BSC 4934: Q'BIC Capstone Workshop

Giri Narasimhan

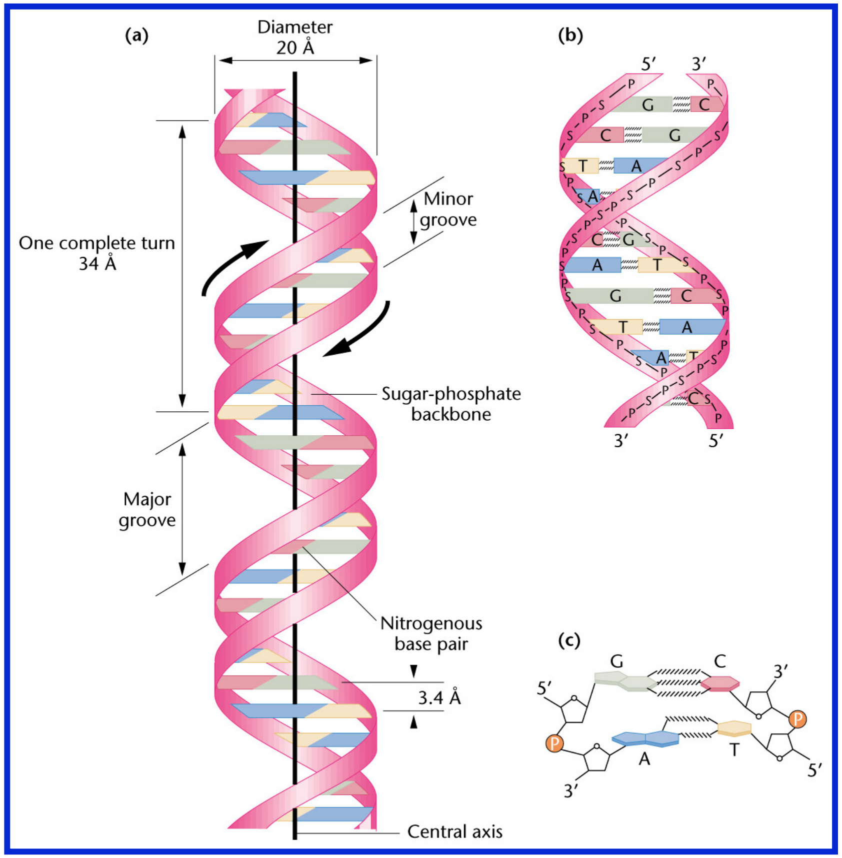
ECS 254A; Phone: x3748

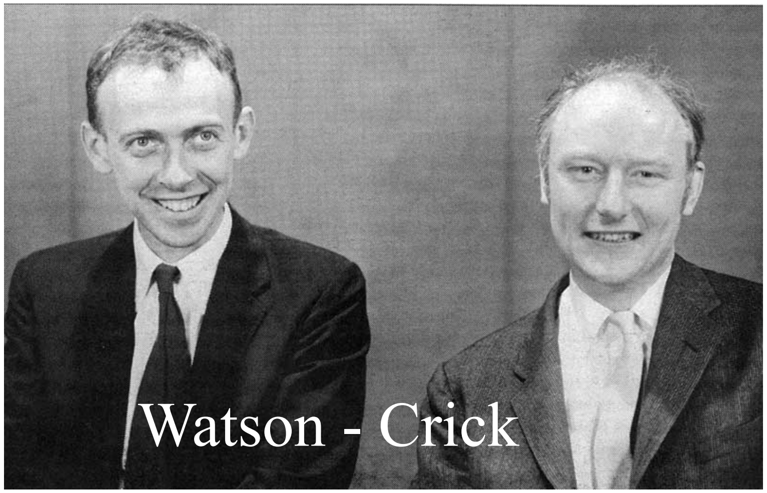
giri@cs.fiu.edu

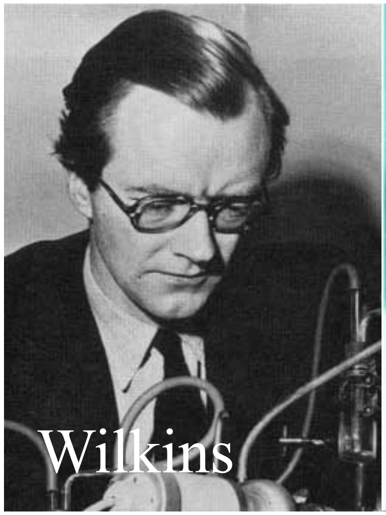
http://www.cs.fiu.edu/~giri/teach/BSC4934_Su11.html

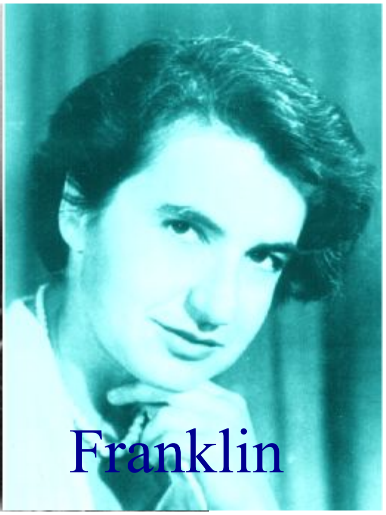
July 2011

DNA Structure - 1953









Courtesy: Dr. Kalai Mathee

DNA Controversy

1. Double Helix by Jim Watson - Personal Account (1968)
2. Rosalind Franklin by Ann Sayre (1975)
3. The Path to the Double Helix by Robert Olby (1974)
4. Rerelease of Double Helix by Jim Watson with Franklin's paper
5. Rosalind Franklin: The Dark Lady of DNA by Brenda Maddox (2003)
6. Secret of Photo 51 - 2003 NOVA Series

Courtesy: Dr. Kalai Mathee

What are the next big Qs?

1. What is order of DNA sequence in a chromosome?
2. How does the DNA replicate?
3. How does the mRNA get transcribed?
4. How does the protein get translated?

Etc.

One of the tools that made a difference

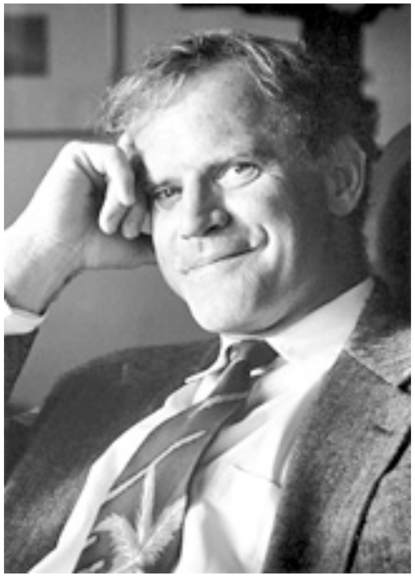
Polymerase Chain Reaction

Courtesy: Dr. Kalai Mathee

Polymerase Chain Reaction

1983 - technique was developed by Kary Mullis & others (1944-)

1993 Nobel prize for Chemistry

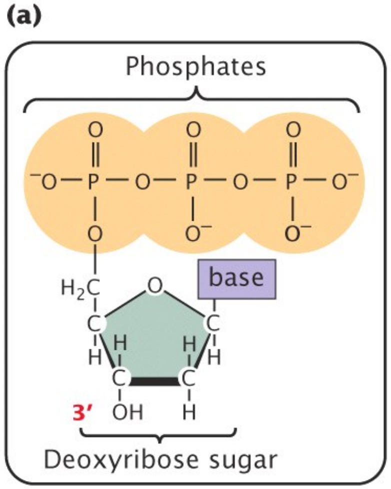


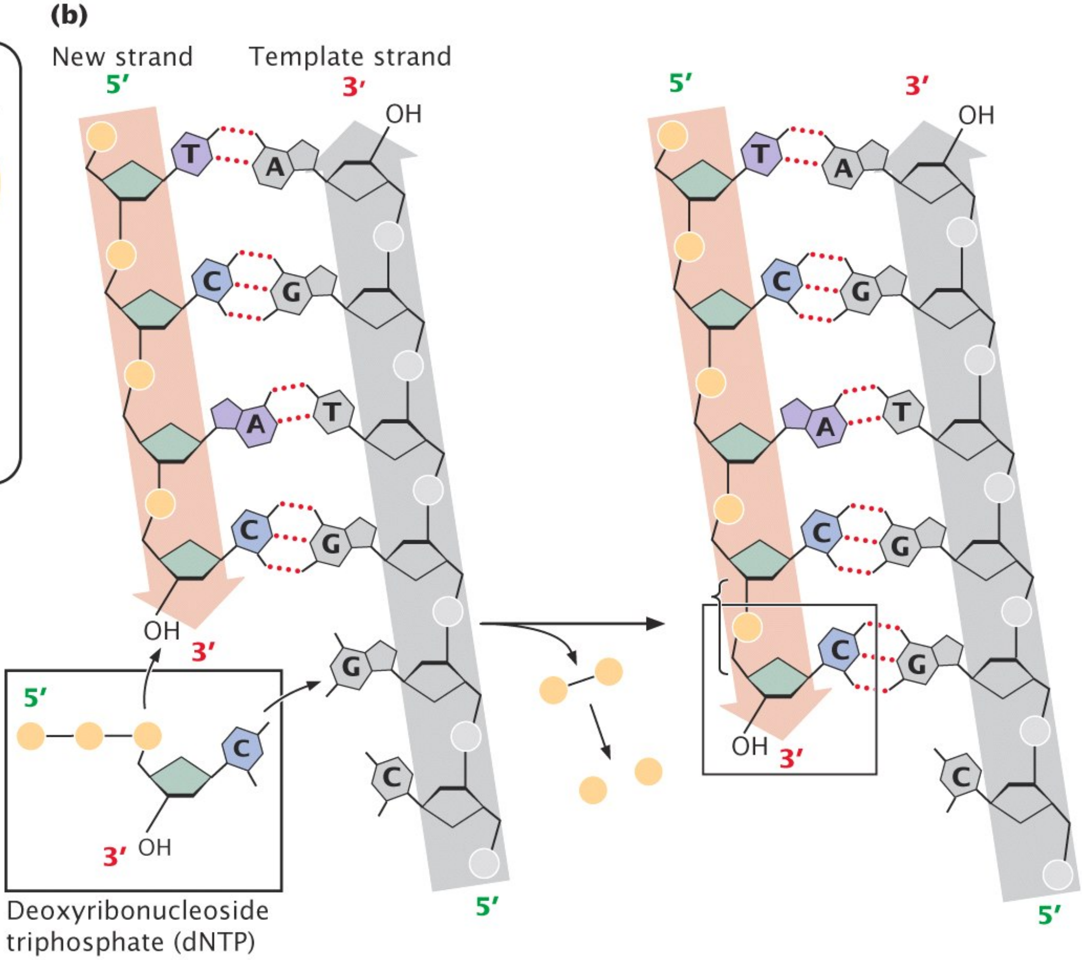
Controversy: Kjell Kleppe, a Norwegian scientist in 1971, published paper describing the principles of PCR

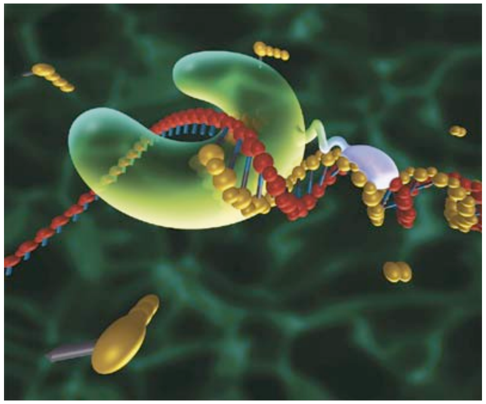
Stuart Linn, professor at University of California, Berkeley, used Kleppe's papers in his own classes, in which Kary Mullis was a student at the time

