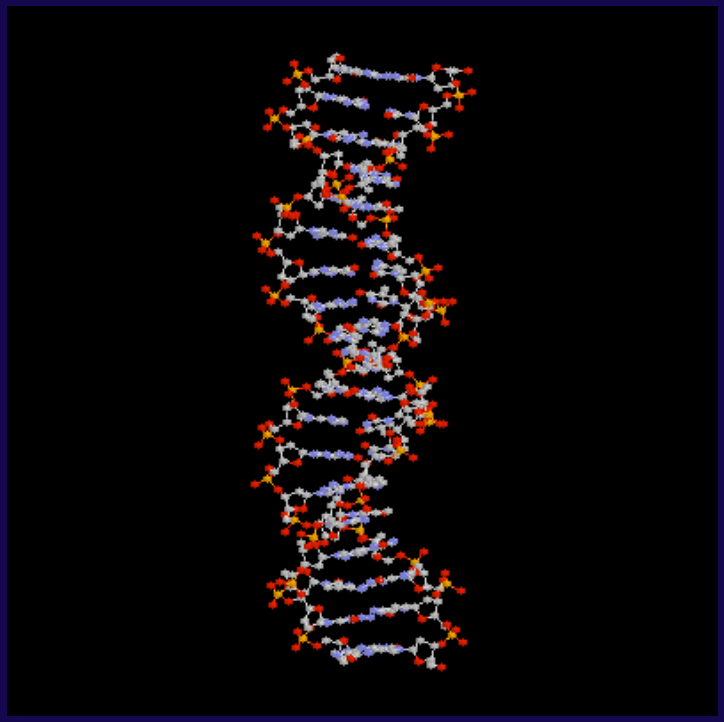
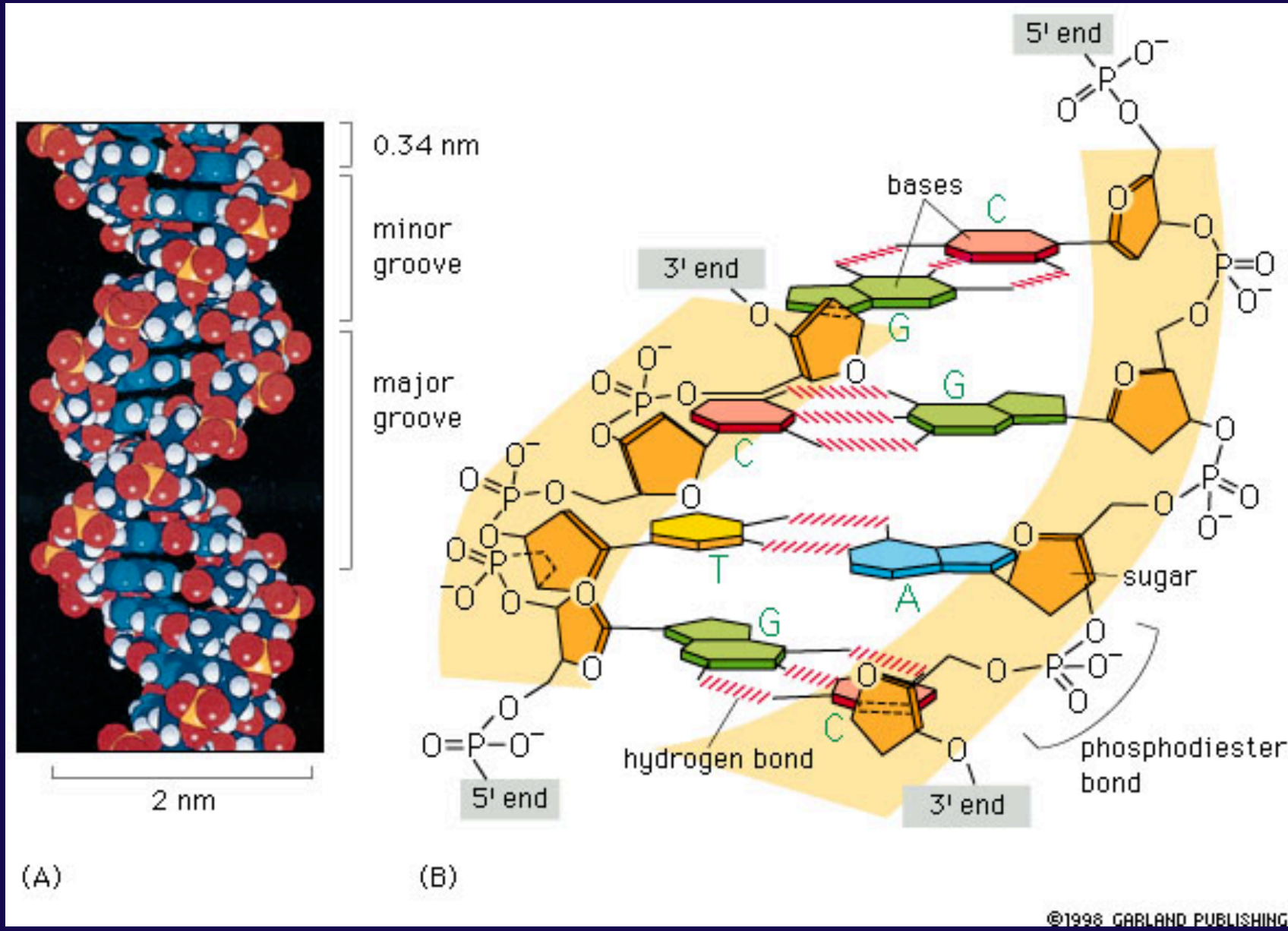
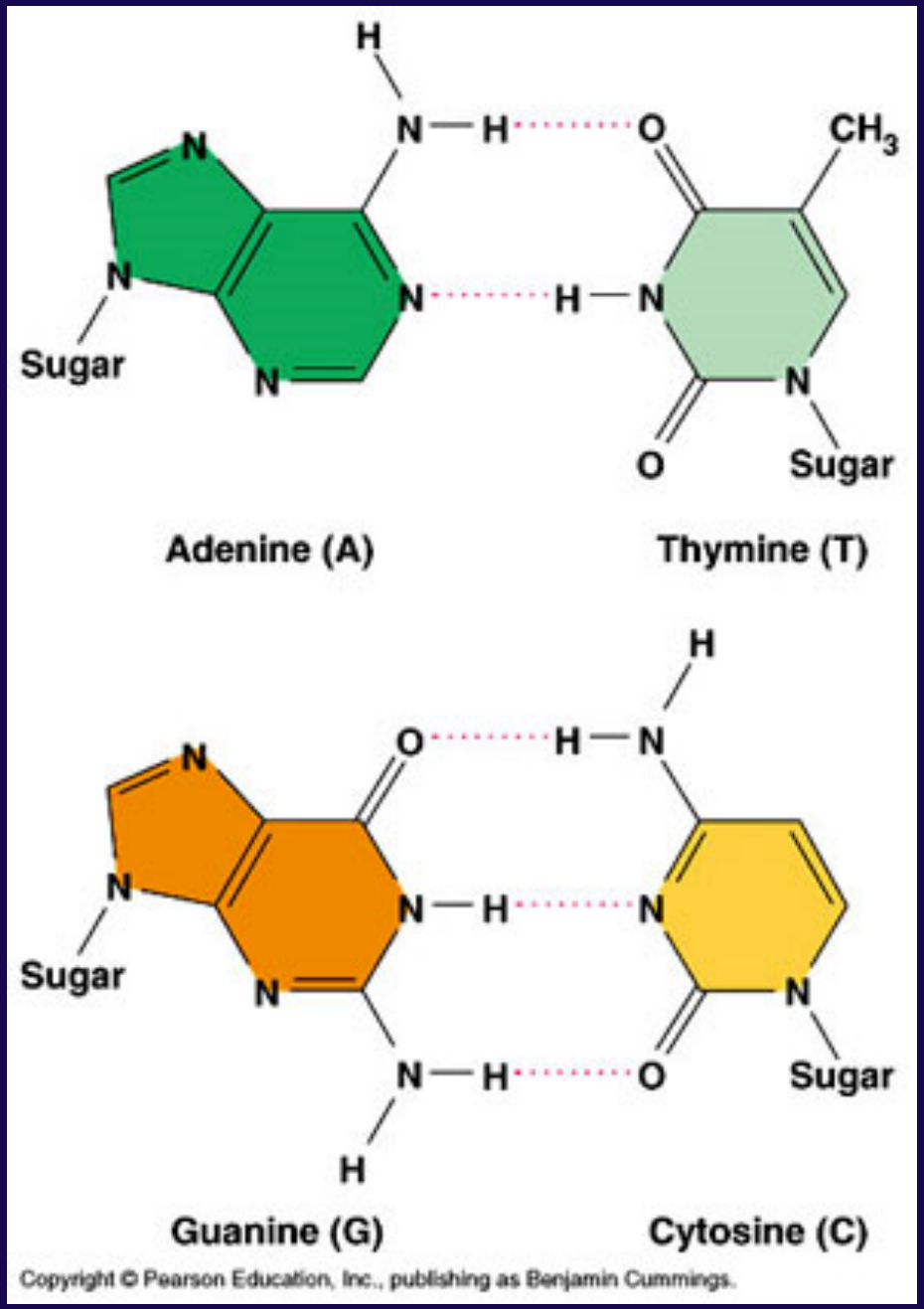
Sequencing Single DNA Molecules



Larry S. Millstein, Ph.D. , J.D.
Foresight Meeting
16 January 2010

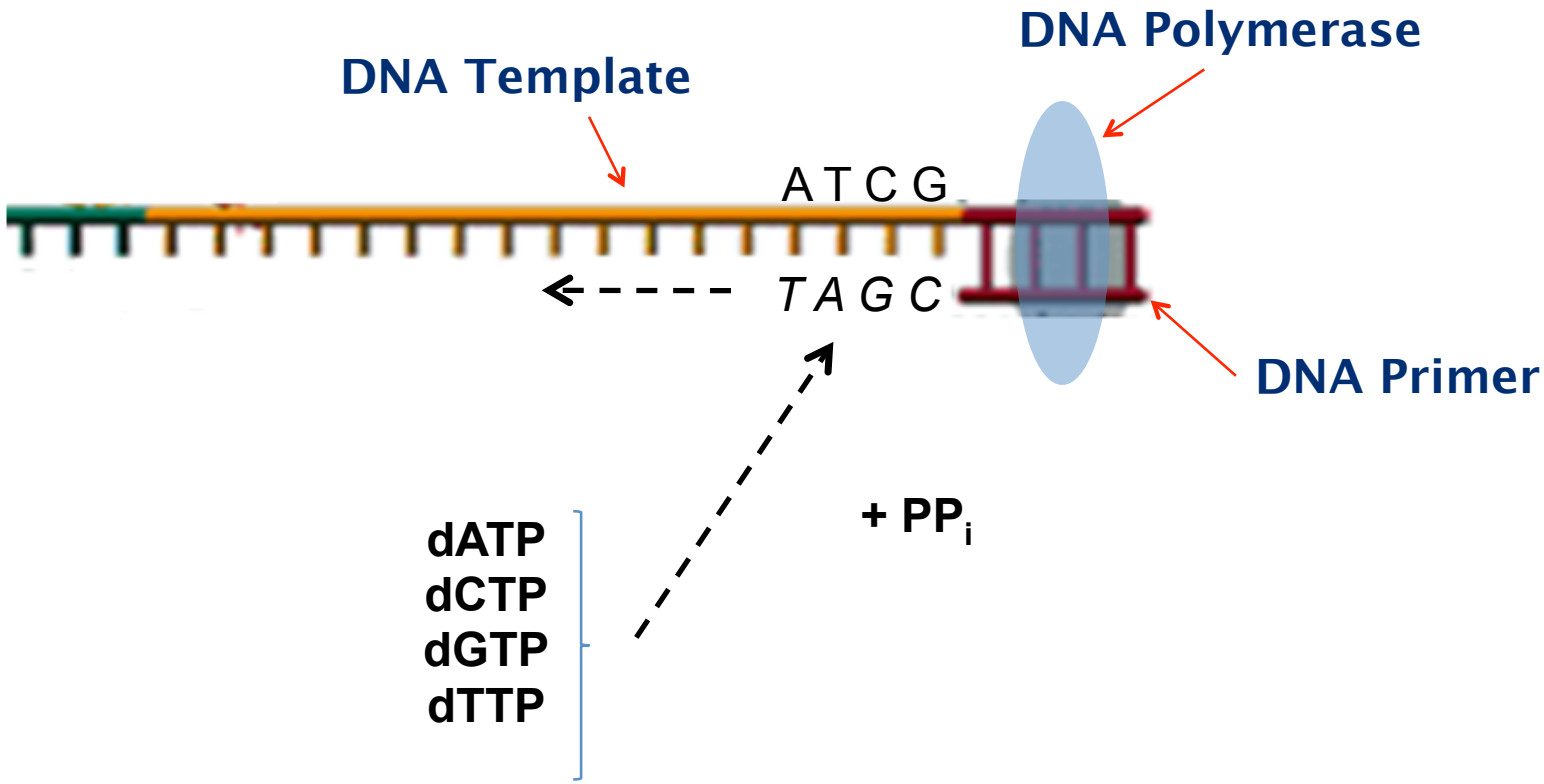

0.34 nm
minor groove
major groove
2 nm
(A)
5' end
bases
3' end
sugar
hydrogen bond
phosphodiester bond
3' end
5' end
(B)
©1998 GARLAND PUBLISHING

