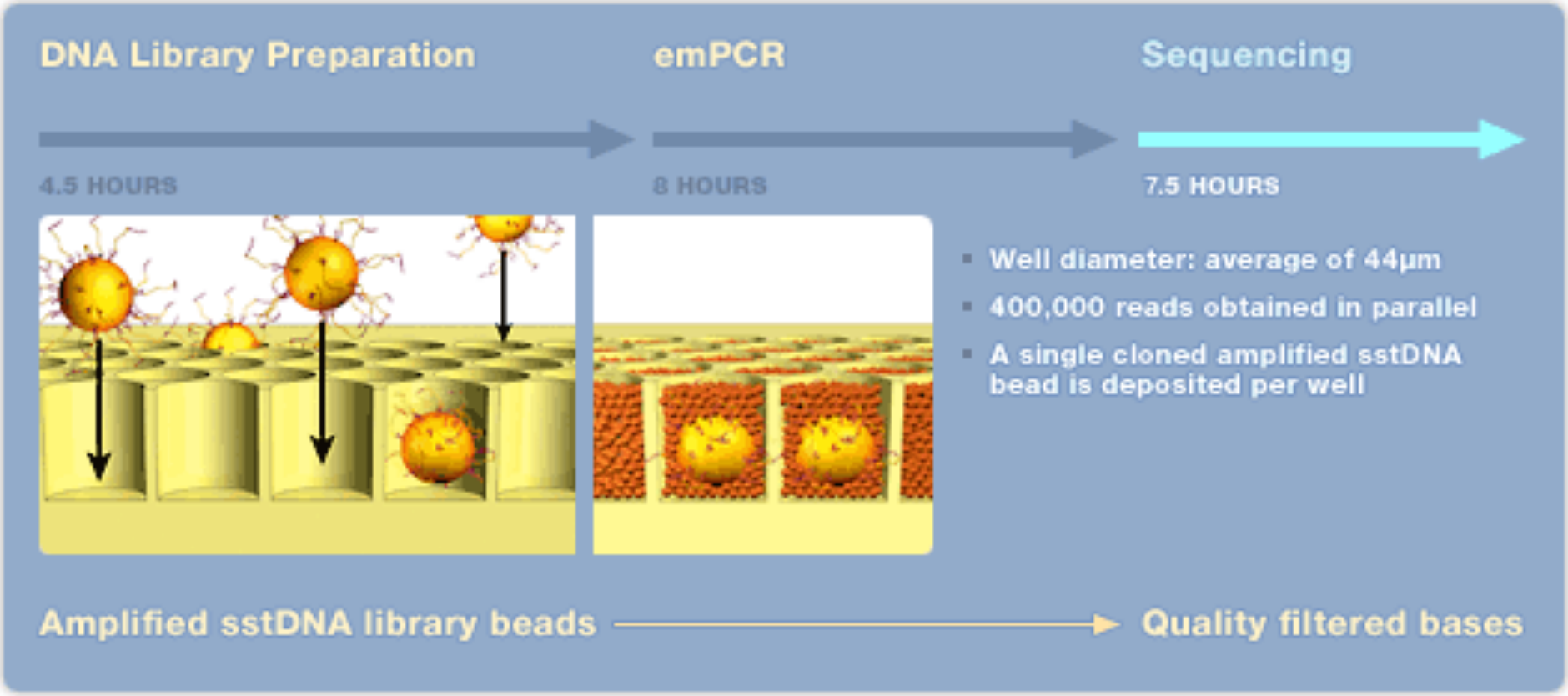


genomic DNA

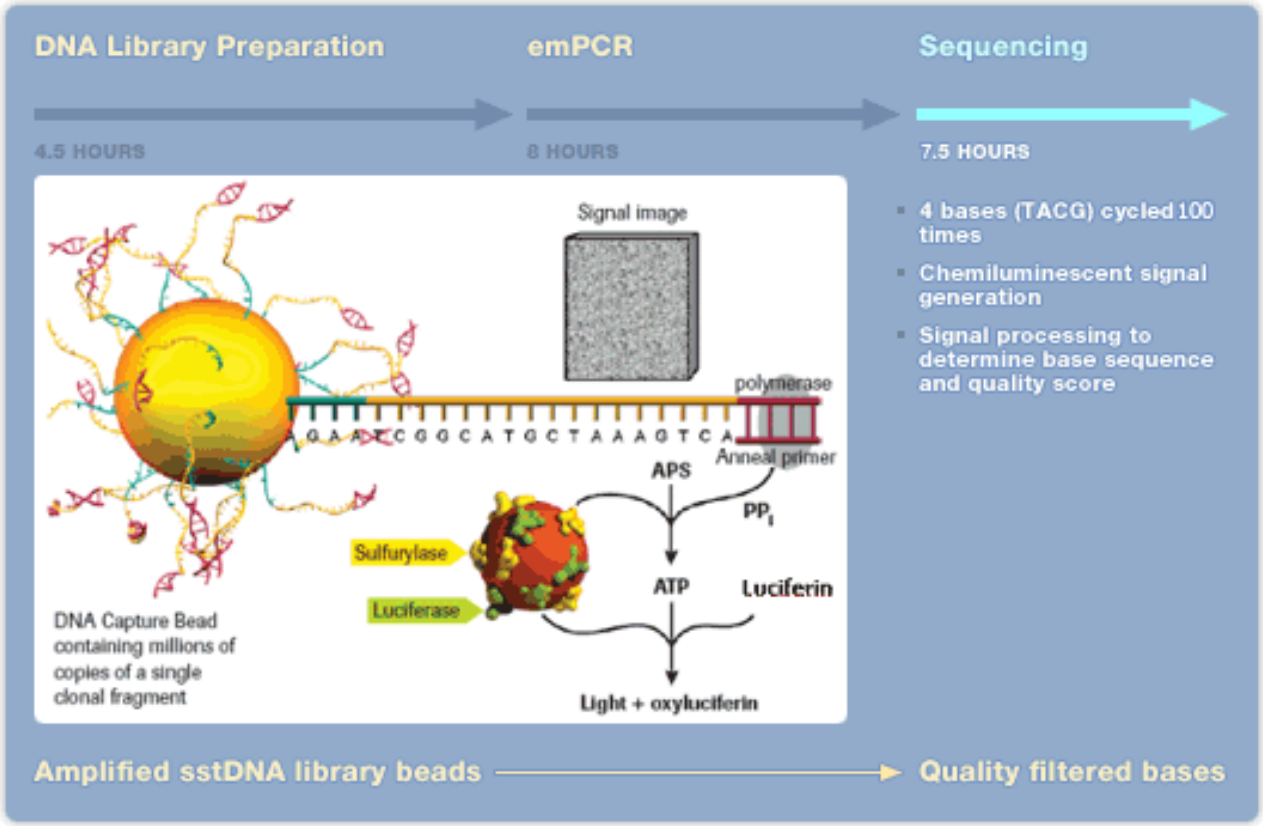
Single stranded template DNA library

Sequencing

FIGURE 9



Sequencing



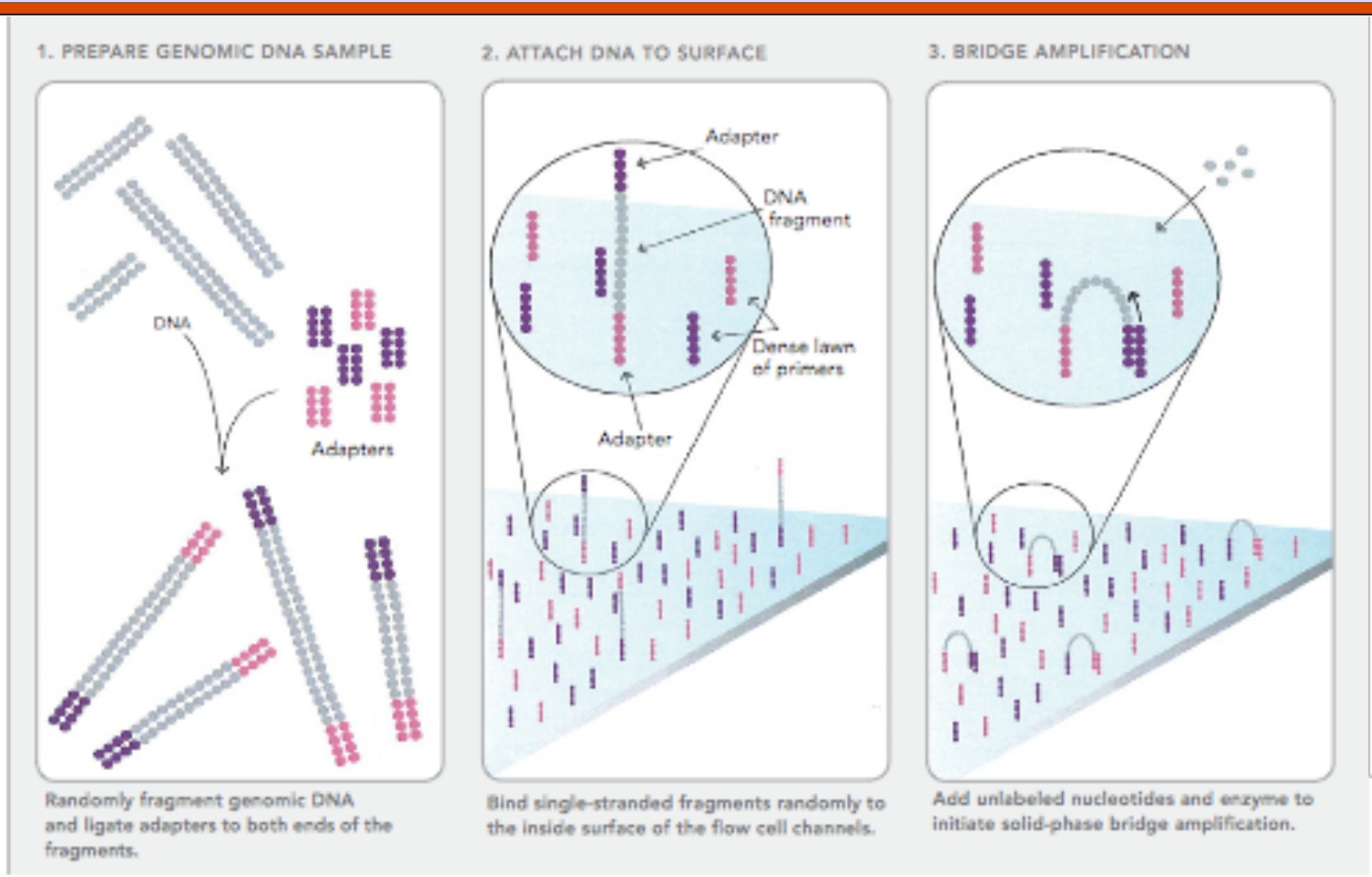
- Hundreds of thousands of beads each carrying millions of copies of unique ssDNA molecule sequenced in parallel
- Sequential flow of nt in fixed order across PicoTiterPlate

- If complementary nt flowed into a well, DNA strand is extended
- Addition reaction releases pyrophosphate molecule & is recorded
- Signal strength proportional to number of nts incorporated