Courtesy: Dr. Kalai Mathee

DNA Replication & Polymerase







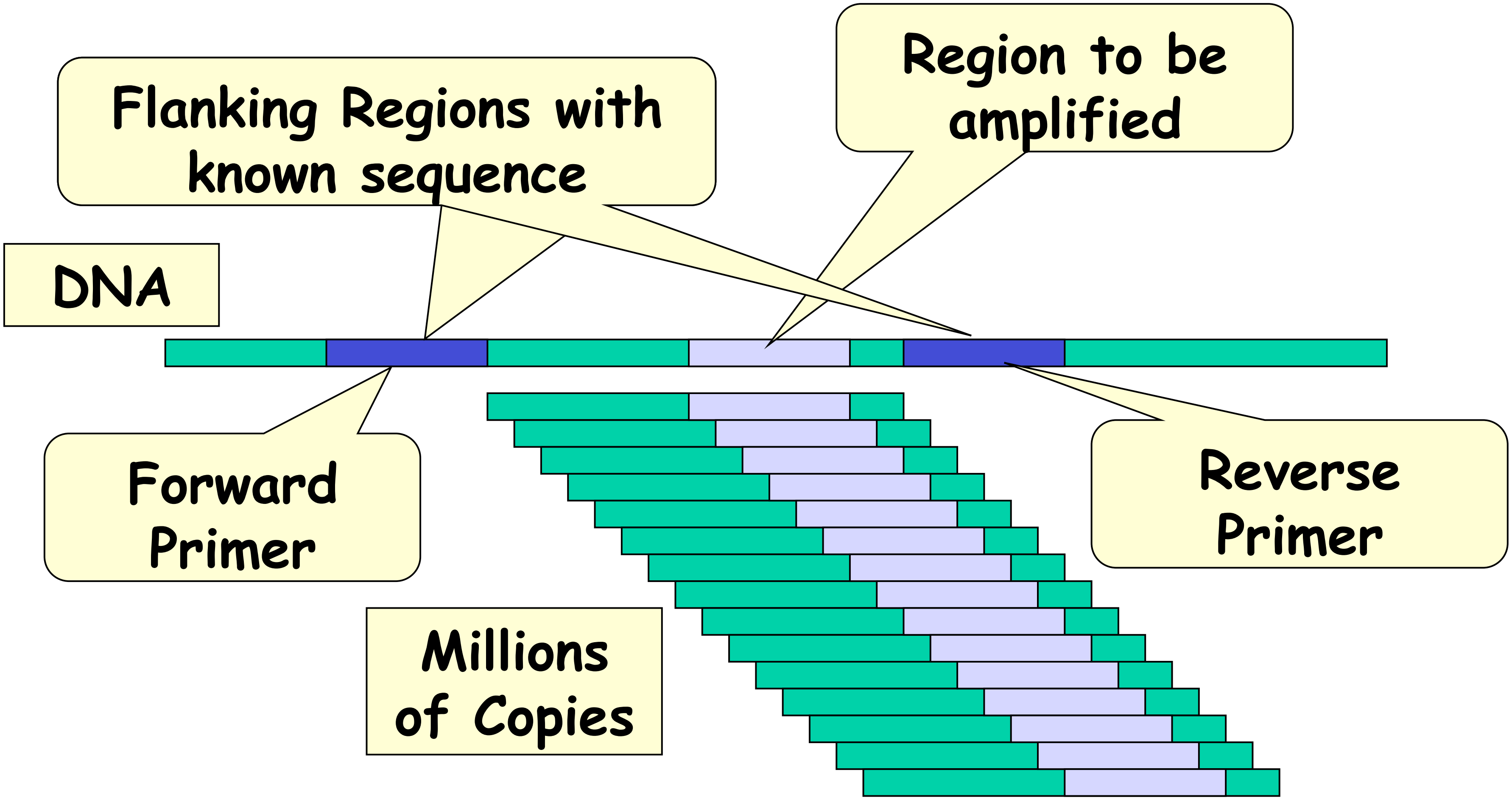
Courtesy: Dr. Kalai Mathee

Polymerase Chain Reaction (PCR)

- ❑ 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
- ❑ If no info known, one can add adapters (short known sequence) then use a primer that recognizes the adapter

Courtesy: Dr. Kalai Mathee

PCR



Flanking Regions with known sequence

Region to be amplified

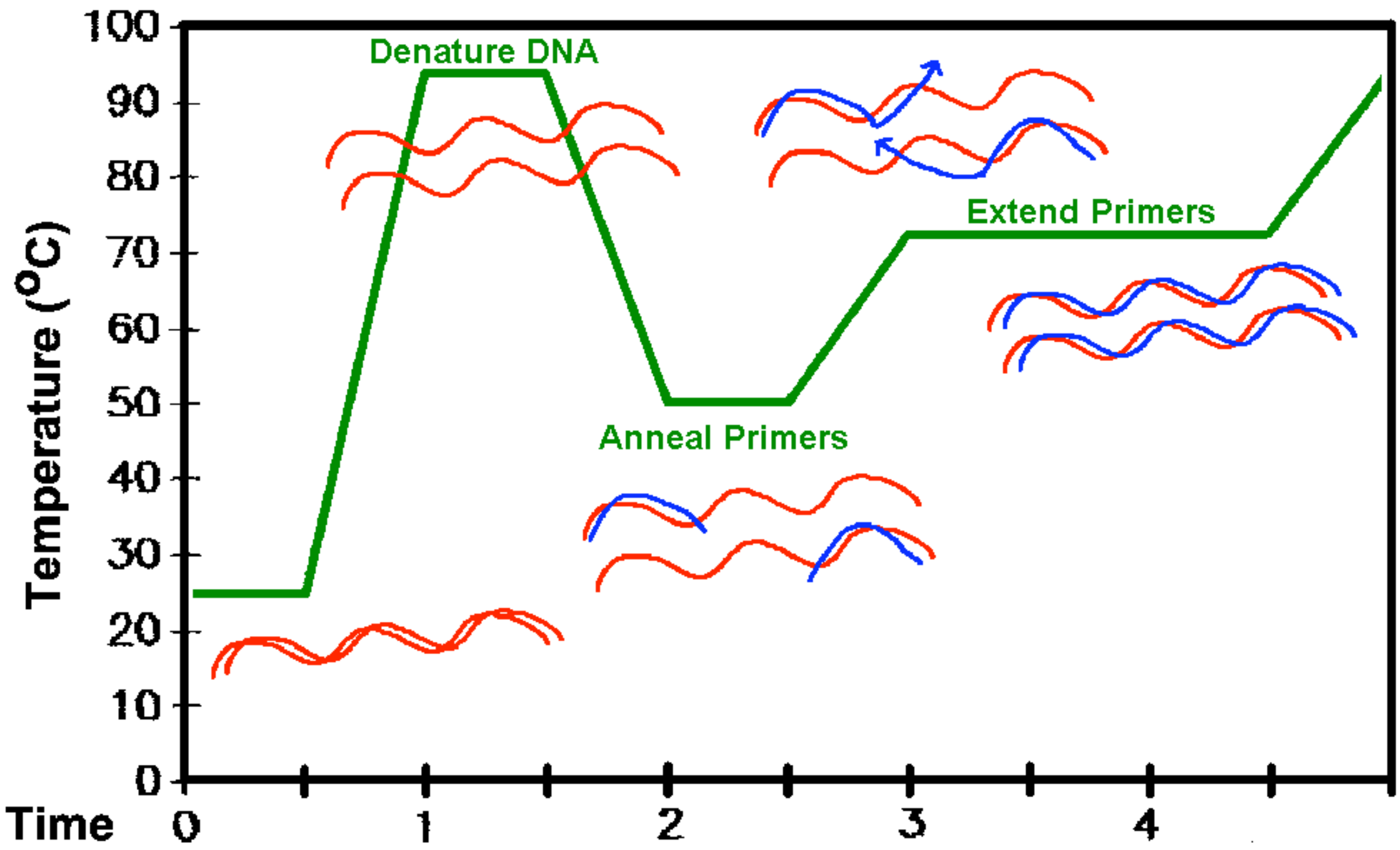
DNA

Forward Primer

Reverse Primer

Millions of Copies

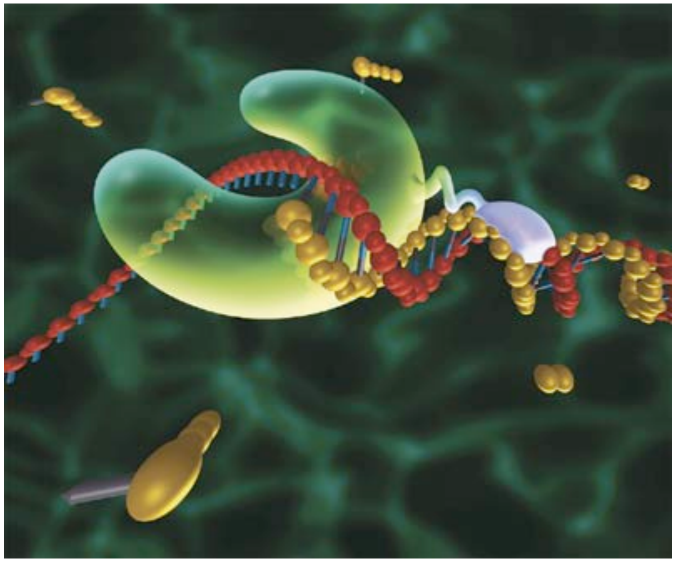
PCR



Courtesy: Dr. Kalai Mathee

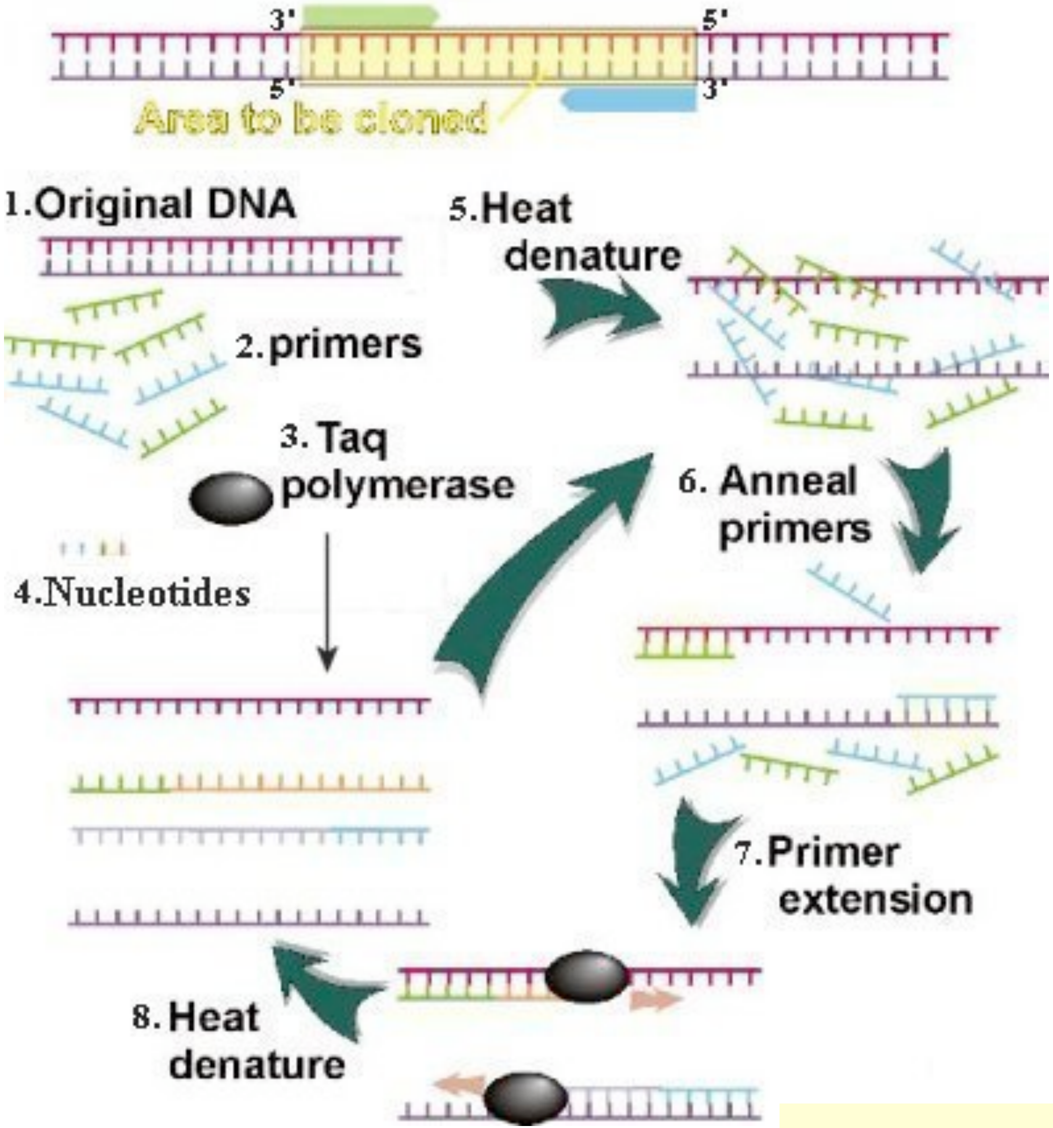
Taq polymerase

- ❑ Thermostable DNA polymerase named after the thermophilic bacterium *Thermus aquaticus*
- ❑ Originally isolated by Thomas D. Brock in 1965
- ❑ Molecule of the 80s
- ❑ Many versions of these polymerases are available
- ❑ Modified for increased fidelity



