Lars Paulin

# New DNA sequencing technologies

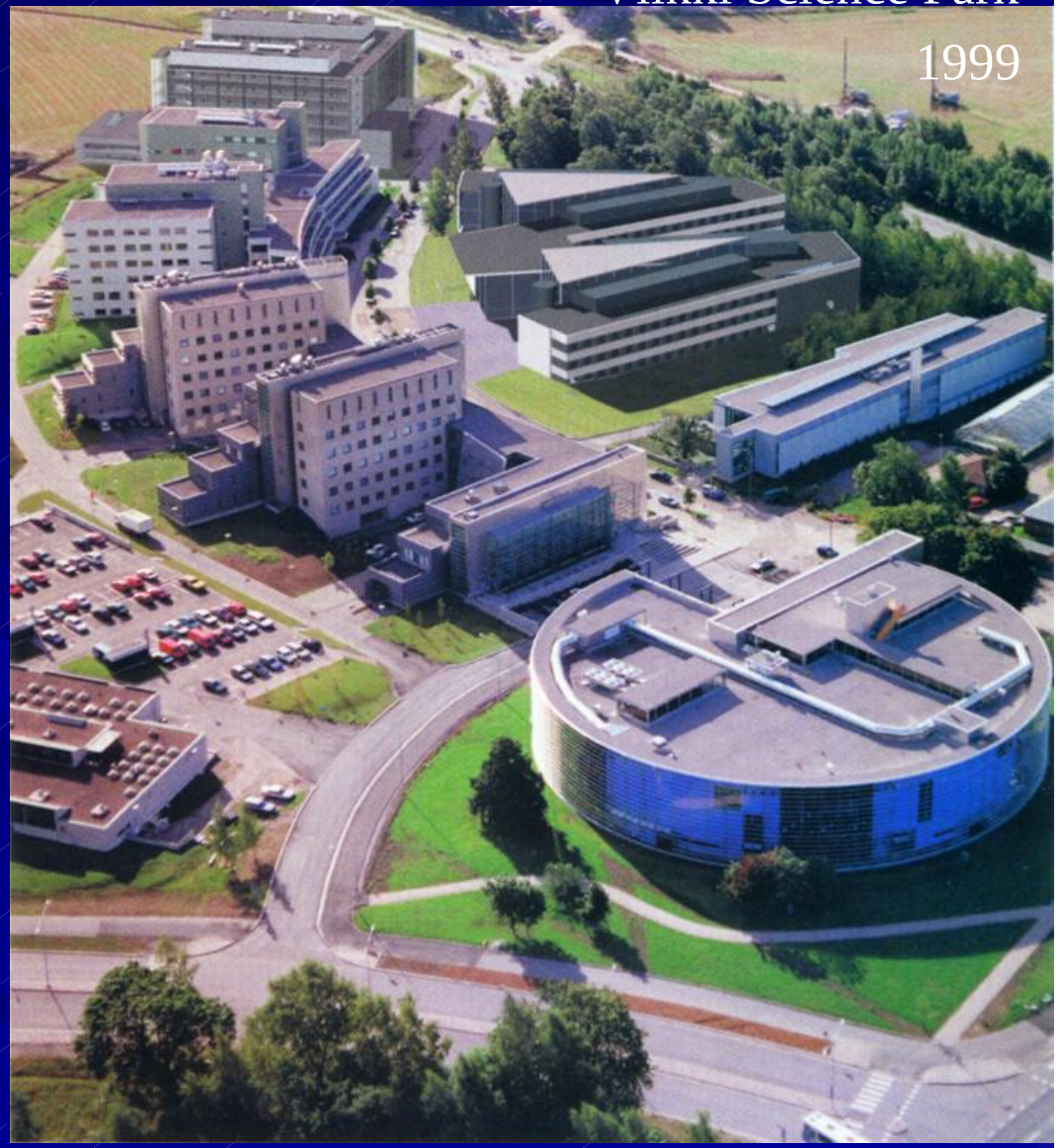
DNA Sequencing and Genomics

Institute of Biotechnology
University of Helsinki

http://www.biocenter.helsinki.fi/bi/DNA/

Viikki Science Park

1999



# DNA Sequencing Laboratory

## Cultivator 2, Viikinkaari 4

- Started in 1990 with DNA Synthesis
- 1991 DNA Sequencing
- 1994 EU Yeast Genome Project
- 1999 - 2000 High-throughput pipeline
- 1999 – 2002 Five EST Sequencing Projects
- 2003 First Microbe Genome Project
  - Move together with Microarray Laboratory to Cultivator 2
- 2006 Genome Sequencer 20

- Core Facility
  - Service DNA sequencing and whole projects
  - Collaborative projects
    - ”Research hotel”
  - Develope high-throughput methods
  - Consulting

# DNA Sequencing and Microarray Pipe-Line

- Bacterial growth
  - QPix Colony Picker
  - QFill 2 Microplate filler
  - 4 Multidrop 96 and 384 well
  - 4 Titramax shakers
- PCR
  - 1 Tetrad MjResearch [1 x 96, 3 x (2x48)]
  - 2 Tetrad MjResearch ( 4 x 96 )
  - 1 Eppendorf Mastercycler (384)
  - 1 ABI 9700 ( 2 x 384 )
  - 4 ABI 9700 ( 96 )
- Robotics
  - Qiagen Biorobot 9600
  - Qiagen Biorobot 8000
  - Qiagen Biorobot 3000 ( 2m )
  - 2 Tecan Genesis RSP100
    - low volume pipetting
  - Tecan Freedom Evo
    - low volume pipetting
  - Biomek NX$^{p}$ , 96-pipettor
- Sequencers
  - ABI 3130 16-Capillary Sequencer
  - ABI 3730 48-Capillary Sequencer
  - Genome Sequencer 20
- Centrifuges
  - 2 Eppendorf 5810
  - 1 Eppendorf 5804
  - 1 Beckman
- Other
  - Tecan SpectraFluor Microplate Reader
  - ImageMaster VDS, GE Healthcare
  - 8 Servers for analysis
    - Phred, Staden Program, AceDB, ARB etc.
  - SUN cluster
  - ABI Prism 7000
  - LightCycler 480, Roche
  - Coulter Counter Z2
  - Bioanalyzer
  - LIMS, Sapphire

Lars Paulin Institute of Biotechnology University of Helsinki

# What can we do?

## All kinds of service

- Sequencing projects
- Fragment analysis
  - T-RLFP
- SNP detection
- Colony picking
- Plasmid, cosmid, fosmid, BAC purification
- PCR, PCR purification
- rDNA sequencing
- EST, SAGE sequencing
- Whole genome sequencing
- Metagenomic sequencing
- MLST-Sequencing
- etc.

## High-throughput

–

- Picking 3 500 colonies / h
  - petri discs, 22 x 22 cm plate
- Growth ~50 x 96 / day
  - 70 µl - 1 ml
- PCR ~5 000 / day
  - 50 - 100 µl
- PCR purification 12 x 96 / day
  - 2 x 100 µl
- Plasmid, Cosmid, Fosmid, BAC purification
  - 96 / 384 well
- Sequencing 0,5 - 1 Mb / 24h
- Gridding 12 membranes / day
  - 1536 spots in duplicate, 384 controls
- Genome Sequencer FLX

# EST Projects 1999 – 2002

Barley
- 44 829 ESTs
- 13 448 clusters
  - 7 062 cluster 1

Birch
Wood, Cold, Oxidative stress, Pathogen
- 73 881 ESTs
- 20 103 clusters
  - 10 554 cluster 1

Gerbera
Genome Res 2005, 15,475-86
- 16 994 ESTs
- 11 266 clusters
  - 8 817 cluster 1


Spruce
Plant Mol Biol 2007, 65, 311-328.
- 8 432 ESTs
- 5 817 clusters
  - 4 343 cluster 1

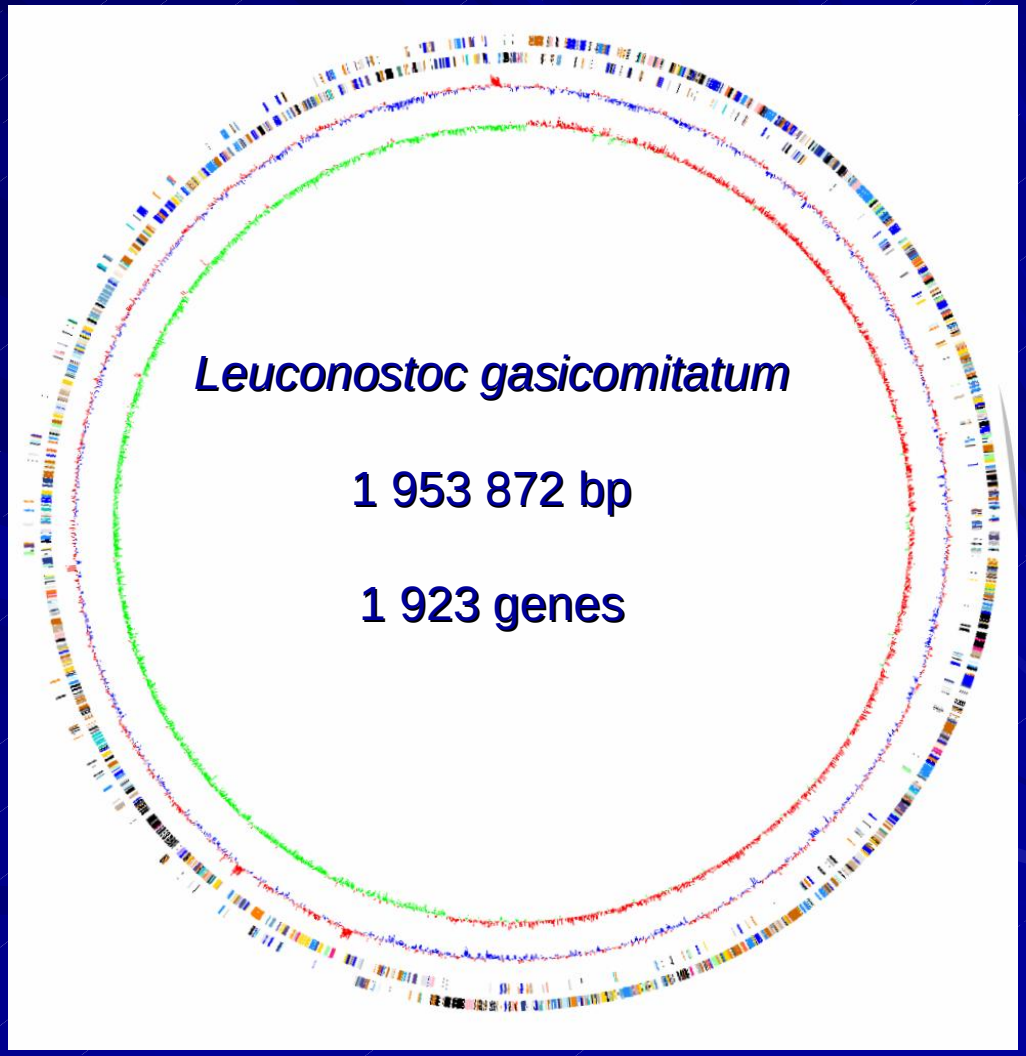
Poplar
Genome Biol, 2005, 6, r101
- 13 845 ESTs
- 8 043 clusters
  - 5 860 cluster 1

Total
- 157 981 ESTs in Databases
- ~260 000 sequencing reactions

# *Leuconostoc gasicomitatum* genome project
## Paulin, Björkroth and Auvinen

- First described by Johanna Björkroth
  - Appl Environ Microbiol. 2000, 66, 3764-72.
- Food spoiling bacteria
  - can grow at + 6 oC
  - marinaded poultry products
  - fish and meat


Björkroth , J.




Lars Paulin Institute of Biotechnology University of Helsinki

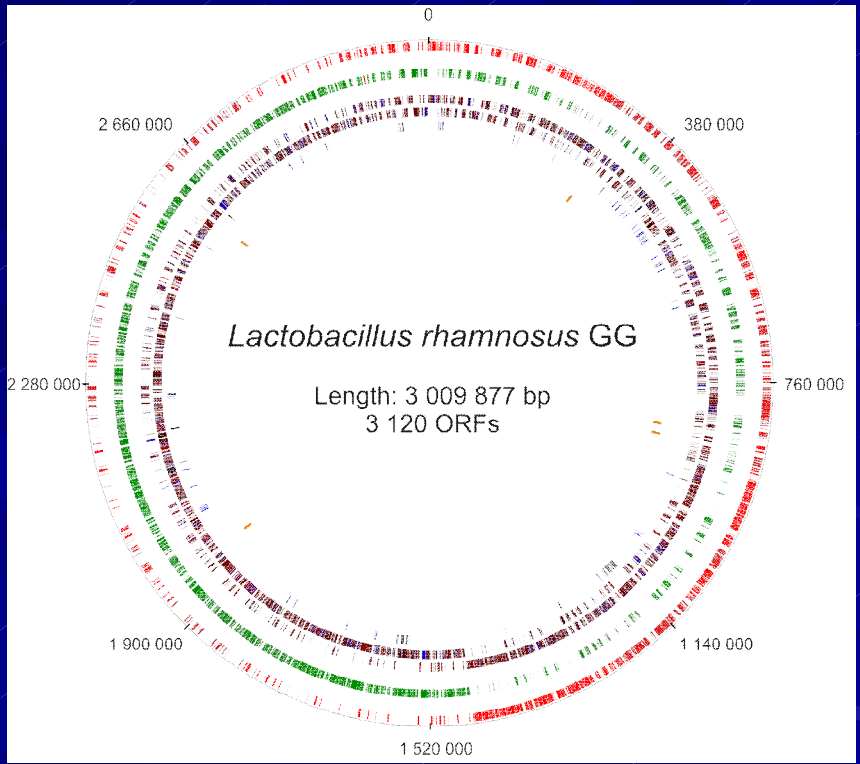
# *Lactobacillus rhamnosus GG*

- Widely used probiote

- Biggest Lactobacillus genome so far
  - 3 009 877 bp
  - 3 120 ORFs
    - 2 129 ORFS with assigned functions
    - 991 ORFs without assigned functions
  - 5 rRNA operons
  - G+C % 46,69
  - Annotation done at ERGO
    http://ergo.integratedgenomics.com/ERGO/
  - Agilent microarrays done

  - The sequence announced at the conference:

    SSA GutImpact 3rd Platform meeting on Foods for Intestinal Health 29.-31.8.2007, Haikko Manor & Spa, Finland



# Genome Projects

*L. gasicomitatum,*1,9 Mb
- Shotgun, fosmids
- Status: finished, annotated

*C. pneumoniae*, 1,3 Mb
- PCR products
- Status: almost finished

*L. rhamnosus* , 3 Mb
- Shotgun, fosmids, 454
- 2 Strains, GG and Lc
- Status: 2 finished, 1 annotated