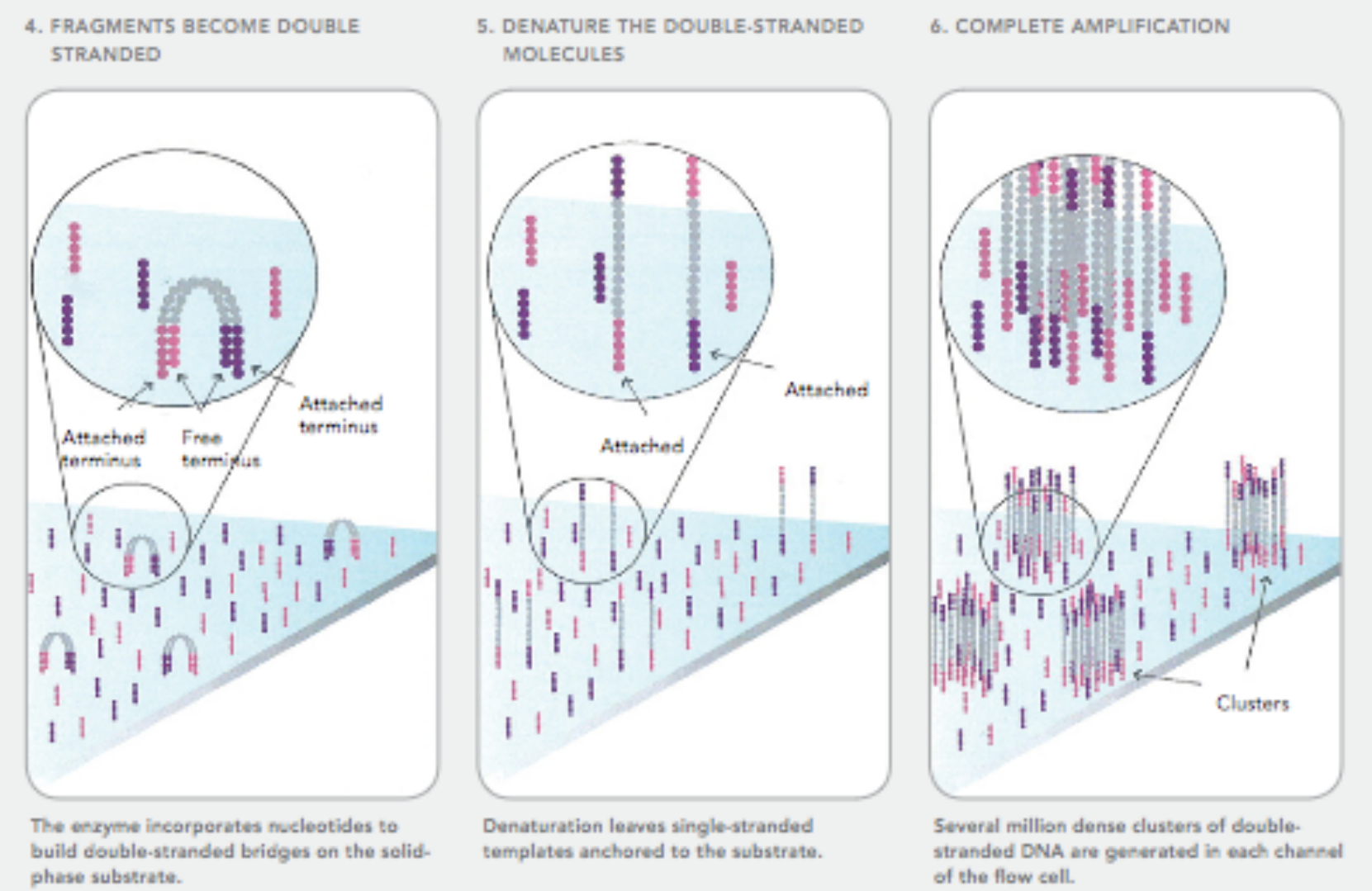
Multimedia presentation

- http://www.roche-applied-science.com/publications/multimedia/genome_sequencer/flx_multimedia/wbt.htm