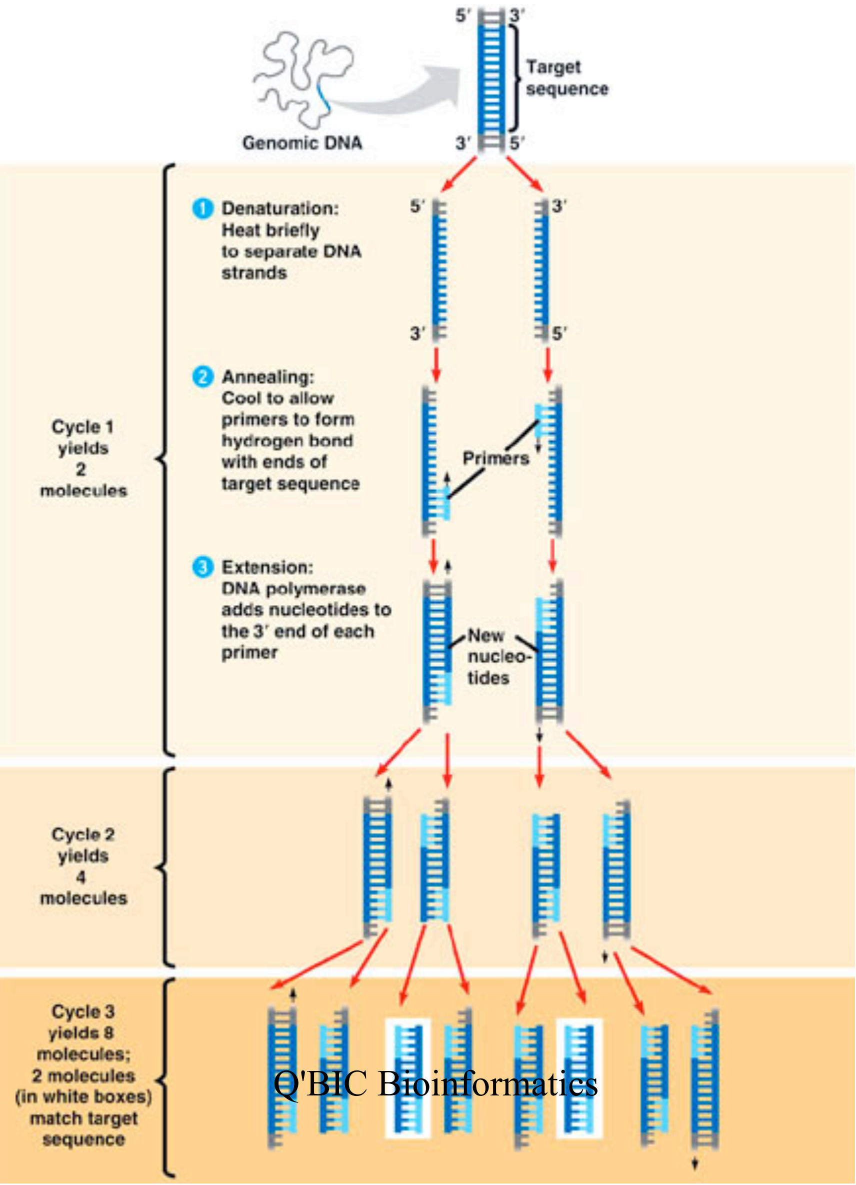
Courtesy: Dr. Kalai Mathee

Schematic outline of a typical PCR cycle



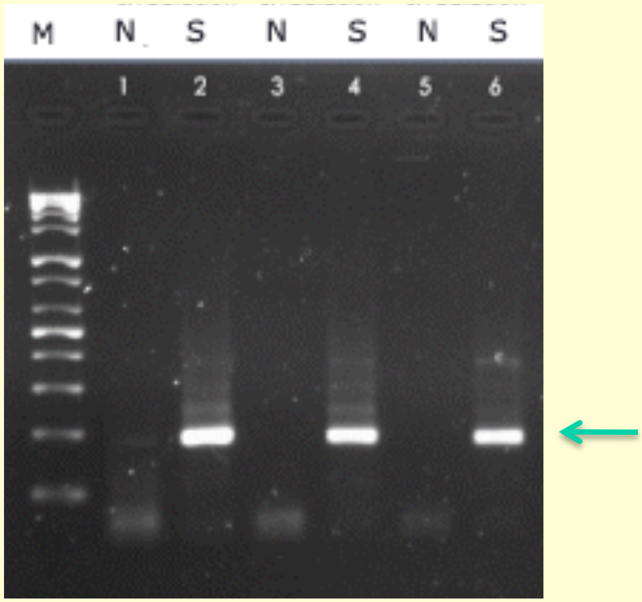
Courtesy: Dr. Kalai Mathee

PCR


5′
3′
Target sequence
Genomic DNA
3′
5′
Cycle 1 yields 2 molecules
1 Denaturation: Heat briefly to separate DNA strands
5′
3′
3′
5′
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).

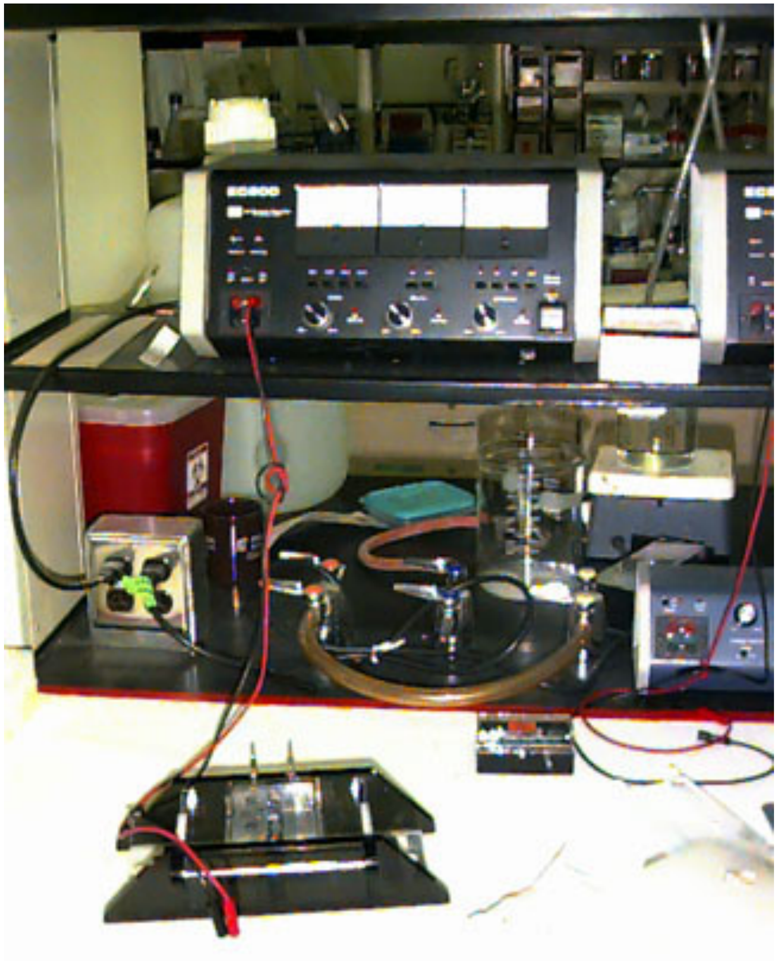


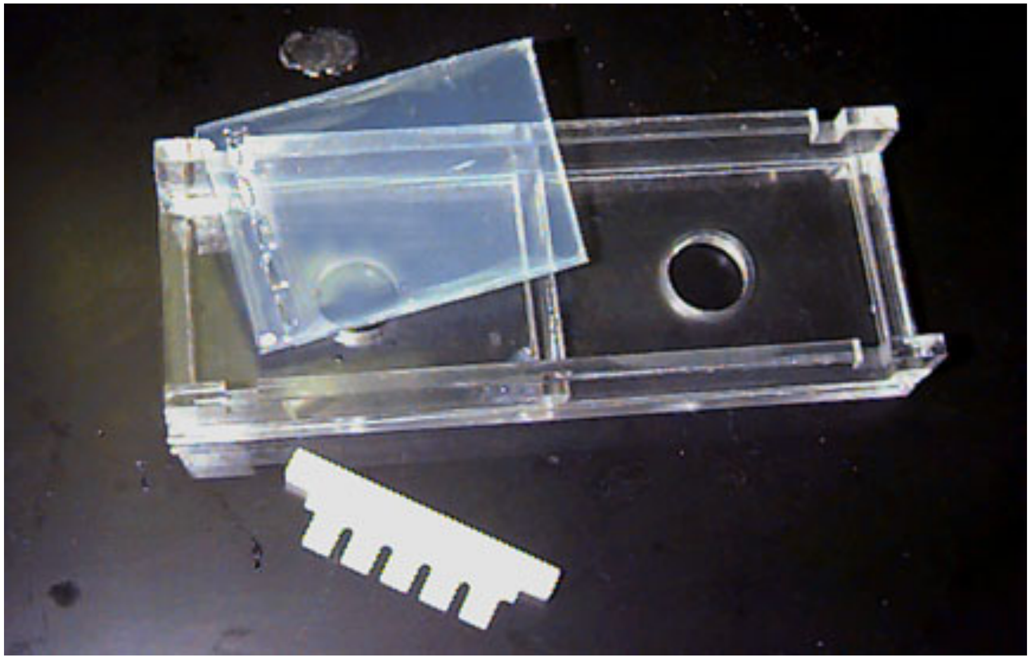
Gel Electrophoresis for DNA

- ❑ DNA is negatively charged – WHY?
- ❑ DNA can be separated according to its size
- ❑ Use a molecular sieve – Gel
- ❑ Varying concentration of agarose makes different pore sizes & results
- ❑ Boil agarose to cool and solidify/polymerize
- ❑ Add DNA sample to wells at the top of a gel
- ❑ Add DNA loading dye (color to assess the speed and make it denser than running buffer)
- ❑ Apply voltage
- ❑ Larger fragments migrate through the pores slower
- ❑ Stain the DNA – EtBr, SyberSafe, etc

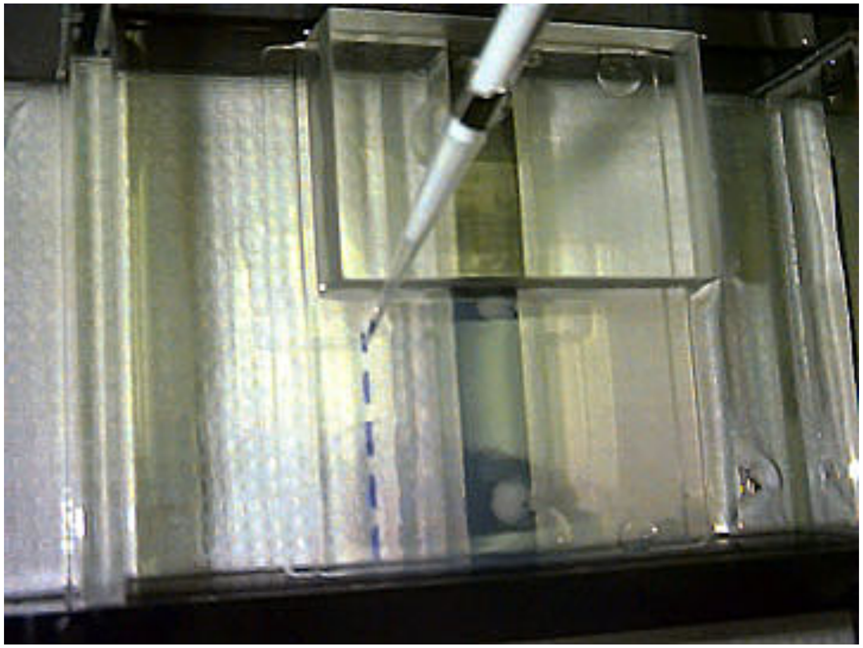
Courtesy: Dr. Kalai Mathee

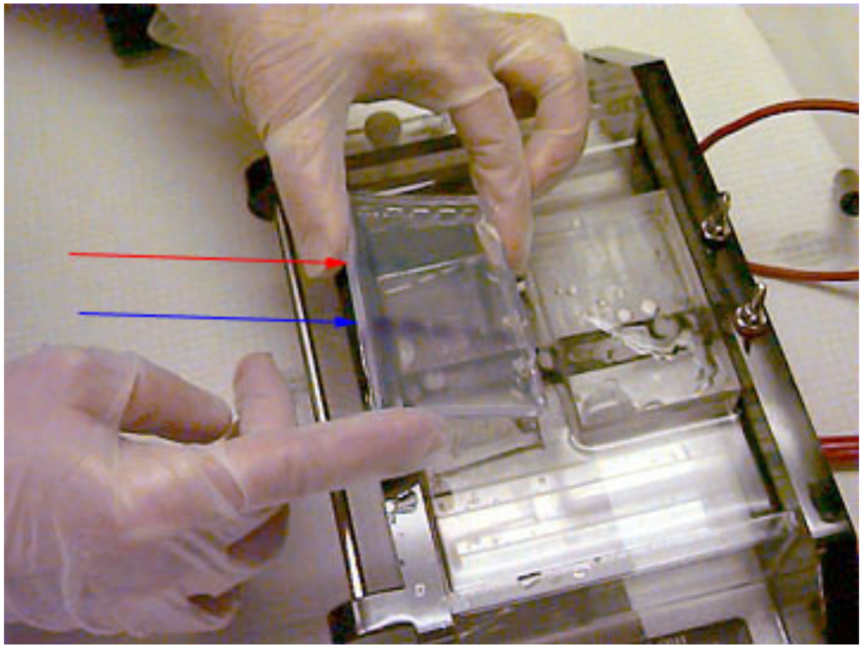
Gel Electrophoresis





Gel Electrophoresis





Sequencing

Why sequencing?