*L. oligofermentans*, 1,8 Mb
- 454 , fosmids
- Status: closing phase

*C. botulinum*, 3,8 Mb
- 454, fosmids
- Status: closing phase

*H. bizzozeronii*, 1,7 Mb
- 454, plasmids
- Status: closing phase

*H. salomonis*, 1,7 Mb
- 454, plasmids
- Status: closing phase

*E. carotovora*, 5 Mb
- 454, fosmids
- Status: closing phase

*S. islandicus*, 2,7 Mb
- shotgun, 454
- Status: closing phase

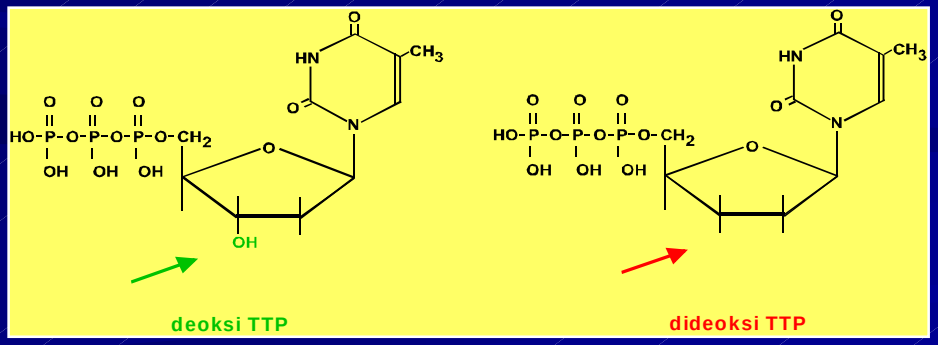
*A. brierleyi*, 2,7 Mb
- Shotgun, 454
- Status: closing phase

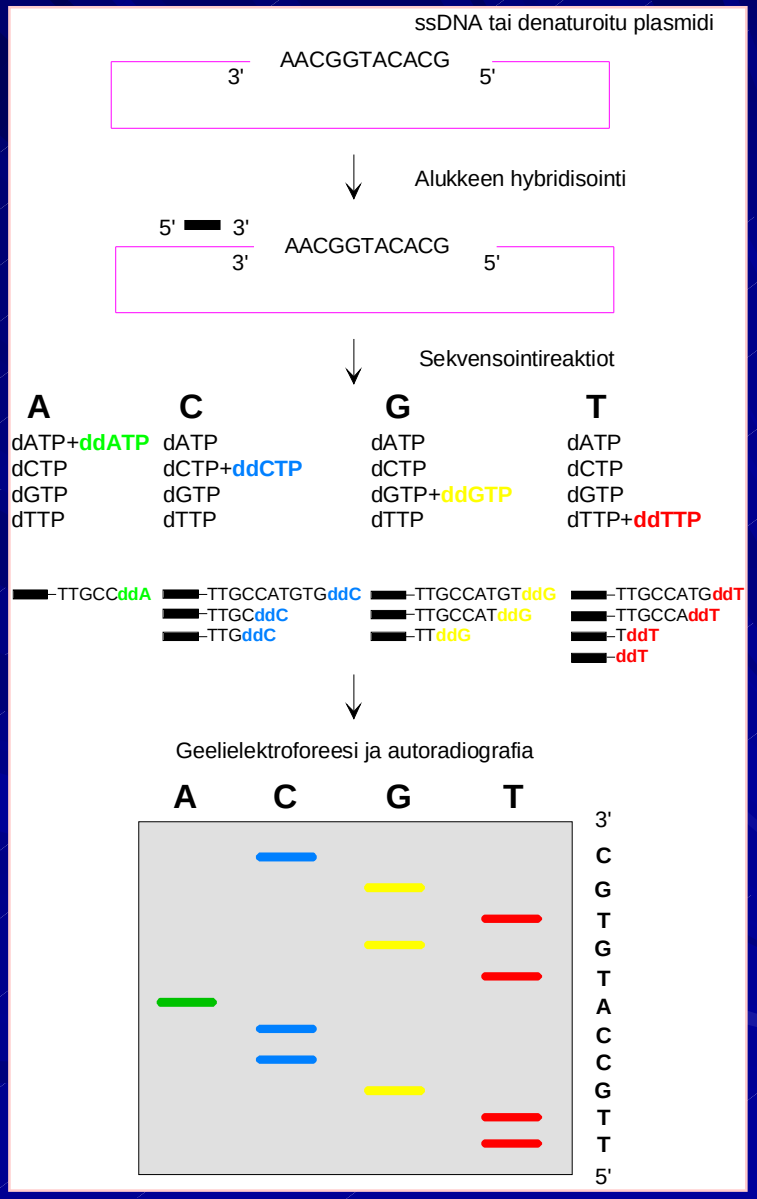
# Short History of DNA Sequencing

- 1977
  - Maxam-Gilbert
  - Sanger
- 1986
  - First Automated DNA Sequencer ABI 370 (373)
- 1988
  - Pharmacia ALF
- 1995
  - ABI 377
    - Up to 96 lanes
- 1996
  - First Capillary DNA Sequencer ABI 310
- 1998
  - First 96 Capillary instruments MegaBace, ABI 3700
- 2000
  - ABI 3100, 16 Capillary
- 2002
  - ABI 3730, 48 or 96 Capillary
- 2005
  - Genome Sequencer GS20
- 2006
  - Solexa (Illumina)
- 2007
  - SOLiD

# Sanger DNA Sequencing

1. Template
   - ssDNA or dsDNA
2. Primer annealing
   - Sequencing primer
3. Elongation
   - DNA polymerase
     - Steps 2 and 3 can be done repeatedly => cycle sequencing
4. Electrophoresis



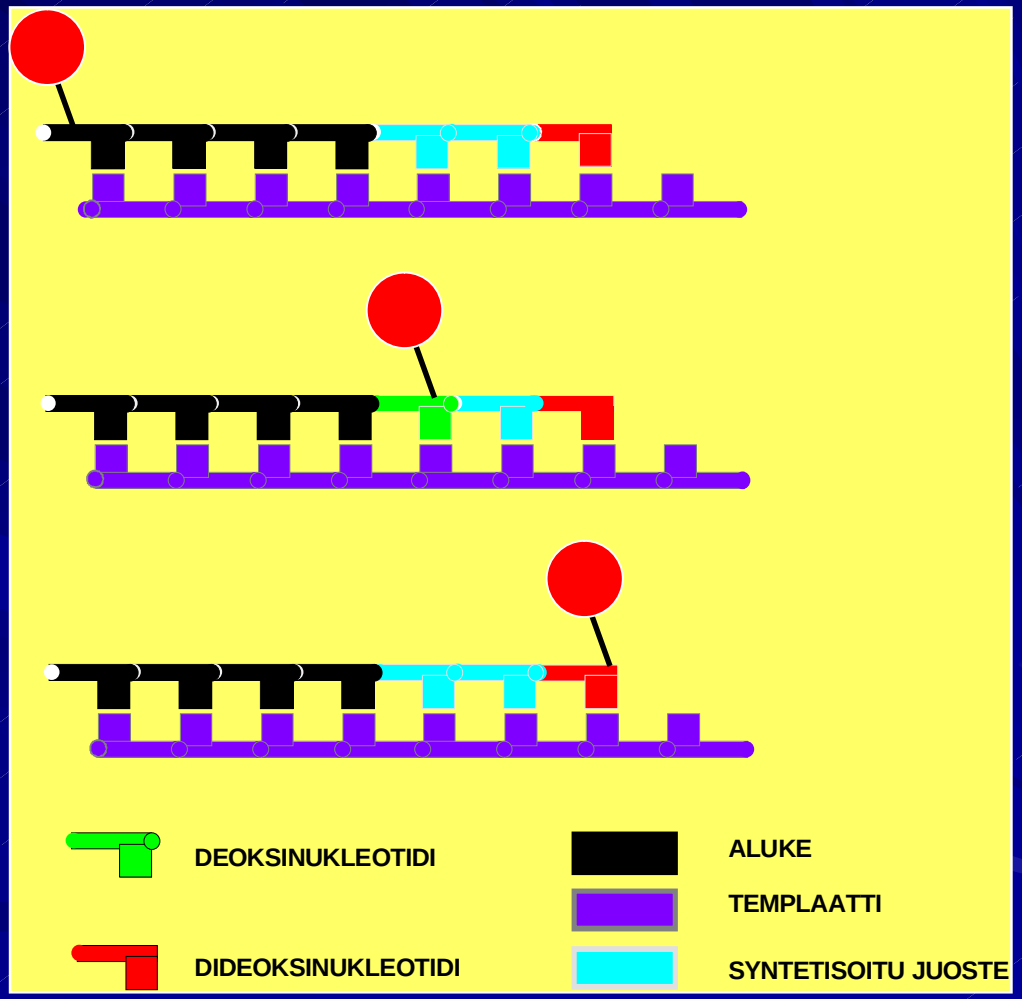

Lars Paulin Institute of Biotechnology University of Helsinki

# Incorporating Labels

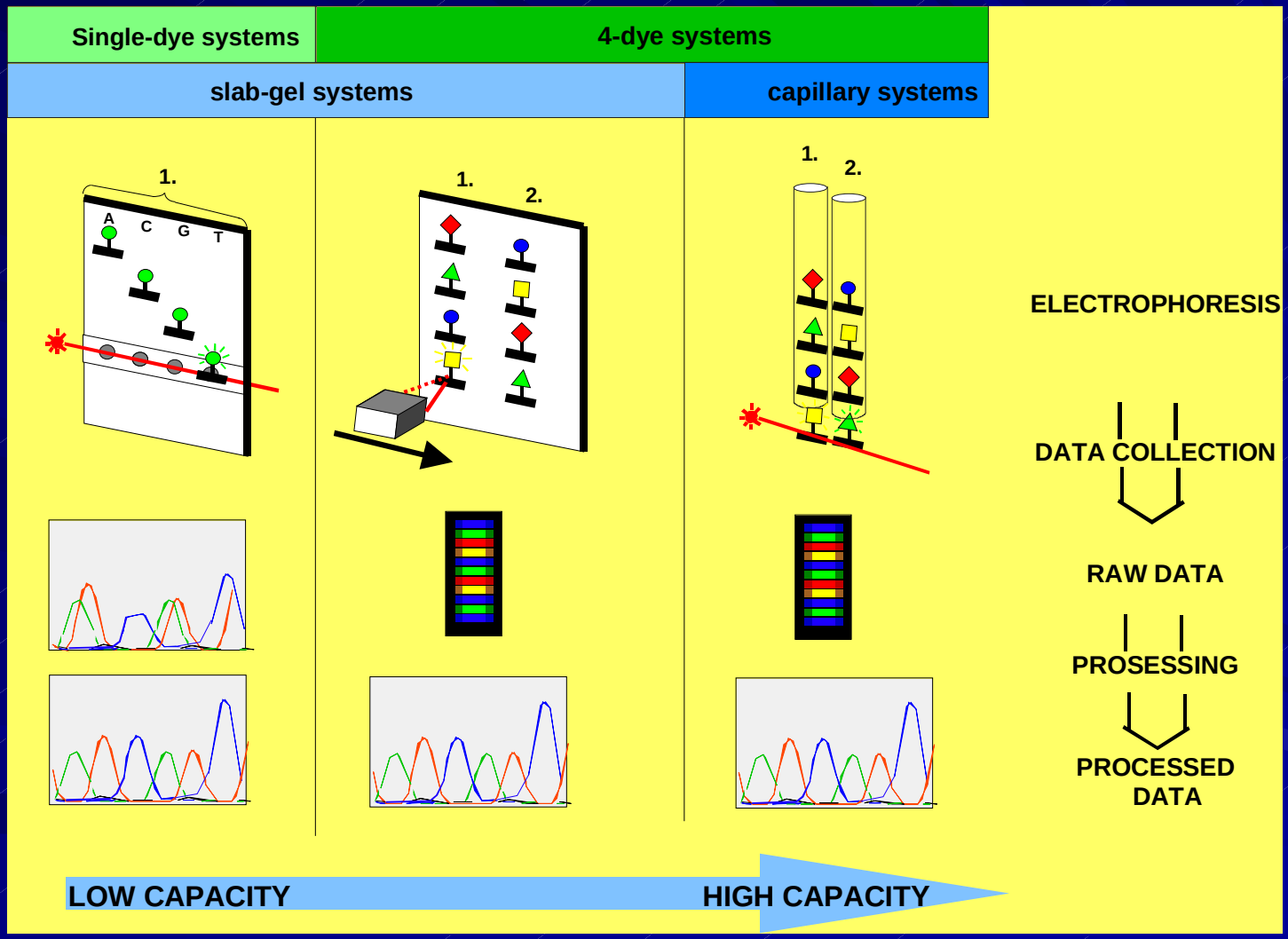
Labelled primers
- 1 or 4 labels

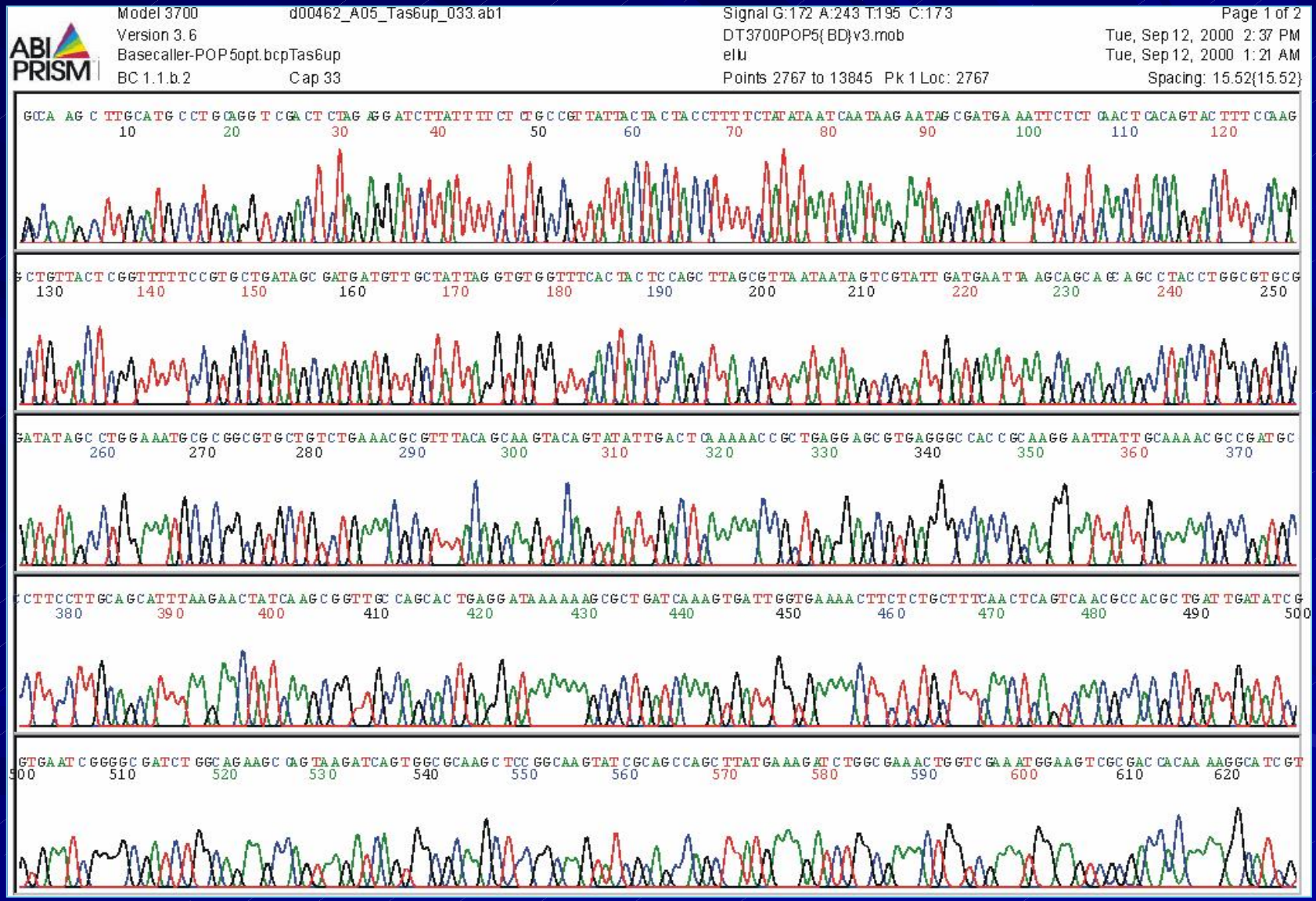
Labelled deoxynucleotides
- 1 label

Labelled dideoxynucleotides
- 1 or 4 labels
- BigDye, ET terminators



# Automated DNA Sequencing




Model 3700
Version 3.6
Basecaller-POP5opt.bcpTas6up
BC 1.1.b.2
d00462_A05_Tas6up_033.ab1
Cap 33
Signal G:172 A:243 T:195 C:173
DT3700POP5{BD}v3.mob
elu
Points 2767 to 13845 Pk 1 Loc: 2767
Page 1 of 2
Tue, Sep 12, 2000 2:37 PM
Tue, Sep 12, 2000 1:21 AM
Spacing: 15.52{15.52}