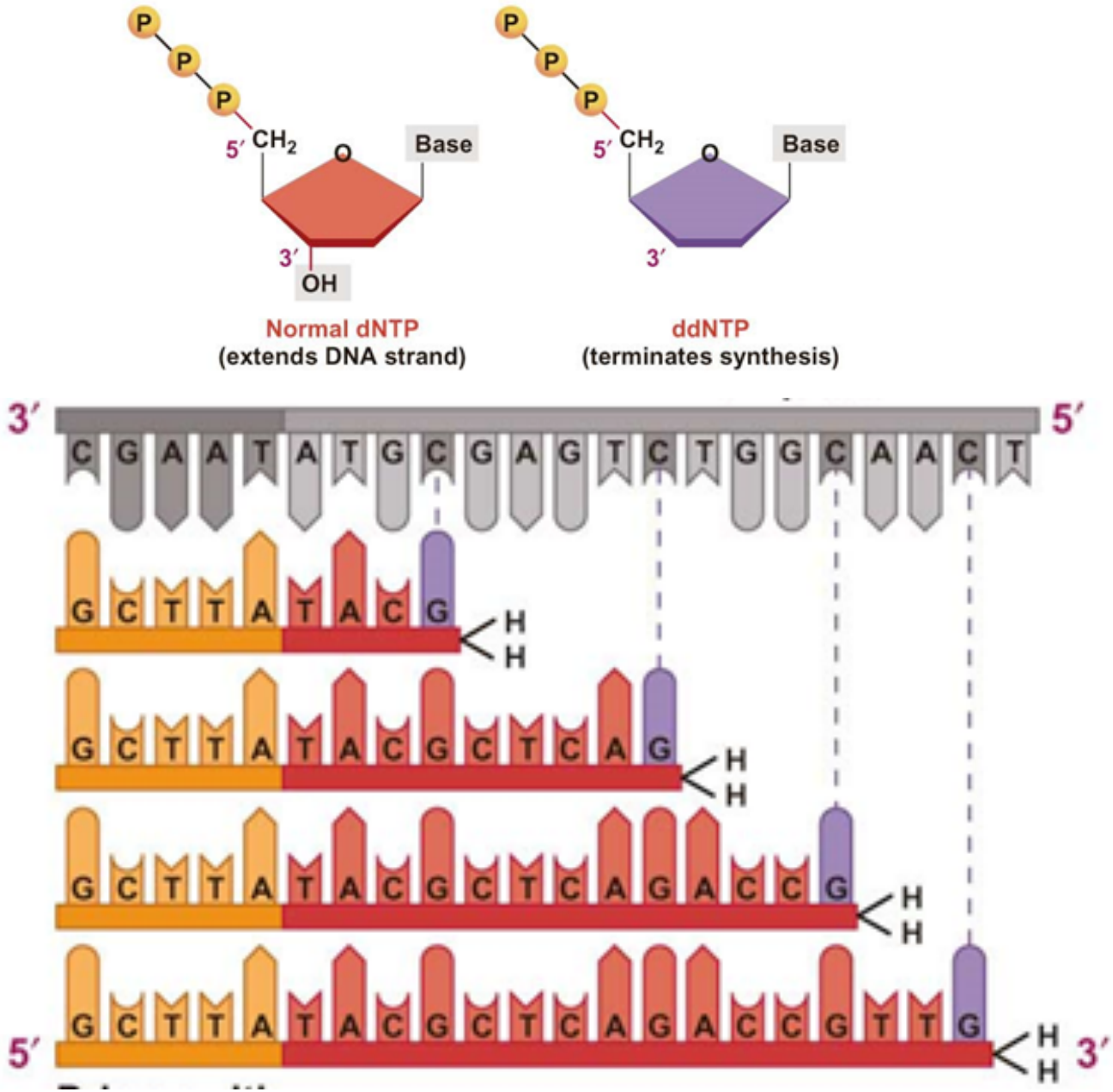


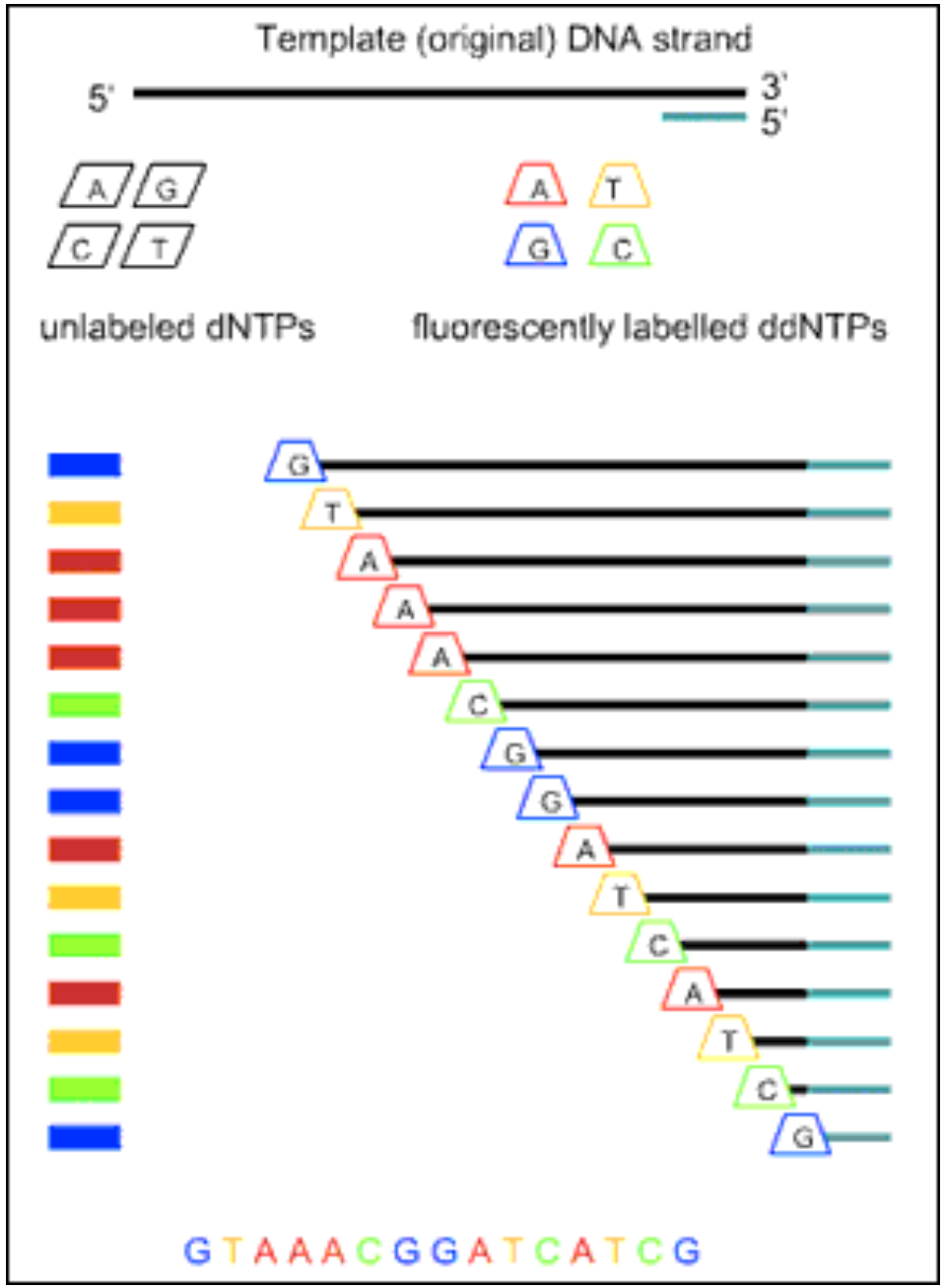
DNA Replication

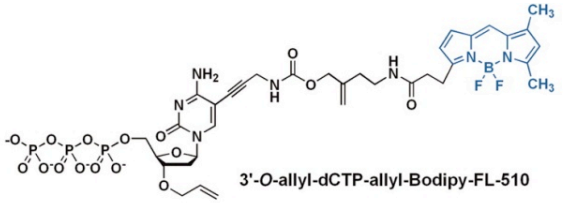
The basis for all current methods of determing DNA seqauences!

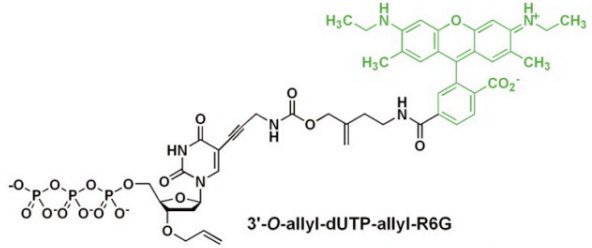
DNA Template, DNA Primer, dNTPs, DNA Polymerase

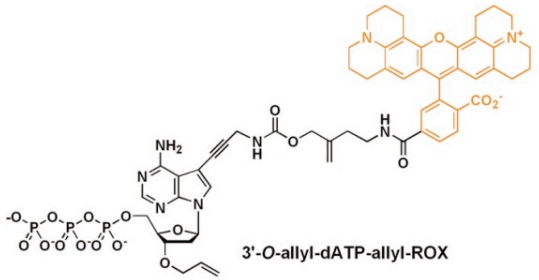


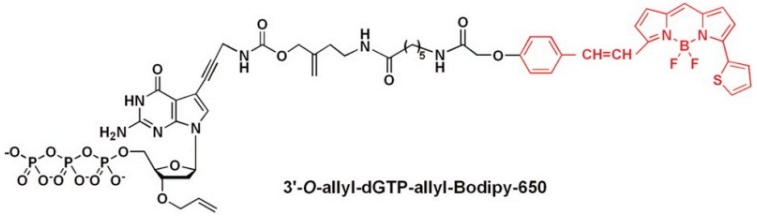

P
P
P
5′ CH2
O
Base
3′
OH
Normal dNTP
(extends DNA strand)
ddNTP
(terminates synthesis)
3′
5′
C G A A T A T G C G A G T C T G G C A A C T
G C T T A T A C G
G C T T A T A C G C T C A G
G C T T A T A C G C T C A G A C C G
G C T T A T A C G C T C A G A C C G T T G
H
H
5′
3′

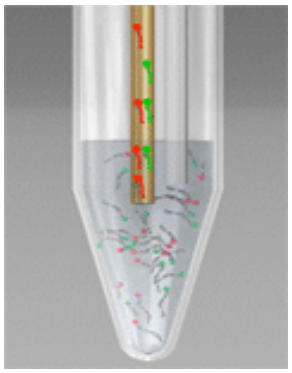

Template (original) DNA strand
5'
3'
5'
A G
C T
A T
G C
unlabeled dNTPs
fluorescently labelled ddNTPs
G T A A A C G G A T C A T C G