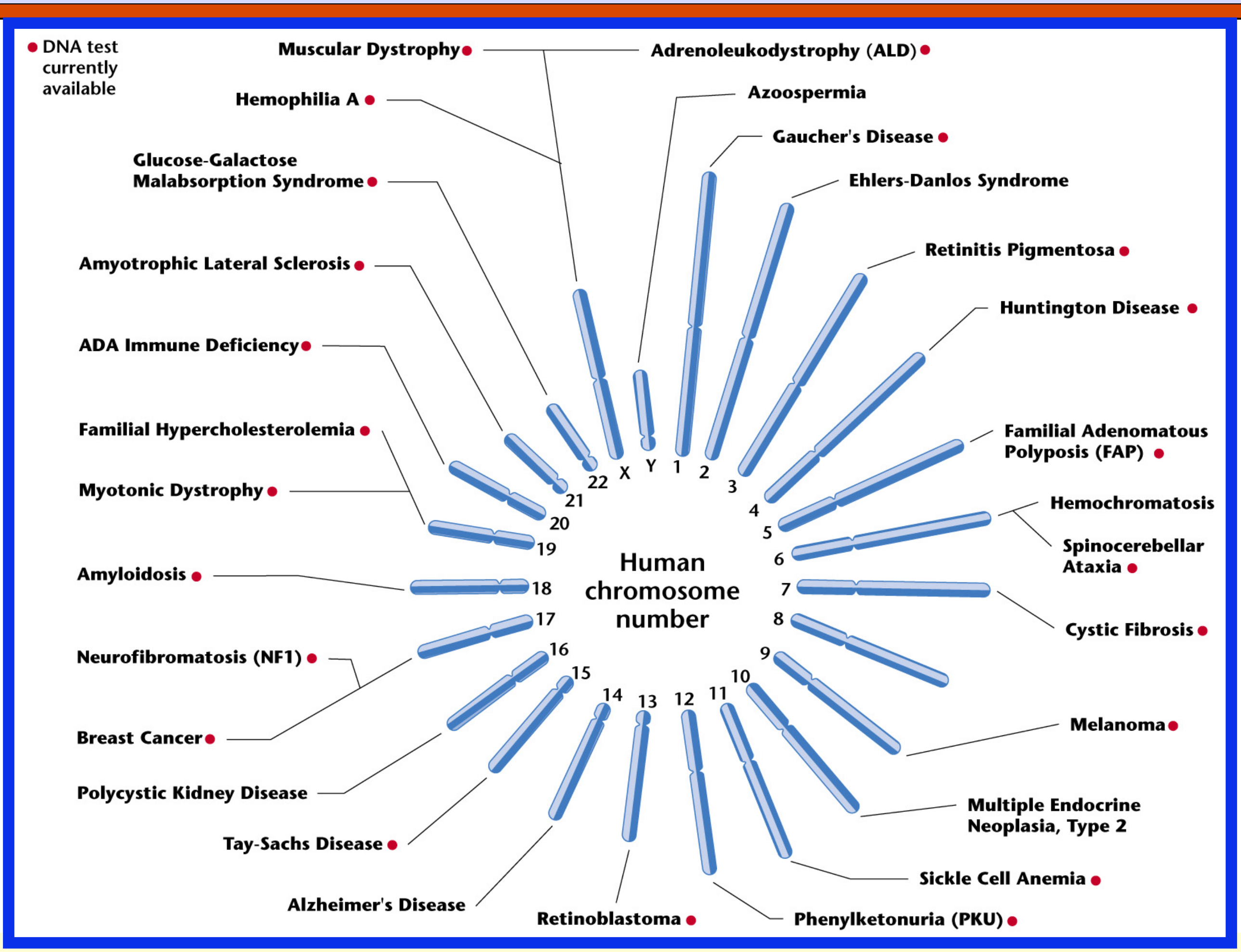
❑ Useful for further study:

- Locate gene sequences, regulatory elements
- Compare sequences to find similarities
- Identify mutations – genetic disorders
- Use it as a basis for further experiments
- Better understand the organism
- Forensics

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

Courtesy: Dr. Kalai Mathee

Human Hereditary Diseases



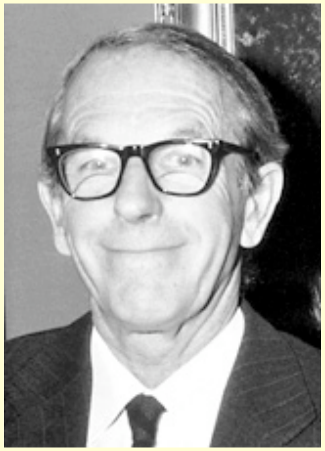
Those inherited conditions that can be diagnosed using DNA analysis are indicated by a (●)