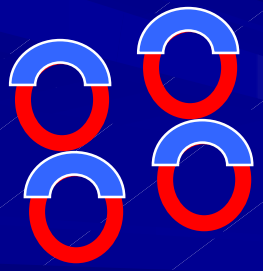
# Strategies for Genome Sequencing

- Shotgun approach
  - random sequencing of different sized libraries
  - assembly using different software
  - closing of gaps using different methods
- Libraries
  - usually made by random shearing of genomic DNA
  - 2 kb, 4-6 kb, 10 kb plasmid libraries
  - fosmid or cosmid libraries with 30 - 50 kb inserts

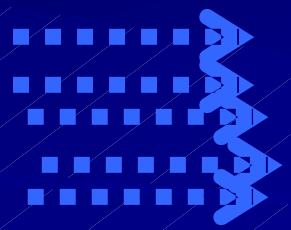
# Whole Genome Shotgun Sequencing

Whole Genome:
~ 3 Mb

Sheared DNA:
~ 2 kb

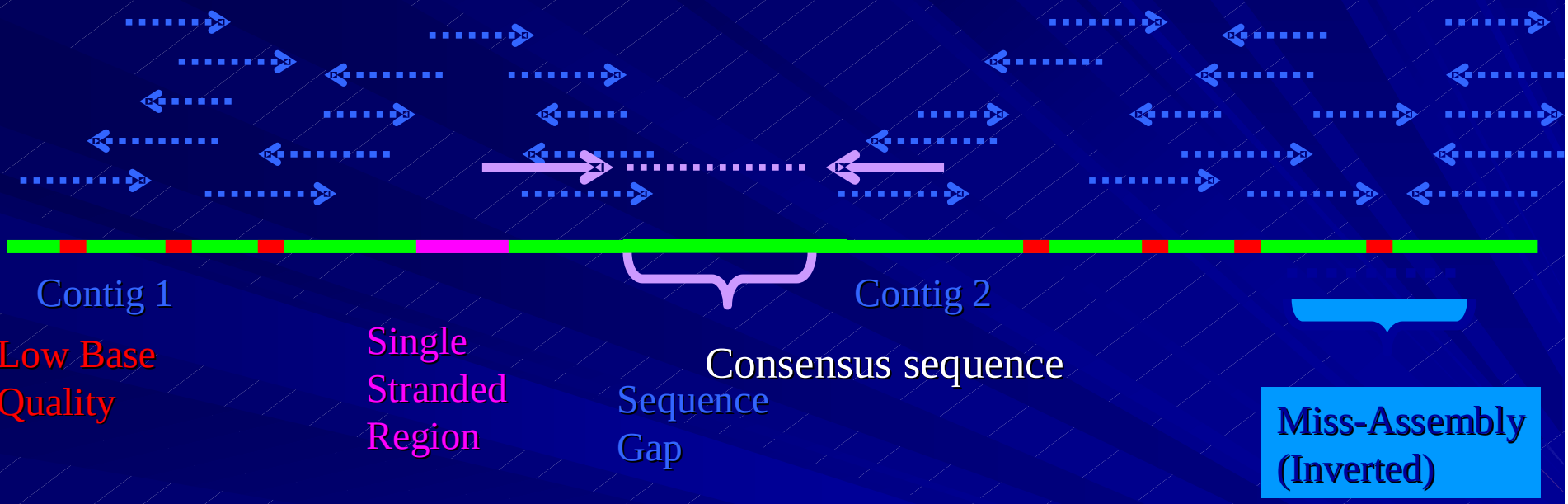


Sequencing
Templates



Random
Reads
Both ends

# Shotgun Sequencing :ASSEMBLY




- 0.5 -1.0 X (2 reads/kb) - 'Skimming'
- 3.5 - 4.0 X (~9 reads/kb) -'half-shotgun'
- 6.5 - 8.0 X (~18 reads/kb) - 'pre-finished'
- 10 X (22-24 reads/kb) - 'deep shotgun'

# Phred, Phrap and Staden Package Program

## Phred and Phrap

- University of Washington
- Phil Green, http://www.phrap.org/

### Phred quality score:

$QV = -10 * \log_{10}( P_e )$

where $P_e$ is the probability that the base call is an error.

| Phred score | $P_e$ | Accuracy of the base call |
|---|---|---|
| 10 | 1 in 10 | 90% |
| 20 | 1 in 100 | 99% |
| 30 | 1 in 1,000 | 99.9% |
| 40 | 1 in 10,000 | 99.99% |
| 50 | 1 in 100,000 | 99.999% |

## Staden Program

- Cambridge, Sanger Center
- Roger Staden, http://staden.sourceforge.net/
- Trace editing
- Phrap assembly and Gap4 editing
  - display of traces from sequencers
  - translations, orfs, RE etc.
  - good capacity

# New DNA Sequencing Technology

Parallel Sequencing Technology

- Massive throughput
- Fast sequencing
- No cloning step
- PCR

- Currently three systems ready
  - Genome Sequencer (http://www.454.com/,http://www.roche.com)
    - 454 Life Sciences, Roche
    - Launched in October 2005
  - Solexa (http://www.illumina.com)
    - Illumina
    - Launched 2006
  - SOLiD (http://www.appliedbiosystems.com)
    - Applied Biosystems
    - Launched in October 2007

nature Vol 437|15 September 2005|doi:10.1038/nature03959

ARTICLES

# Genome sequencing in microfabricated high-density picolitre reactors

Marcel Margulies[1]*, Michael Egholm[1]*, William E. Altman[1], Said Attiya[1], Joel S. Bader[1], Lisa A. Bemben[1], Jan Berka[1], Michael S. Braverman[1], Yi-Ju Chen[1], Zhoutao Chen[1], Scott B. Dewell[1], Lei Du[1], Joseph M. Fierro[1], Xavier V. Gomes[1], Brian C. Godwin[1], Wen He[1], Scott Helgesen[1], Chun He Ho[1], Gerard P. Irzyk[1], Szilveszter C. Jando[1], Maria L. I. Alenquer[1], Thomas P. Jarvie[1], Kshama B. Jirage[1], Jong-Bum Kim[1], James R. Knight[1], Janna R. Lanza[1], John H. Leamon[1], Steven M. Lefkowitz[1], Ming Lei[1], Jing Li[1], Kenton L. Lohman[1], Hong Lu[1], Vinod B. Makhijani[1], Keith E. McDade[1], Michael P. McKenna[1], Eugene W. Myers[2], Elizabeth Nickerson[1], John R. Nobile[1], Ramona Plant[1], Bernard P. Puc[1], Michael T. Ronan[1], George T. Roth[1], Gary J. Sarkis[1], Jan Fredrik Simons[1], John W. Simpson[1], Maithreyan Srinivasan[1], Karrie R. Tartaro[1], Alexander Tomasz[3], Kari A. Vogt[1], Greg A. Volkmer[1], Shally H. Wang[1], Yong Wang[1], Michael P. Weiner[4], Pengguang Yu[1], Richard F. Begley[1] & Jonathan M. Rothberg[1]

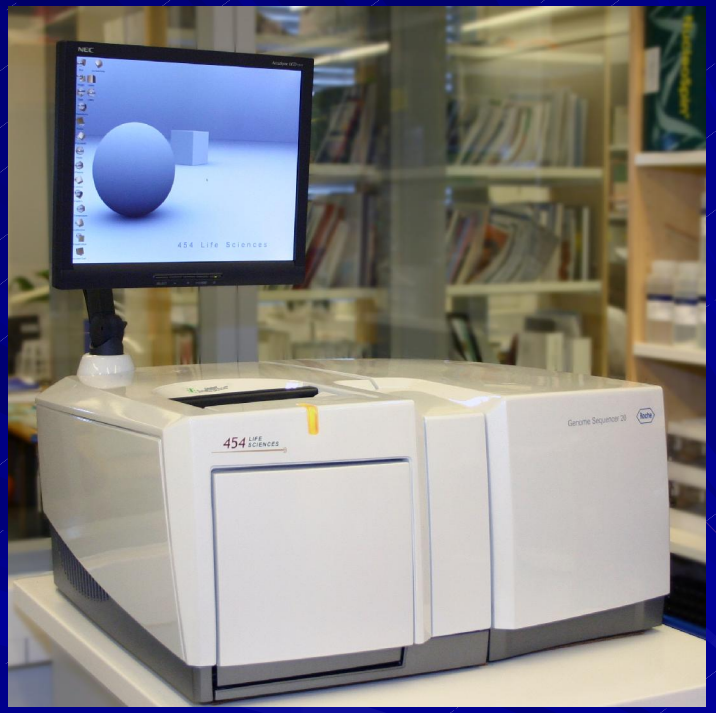
[1]454 Life Sciences Corp., 20 Commercial Street, Branford, Connecticut 06405, USA. [2]University of California, Berkeley, California 94720, USA. [3]Laboratory of Microbiology, The Rockefeller University, New York, New York 10021, USA. [4]The Rothberg Institute for Childhood Diseases, 530 Whitfield Street, Guilford, Connecticut 06437, USA.
*These authors contributed equally to this work.

# Genome Sequencer

(http://www.454.com/,http://www.roche.com)

- Genome Sequencer GS20;FLX
  - Manufacturer 454 Life Science
  - Marketing Roche
- Parallel Sequencing
  - Shotgun sequencing
    - No plasmid libraries
    - Linkers ligated to fragments
    - Emulsion PCR
    - Picotiter plate, 1 600 000 wells
  - Pyrosequencing

    (Nyren, P. et al Anal Biochem. 1993, 208,171-5)
    - Detection with sensitive CCD camera
    - Run time ca. 4,5 h; 7,5 h
    - Read lenght 100 -120 bp; 250 – 300 bp
    - Raw sequence ca. 25 – 35 Mb/run; 80 – 100 Mb/run



# Genome Sequencer GS 20/FLX


### DNA Library Preparation

1. DNA fragmentation
2. Fragment end polishing
3. Adaptor ligation
4. Library immobilisation
5. Fill-in reaction
6. Single-stranded template DNA (sstDNA) library isolation
7. sstDNA library quality assessment and quantitation

### Emulsion PCR Amplification

1. Preparation of the live amplification mix
2. sstDNA library capture
3. Emulsification
4. Amplification
5. Bead recovery
6. sstDNA library bead enrichment
7. Sequencing primer annealing

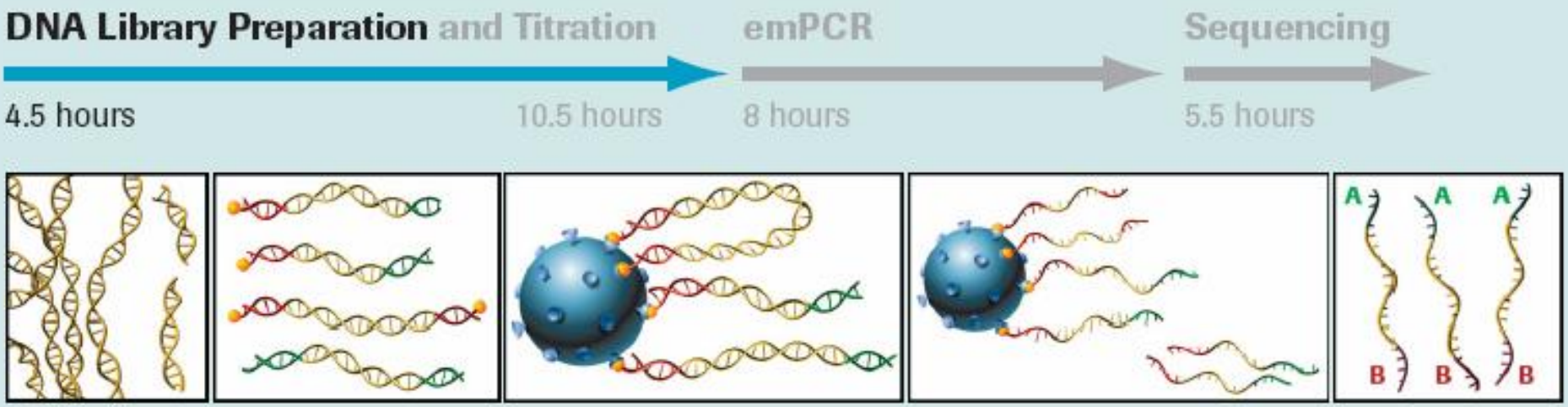
### Sequencing/ Genome Sequencer 20 Operation

1. The pre-wash Run
2. PicoTiterPlate™ preparation
3. The PREP Run
4. The Sequencing Run

### Output

1. FASTA file
2. Assembly
3. Mapping

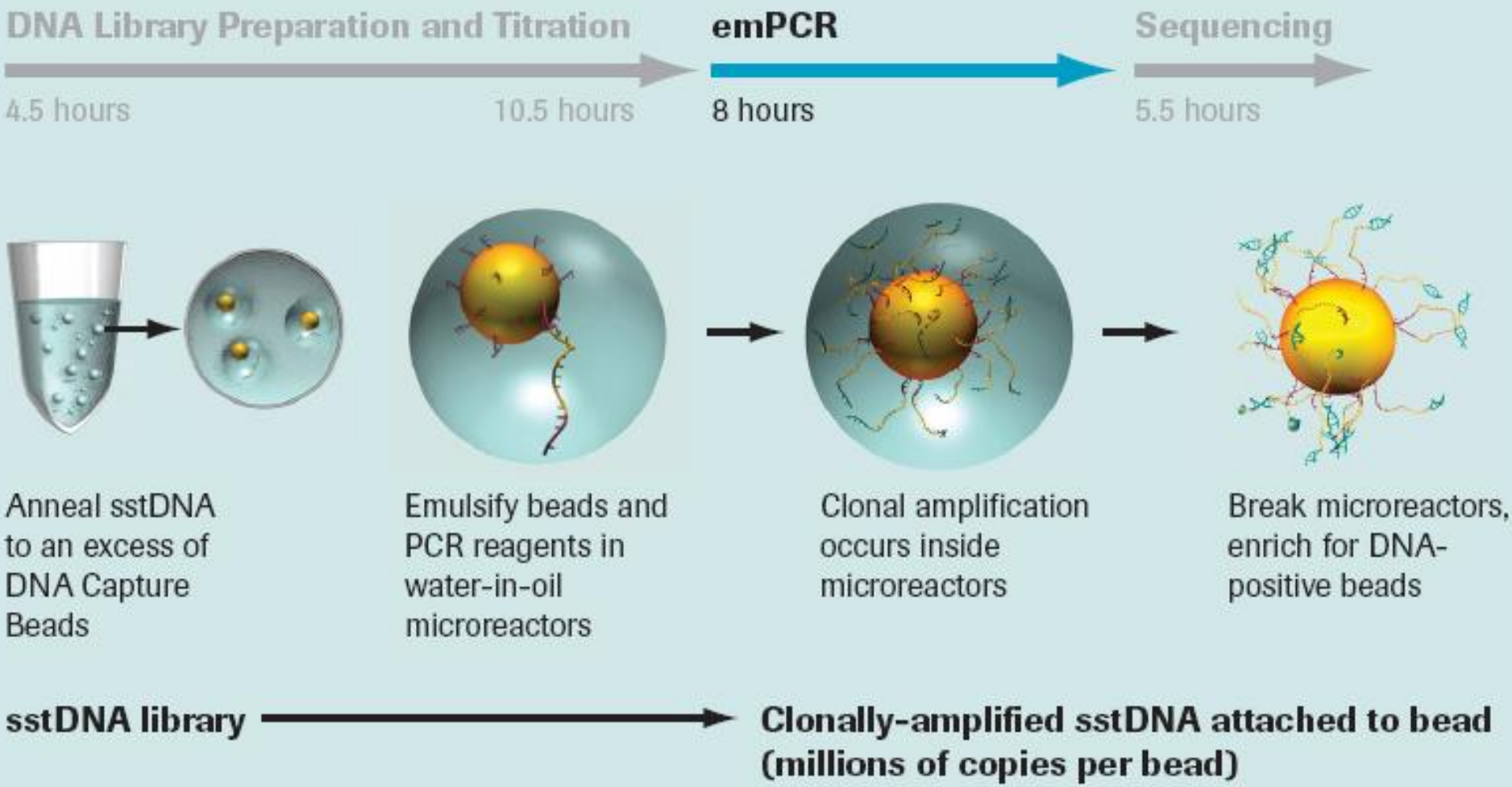
# Library preparation




**gDNA** → **sstDNA library**

- Genome fragmented by nebulization
- No cloning; no colony picking
- sstDNA library created with adaptors. The adaptors are used as primers, and for binding to beads.
- A/B fragments selected using streptavidin-biotin purification