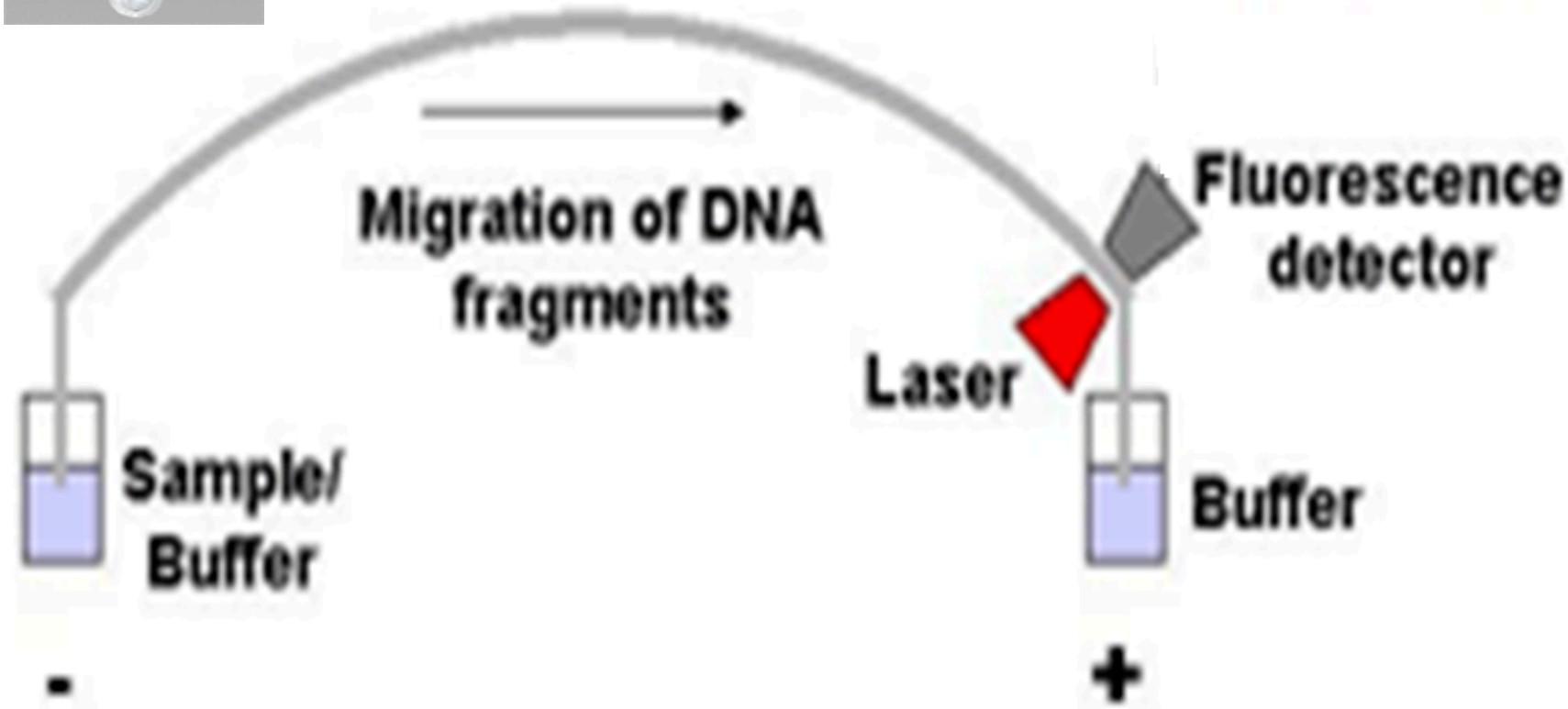

3'-O-allyl-dCTP-allyl-Bodipy-FL-510

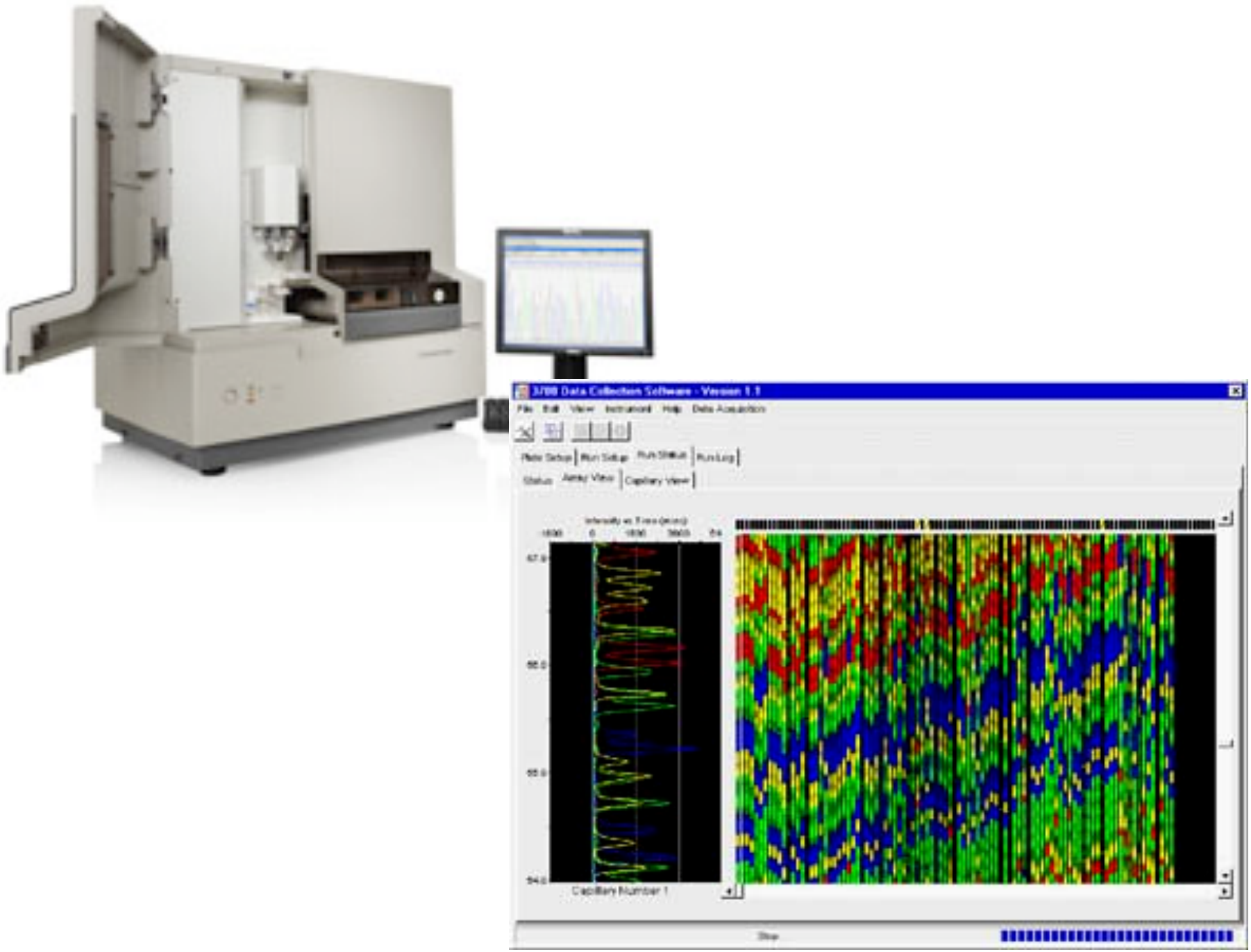

3'-O-allyl-dUTP-allyl-R6G

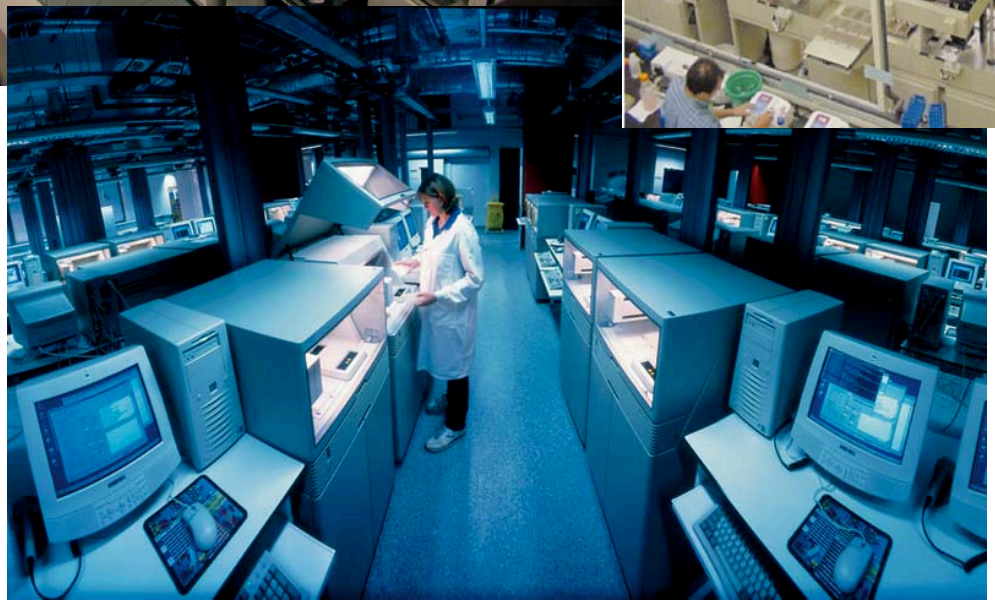

3'-O-allyl-dATP-allyl-ROX

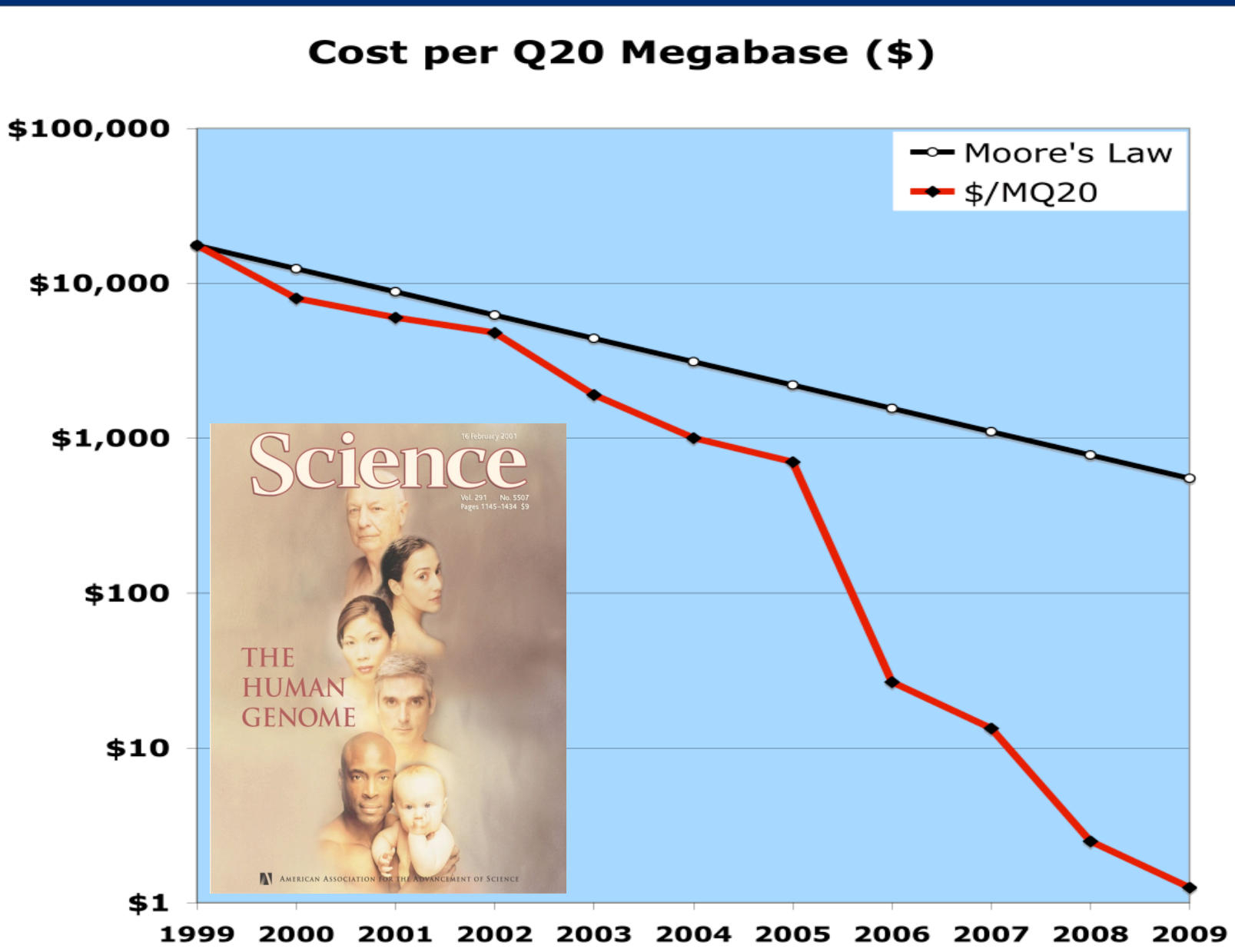

3'-O-allyl-dGTP-allyl-Bodipy-650




Migration of DNA fragments
Fluorescence detector
Laser
Sample/ Buffer
Buffer
-
+


3700 Data Collection Software - Version 1.1
Capillary Number 1




Cost per Q20 Megabase ($)
$100,000
$10,000
$1,000
$100
$10
$1
1999
2000
2001
2002
2003
2004
2005
2006
2007
2008
2009
Moore's Law
$/MQ20
Science
16 February 2001
Vol. 291 No. 5507
Pages 1145–1434 $9
THE HUMAN GENOME
American Association for the Advancement of Science

Year	Estimated cost	Technology	Reference	Machine runs	Authors	Coverage
2001	$300,000,000	Sanger (ABI)	1		251	4
2001	$100,000,000	Sanger (ABI)	2	100,000	274	5
2007	$10,000,000	Sanger (ABI)	3	100,000	31	7
2008	$2,000,000	Roche(454)	4	234	27	7
2008	$1,000,000	Illumina	5	98	48	33
2008	$500,000	Illumina	6	35	77	36
2008	$250,000	Illumina	7	40	196	30
2009	$48,000	Helicos	This work	4	3	28

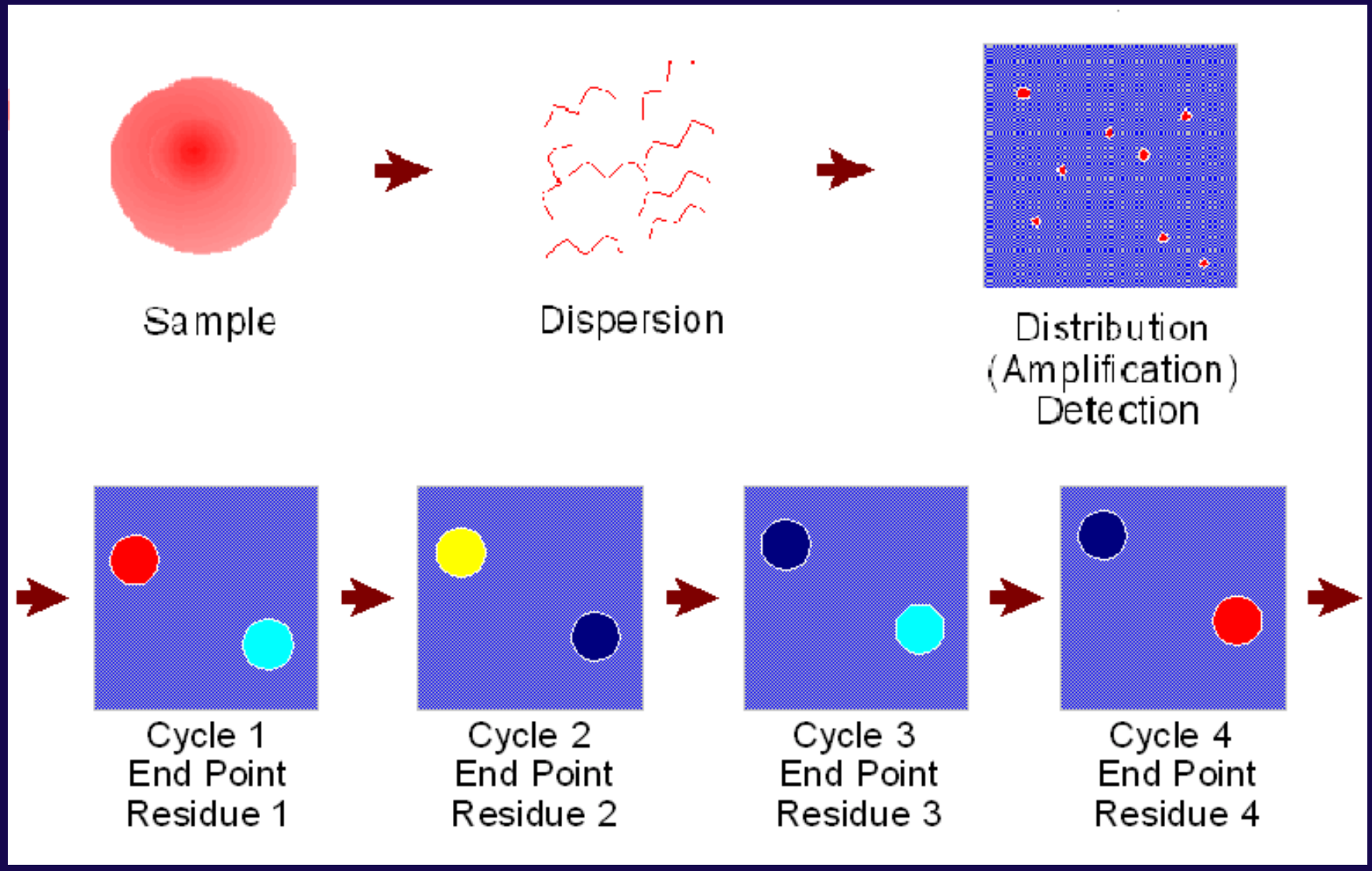
NexGen Sequencers
(Gen 2a - c)

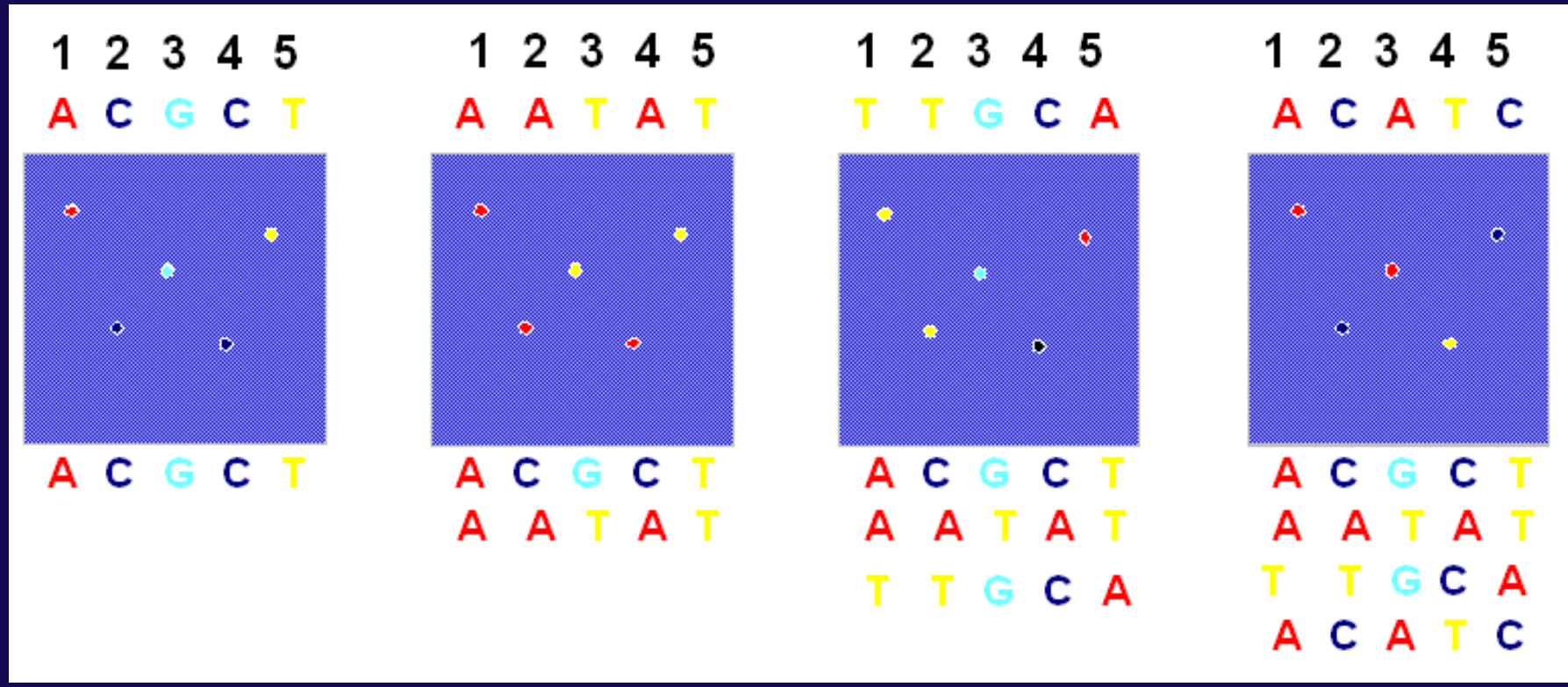
Paradigm Shift:

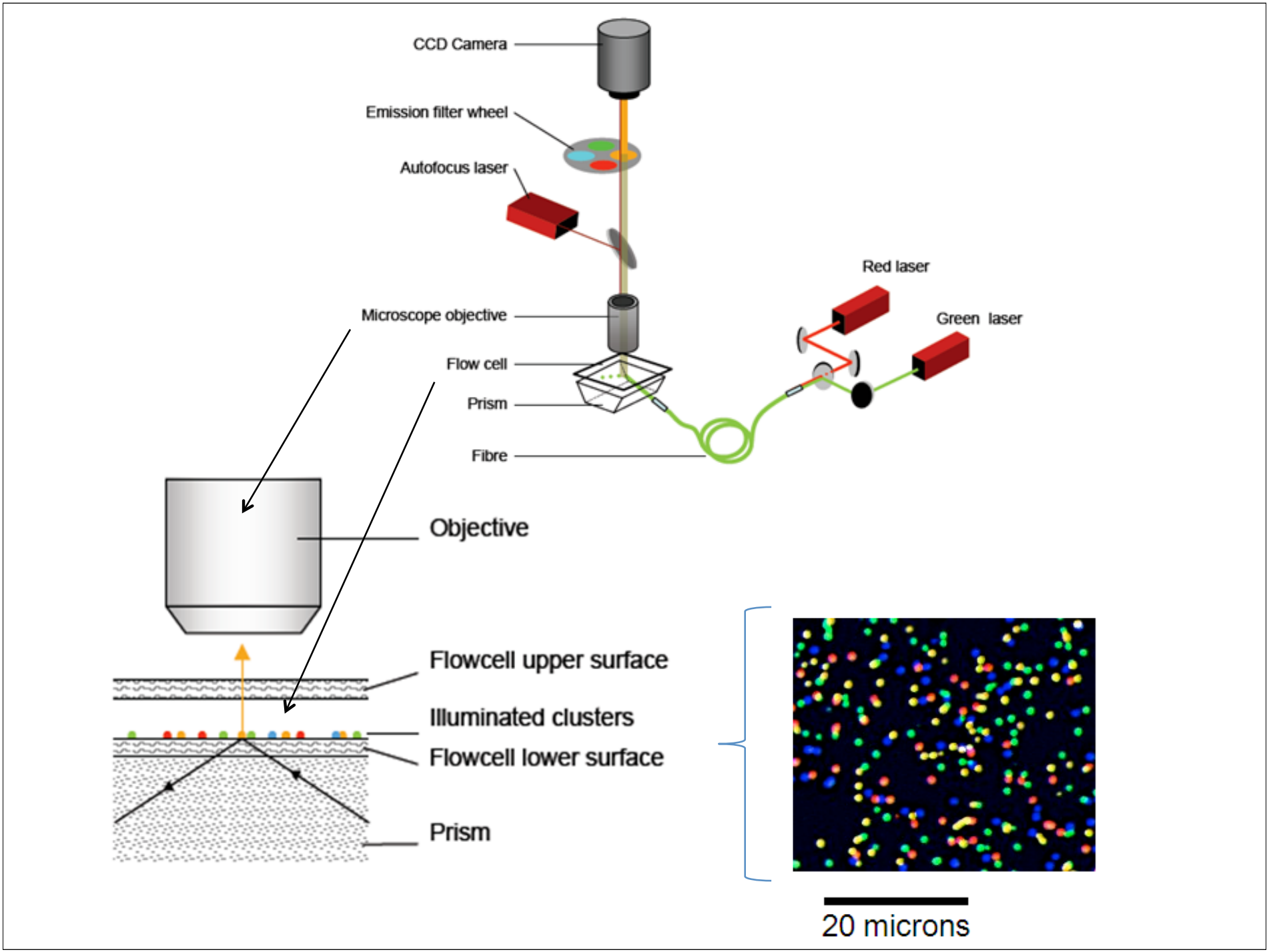
Sequencing by Synthesis

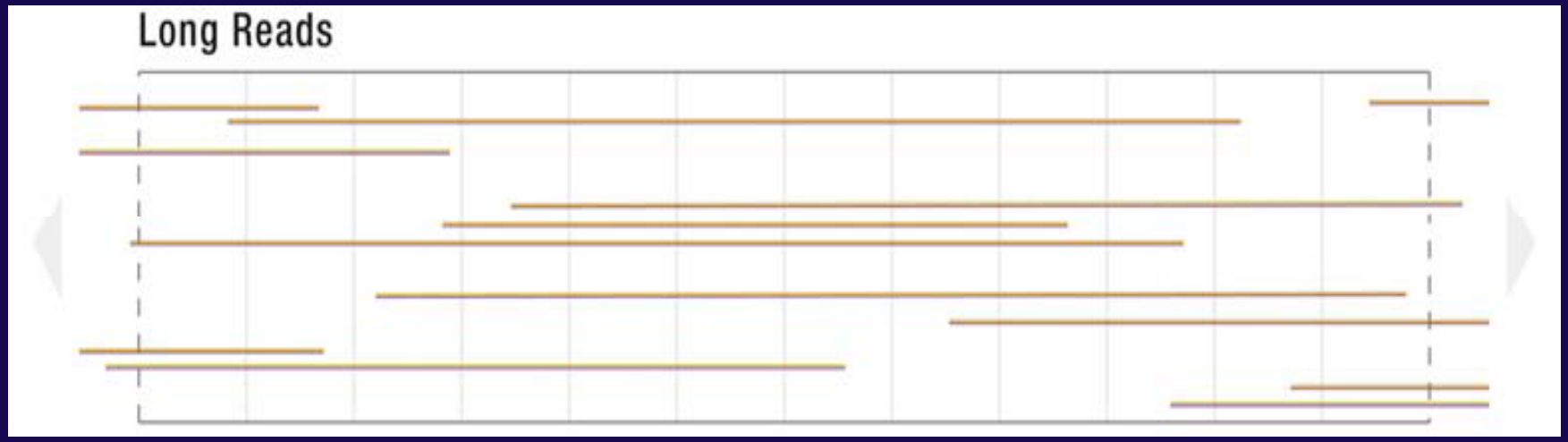
No separation step

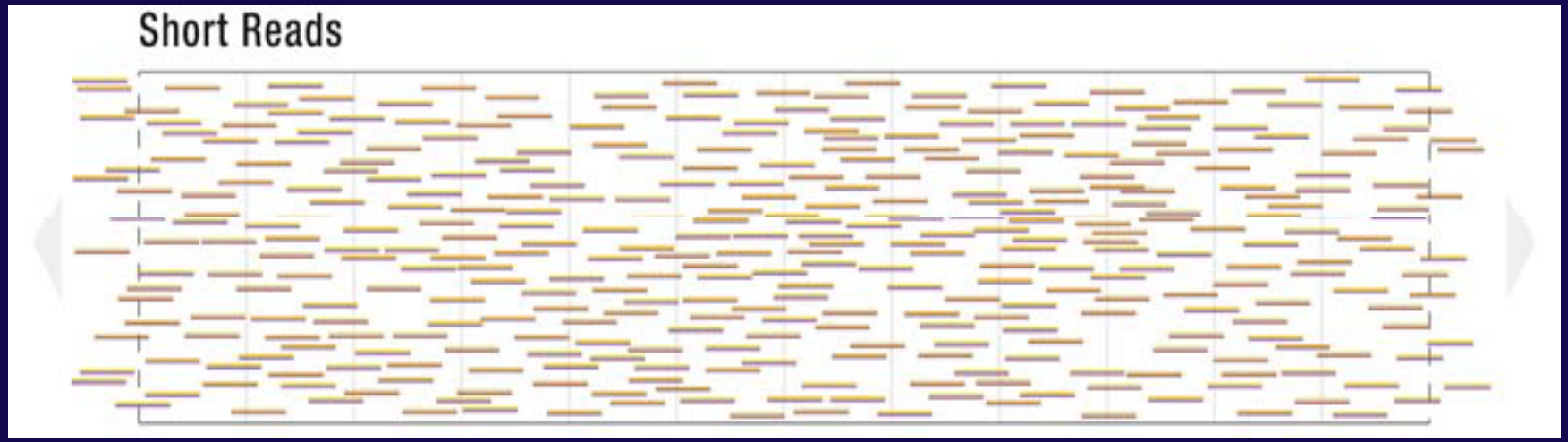
Massive parallelism


Sample
Dispersion
Distribution
(Amplification)
Detection
Cycle 1
End Point
Residue 1
Cycle 2
End Point
Residue 2
Cycle 3
End Point
Residue 3
Cycle 4
End Point
Residue 4