Courtesy: Dr. Kalai Mathee

History

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

Insulin; Sanger, 1958



Sanger

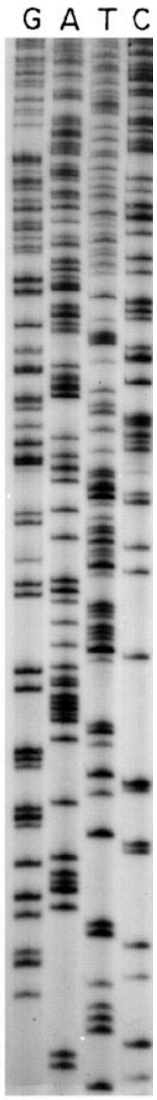
Gilbert

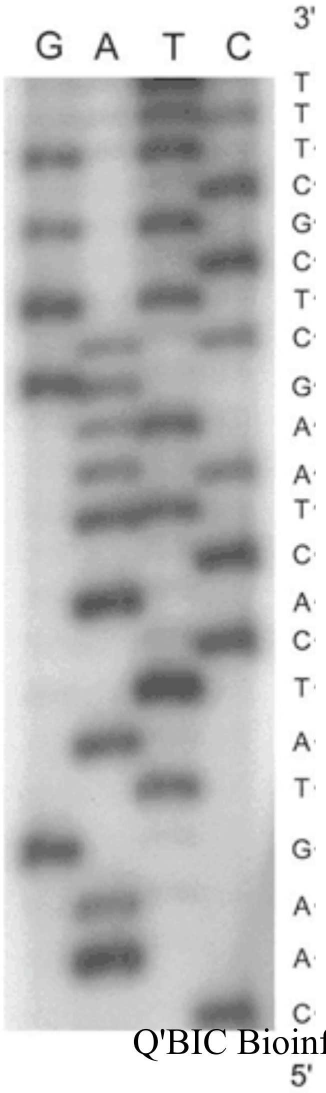
Courtesy: Dr. Kalai Mathee

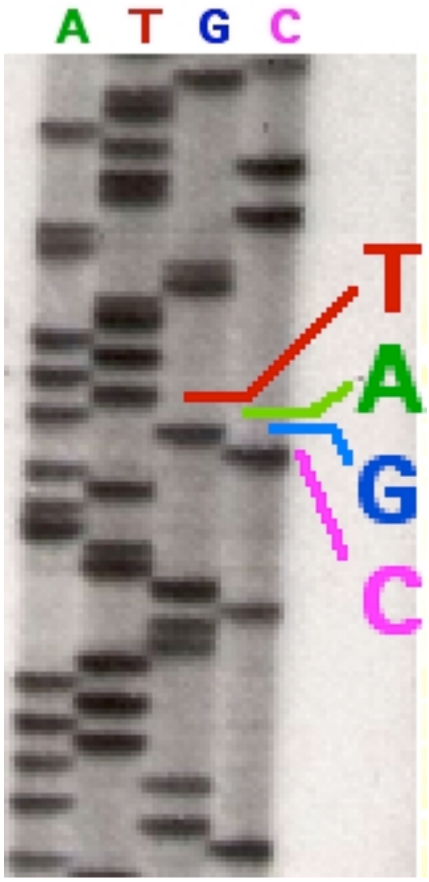
Original Sanger Method

- (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

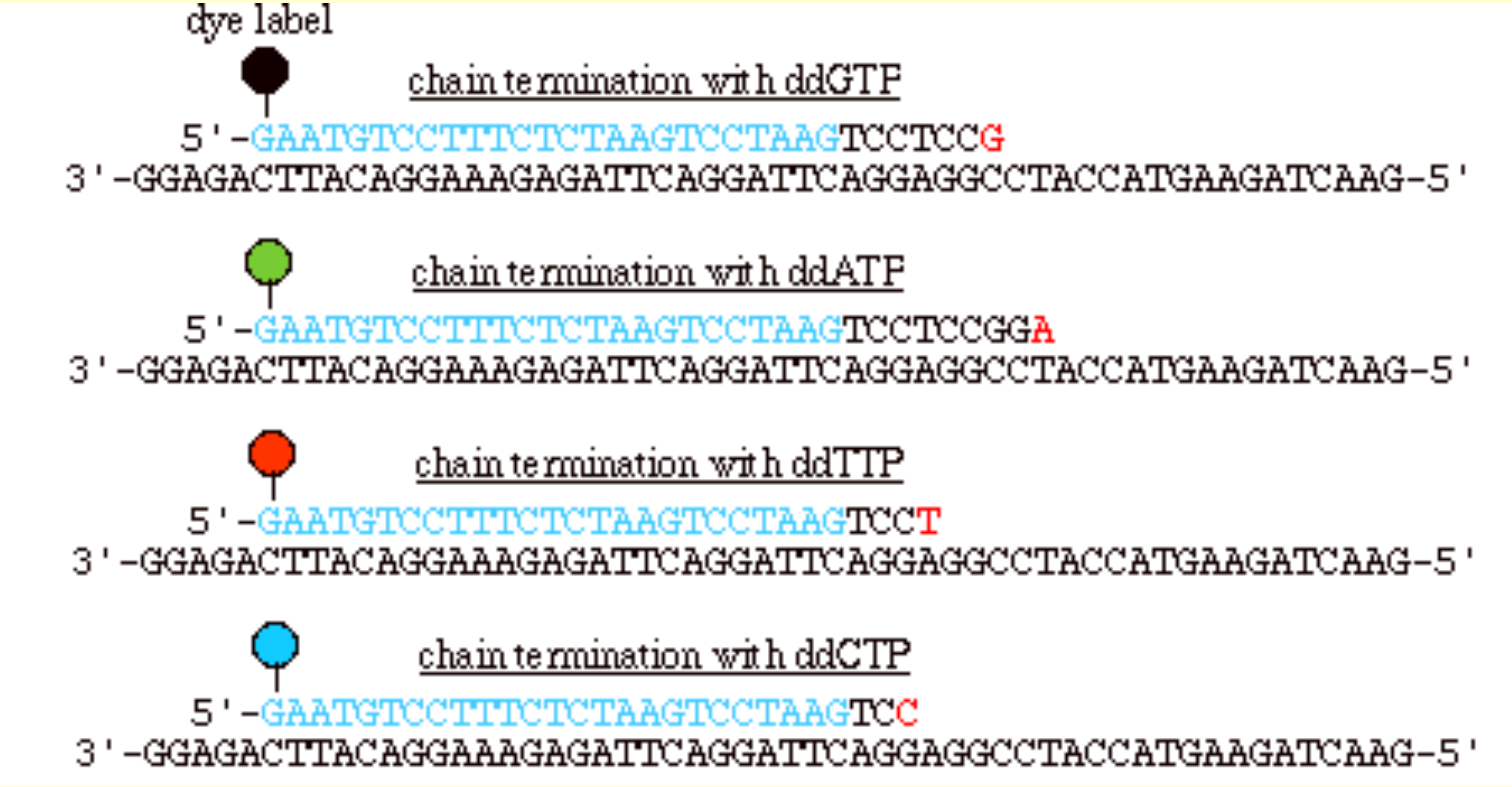






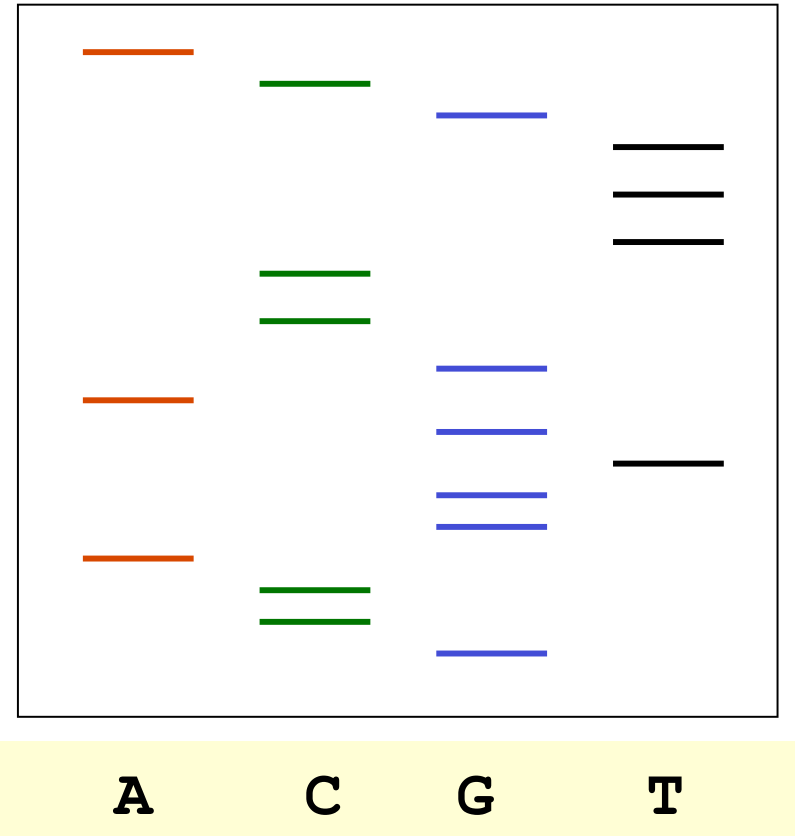
Modified Sanger

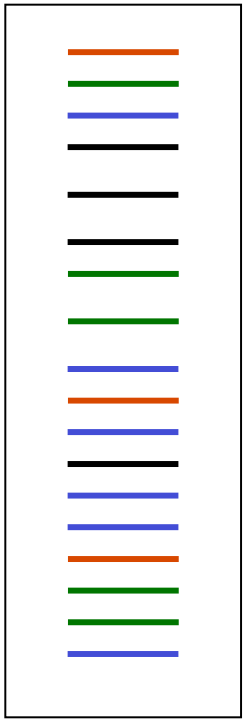
- Reactions performed in a single tube containing all four ddNTP's, each labeled with a different color fluorescent dye


dye label
chain termination with ddGTP
5'-GAATGTCCTTTCTCTAAGTCCTAAGTCCTCCG
3'-GGAGACTTACAGGAAAGAGATTCAGGATTCAGGAGGCCTACCATGAAGATCAAG-5'
chain termination with ddATP
5'-GAATGTCCTTTCTCTAAGTCCTAAGTCCTCCGGA
3'-GGAGACTTACAGGAAAGAGATTCAGGATTCAGGAGGCCTACCATGAAGATCAAG-5'
chain termination with ddTTP
5'-GAATGTCCTTTCTCTAAGTCCTAAGTCCT
3'-GGAGACTTACAGGAAAGAGATTCAGGATTCAGGAGGCCTACCATGAAGATCAAG-5'
chain termination with ddCTP
5'-GAATGTCCTTTCTCTAAGTCCTAAGTCC
3'-GGAGACTTACAGGAAAGAGATTCAGGATTCAGGAGGCCTACCATGAAGATCAAG-5'

Sequencing Gels: Separate vs Single Lanes

GCCAGGTGAGCCTTTGCA

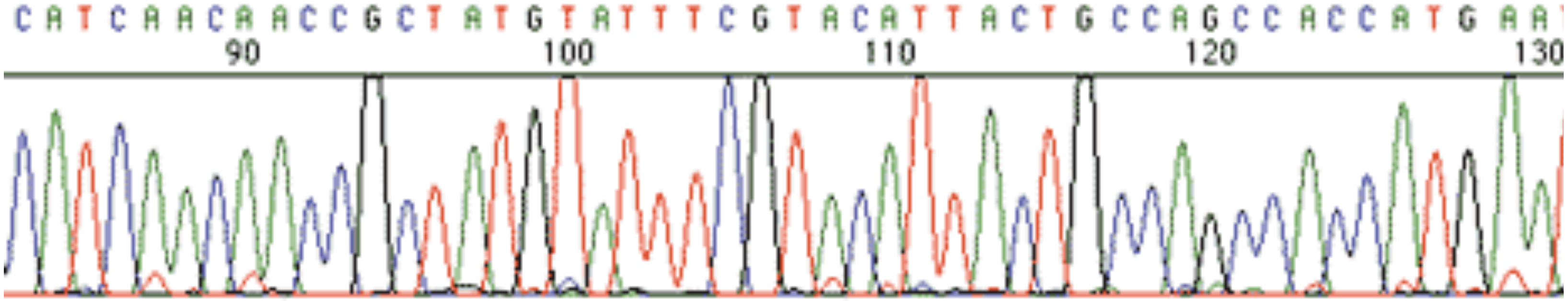




Automated Sequencing Instruments

Sequencing

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



Courtesy: Dr. Kalai Mathee

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.

Courtesy: Dr. Kalai Mathee

Human Genome Project

Play the Sequencing Video:

- Download Windows file from http://www.cs.fiu.edu/~giri/teach/6936/Papers/Sequence.exe
- Then run it on your PC.

Human Genome Project

1980 The sequencing methods were sufficiently developed

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

Estimated expense: $3B and 15 years

Part of this project is to sequence: *E. coli, Sacchromyces cerevisiae, Drosophila melanogaster, Arabidopsis thaliana, Caenorhabdidtis elegans*

- Allow development of the sequencing methods

Got underway in October 1990

Automated sequencing and computerized analysis

Public effort: 150,000 bp fragments into artificial chromosomes (unstable - but progressed)

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 TIGR's discoveries

PE developed the automated sequencer & Venter - Whole-genome short-gun approach

“While the NIH is not very good at funding new ideas, once an idea is established they are extremely good,” Venter

In May 1998, Venter, in collaboration with Michael Hunkapiller at PE Biosystems (aka Perkin Elmer / Applied Biosystems / Applera), formed Celera Genomics

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

April 6, 2000 - Celera announces the completion “Cracks the human code”

Agrees to wait for HGP

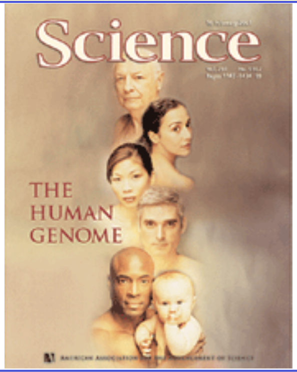
Summer 2000 - both groups announced the rough draft is ready

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

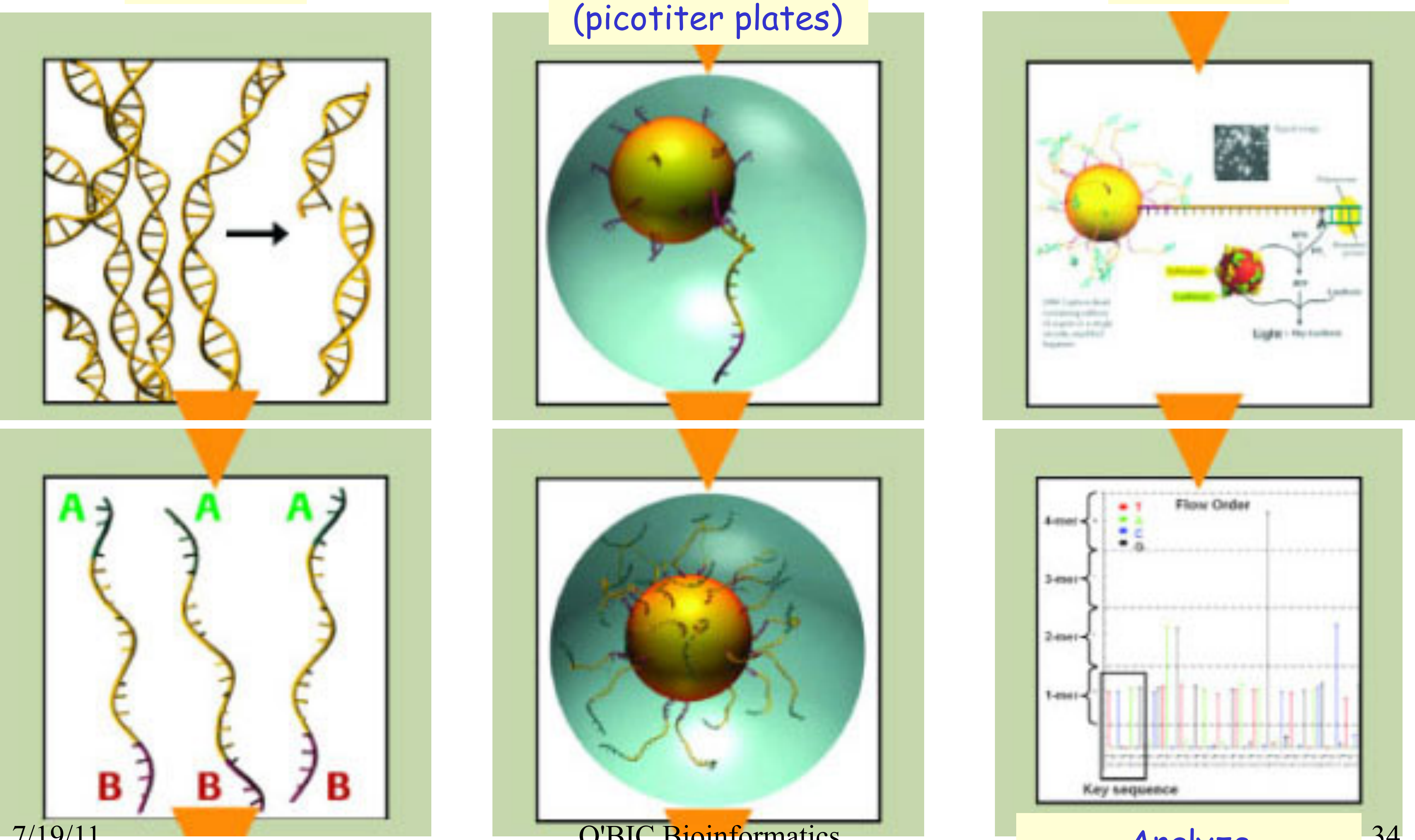
❑ Sanger Method (70Kbp/run)