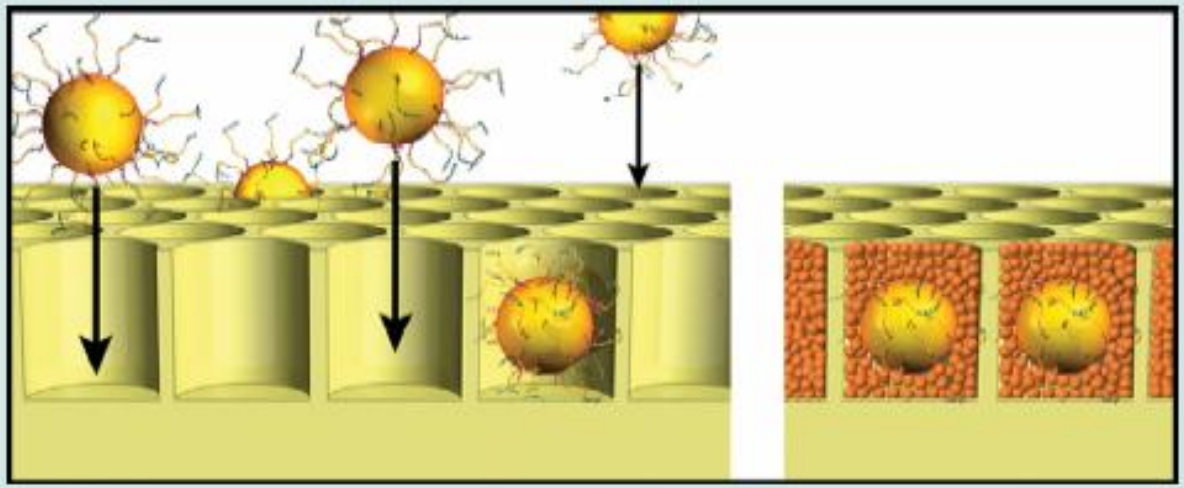
# Emulsion PCR



# PicoTiterPlate (PTP)

DNA Library Preparation and Titration → emPCR → **Sequencing**

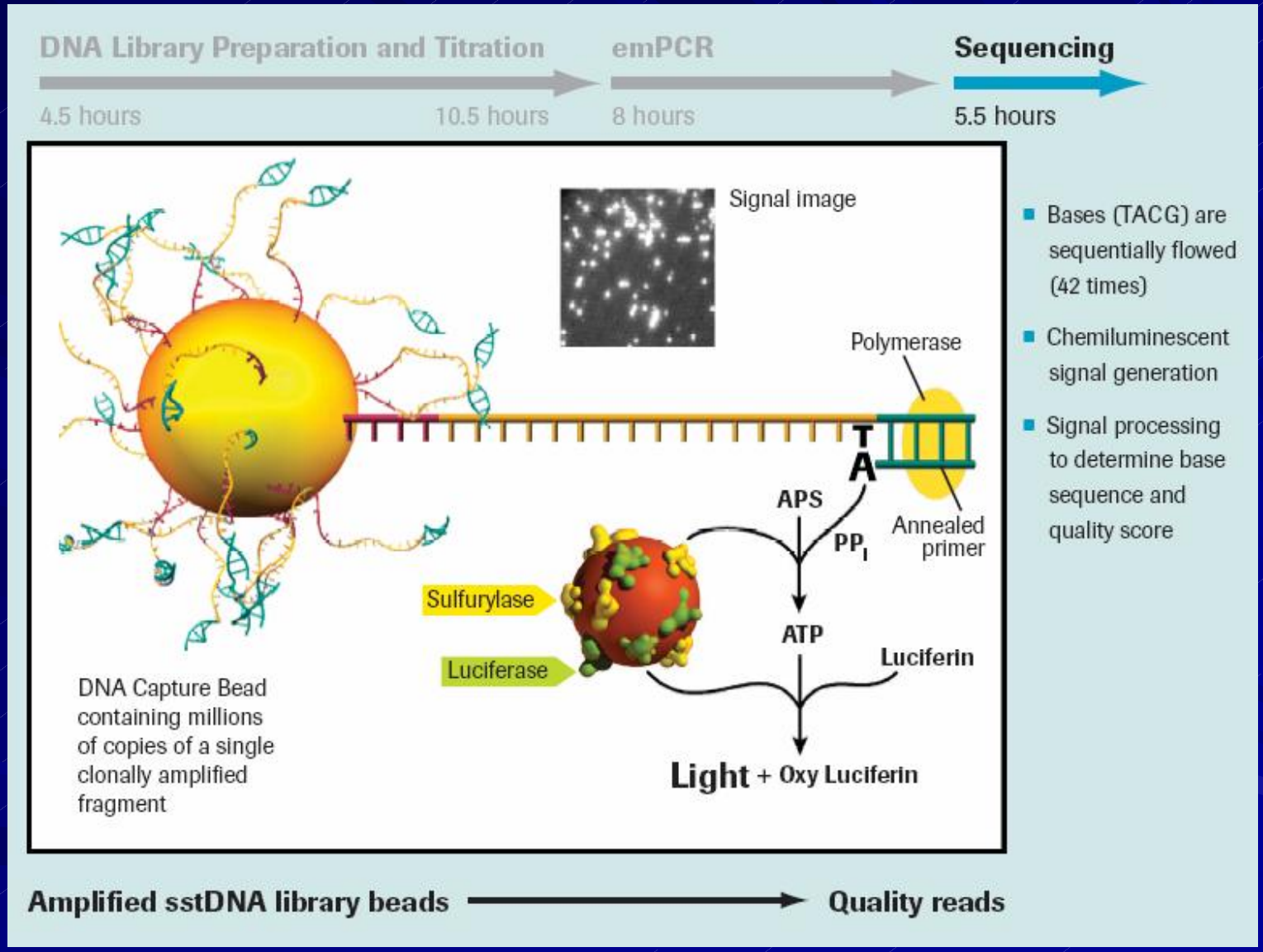
4.5 hours – 10.5 hours | 8 hours | 5.5 hours



- Well diameter: average of 44 µm
- A single clonally amplified sstDNA bead is deposited per well
- 200,000 reads obtained in parallel on large-format PicoTiterPlate device

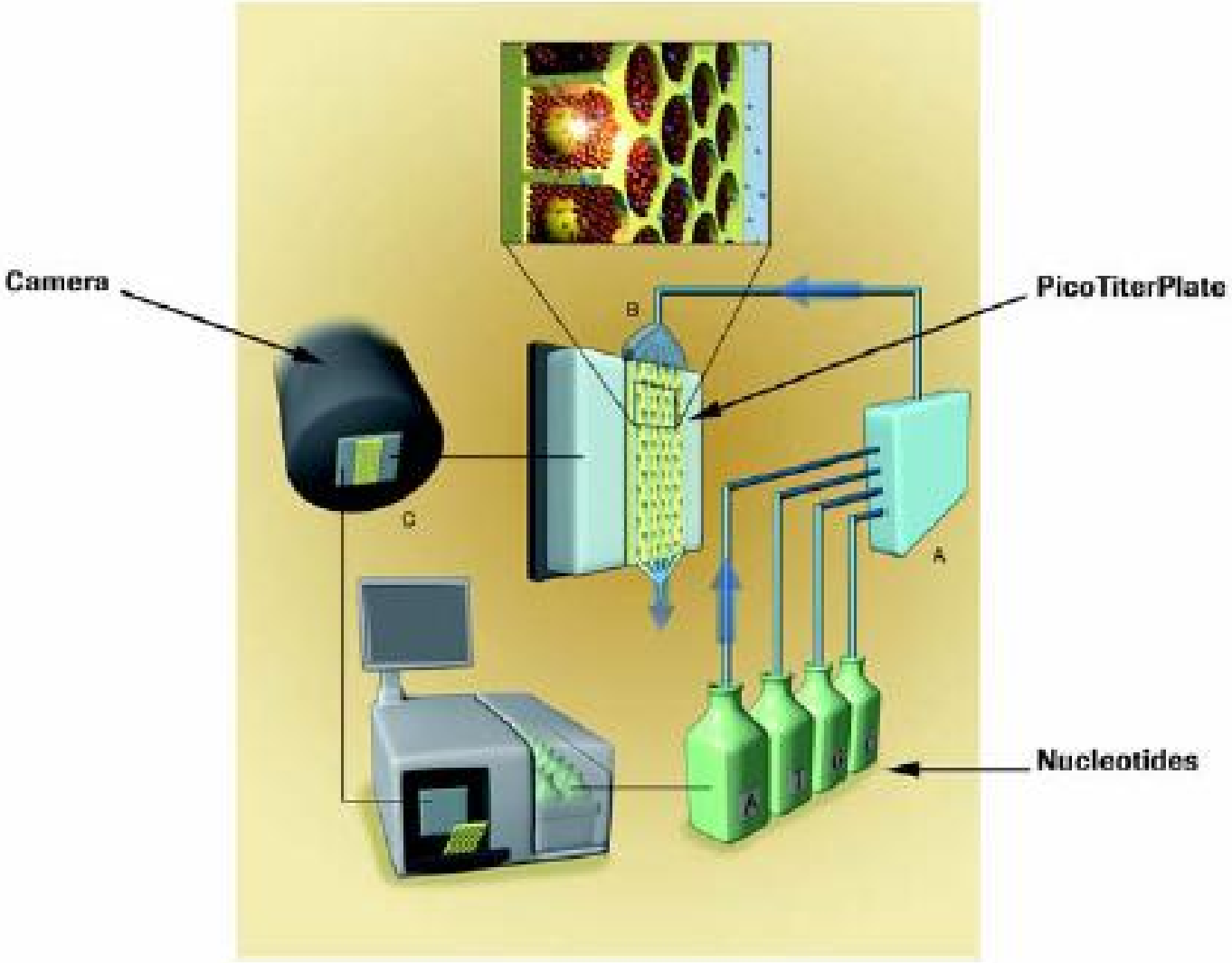
**Amplified sstDNA library beads** → **Quality reads**