1 2 3 4 5
A C G C T
A C G C T
1 2 3 4 5
A A T A T
A C G C T
A A T A T
1 2 3 4 5
T T G C A
A C G C T
A A T A T
T T G C A
1 2 3 4 5
A C A T C
A C G C T
A A T A T
T T G C A
A C A T C


CCD Camera
Emission filter wheel
Autofocus laser
Red laser
Green laser
Microscope objective
Flow cell
Prism
Fibre
Objective
Flowcell upper surface
Illuminated clusters
Flowcell lower surface
Prism
20 microns

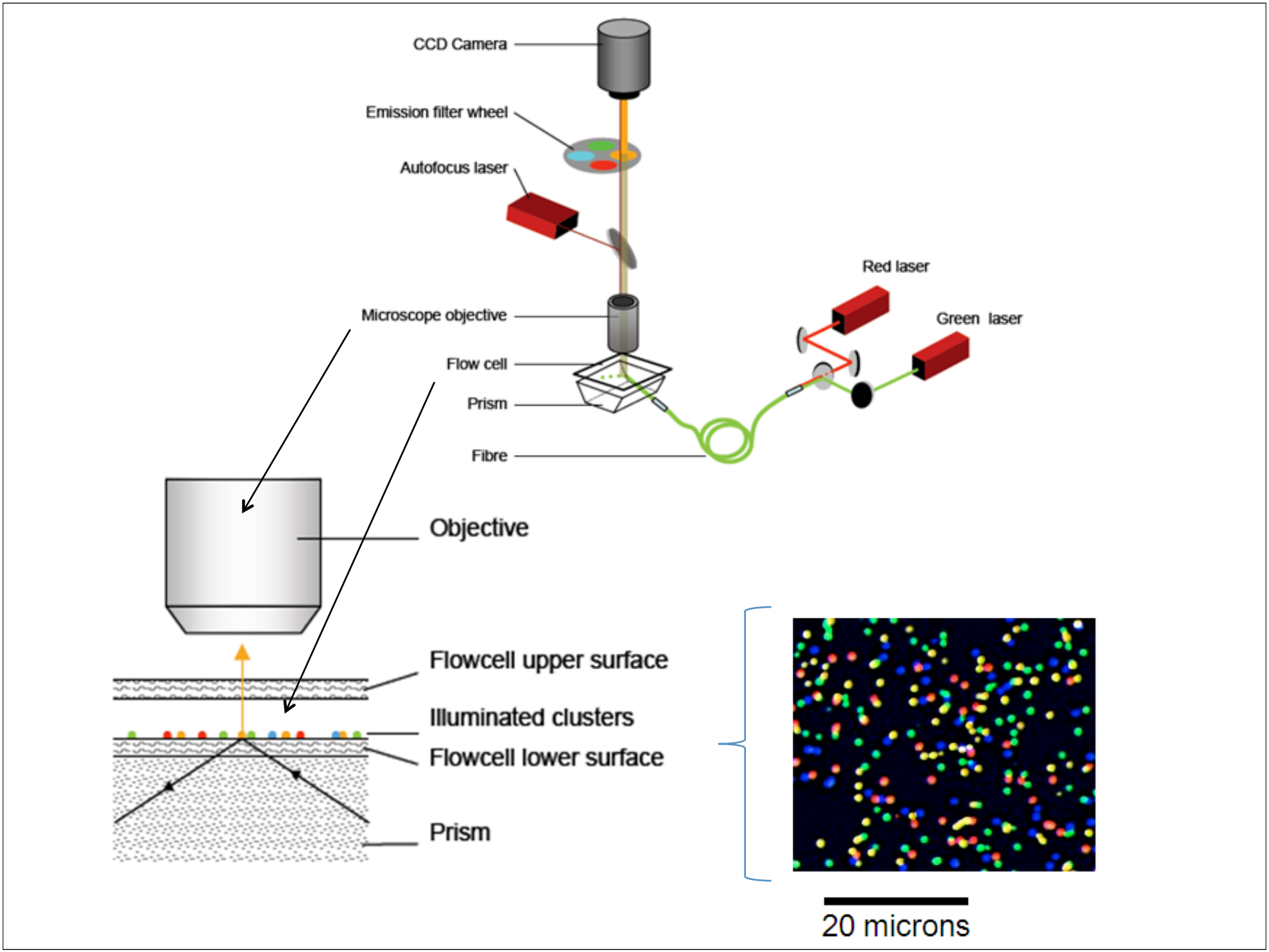

Long Reads


Short Reads

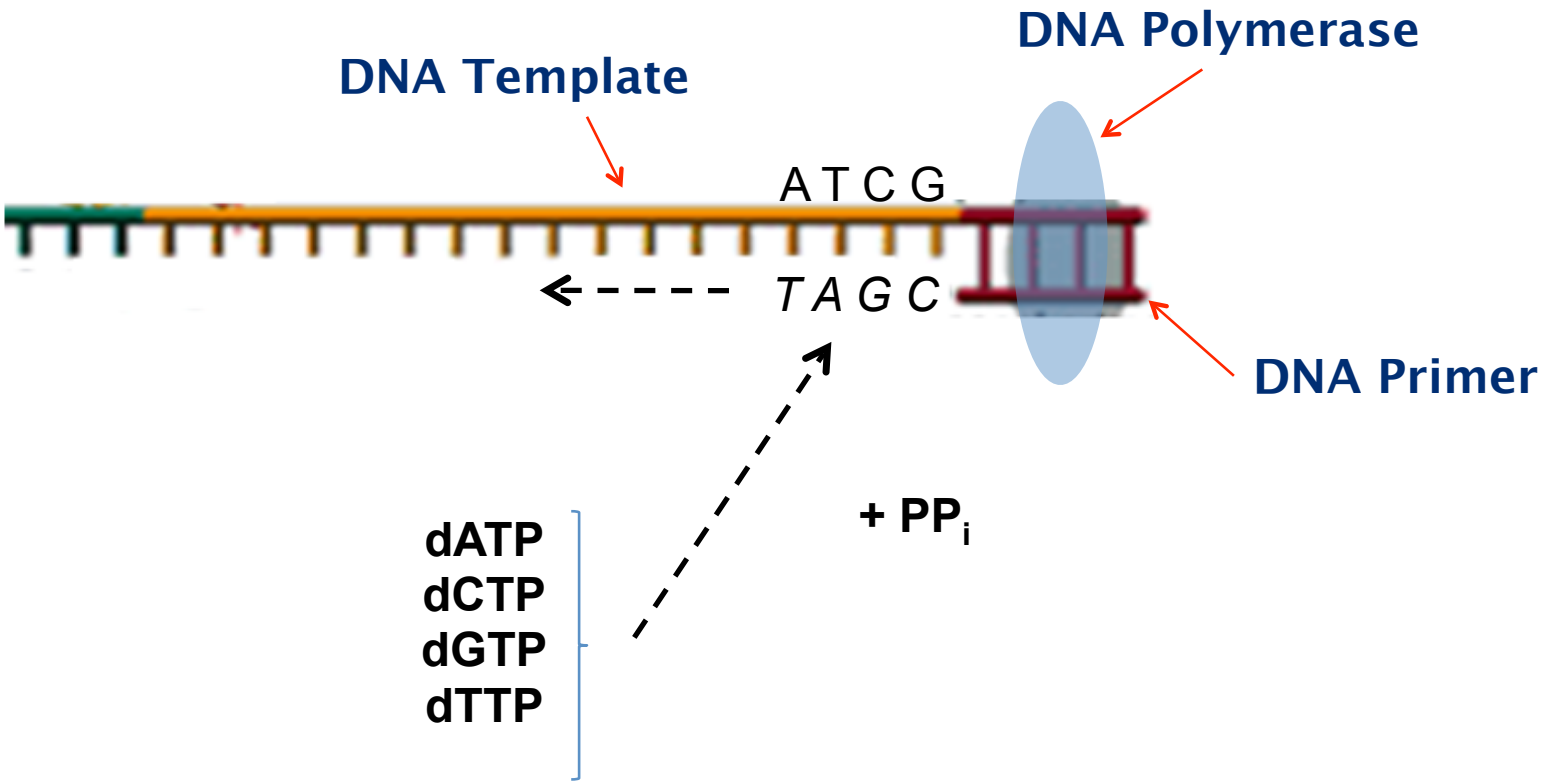
Illumina

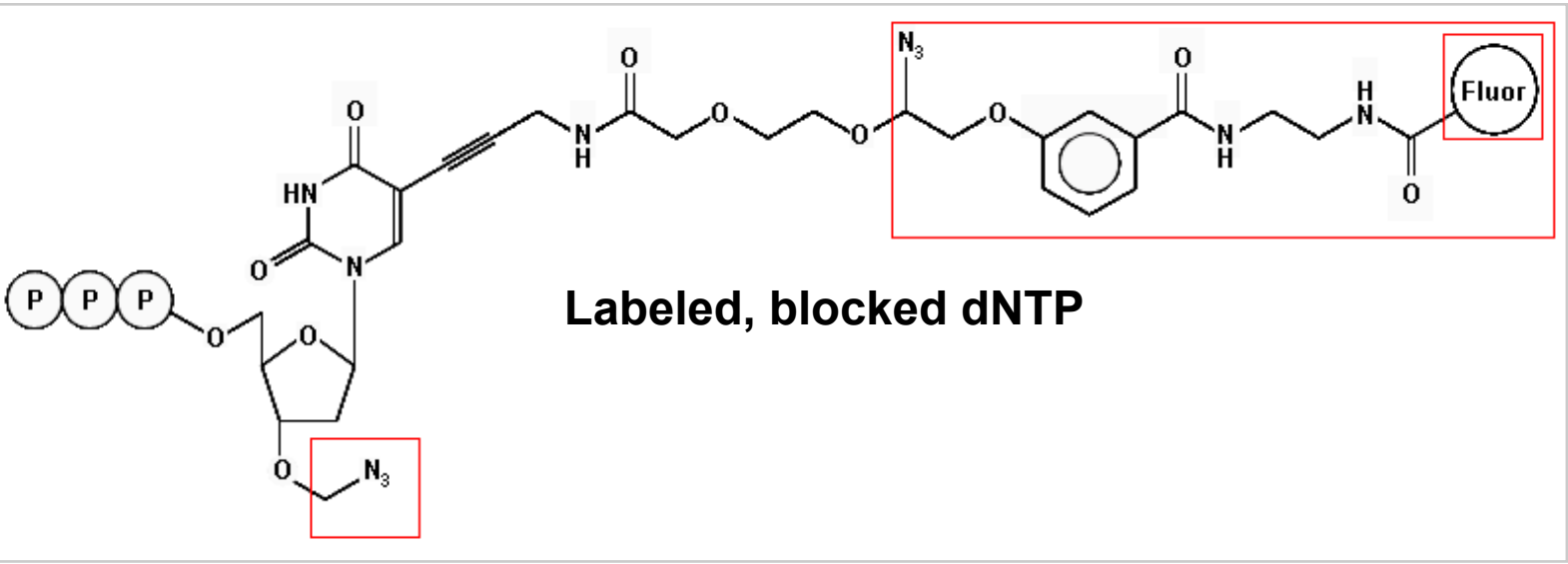
Single molecular amplified into clusters on a surface
Discontinuous (iterative) synthesis using four labels

Picowell plates not required
All four dNTPs at once
One base at a time additon

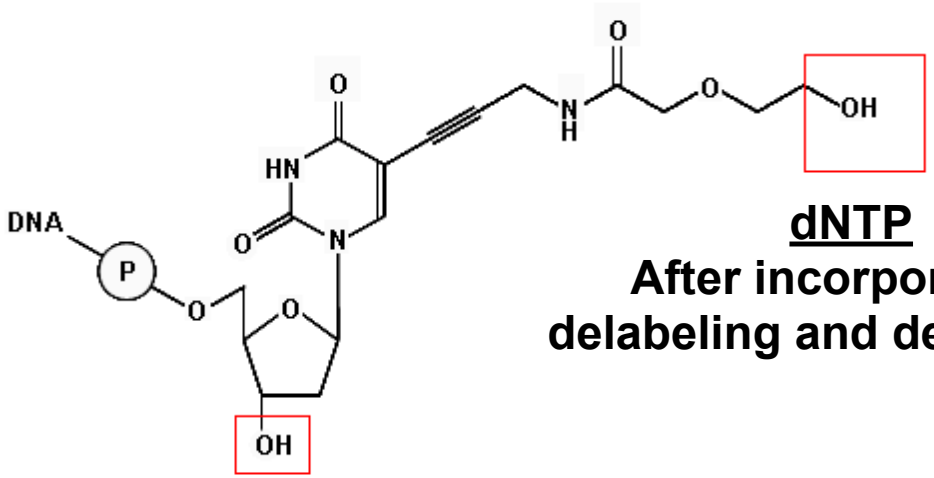

CCD Camera
Emission filter wheel
Autofocus laser
Red laser
Green laser
Microscope objective
Flow cell
Prism
Fibre
Objective
Flowcell upper surface
Illuminated clusters
Flowcell lower surface
Prism
20 microns

DNA Replication
DNA Template, DNA Primer, dNTPs, DNA Polymerase

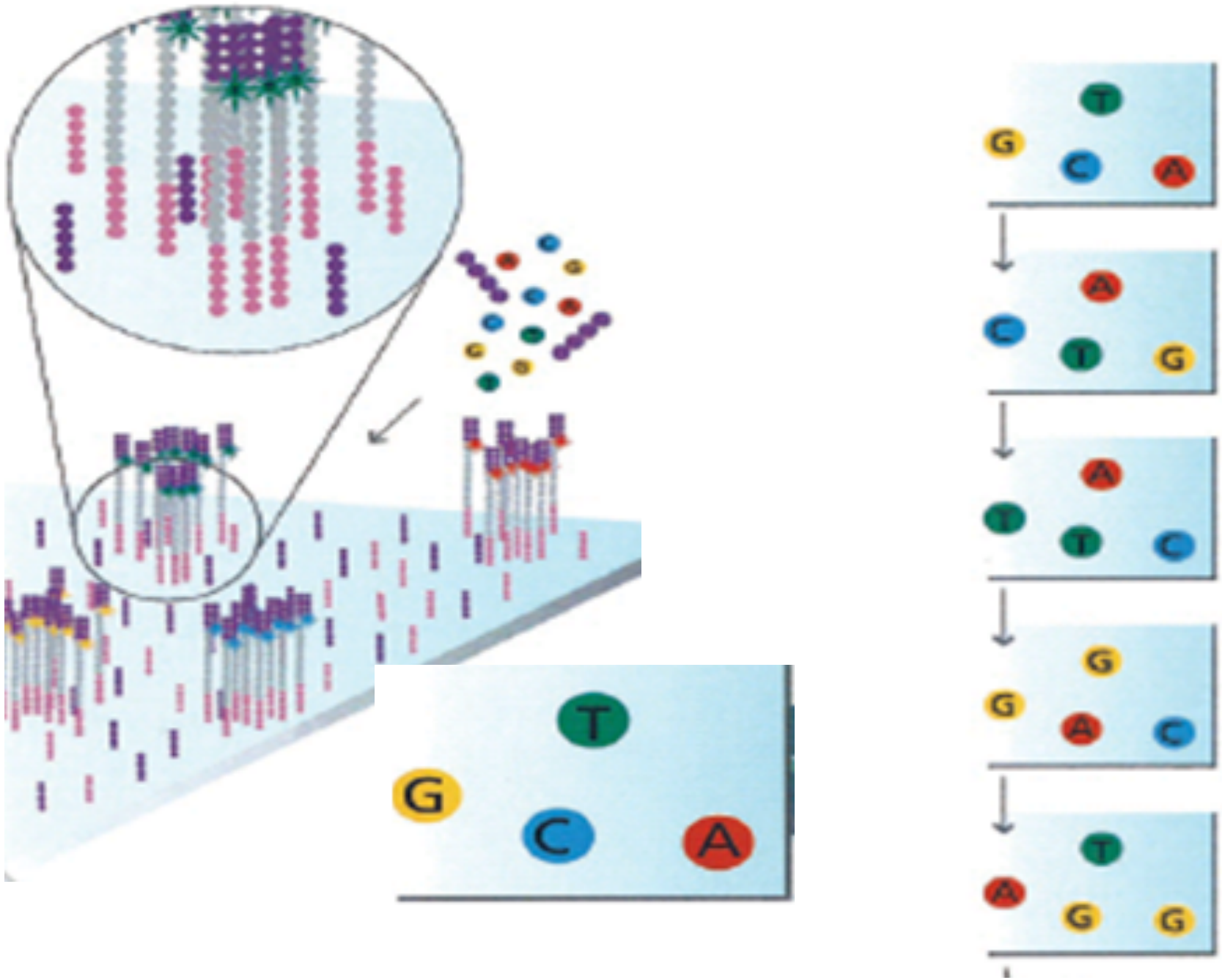




Labeled, blocked dNTP



dNTP
After incorporation
delabeling and deblocking


T
G
C
A
T
G
C
A
A
C
T
G
A
T
T
C
G
G
A
C
T
A
G
G

ILLUMINA – Just Released Model

Read Length	Run Time	Output
1 x 35 bp	~1.5 days	26–35 Gb
2 x 50 bp	~4 days	75–100 Gb
2 x 100 bp	~8 days	150–200 Gb

*Sequencing output generated with a PhiX library and cluster densities between 260,000–347,000 clusters/mm² that pass filtering on a HiSeq 2000.