❑ 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

emPCR on bead

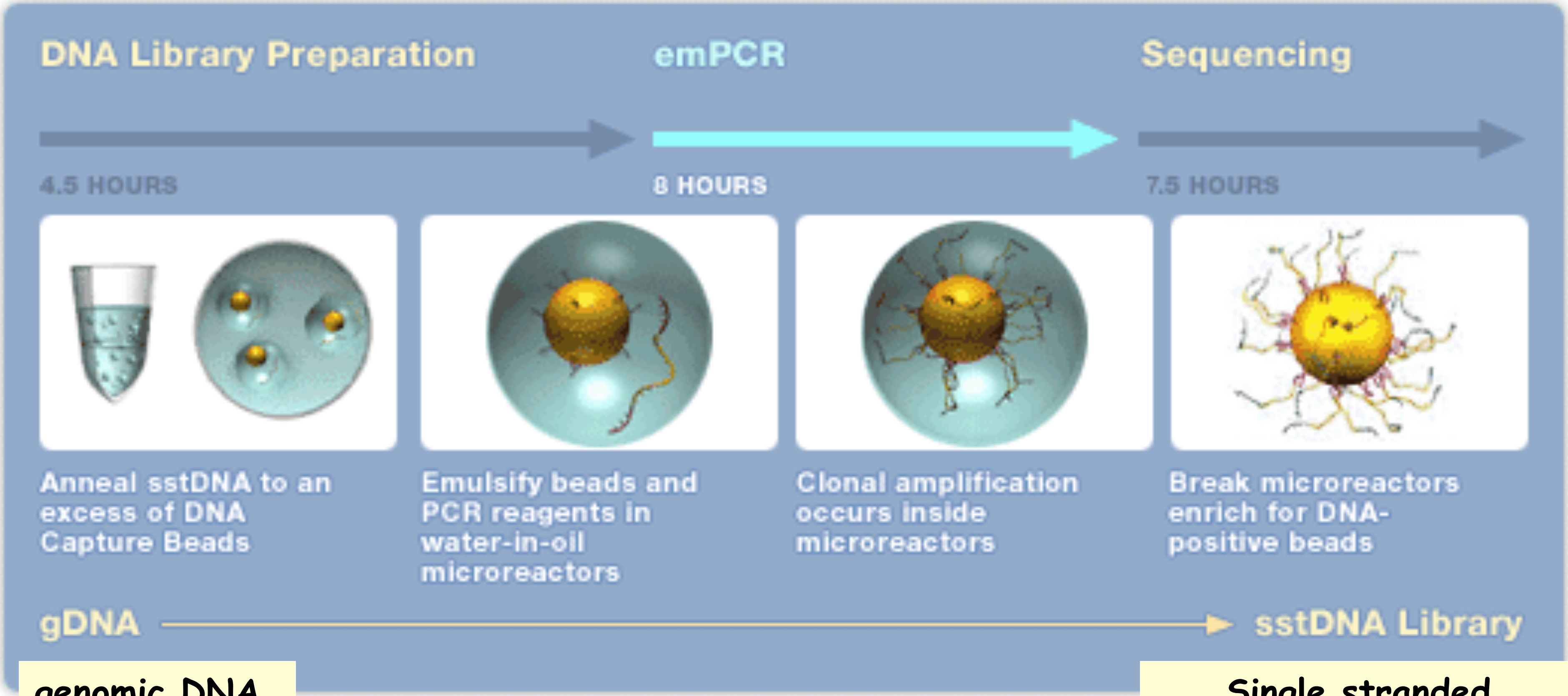
1 fragment-1 bead (picotiter plates)

Sequence

Analyze
one bead - one read

emPCR

FIGURE 8

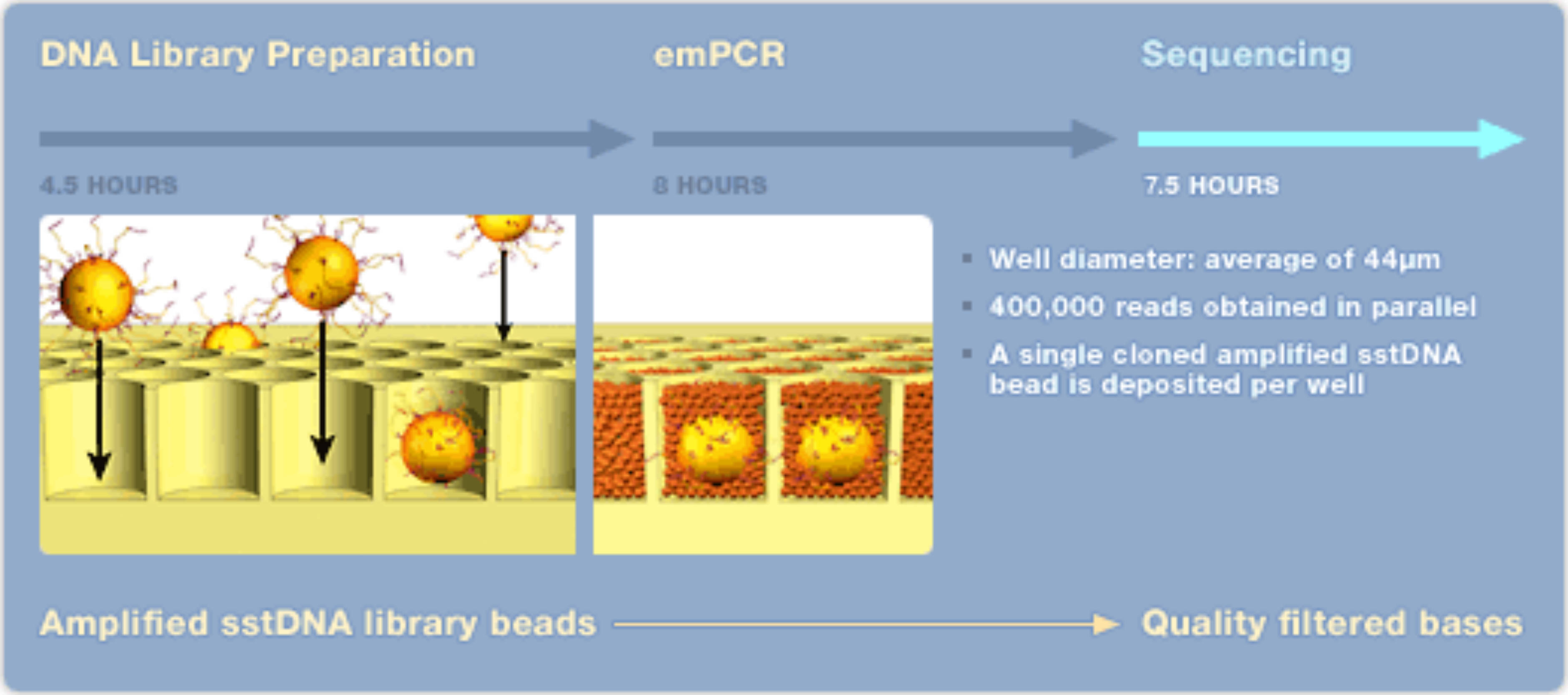

DNA Library Preparation
emPCR
Sequencing
4.5 HOURS
8 HOURS
7.5 HOURS
Anneal sstDNA to an excess of DNA Capture Beads
Emulsify beads and PCR reagents in water-in-oil microreactors
Clonal amplification occurs inside microreactors
Break microreactors enrich for DNA-positive beads
gDNA
sstDNA Library

genomic DNA

Single stranded template DNA library

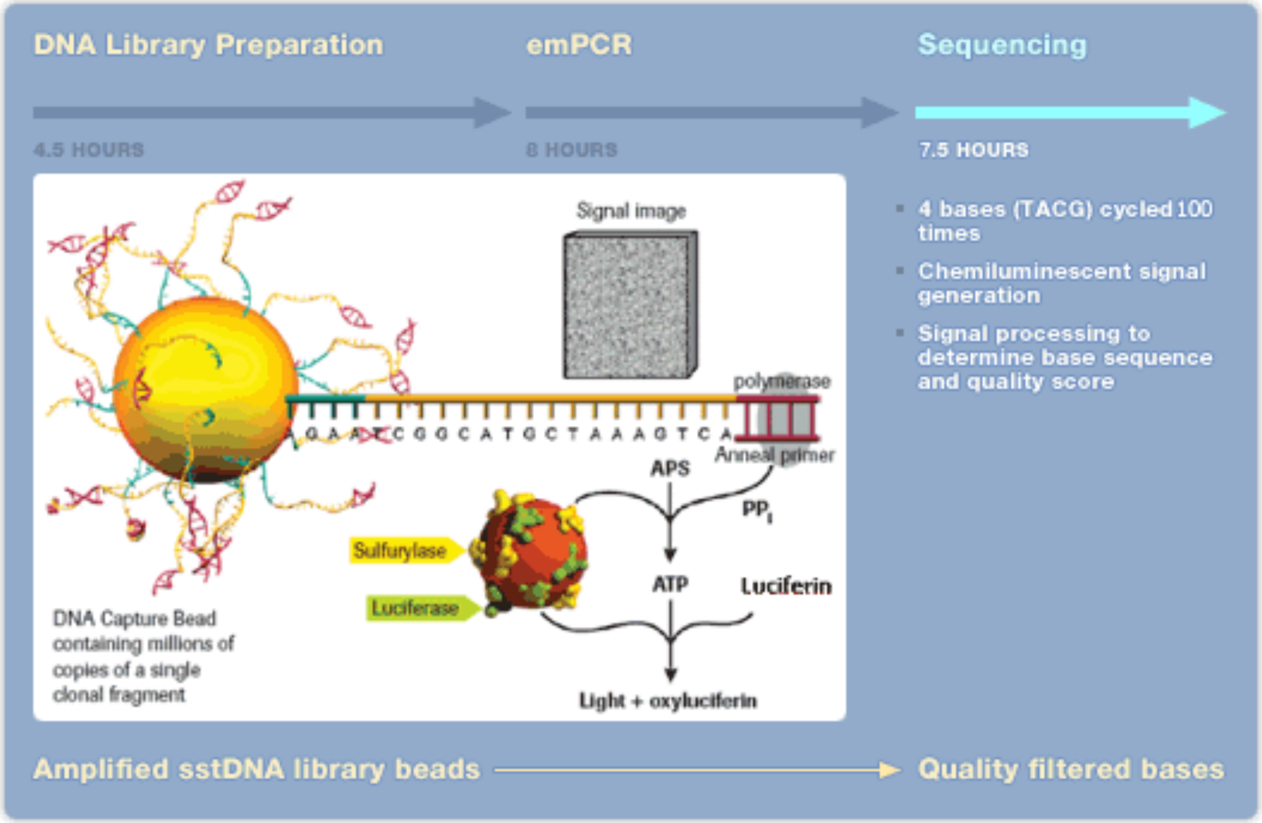
Sequencing

FIGURE 9



- Well diameter: average of 44µm
- 400,000 reads obtained in parallel
- A single cloned amplified sstDNA bead is deposited per well

Sequencing



- 4 bases (TACG) cycled 100 times
- Chemiluminescent signal generation
- Signal processing to determine base sequence and quality score