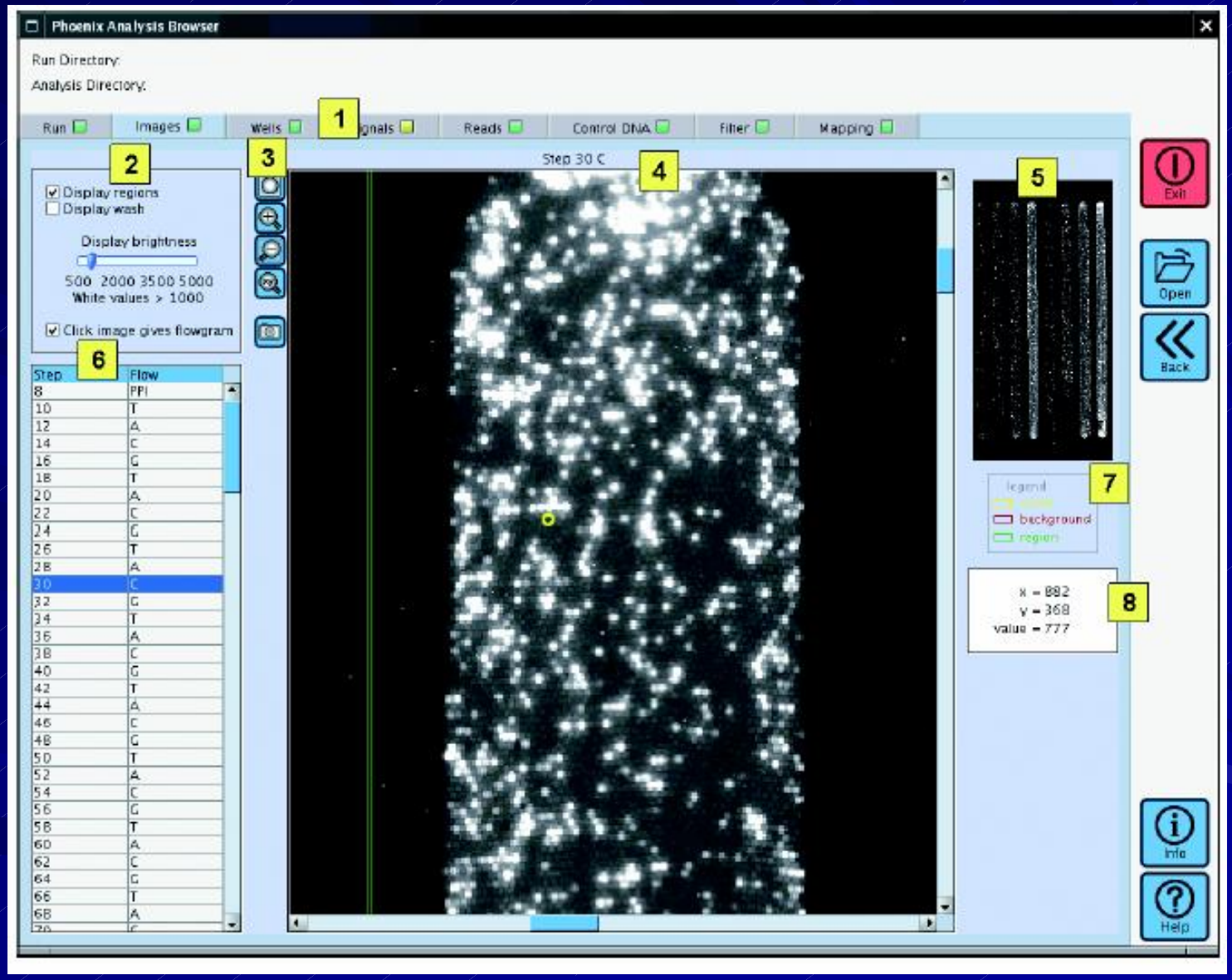
# Pyrosequencing



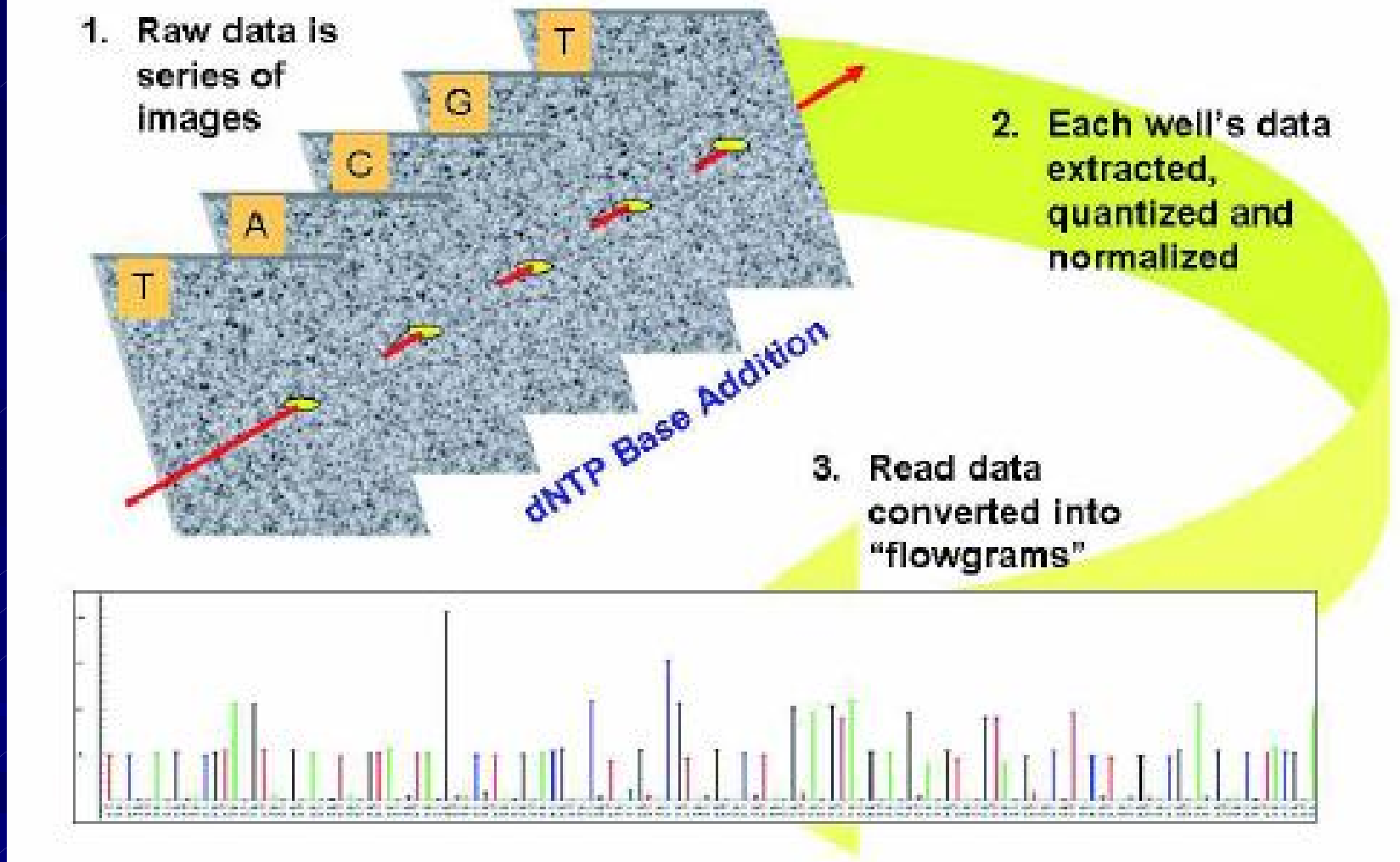

Adaptor Taq TCAG -- CTGA

# Genome Sequencer GS20/FLX



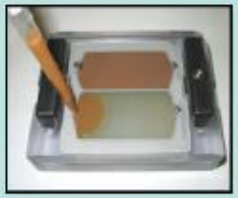

Phoenix Analysis Browser
Run Directory:
Analysis Directory:
Run
Images
Wells
1
Reads
Control DNA
Filter
Mapping
2
3
Step 30 C
4
5
Display regions
Display wash
Display brightness
500 2000 3500 5000
White values > 1000
Click image gives flowgram
6
Step
Flow
8 PPi
10 T
12 A
14 C
16 G
18 T
20 A
22 C
24 G
26 T
28 A
30 C
32 G
34 T
36 A
38 C
40 G
42 T
44 A
46 C
48 G
50 T
52 A
54 C
56 G
58 T
60 A
62 C
64 G
66 T
68 A
7
background
region
x = 882
y = 368
value = 777
8
Exit
Open
Back
Info
Help

# Flowgram



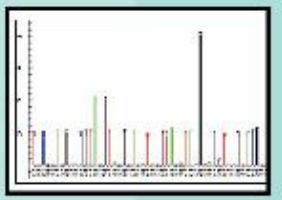

Adaptor Taq TCAG -- CTGA

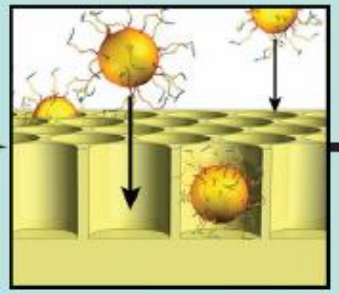
PicoTiterPlate device



Genome Sequencer 20 Instrument

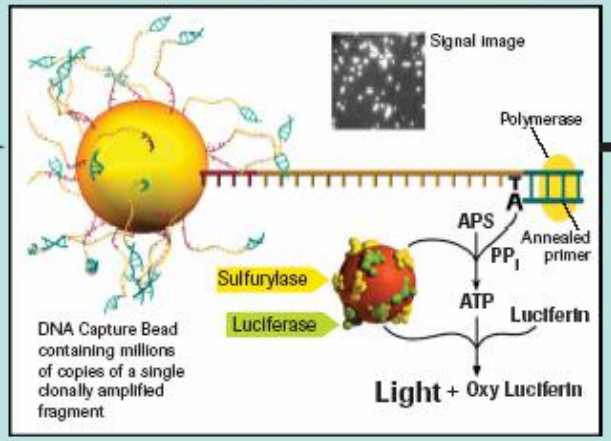
Flowgram





## Sequencing

- Loading PicoTiterPlate device
- One bead per well

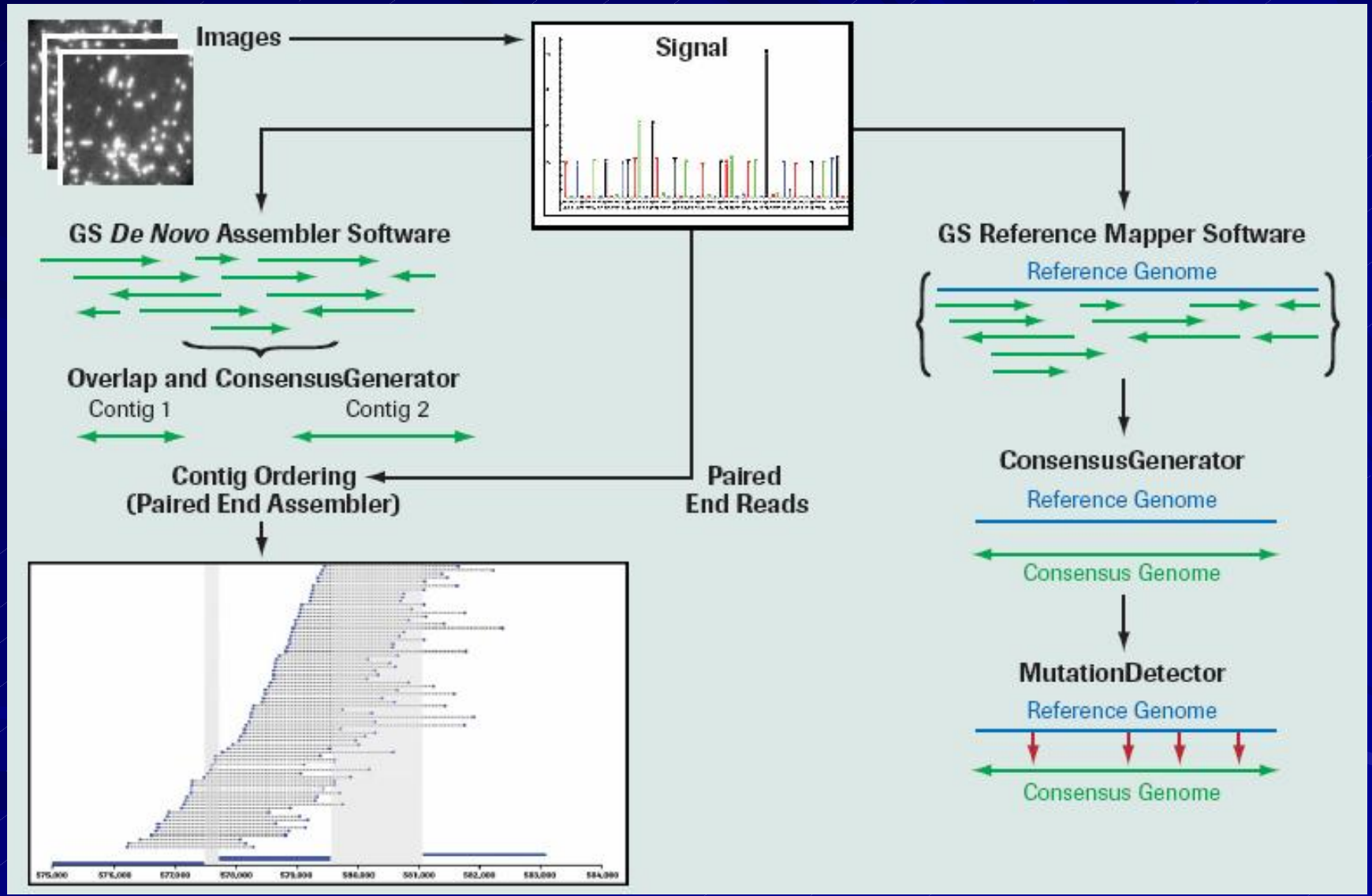



- Sequencing-by-synthesis
- Light signal proportional to incorporated nucleotide is captured by CCD camera
- Image and signal processing to determine sequence and quality score