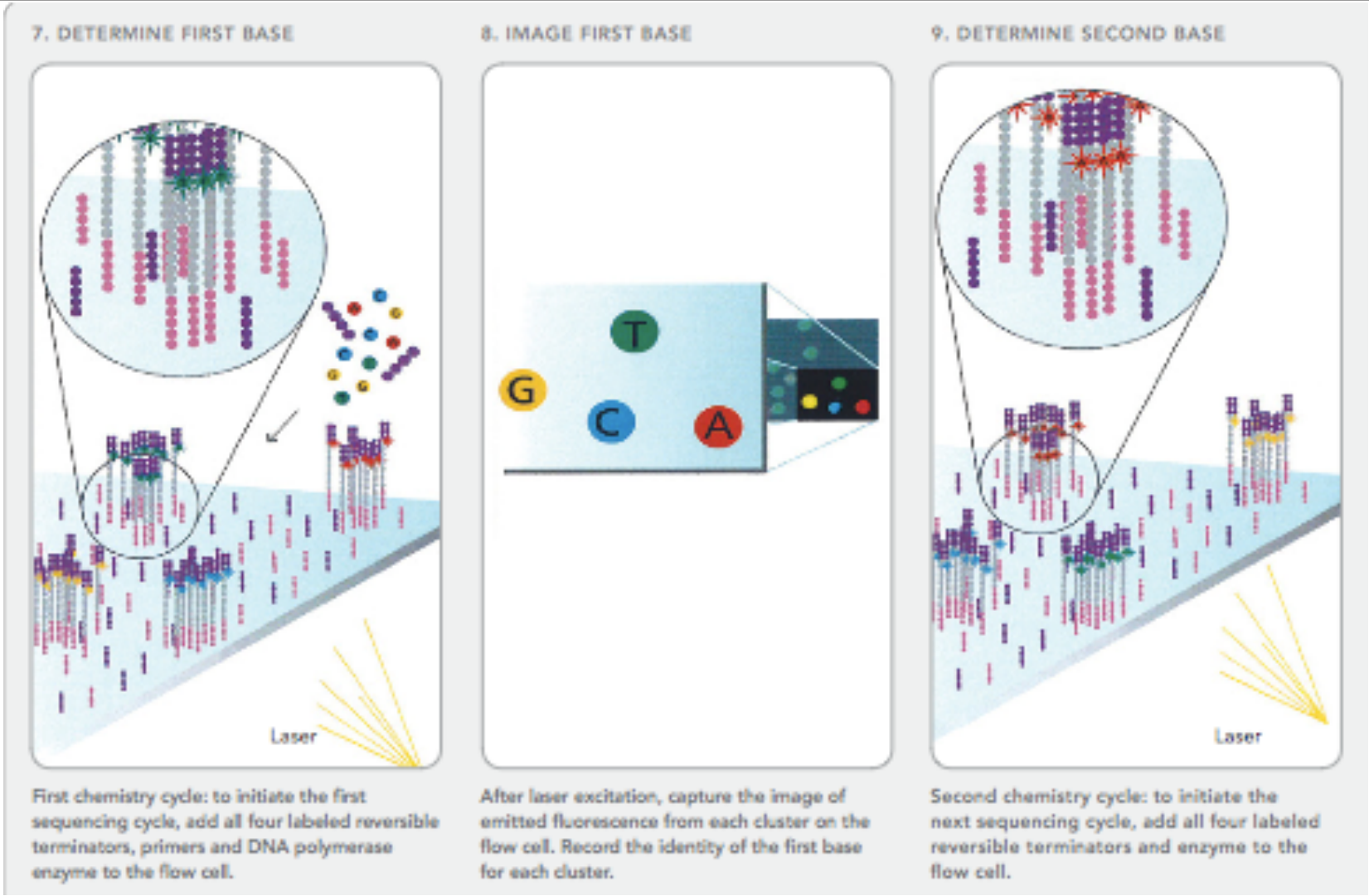
Solexa Sequencing



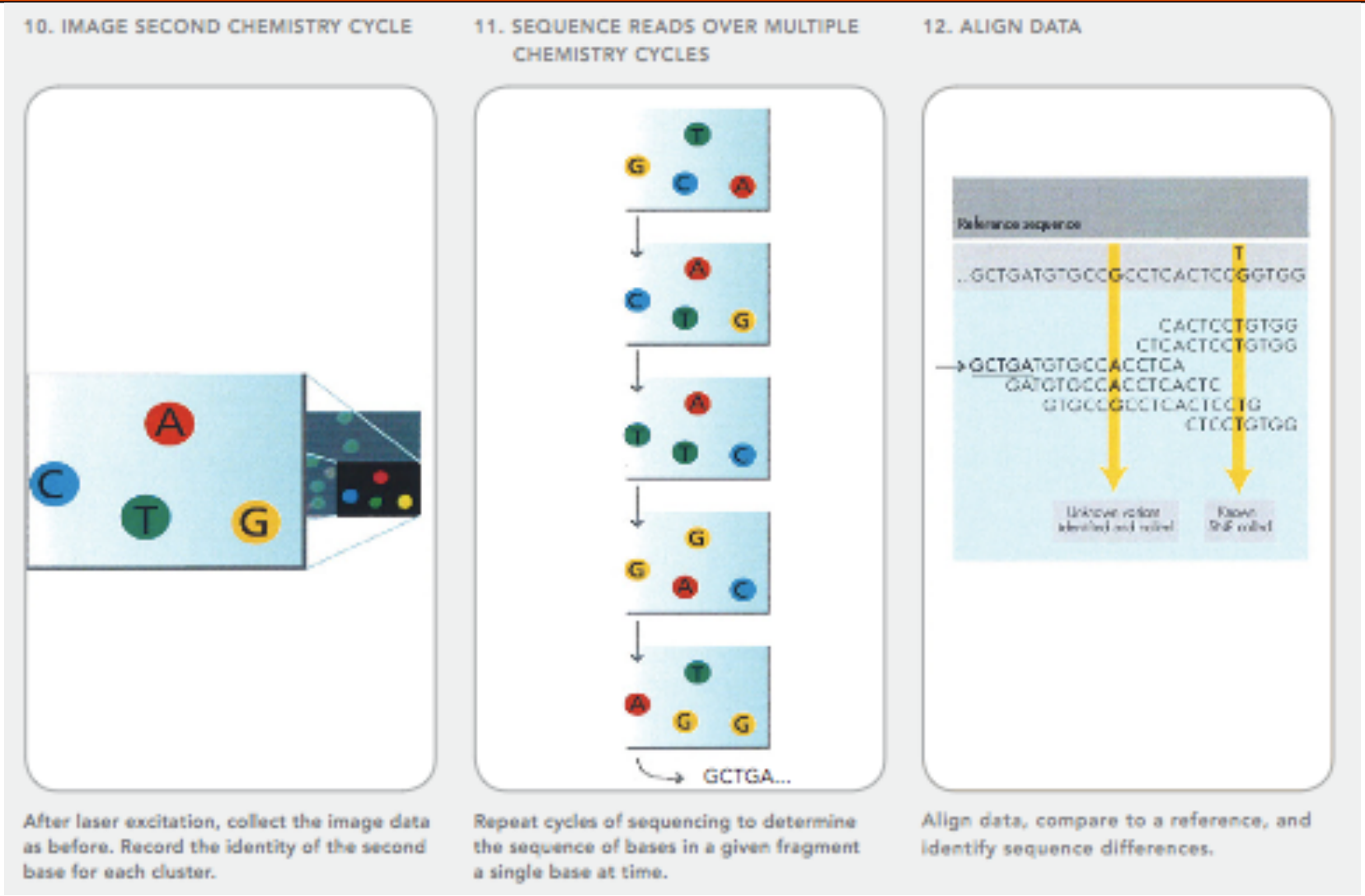
Solexa Sequencing



Solexa Sequencing

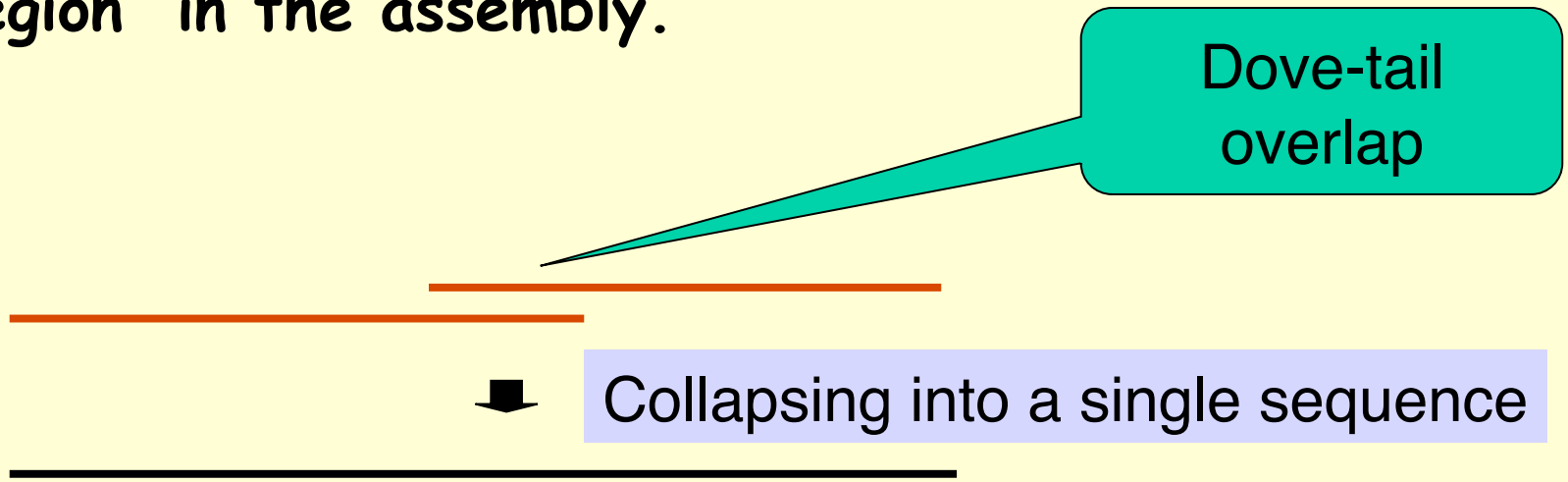


Solexa Sequencing



Sequencing: Generate Contigs

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



- 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.