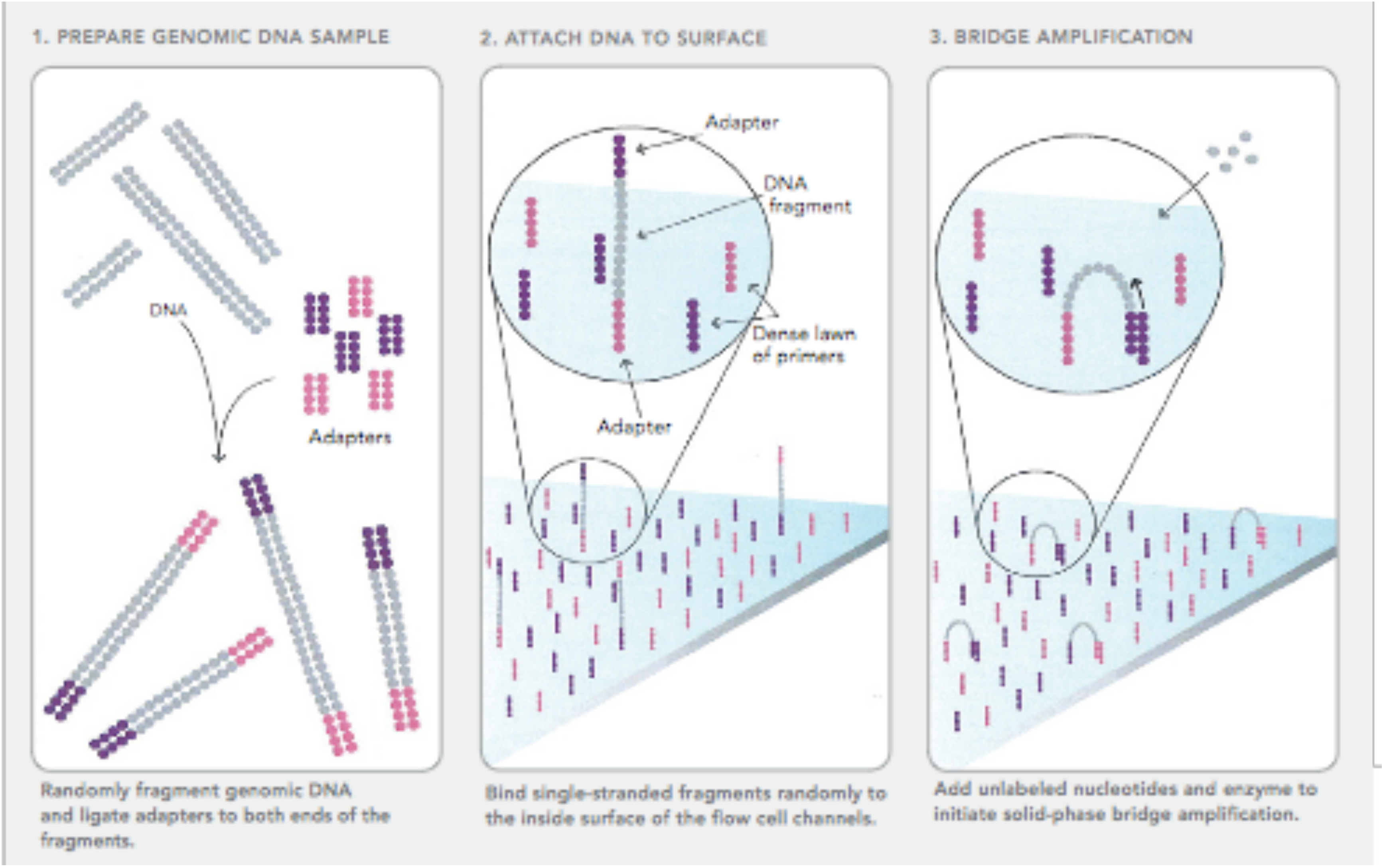
- Hundreds of thousands of beads each carrying millions of copies of unique ssDNA molecule sequenced in parallel
- Sequential flow of nt in fixed order across PicoTiterPlate

- 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



1. PREPARE GENOMIC DNA SAMPLE

Randomly fragment genomic DNA and ligate adapters to both ends of the fragments.

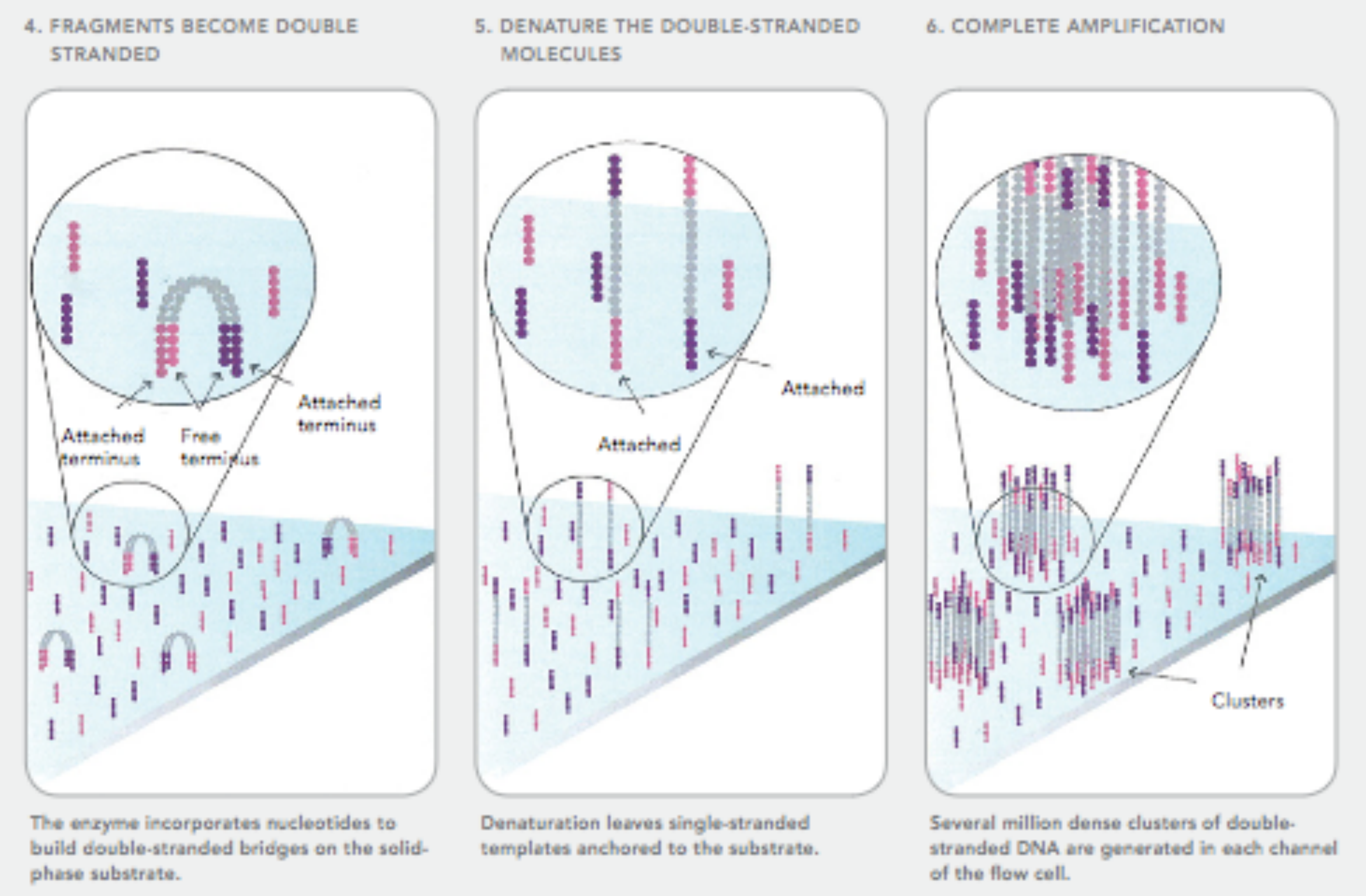
2. ATTACH DNA TO SURFACE

Bind single-stranded fragments randomly to the inside surface of the flow cell channels.

3. BRIDGE AMPLIFICATION

Add unlabeled nucleotides and enzyme to initiate solid-phase bridge amplification.

Solexa Sequencing



4. FRAGMENTS BECOME DOUBLE STRANDED

The enzyme incorporates nucleotides to build double-stranded bridges on the solid-phase substrate.

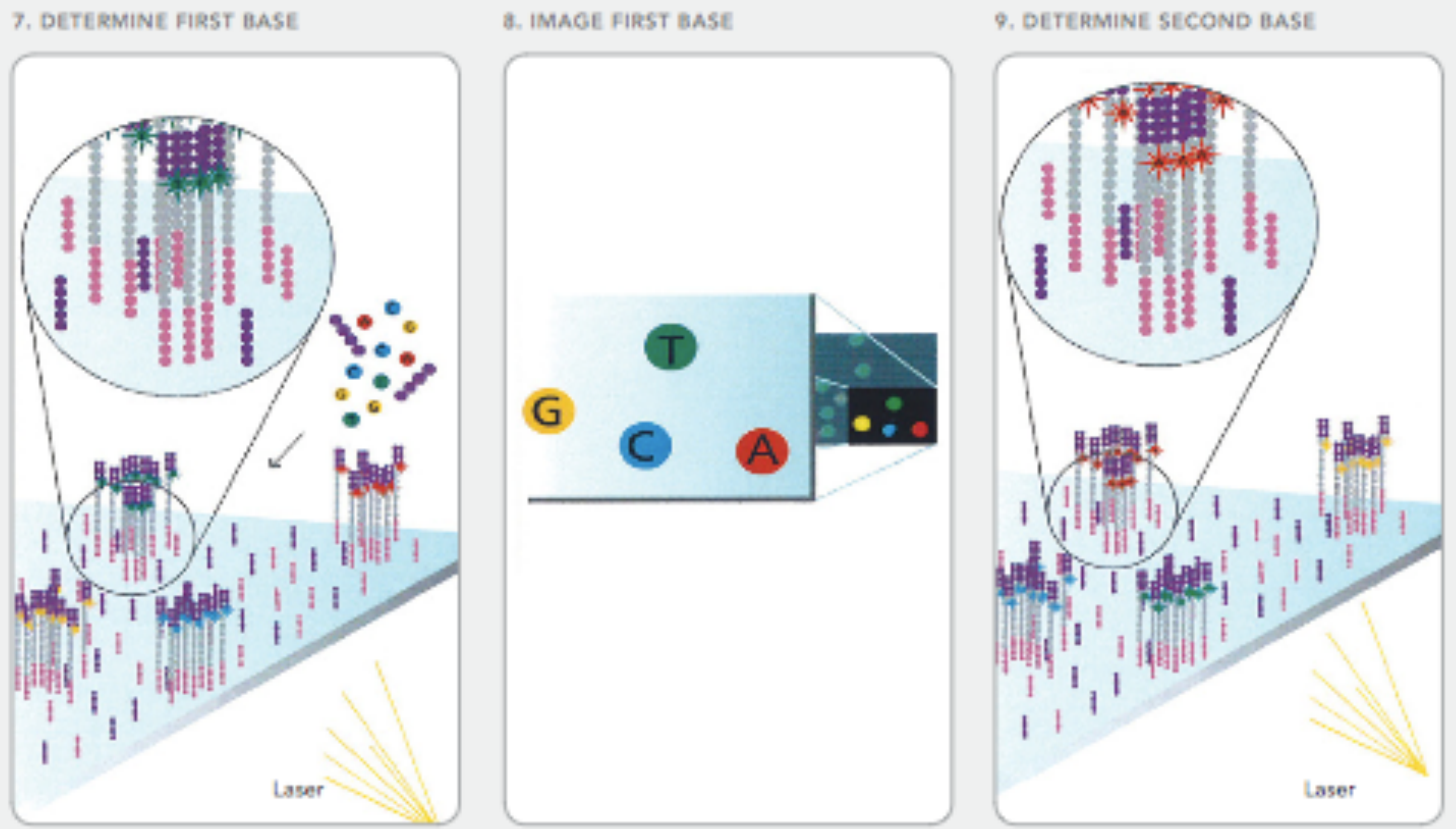
5. DENATURE THE DOUBLE-STRANDED MOLECULES

Denaturation leaves single-stranded templates anchored to the substrate.

6. COMPLETE AMPLIFICATION

Several million dense clusters of double-stranded DNA are generated in each channel of the flow cell.

Solexa Sequencing



7. DETERMINE FIRST BASE

First chemistry cycle: to initiate the first sequencing cycle, add all four labeled reversible terminators, primers and DNA polymerase enzyme to the flow cell.