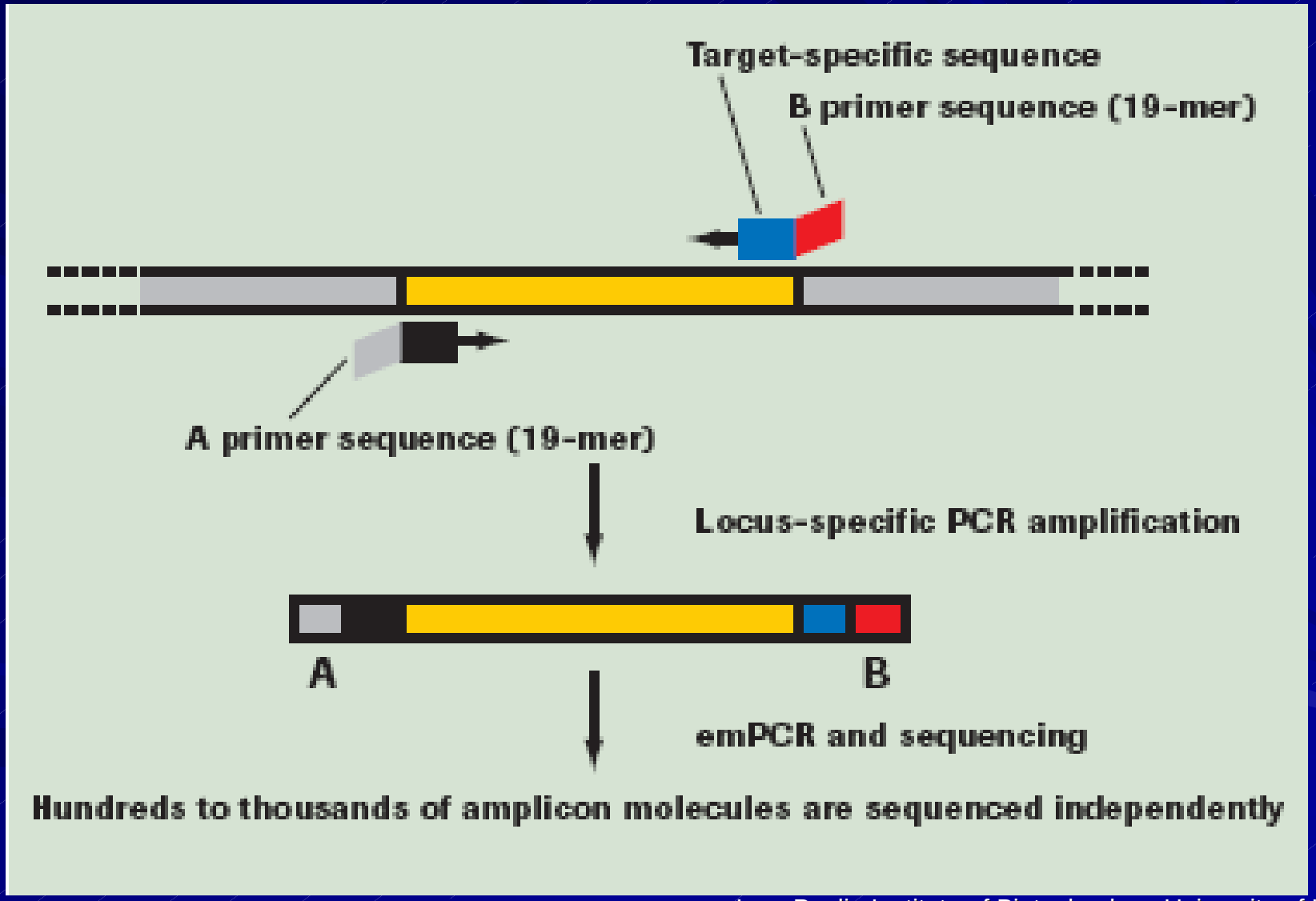
## Bioinformatics Tools

- Resequencing
- *De novo* sequencing
- Amplicon Sequencing

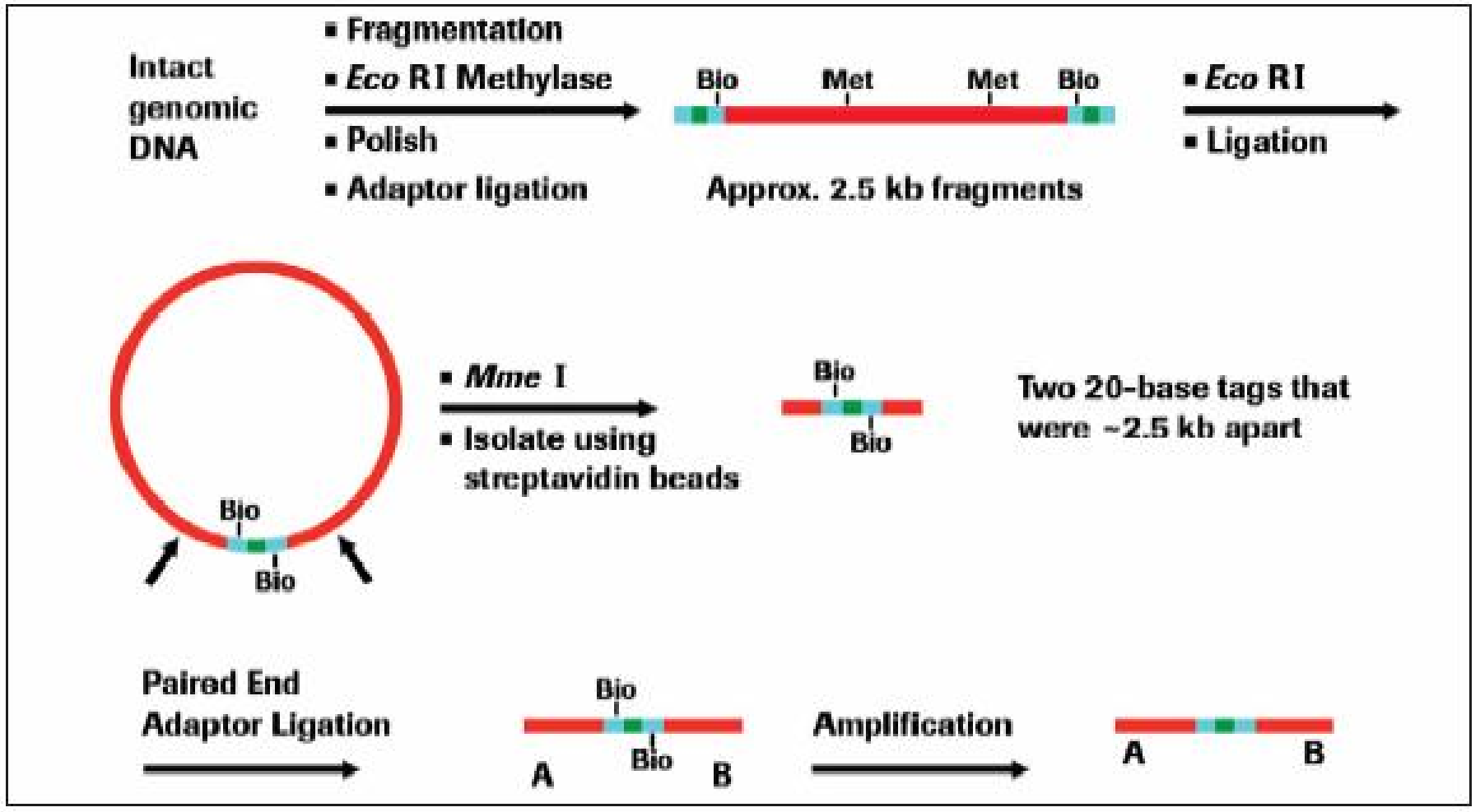
Publicly available tools


Images
Signal
GS De Novo Assembler Software
Overlap and ConsensusGenerator
Contig 1
Contig 2
Contig Ordering
(Paired End Assembler)
Paired
End Reads
GS Reference Mapper Software
Reference Genome
ConsensusGenerator
Reference Genome
Consensus Genome
MutationDetector
Reference Genome
Consensus Genome

# Amplicon sequencing



# Paired-end Sequencing



# Illumina/Solexa Genome Analyzer

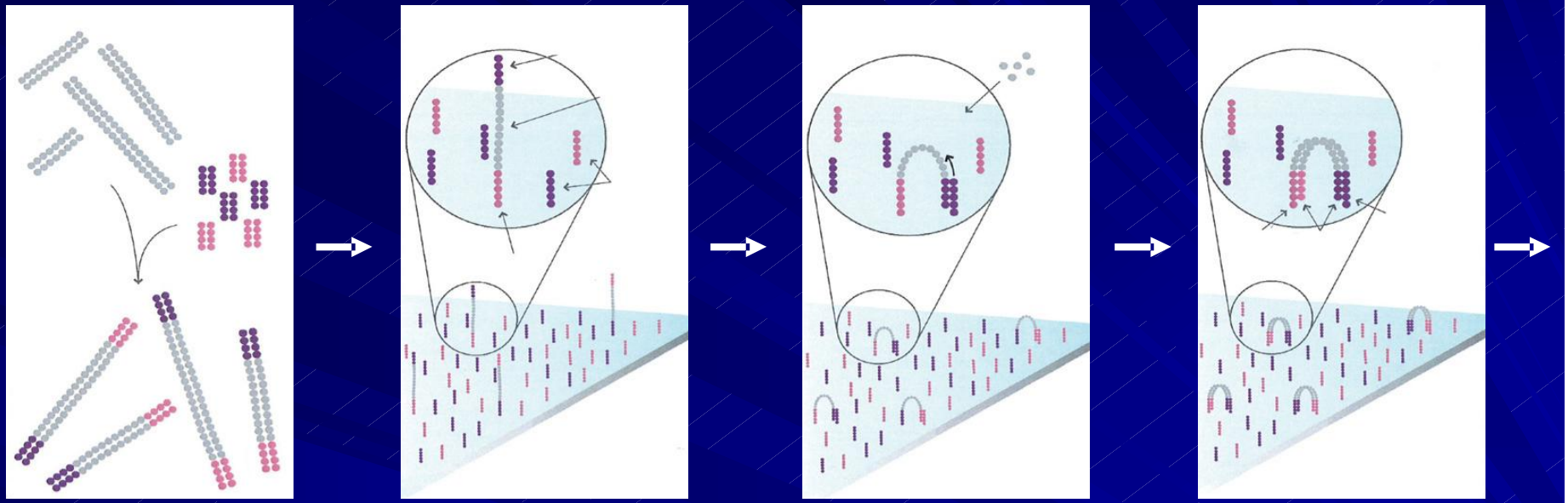
(http://www.illumina.com)

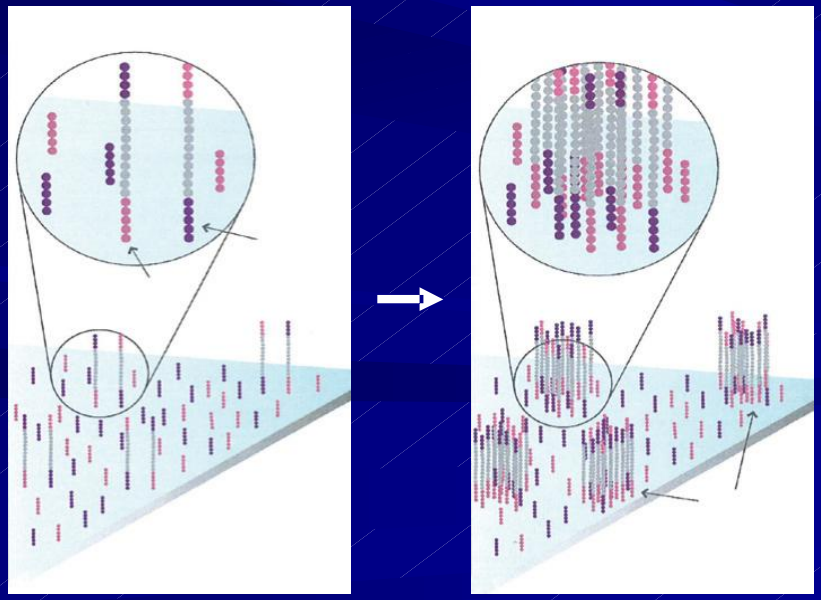
- Clonal Single Molecule Array technology
  - Sequencing-by-synthesis technology
  - Reversible terminator-based sequencing
    - removable fluorescence
  - Flow cell with > 10 million clusters
    - Each cluster ~1,000 copies of template /cm$^2$
  - 1–8 samples / run
  - 3 laser system (660, 635, and 532 nm)
  - Read length 35 - 50 bp, 1- 2 Gb / run
    - Run time 3 – 6 days,

Cluster Station

Flow cell

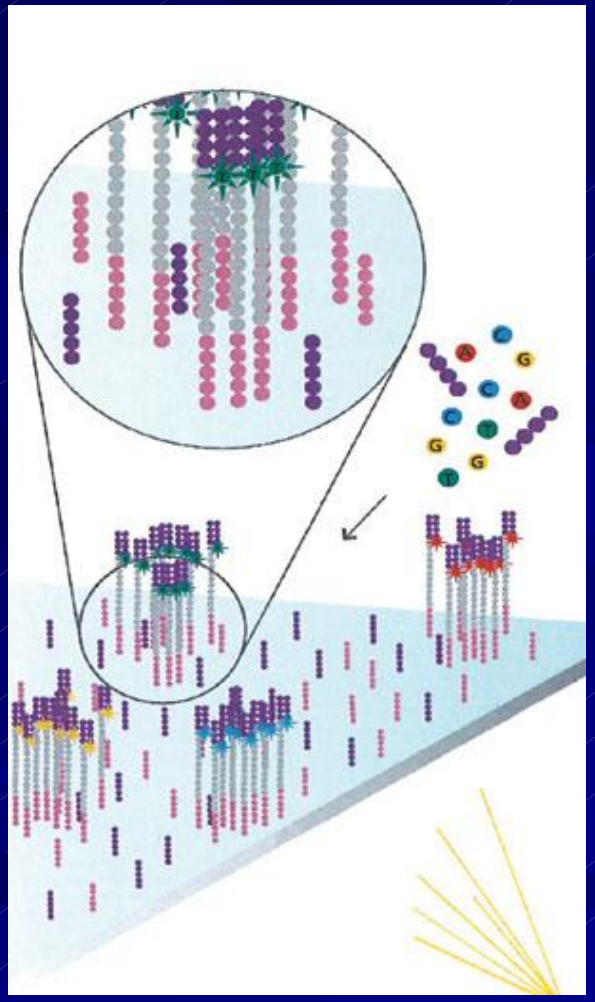
# Illumina/Solexa

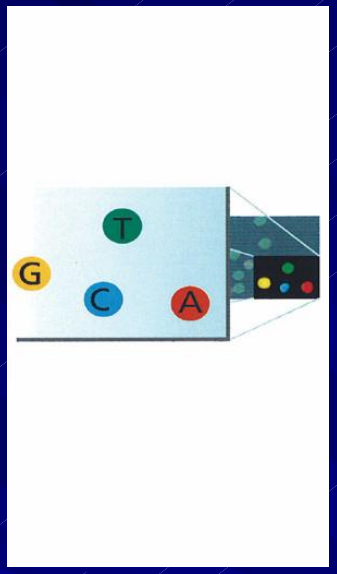




- Sample preparation
  - 100ng–1µg
  - Attaching to Flow cell
  - Bridging
  - PCR
    - Elongation
    - Denaturation
    - Clonal amplification

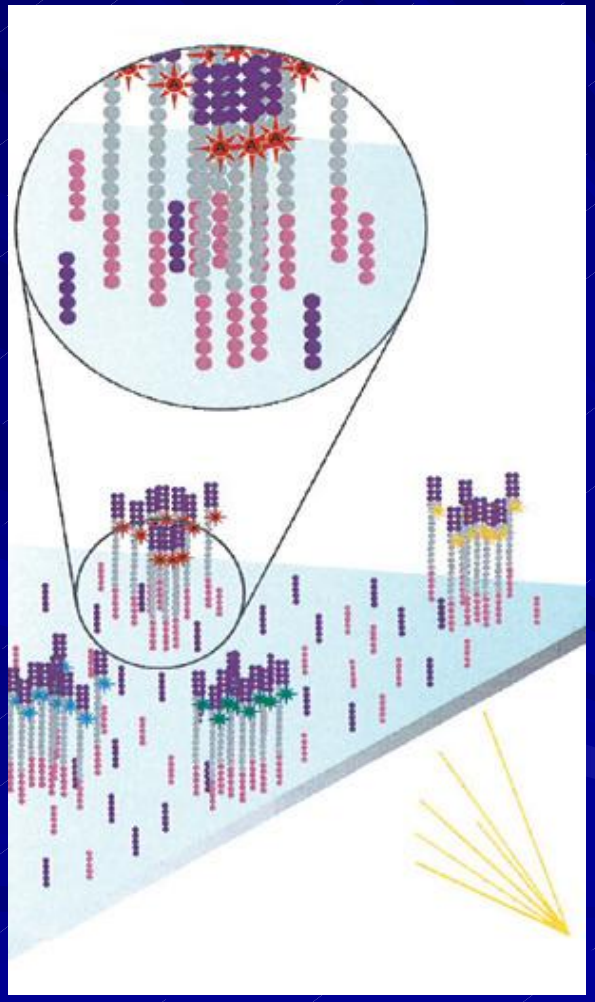
# Illumina/Solexa sequencing






Sequencing

- First bases
- Fluorescent reversible terminators
- Detection with laser and CCD camera




Sequencing

- Second bases detected after removal of label and blocking

# SOLiD, Applied Biosystems

(http://www.appliedbiosystems.com)

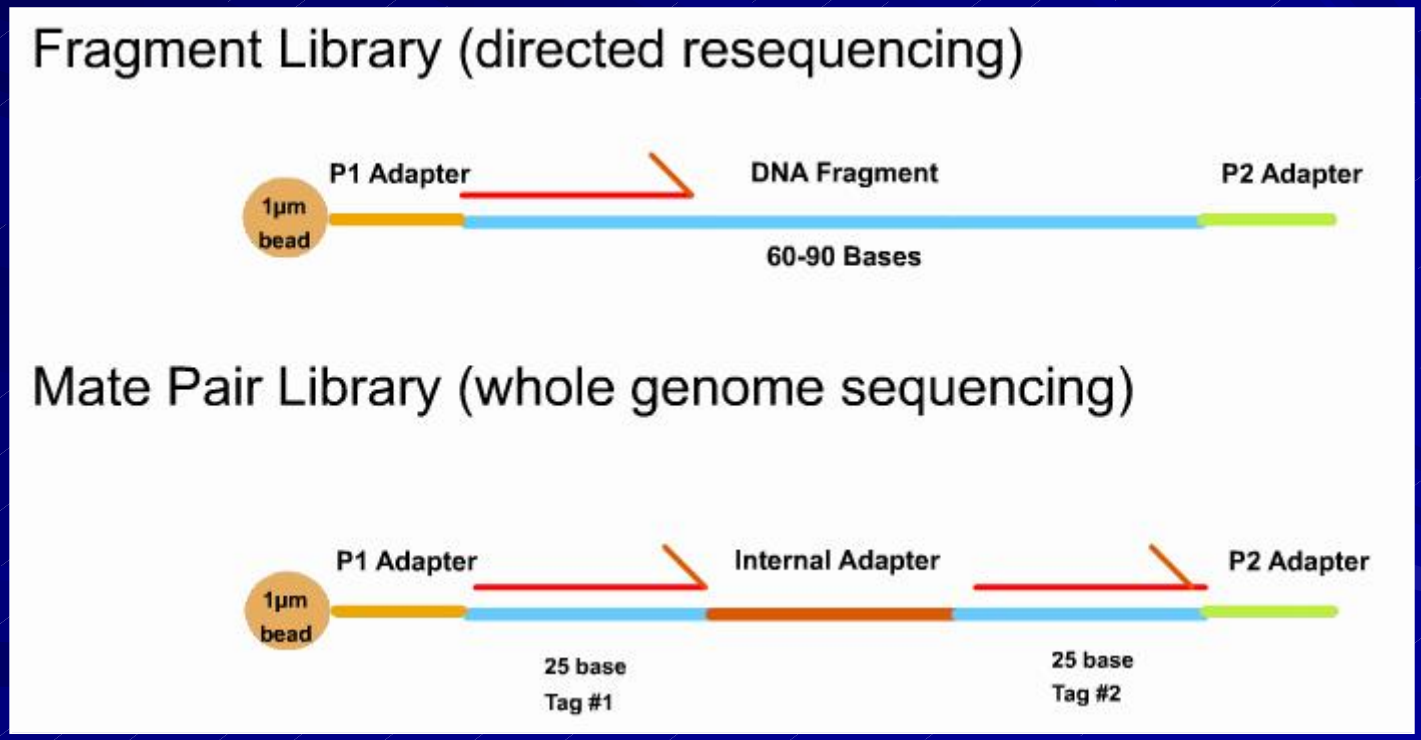
- Sequencing by Ligation
  - emPCR
    - Small beads, 1µm
  - Attaching to glass slides
  - Labelled probes
    - Fuor colours
    - 2 base encoding system
  - Repeated ligation steps
  - Detection with 4 Mpixel camera
  - Read lenght 25-30 bp
  - 1-2 slides / run
  - 1-2 Gb / run
  - Run time 5 -10 days