Throughput

Up to 25 Gb per day for a 2 x 100 bp run.

Reads

Up to one billion clusters passing filter, and up to two billion paired-end reads.

ILLUMINA

Just Released Model

Cost: ~ $650,000 each
Capacity: ~ 10 Terabases / machine-year
~ 200 HGE / year @ 20x

BGI (Beijing Genomics Institute)
Ordered 200 machines
Total capacity: ~ 2 Petabases / year
~ 40,000 HGE / year @ 20x

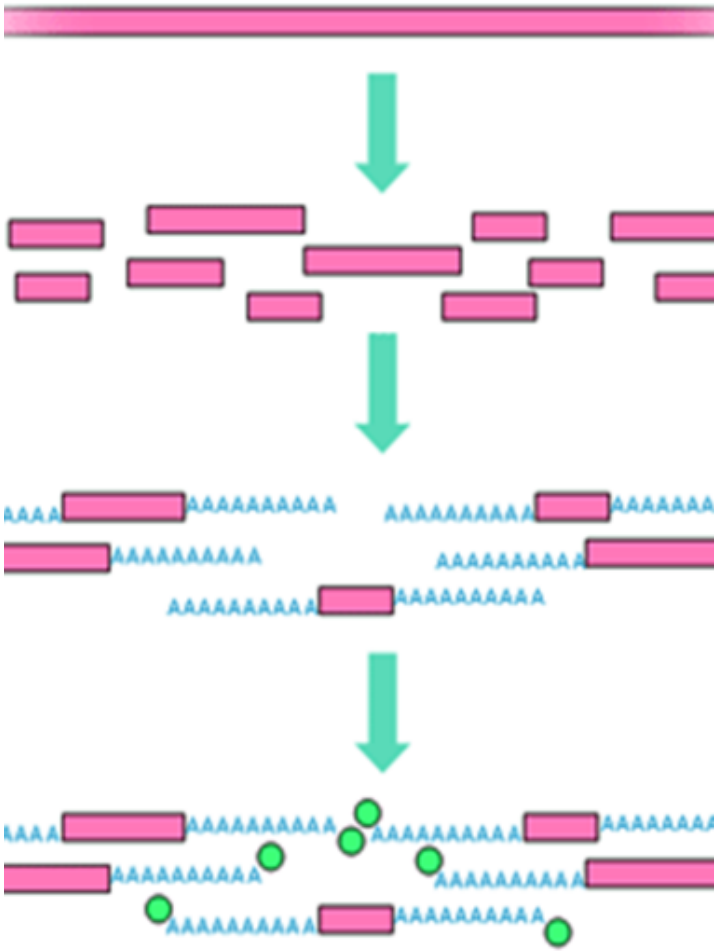
Year	Estimated cost	Technology	Reference	Machine runs	Authors	Coverage
2001	$300,000,000	Sanger (ABI)	1		251	4
2001	$100,000,000	Sanger (ABI)	2	100,000	274	5
2007	$10,000,000	Sanger (ABI)	3	100,000	31	7
2008	$2,000,000	Roche(454)	4	234	27	7
2008	$1,000,000	Illumina	5	98	48	33
2008	$500,000	Illumina	6	35	77	36
2008	$250,000	Illumina	7	40	196	30
2009	$48,000	Helicos	This work	4	3	28

Helicos

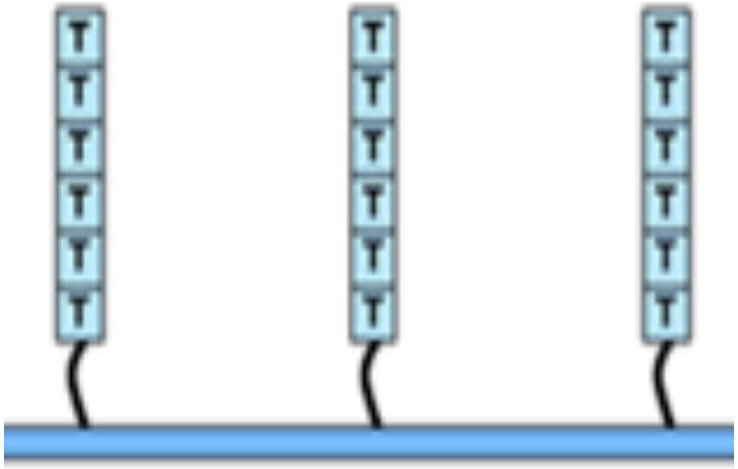
Single Molecule Sequencing
One fluorescent label
One base at a time

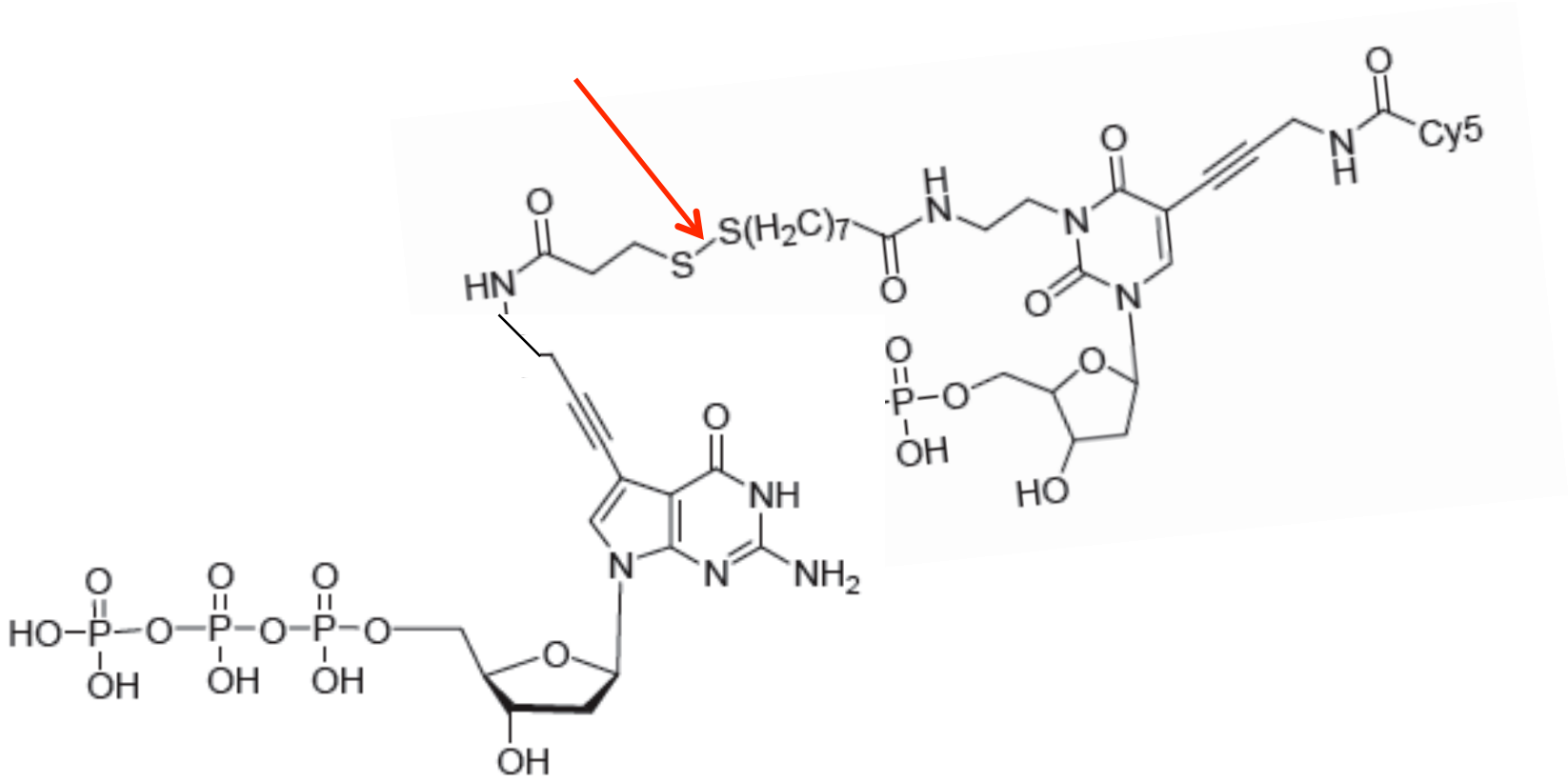
No picowells
No amplification

Sample preparation

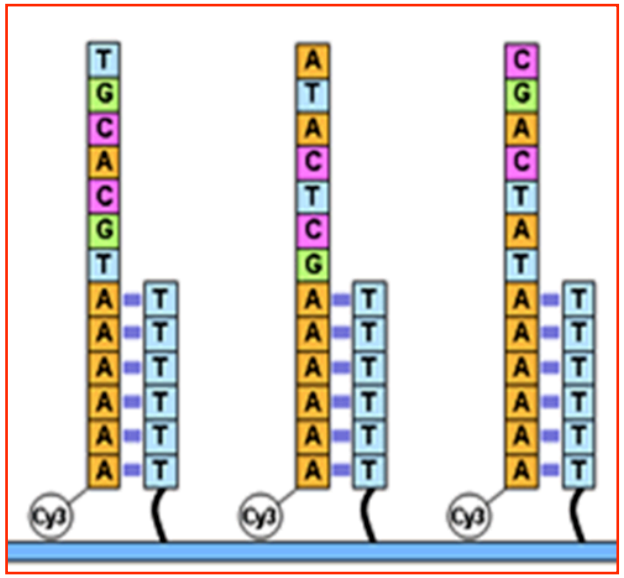


Prepared surface

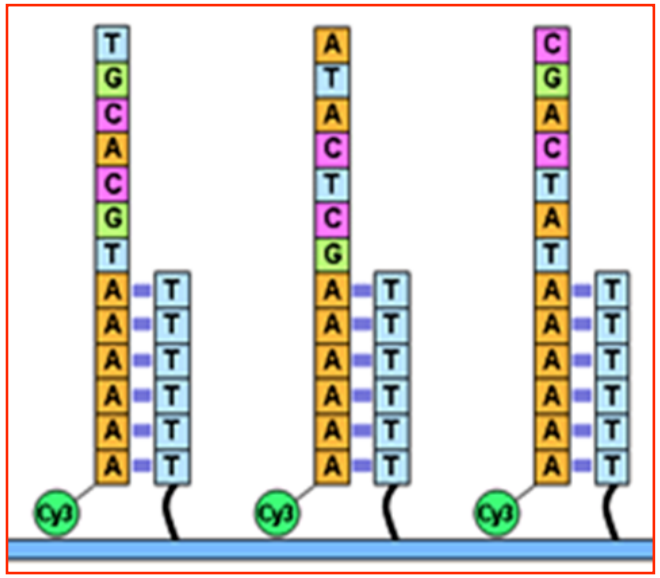



Cy5
S(H2C)7
HN
NH
NH2
HO
OH

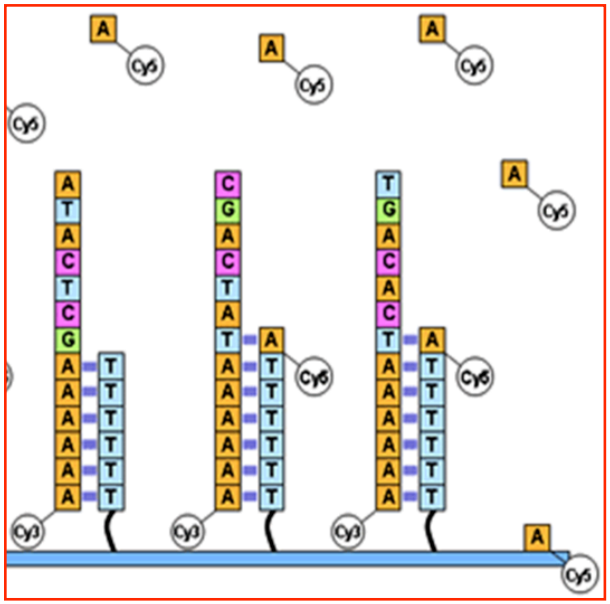
1


T G C A C G T A A A A A A A
T T T T T T
A T A C T C G A A A A A A A
T T T T T T
C G A C T A T A A A A A A A
T T T T T T
Cy3
Cy3
Cy3

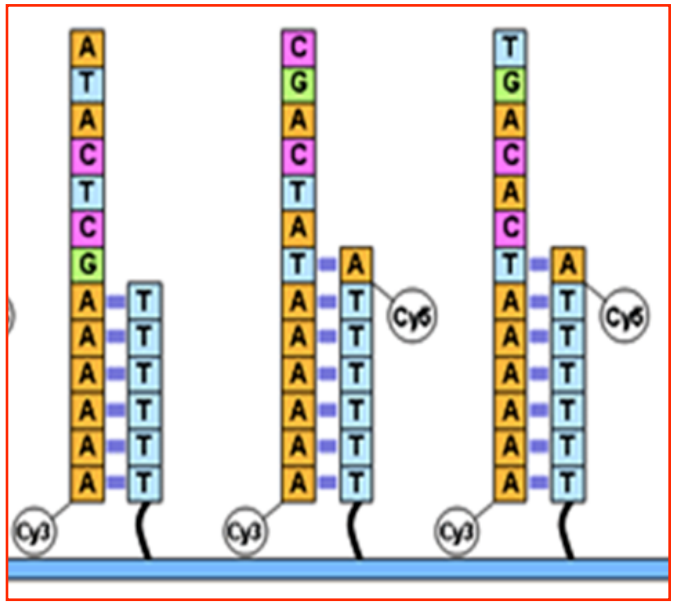
2


T G C A C G T A A A A A A A
T T T T T T
A T A C T C G A A A A A A A
T T T T T T
C G A C T A T A A A A A A A
T T T T T T
Cy3
Cy3
Cy3