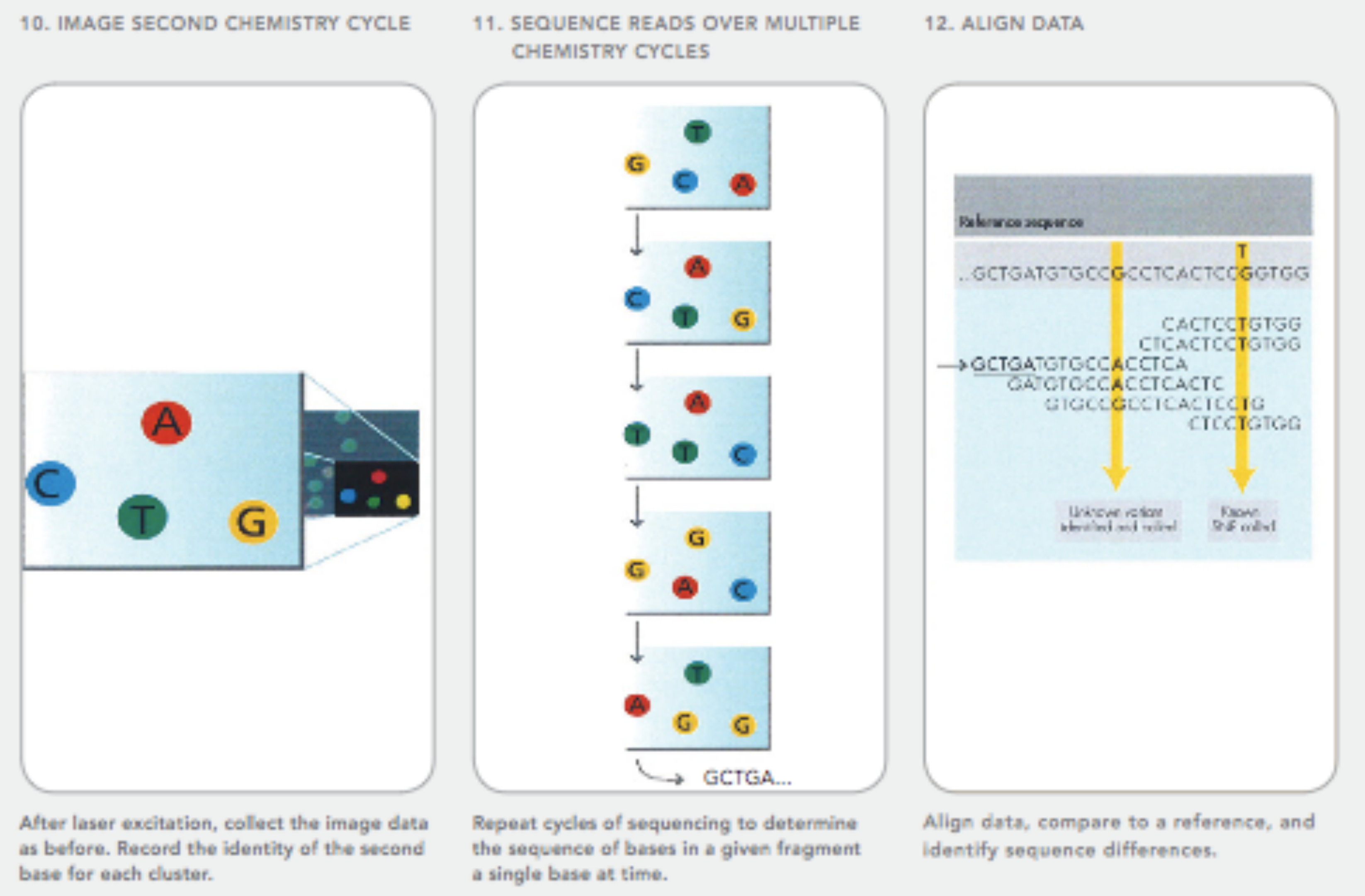
8. IMAGE FIRST BASE

After laser excitation, capture the image of emitted fluorescence from each cluster on the flow cell. Record the identity of the first base for each cluster.

9. DETERMINE SECOND BASE

Second chemistry cycle: to initiate the next sequencing cycle, add all four labeled reversible terminators and enzyme to the flow cell.

Solexa Sequencing



10. IMAGE SECOND CHEMISTRY CYCLE

11. SEQUENCE READS OVER MULTIPLE CHEMISTRY CYCLES

12. ALIGN DATA

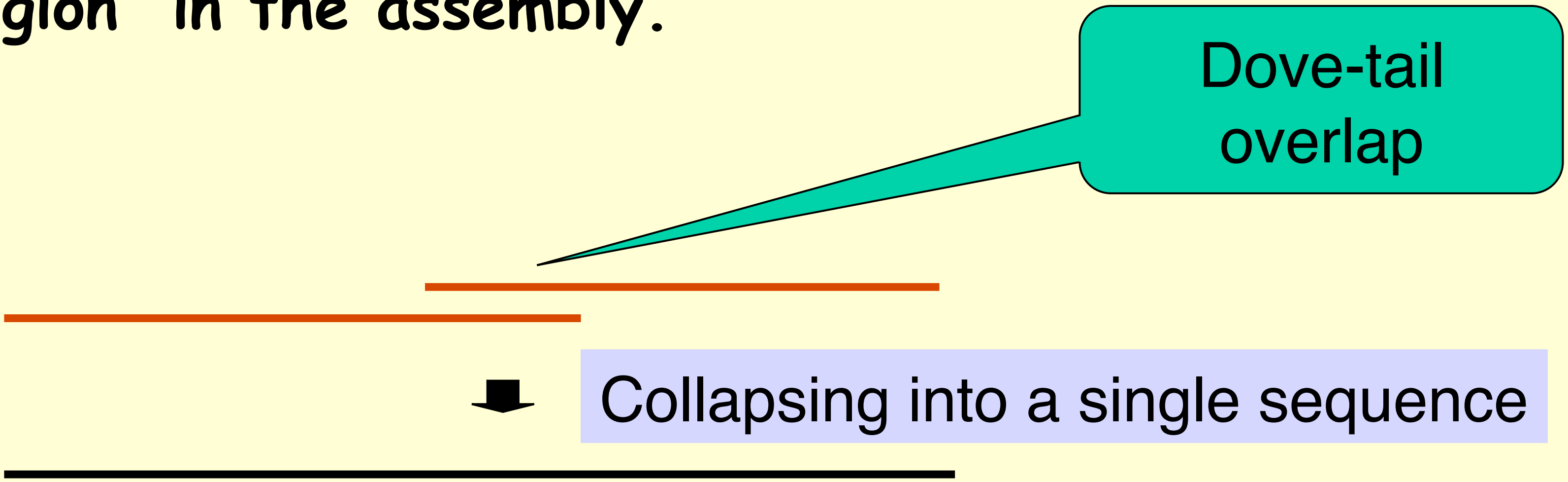
After laser excitation, collect the image data as before. Record the identity of the second base for each cluster.

Repeat cycles of sequencing to determine the sequence of bases in a given fragment a single base at time.

Align data, compare to a reference, and identify sequence differences.

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

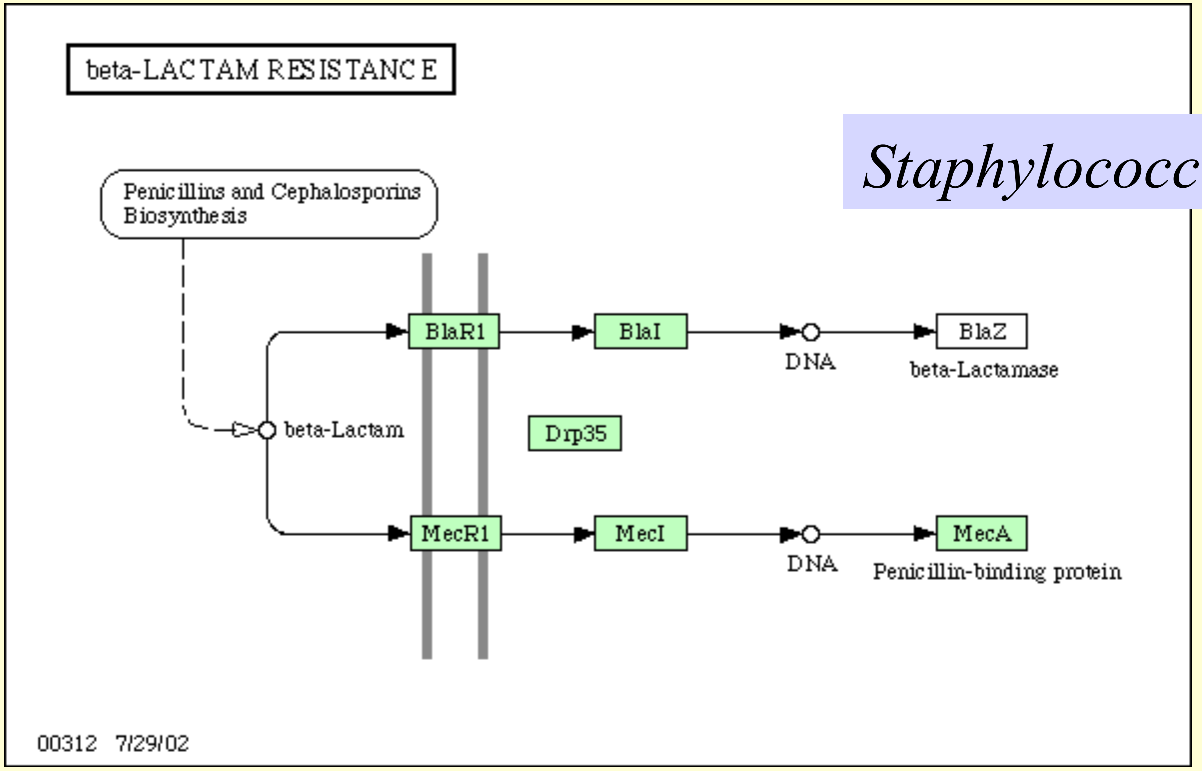
- True Single Molecule Sequencing
- DNA is fragmented and polyA added to end and fluorescent tag added
- DNA hybridized to flow cell with polyT immobilized on it
- Templates packed very closely
- Sequence extension happens one base at a time and a CCD camera takes pictures to produce images after each 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

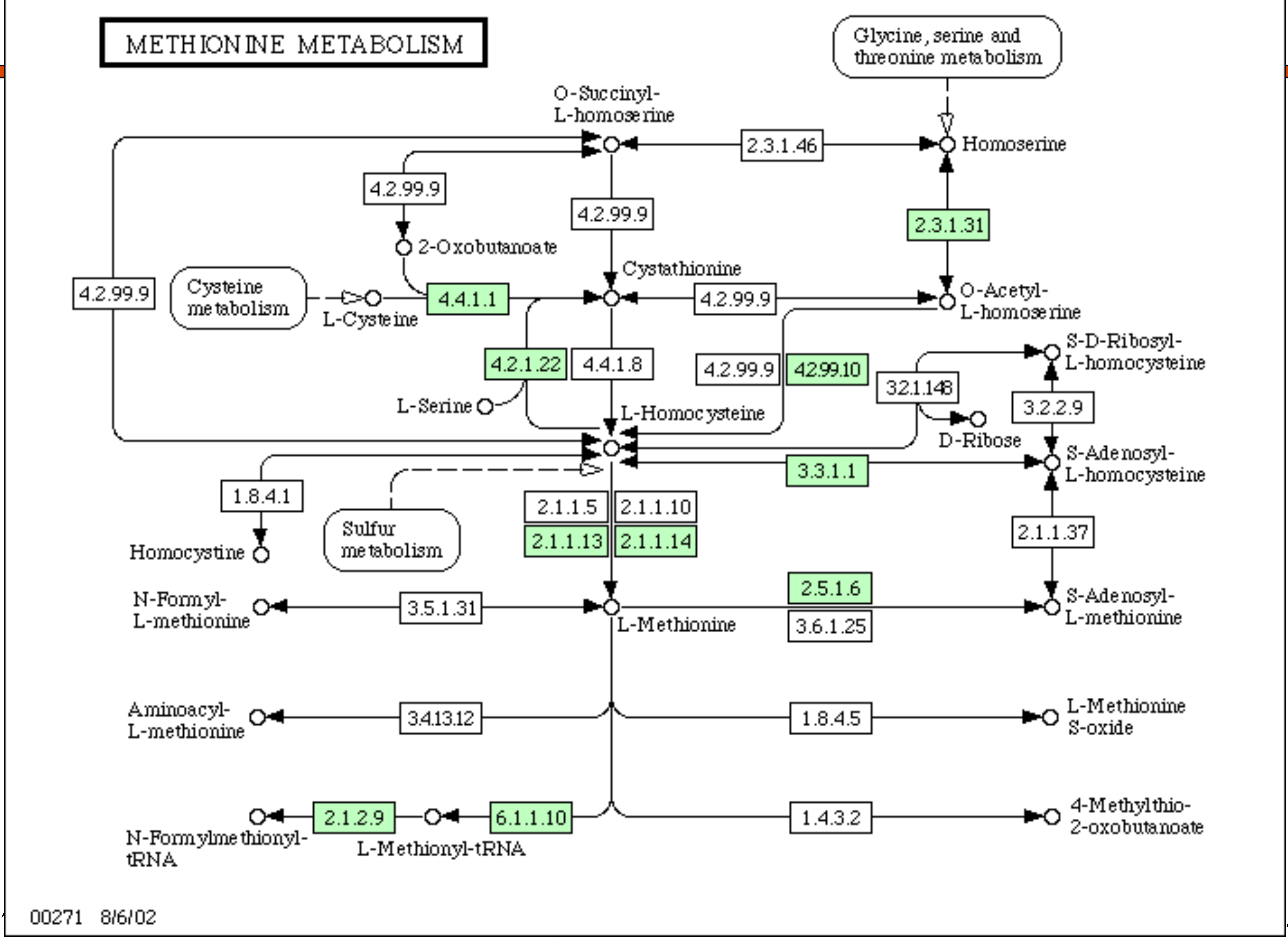
Gene Networks & Pathways

- 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