Shendure, J. *et.al.* Science 2005, 309, 1728-1732

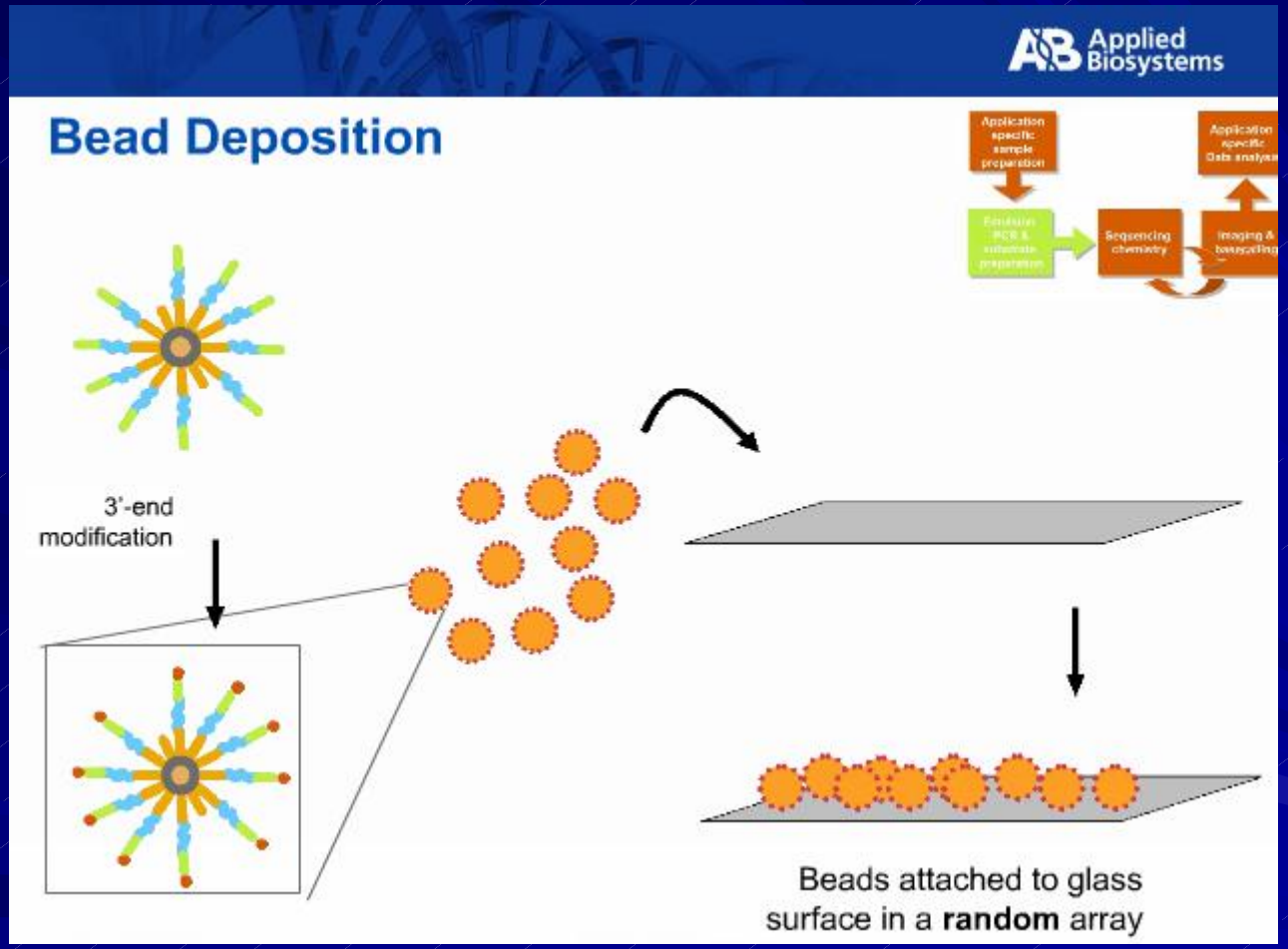
SOLiD

# SOLiD

- Library preparation



# SOLiD

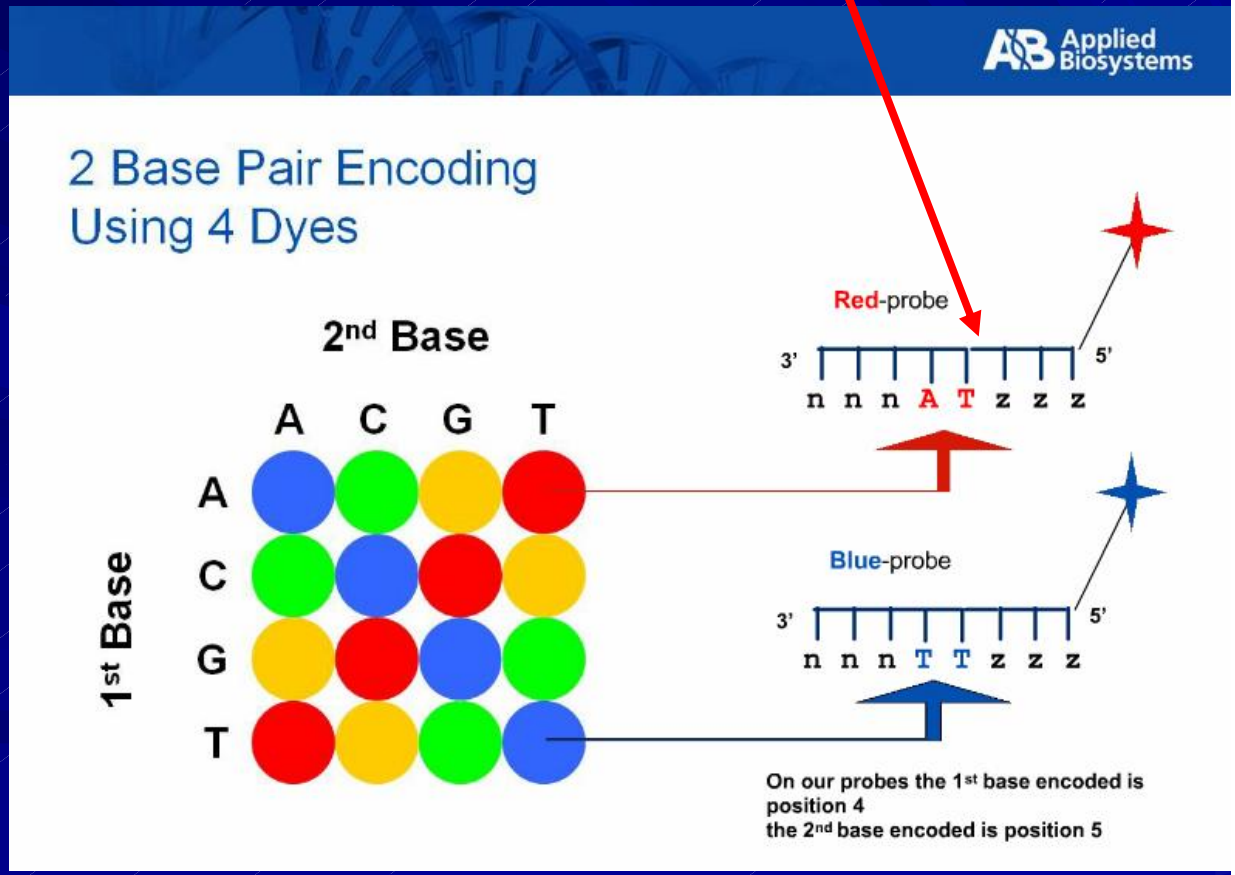

Applied Biosystems
Bead Deposition
Application specific sample preparation
Sequencing chemistry
3'-end modification
Beads attached to glass surface in a **random** array

# SOLiD

- Probes
  - 1 024 Octamer Probes
  - 4 Dyes
  - 4 dinucleotides
  - 256 probes / dye

N = degenerate bases

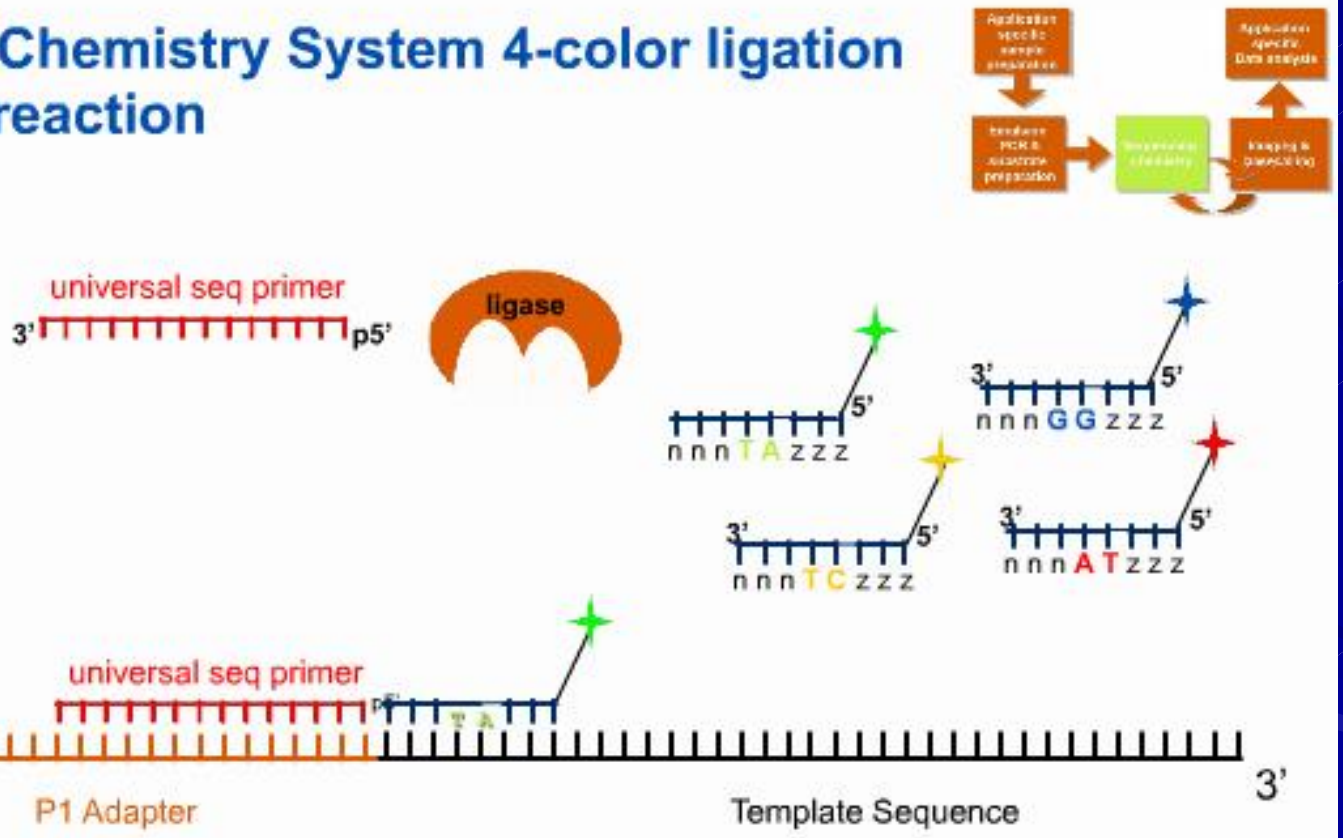
Z = universal base



# SOLiD

Applied Biosystems

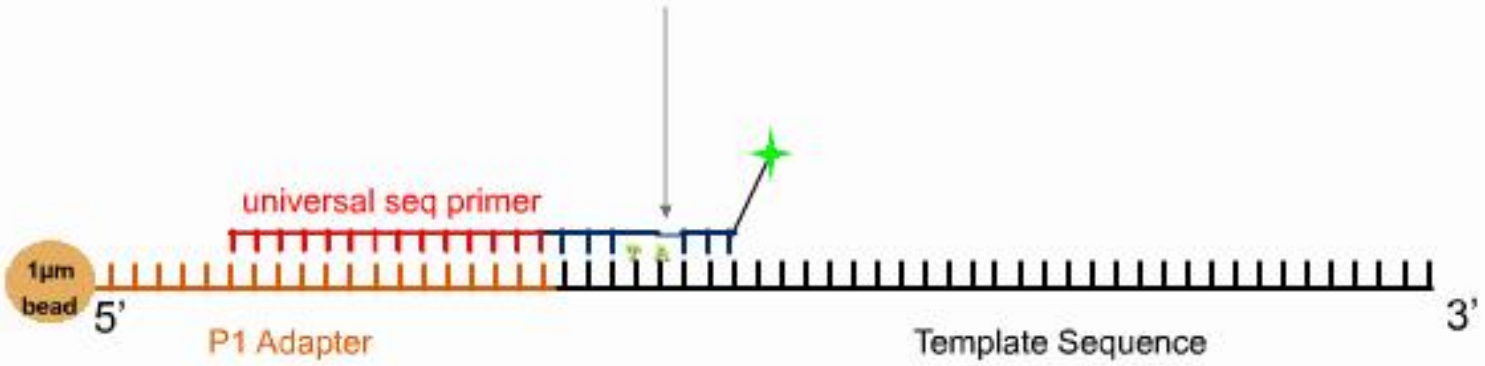
## SOLiD™ Chemistry System 4-color ligation Ligation reaction