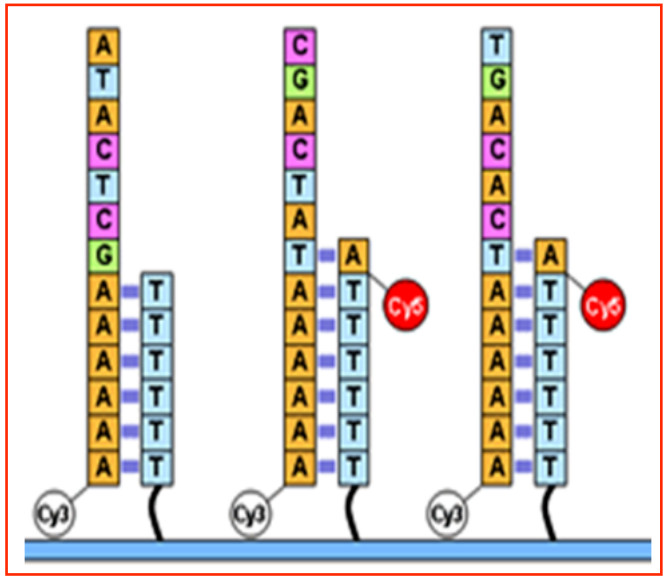
3

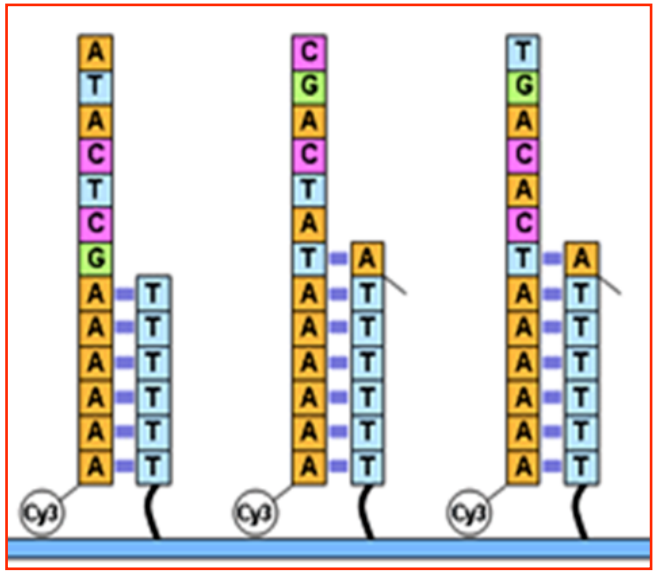

A
Cy5
A
Cy5
A
Cy5
Cy5
A
Cy5
A T A C T C G A A A A A A A
T T T T T T
C G A C T A T A A A A A A A
T T T T T T
T G A C A C T A A A A A A A
T T T T T T
A
Cy5
A
Cy5
A
Cy5
A
Cy5
Cy3
Cy3
Cy3

4


A T A C T C G A A A A A A A
T T T T T T
C G A C T A T A A A A A A A
T T T T T T
T G A C A C T A A A A A A A
T T T T T T
A
T T
Cy5
A
T T
Cy5
A
T T
Cy5
Cy3
Cy3
Cy3

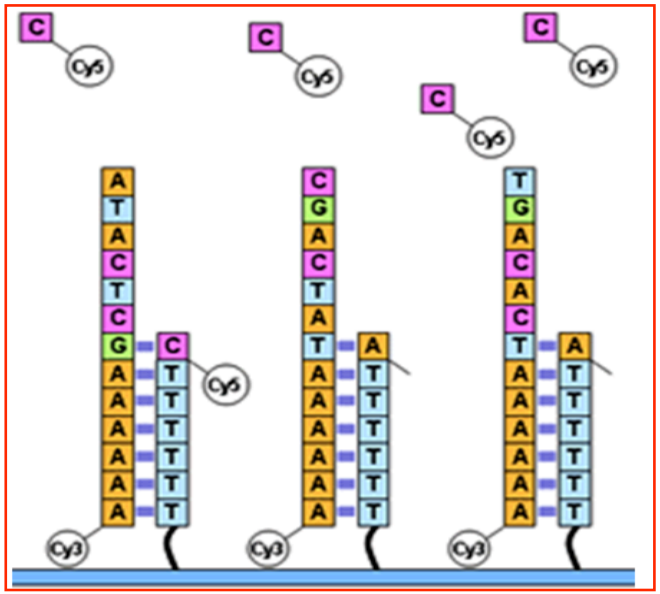
5





6

7



Detect incorporation

Repeat cycle for G and T

Repeat in turn for A, C, G and T in turn continuing to end of run

Etc.

System Configuration

A real production-level genetic analyzer

Helicos

Flow Cell (X 2)
- 2.8B strands
- 50 channels (50 samples) 56M strands

HeliScope Sequencer (>1GB/hr)
- Laser Illumination
- CCD Camera
- Microfluidics
- High-speed stage
- Instrument-control computer
- System UPS

60 tbytes

HeliScope Analysis Engine
- Multi-blade tower
- 28 terabytes data storage
- Near real-time data processing
- No MB/hr penalty to acquire 'alignable data'
- Scalable to $1,000 genome performance

Helicos SMS Performance

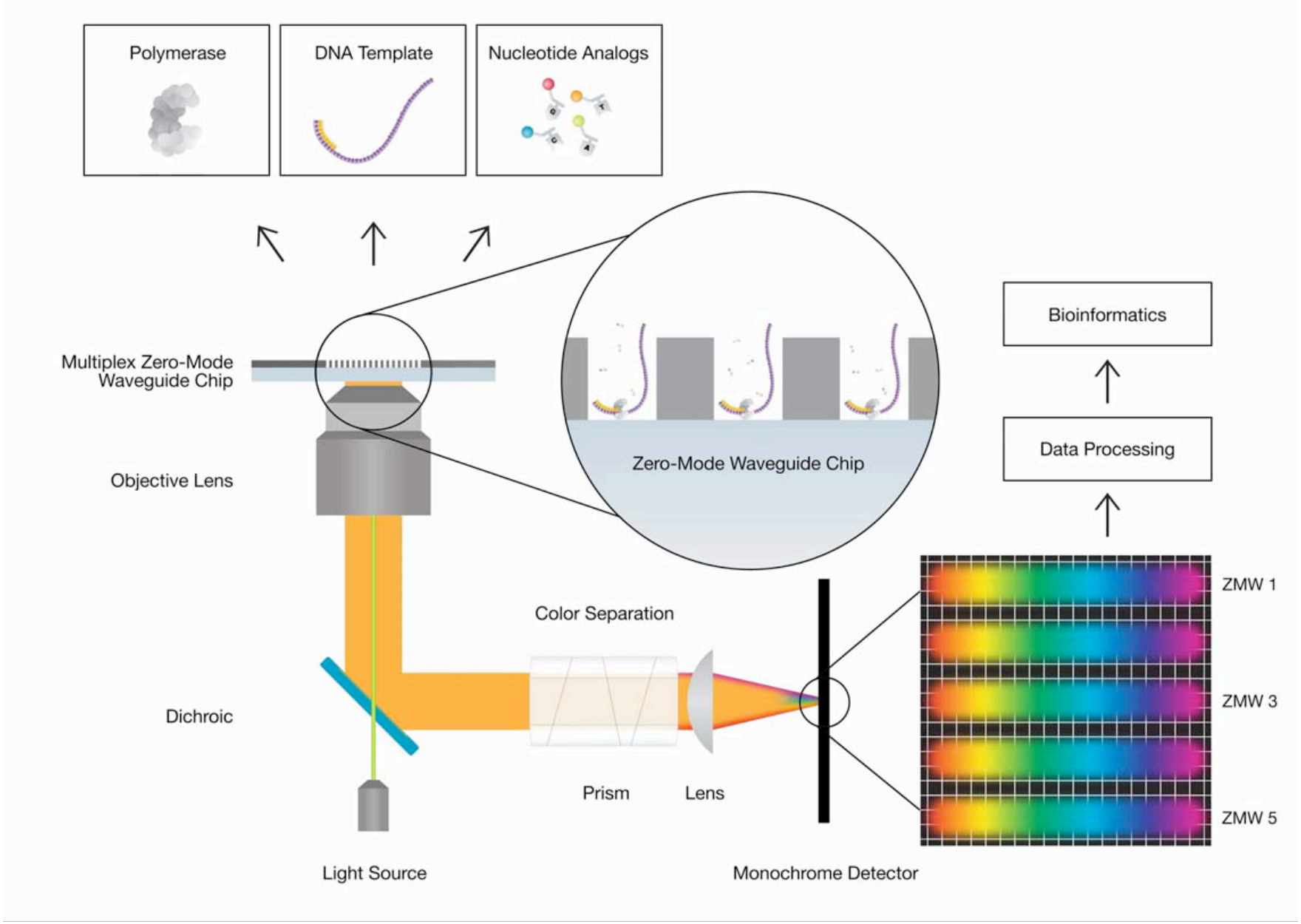
- 2 Flow Cells / 50 channels
- Up to >100,000 samples per run (multiplexed)
- > 1 Gigabases per hour (imaging system design)
- 600M to 800M usable strands / run (25 – 5,000 bases each)
- (> 100,000,000 strands / cm^2)
- **21 to 28 Gigabases / run**
- **105 to 140 Megabases / hour**
- Read length - 25 to 55 bases (30-35 bases, average)
- Raw Error Rate - <5% (~0.5% for substitutions)
- Consensus accuracy at >20X coverage - 99.995%
- ~ 1 Gigabase / run at 99.995% accuracy
- **8 days for a "30 quad" run**

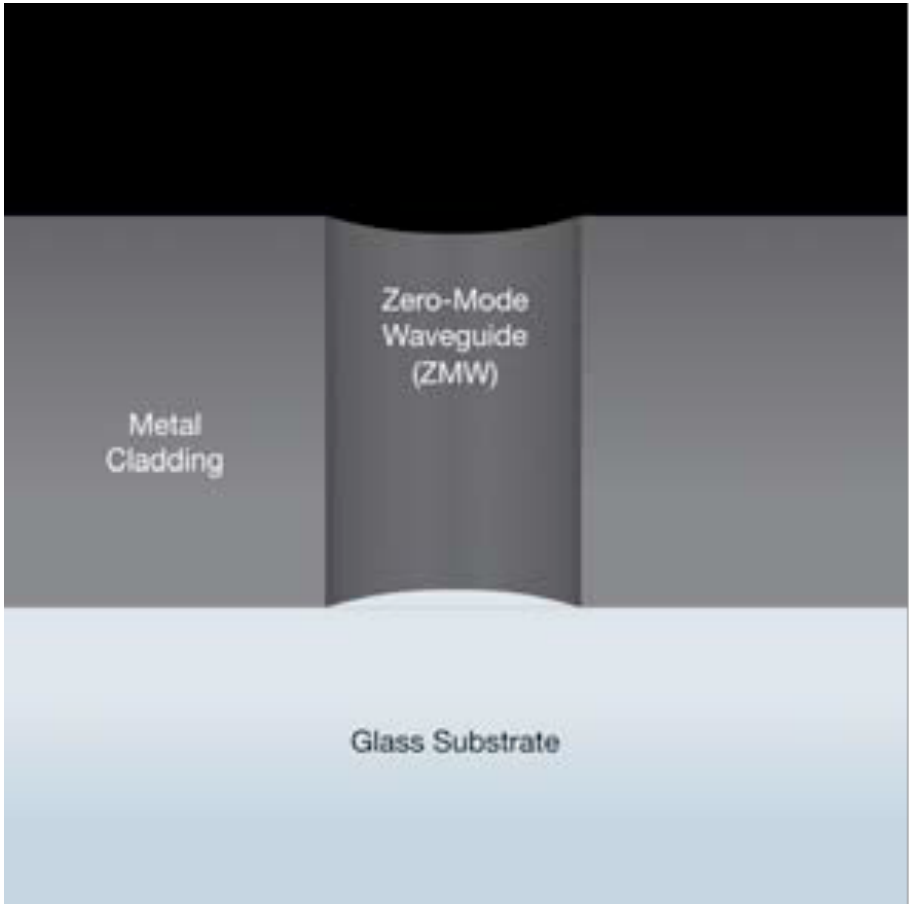
Year	Estimated cost	Technology	Reference	Machine runs	Authors	Coverage
2001	$300,000,000	Sanger (ABI)	1		251	4
2001	$100,000,000	Sanger (ABI)	2	100,000	274	5
2007	$10,000,000	Sanger (ABI)	3	100,000	31	7
2008	$2,000,000	Roche(454)	4	234	27	7
2008	$1,000,000	Illumina	5	98	48	33
2008	$500,000	Illumina	6	35	77	36
2008	$250,000	Illumina	7	40	196	30
2009	$48,000	Helicos	This work	4	3	28

Pacific Biosciences

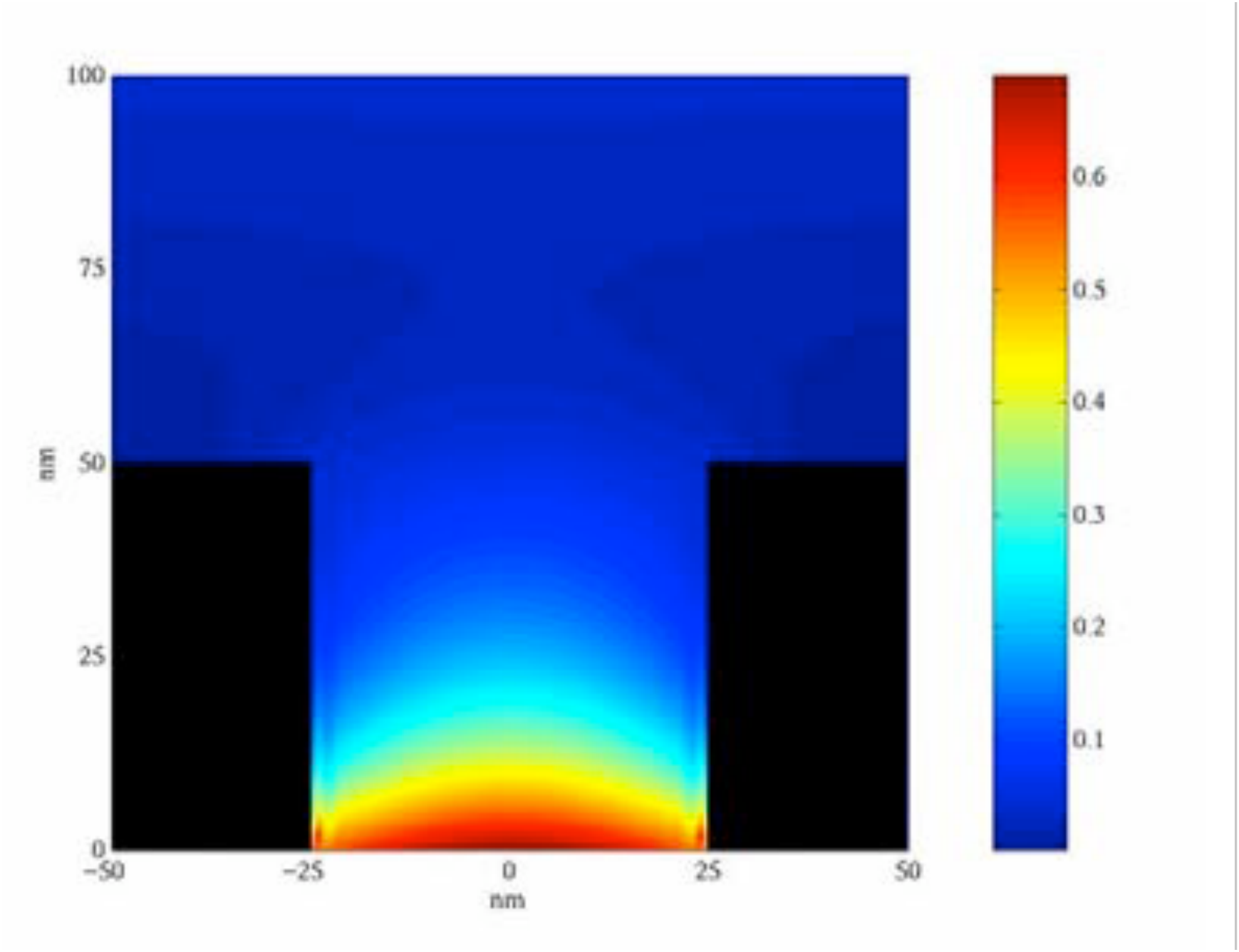
Real Time Single Molecule Sequencing
In Zero Mode Waveguides
Four colors, labels on the gamma phosphates

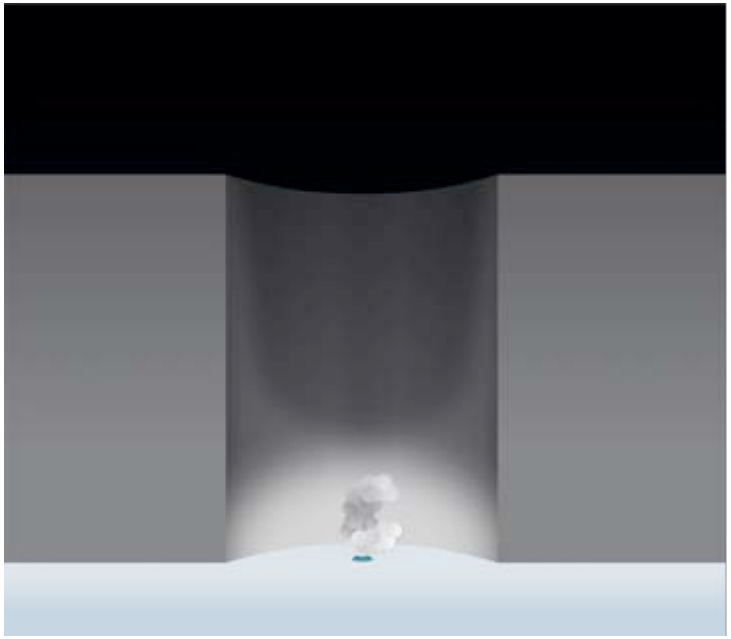
No amplification
No step wise chemistry
No wash steps
Continuous synthesis