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

Helicos Technology

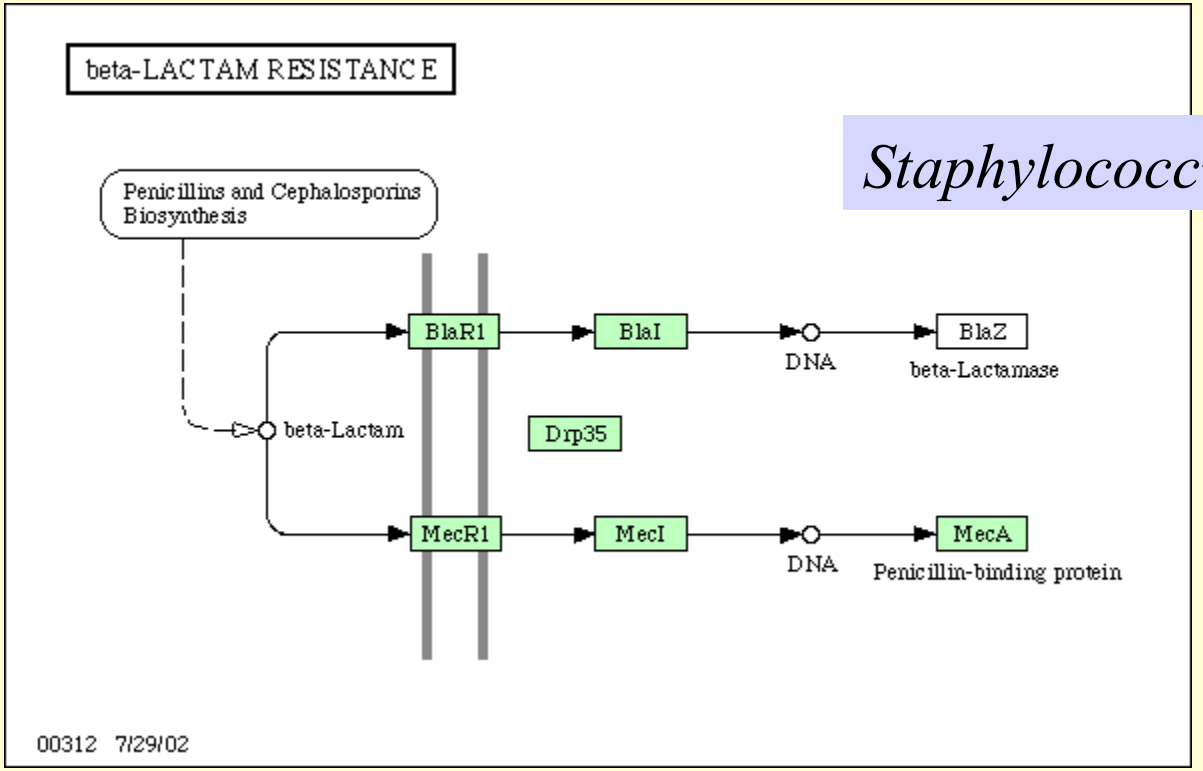
- True Single Molecule Sequencing
- DNA is fragmented and polyA added to end and fluorescent tag added
- DNA hybridized to flow cell with polyT immobilized on it
- Templates packed very closely
- Sequence extension happens one base at a time and a CCD camera takes pictures to produce images after each round
- Every strand is unique and is sequenced independently
- Very fast (1GB/hour)
- Tremendous throughput and is expected to deliver $1000 and 1-day sequencing target
- Very little preparation; No ligations needed
- No amplification
- No cluster picking

Applications of NGS

- Sequencing: Study new genomes
- RNA-Seq: Study transcriptomes and gene expression by sequencing RNA mixture
- ChIP-Seq: Analyze protein-binding sites by sequencing DNA precipitated with TF
- Metagenomics: Sequencinng metagenoms
- SNP Analysis: Study SNPs by deep sequencing of regions with SNPs
- Resequencing: Study variations, close gaps, etc.
- Misc applications: DNA barcoding, CNV, sRNA