# SOLiD

Applied Biosystems

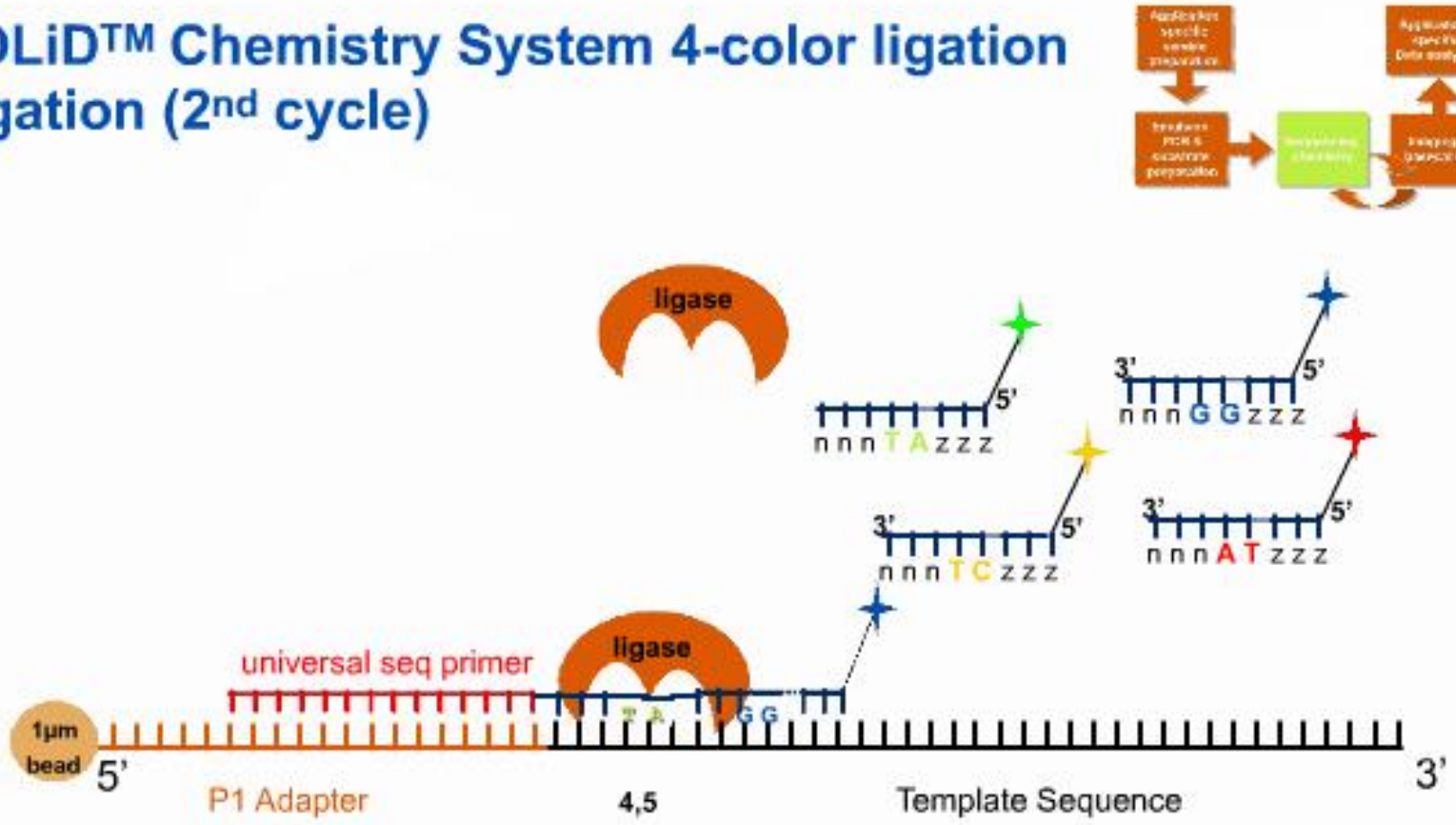
## SOLiD™ Chemistry System 4-color ligation Cleavage



# SOLiD

Applied Biosystems

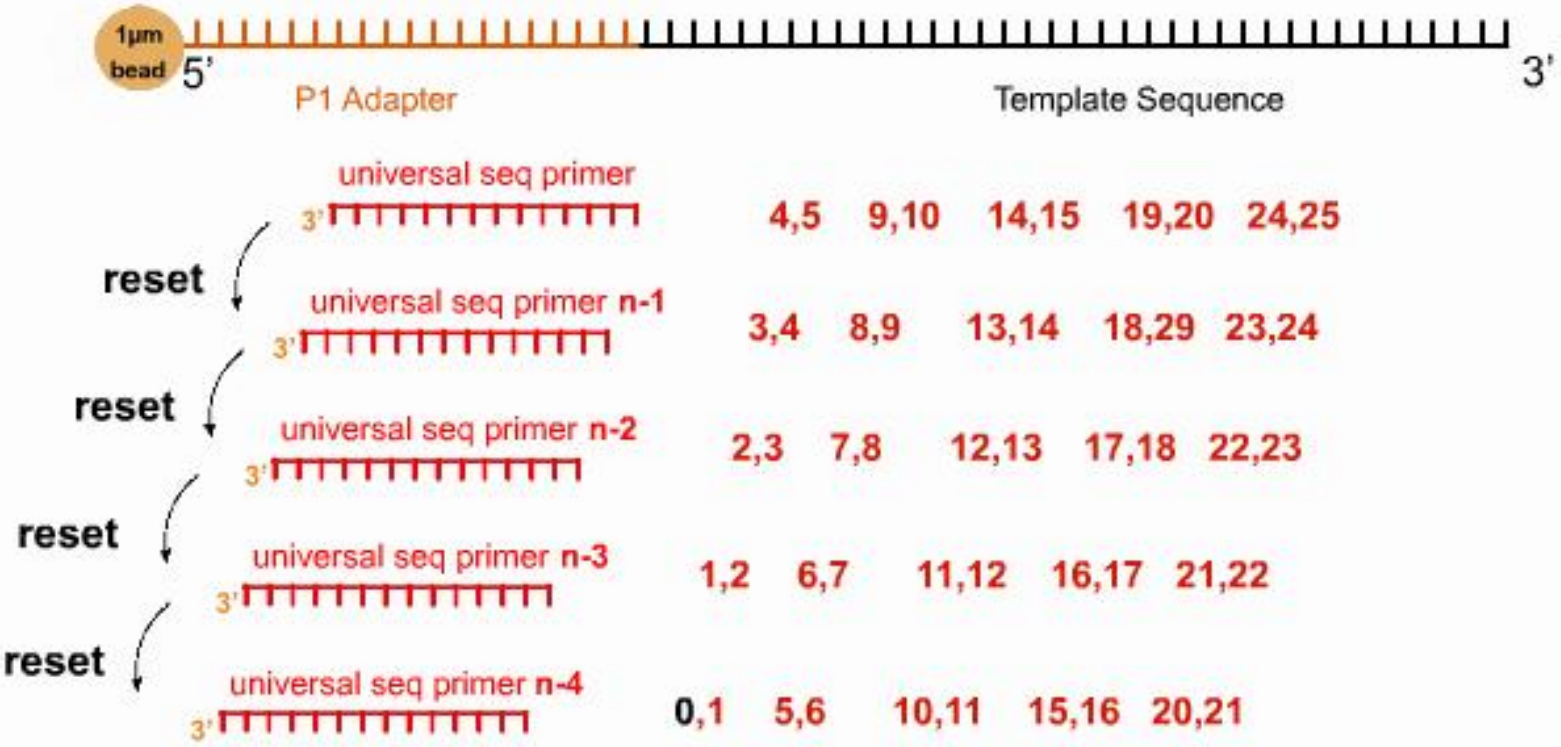
## SOLiD™ Chemistry System 4-color ligation
## Ligation (2nd cycle)



# SOLiD

Applied Biosystems

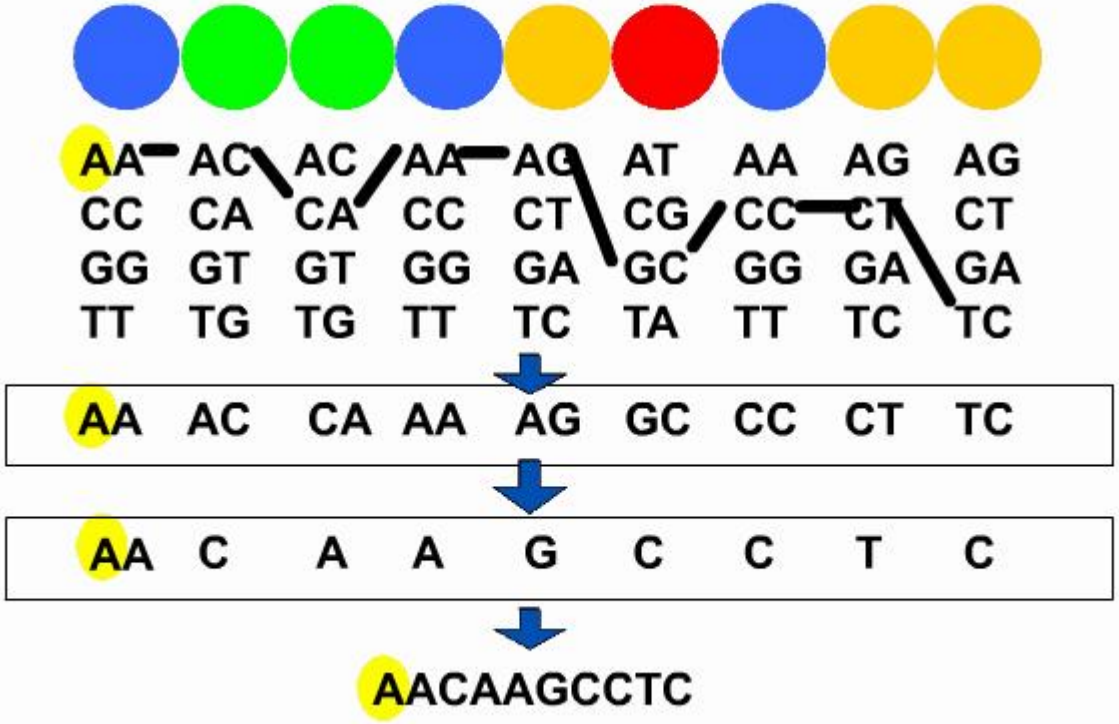
## Sequential rounds of sequencing
## Multiple cycles per round

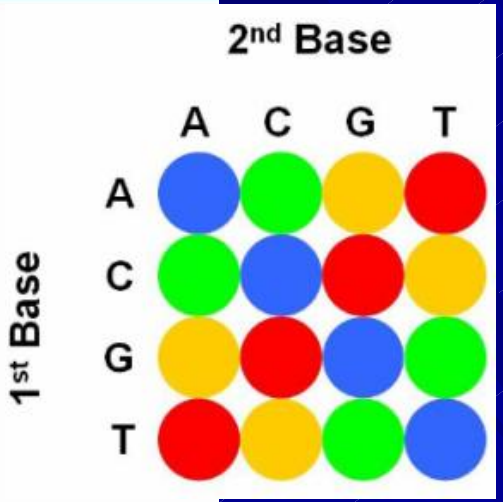


# SOLiD

Applied Biosystems

## Summary of decoding



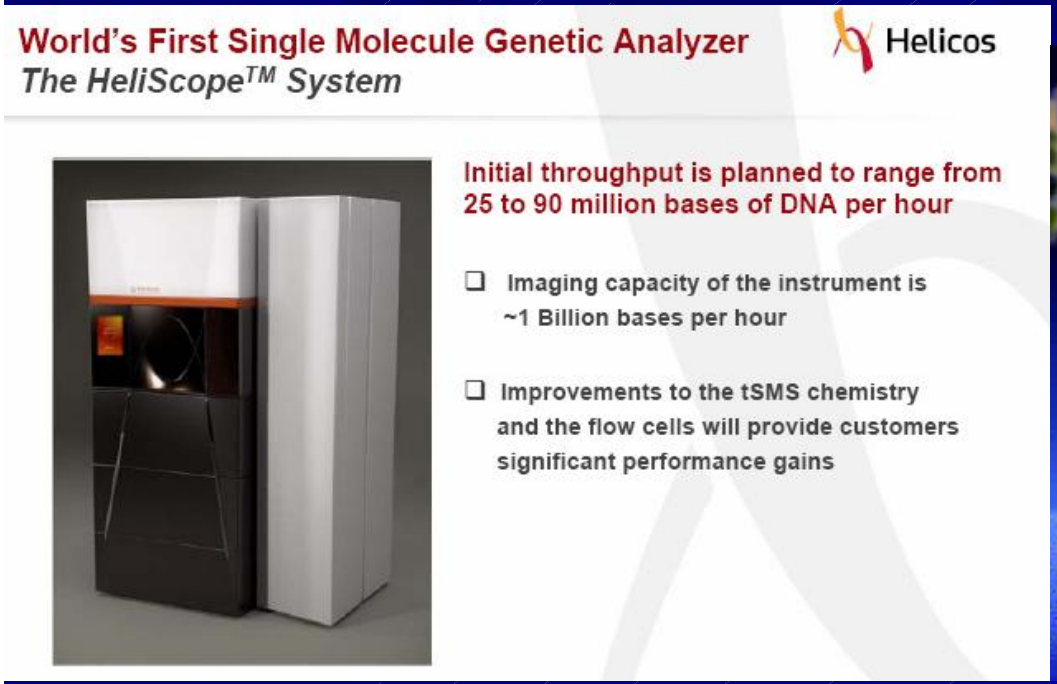



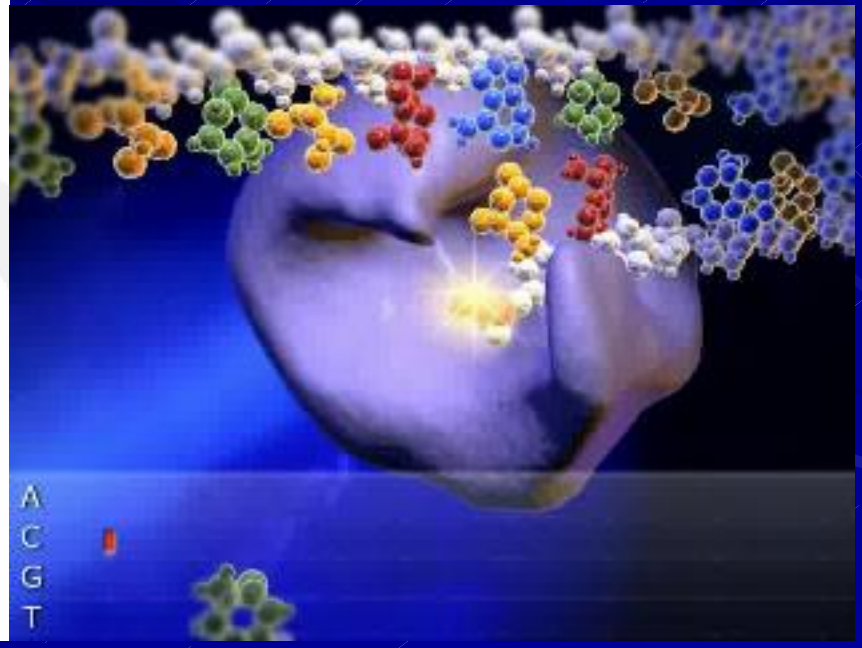
# Applications

- Whole genome sequencing
  - *de novo* sequencing
    - Genome Sequencer FLX
- Comparative sequencing
  - All three systems
- Metagenomics
  - Genome Sequencer FLX
- Amplicon sequencing
  - Mutations / SNP
  - All three systems
- Transcriptome sequencing
  - cDNA
    - All three systems
  - Small RNA
    - All three systems
- ChIP sequencing
  - All three systems
- Methylation sequencing
  - All three systems

# Other technologies

- Helicos (www.helicosbio.com)
  - Sequencin-by-synthesis
  - No PCR amplification
  - 25-90 Mb/h
- VisiGen (www.visigenbio.com)
  - Real-time detection of DNA synthesis, FRET
  - Intact DNA fragments
  - 1Mb/sec/machine



ARCHON GENOMICS XPRIZE

http://genomics.xprize.org/genomics

# $10M to the First Team to Sequence 100 Human Genomes in 10 Days

RegisteredTeams

- 454 Life Sciences (Roche) (www.454.com )
- VisiGen (www.visigenbio.com )
- FfAME (www.ffame.org )
- Reveo (www.reveo.com )
- Base4innovation (www.base4innovation.co.uk/ )
- Personal Genome X-Team (PGx), George Church

# Assemblers

- Phrap, Phil Green
  - University of Washington
- TIGR assembler
  - TIGR
- Celera
  - Celera Corporation
- Euler
  - University of California
- Arachne
  - Broad Institute
- CAP3, PCAP
  - Iowa State University
- gsAssembler
  - Roche, 454 Life Science
- Amos
  - University of Maryland
- SHARCGS
  - Genome Res,2007,17,1697
- SAKE
  - Bioinformatics,2007,23,500
- SHRAP
  - PlosONE,2007,2:e484
- New Euler
  - Genome Res,2007,Dec14