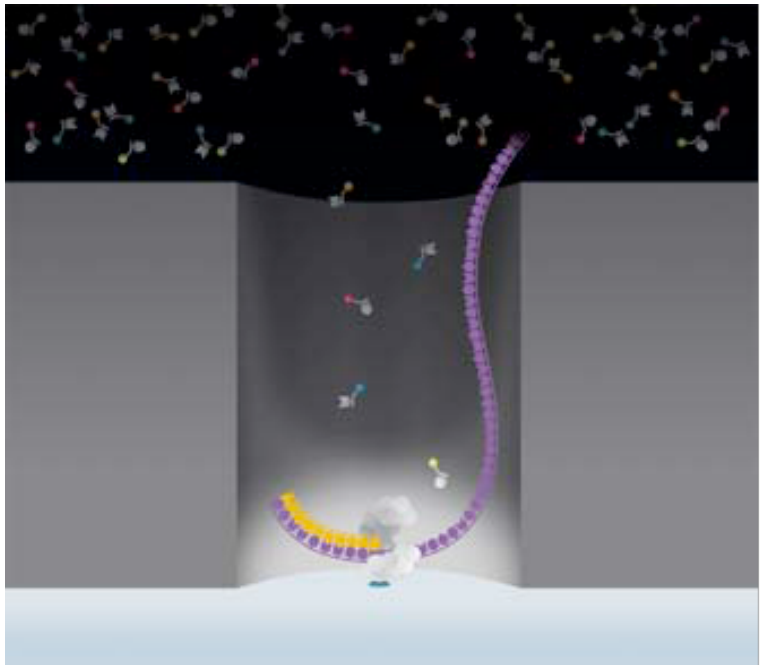

Polymerase
DNA Template
Nucleotide Analogs
Multiplex Zero-Mode Waveguide Chip
Objective Lens
Zero-Mode Waveguide Chip
Bioinformatics
Data Processing
Color Separation
Dichroic
Prism
Lens
ZMW 1
ZMW 3
ZMW 5
Light Source
Monochrome Detector

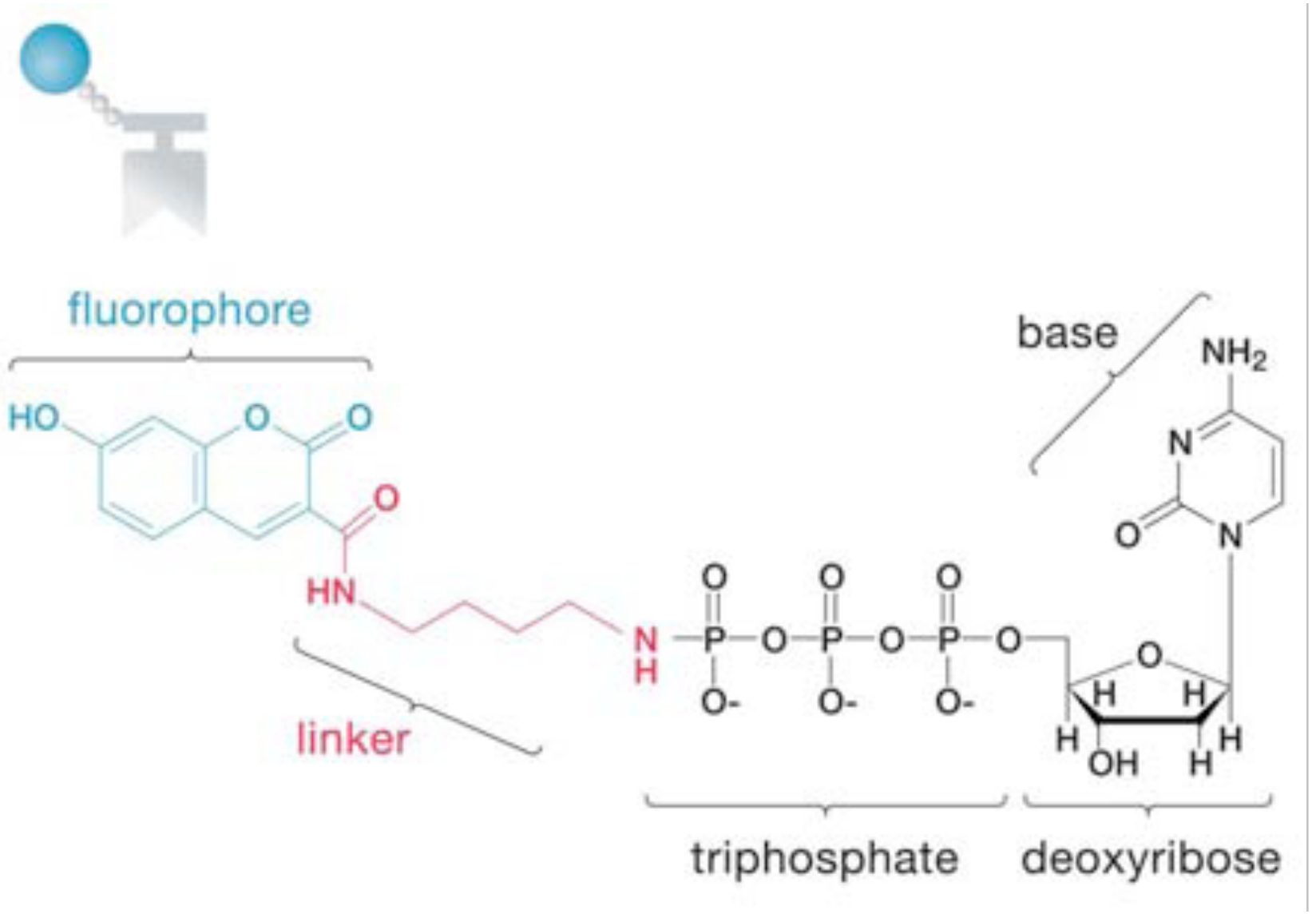


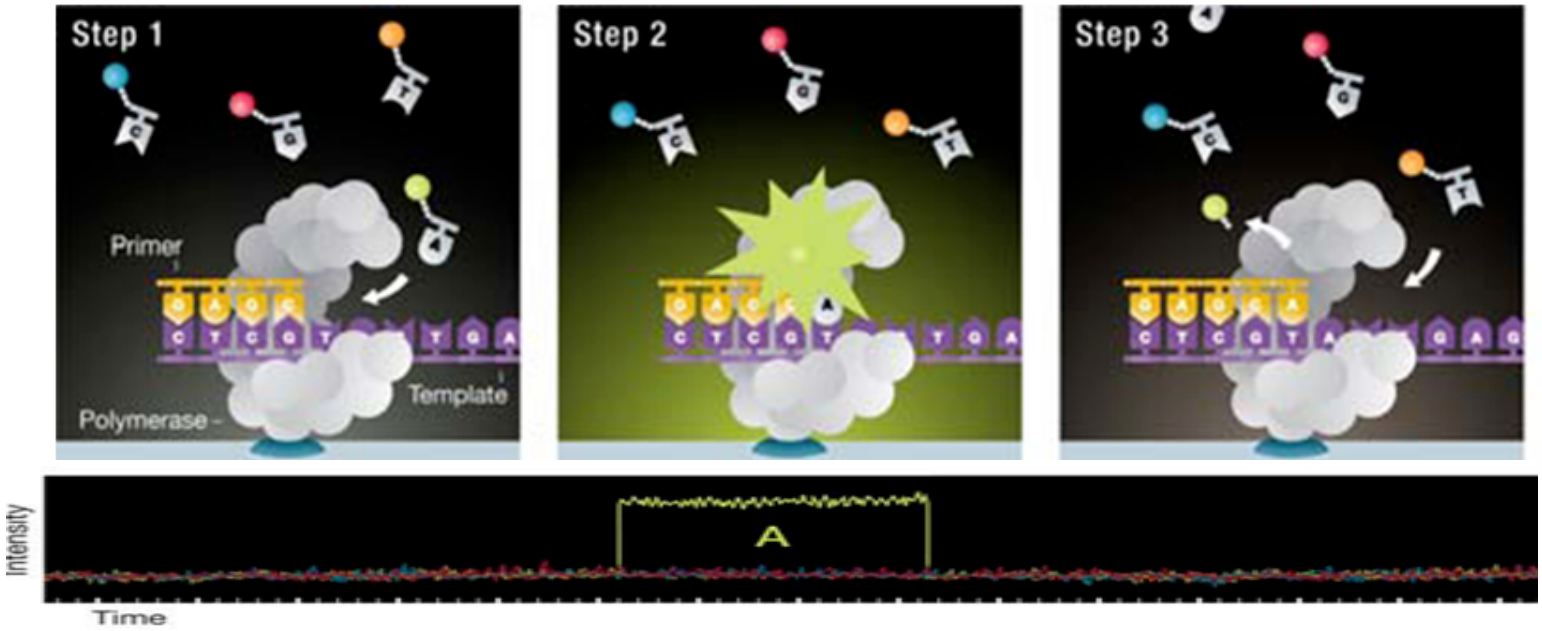
ZMW are cylinders on the order of 50 nanometers in diameter
They are fabricated in a 100nm metal film pn a silicon dioxide substrate.
Each ZMW has a detection volume of roughly 20 zeptoliters (10^{-21} liters).
Enough room for 600,000 molecules of liquid water at room temperature
They are in essence nanophotonic single molecule visualization chambers

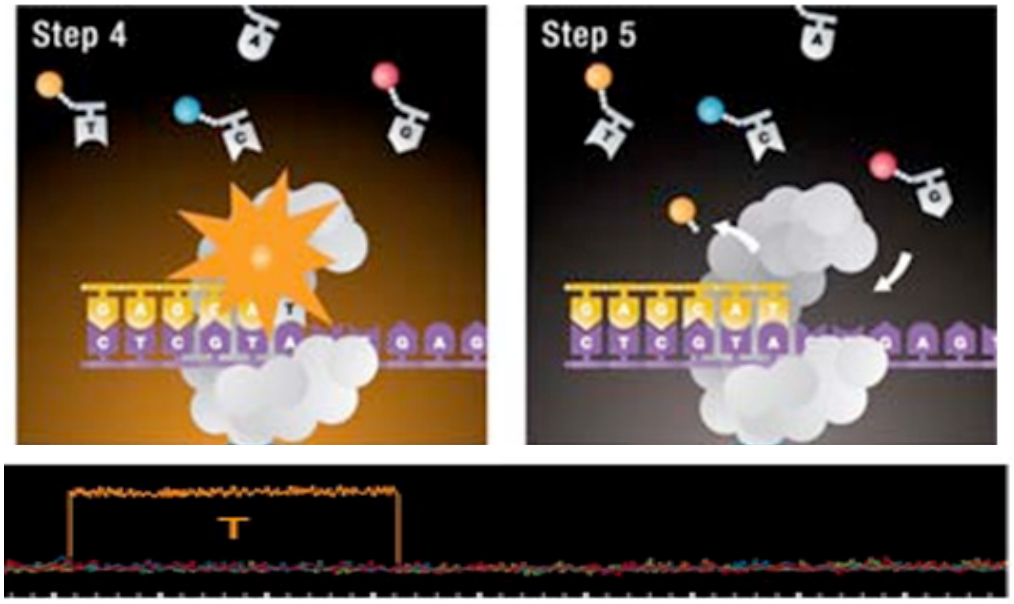

100
75
50
25
0
nm
−50
−25
0
25
50
nm
0.6
0.5
0.4
0.3
0.2
0.1






fluorophore
base
linker
triphosphate
deoxyribose

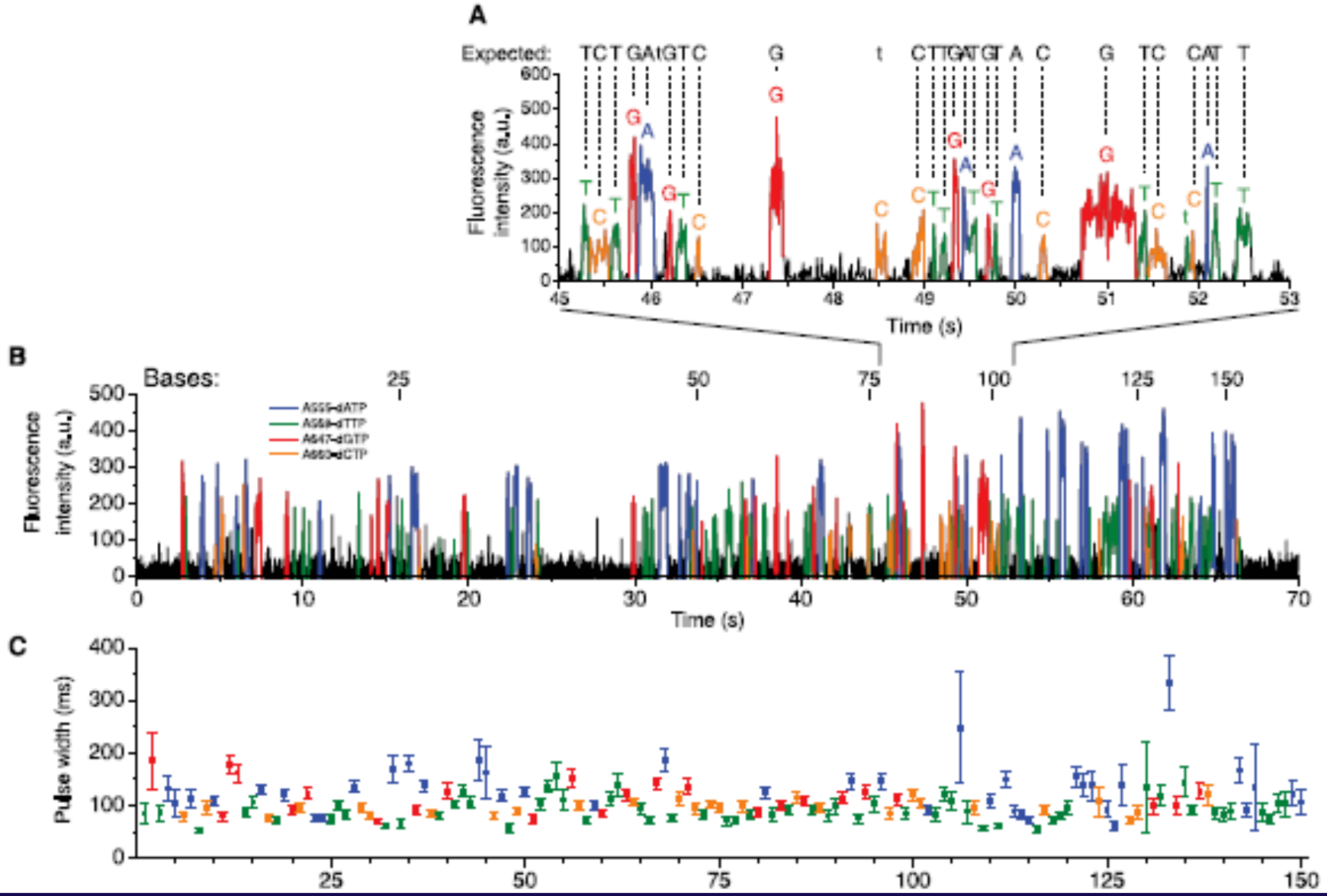

Step 1
Primer
Polymerase
Template
Step 2
Step 3
Intensity
A
Time


Step 4
Step 5
T

Etc.