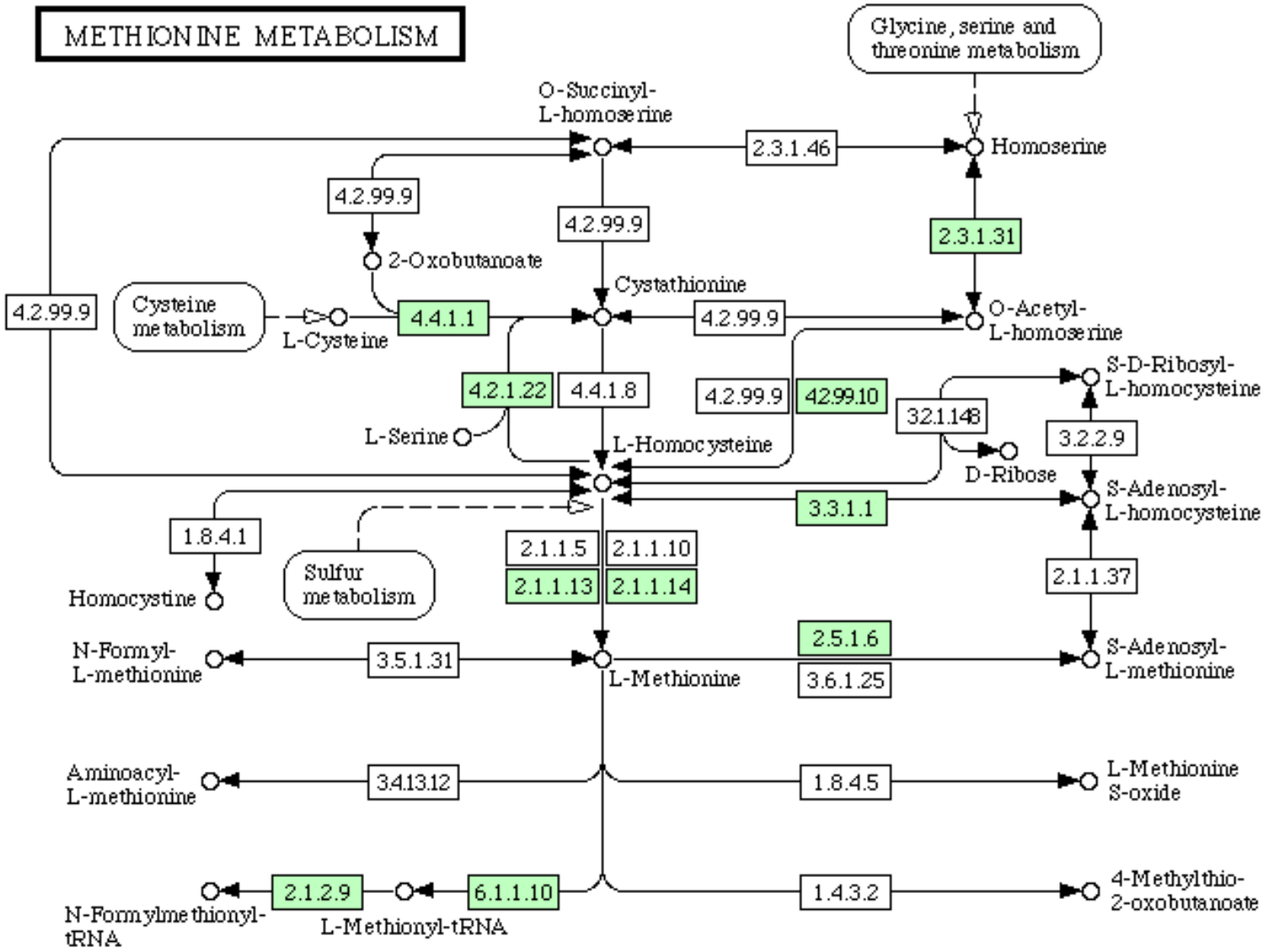
Gene Networks & Pathways

- **Genes & Proteins act in concert and therefore form a complex network of dependencies.**



Staphylococcus aureus

Pseudomonas aeruginosa



Omics

- Genomics: Study of all genes in a genome, or comparison of whole genomes.
  - Whole genome sequencing
- Metagenomics
  - Study of total DNA from a community (sample without separation or cultivation)
- Proteomics: Study of all proteins expressed by a genome
  - What is expressed at a particular time
  - 2D gel electrophoresis & Mass spectrometry
- Transcriptomics
  - Gene expression – mRNA (Microarray)
  - RNA sequencing
- Glycomics
  - Study of carbohydrates/sugars