PB Animation

http://www.pacificbiosciences.com/
index.php?q=smrt-technology-at-a-glance


A
Expected:
TCT GAtGTC
G
t
CTTGATGTA C
G
TC
CAT
T
Fluorescence intensity (a.u.)
Time (s)
B
Bases:
A555-dATP
A568-dTTP
A647-dGTP
A660-dCTP
Fluorescence intensity (a.u.)
Time (s)
C
Pulse width (ms)

Pac Bio

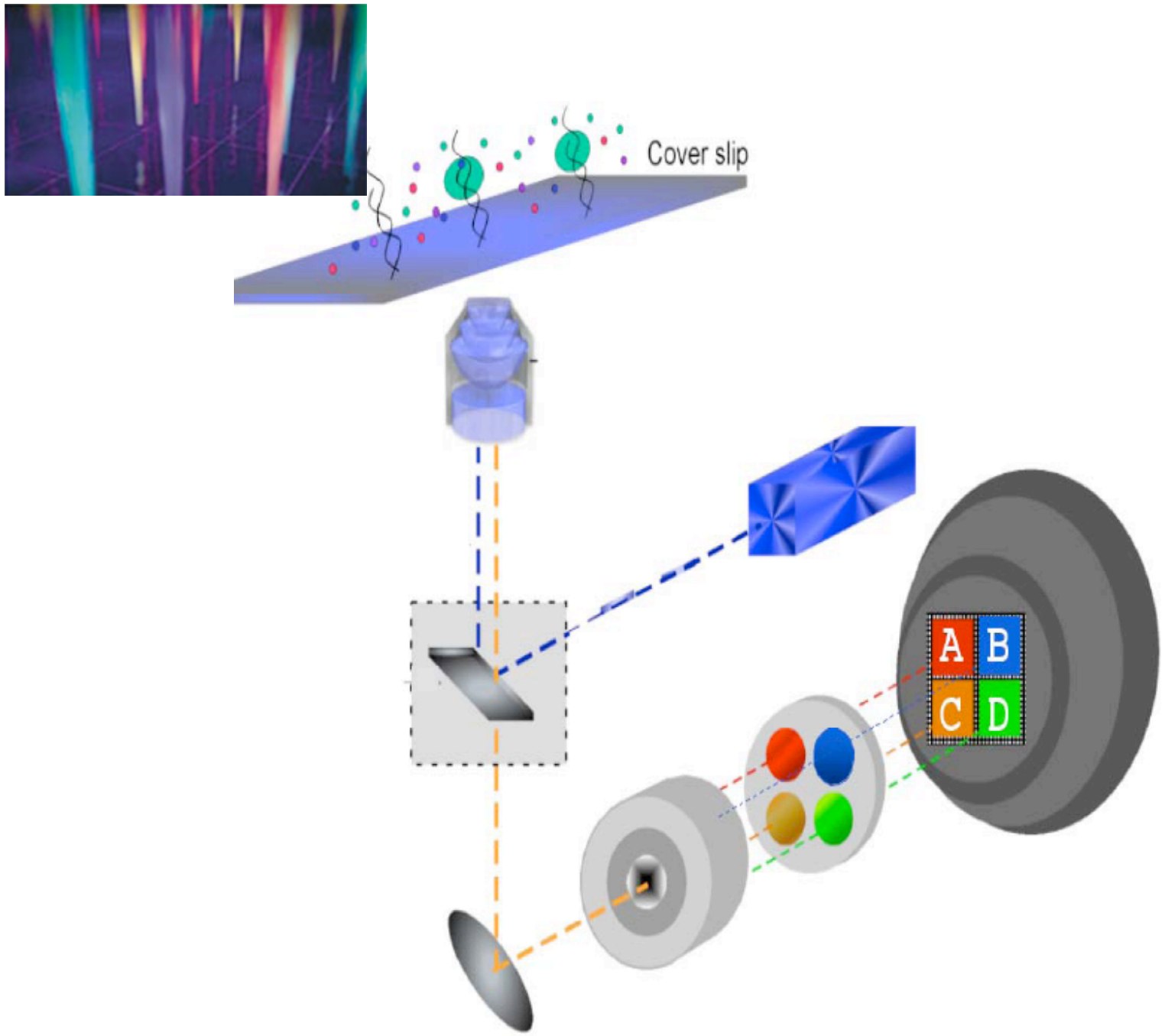
- *No commercial device as yet*
- *Sub $1,000 HGE expected*
- *Sub 1 HGE/4 hours expected*
- ***Maybe much better***

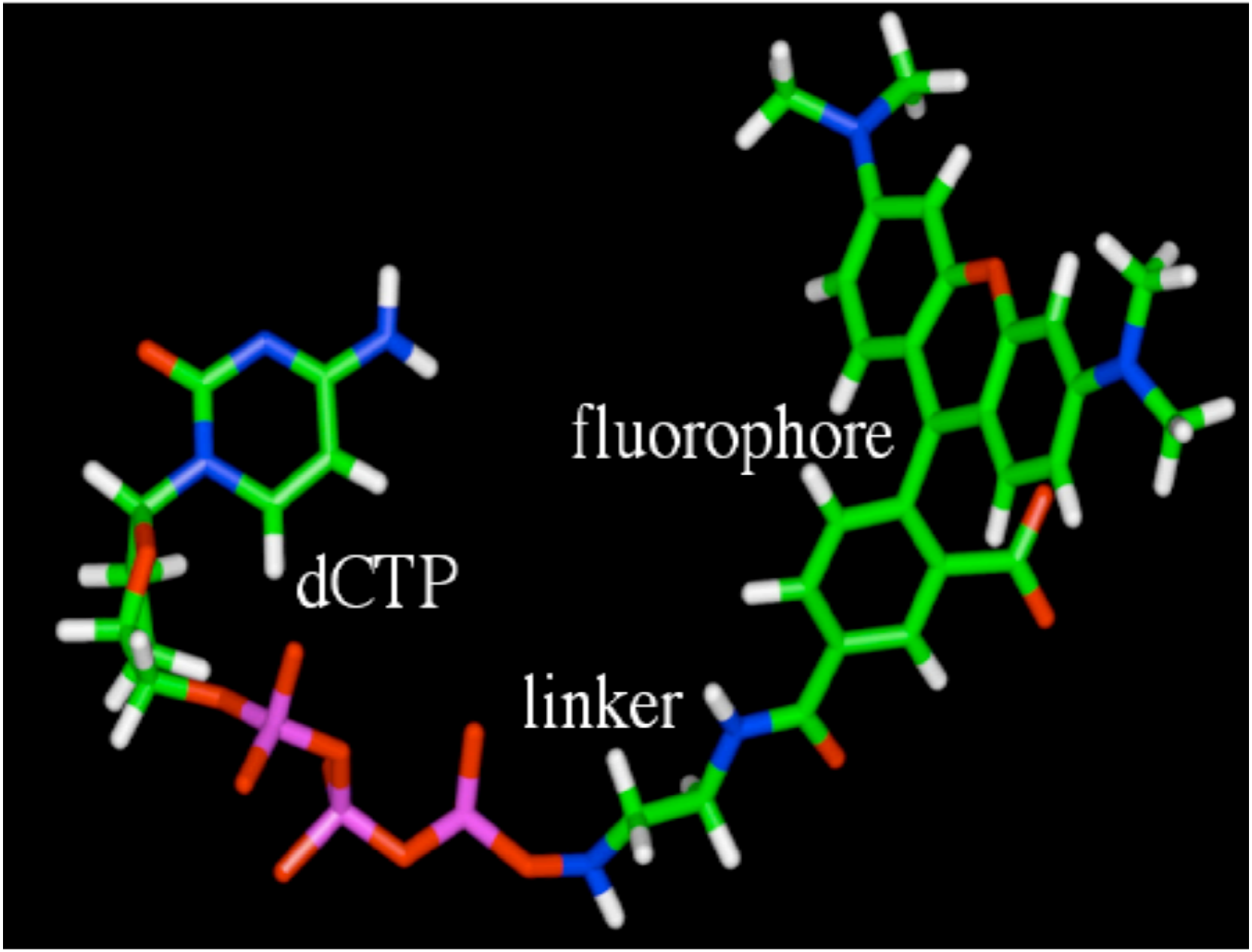
VisiGen

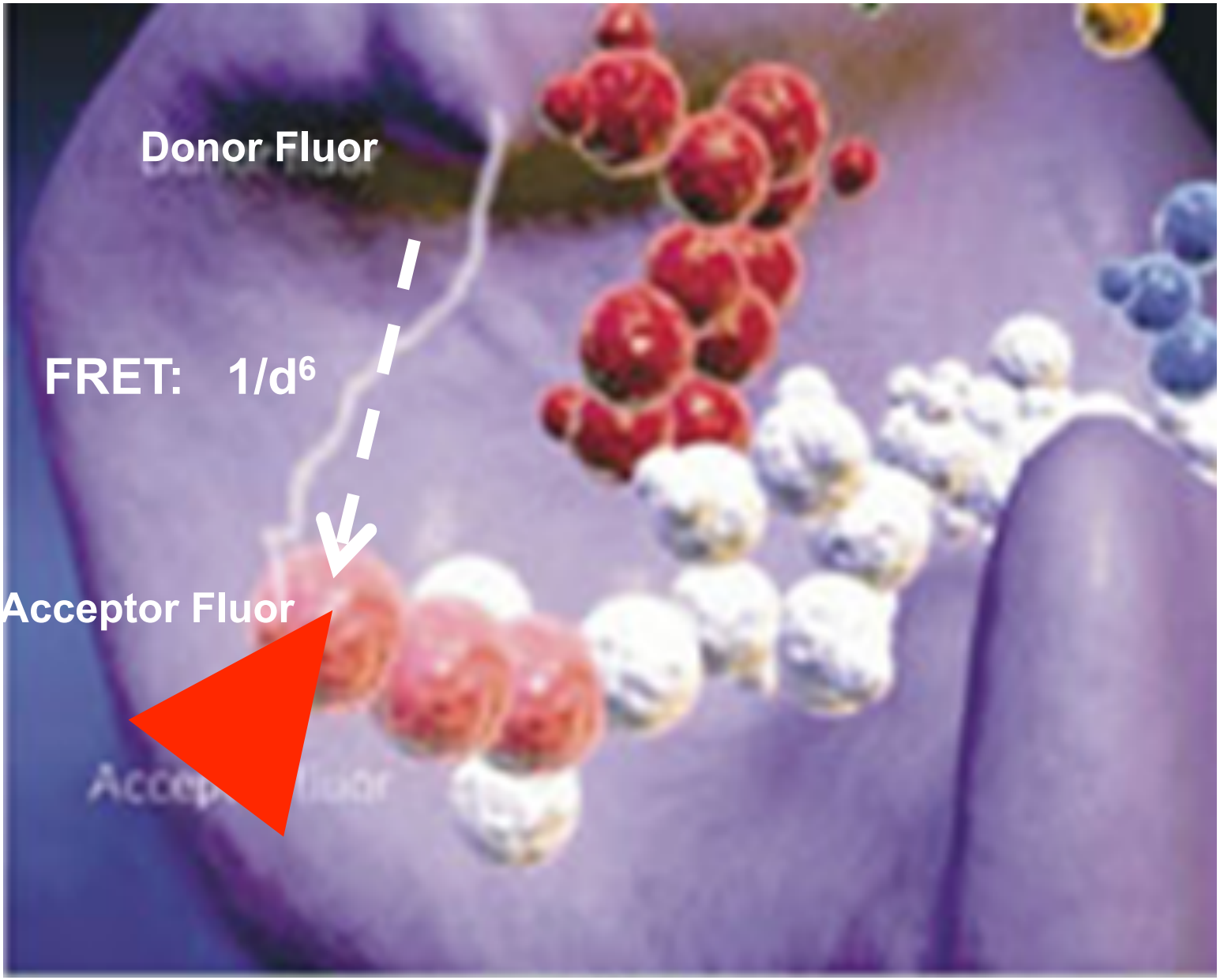
(Now part of Invitrogen)

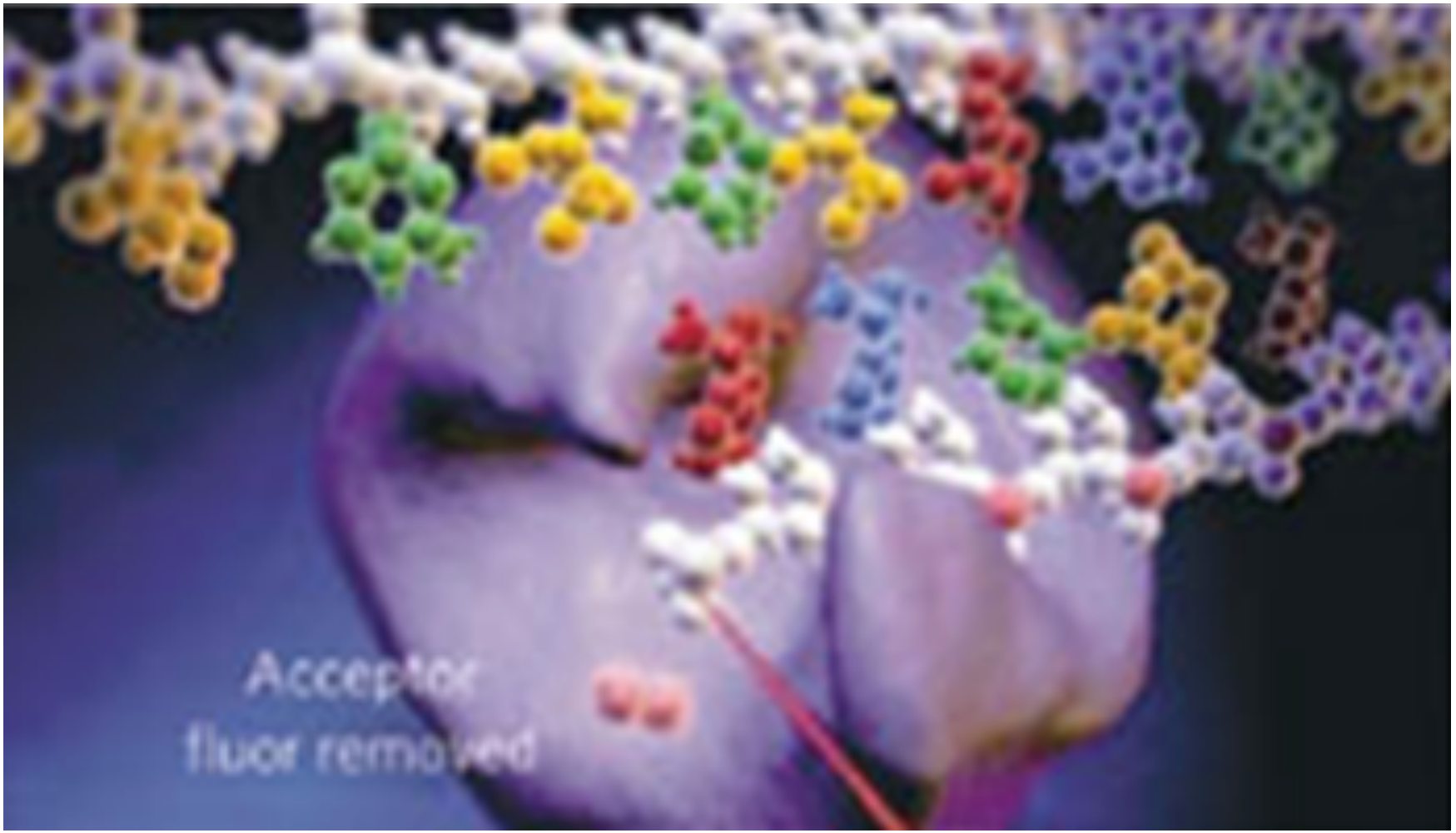
- *Real time single molecular sequencing*
- *Without fabricated "wells" such as ZMGs*
- *Forster energy transfer (FRET)*

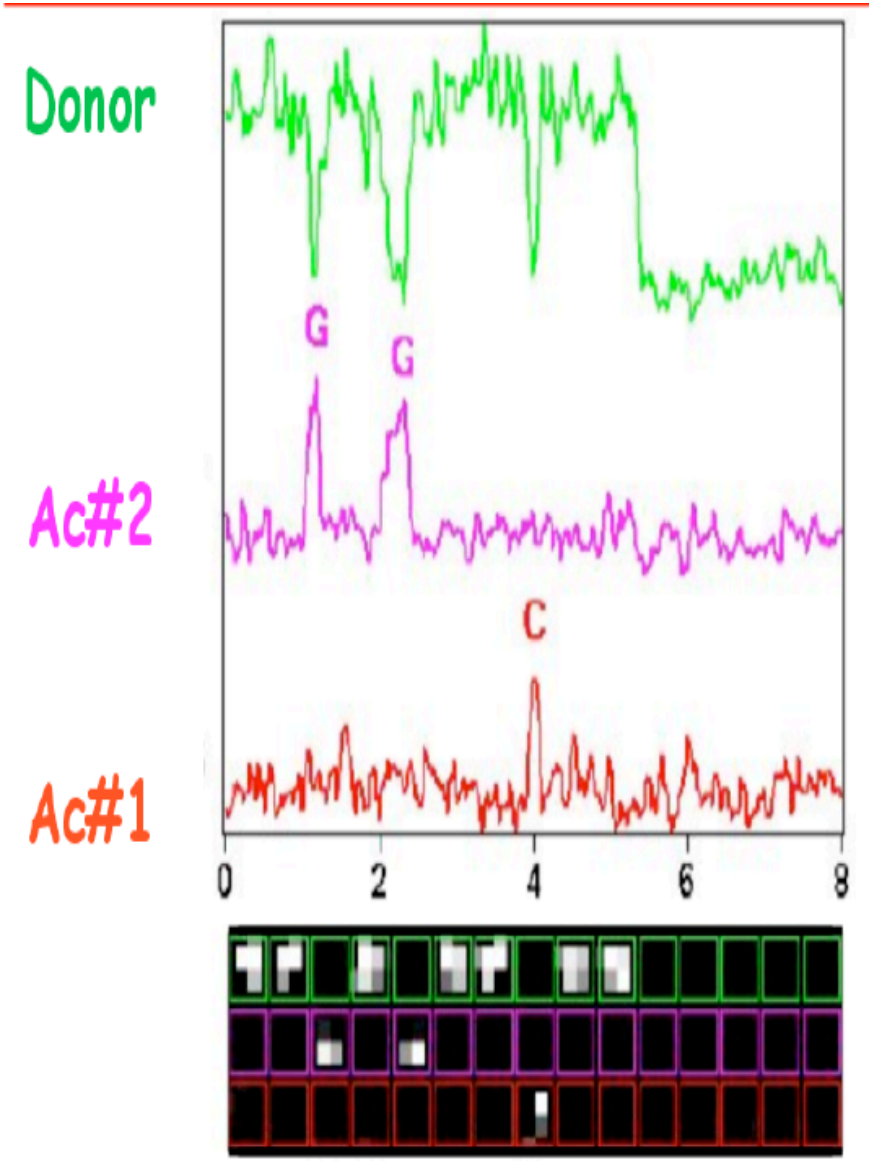
  10^{-6} D-R dependence


Cover slip
A
B
C
D


fluorophore
dCTP
linker


Donor Fluor
FRET: 1/d6
Acceptor Fluor


Acceptor
fluor removed


Donor
G
G
Ac#2
C
Ac#1
0
2
4
6
8

Visigen / Invitrogen

- *Commercial device announced for 2011*
- *Sub $1,000 HGE expected*
- *Sub 1 HGE/4 hours expected*
- ***Maybe much better***

Cracker

- *Real time single molecule sequencing*
- *CMOS technology*
- *LED wells on photodiodes on IC*

Gen 3?
Direct Sequencing

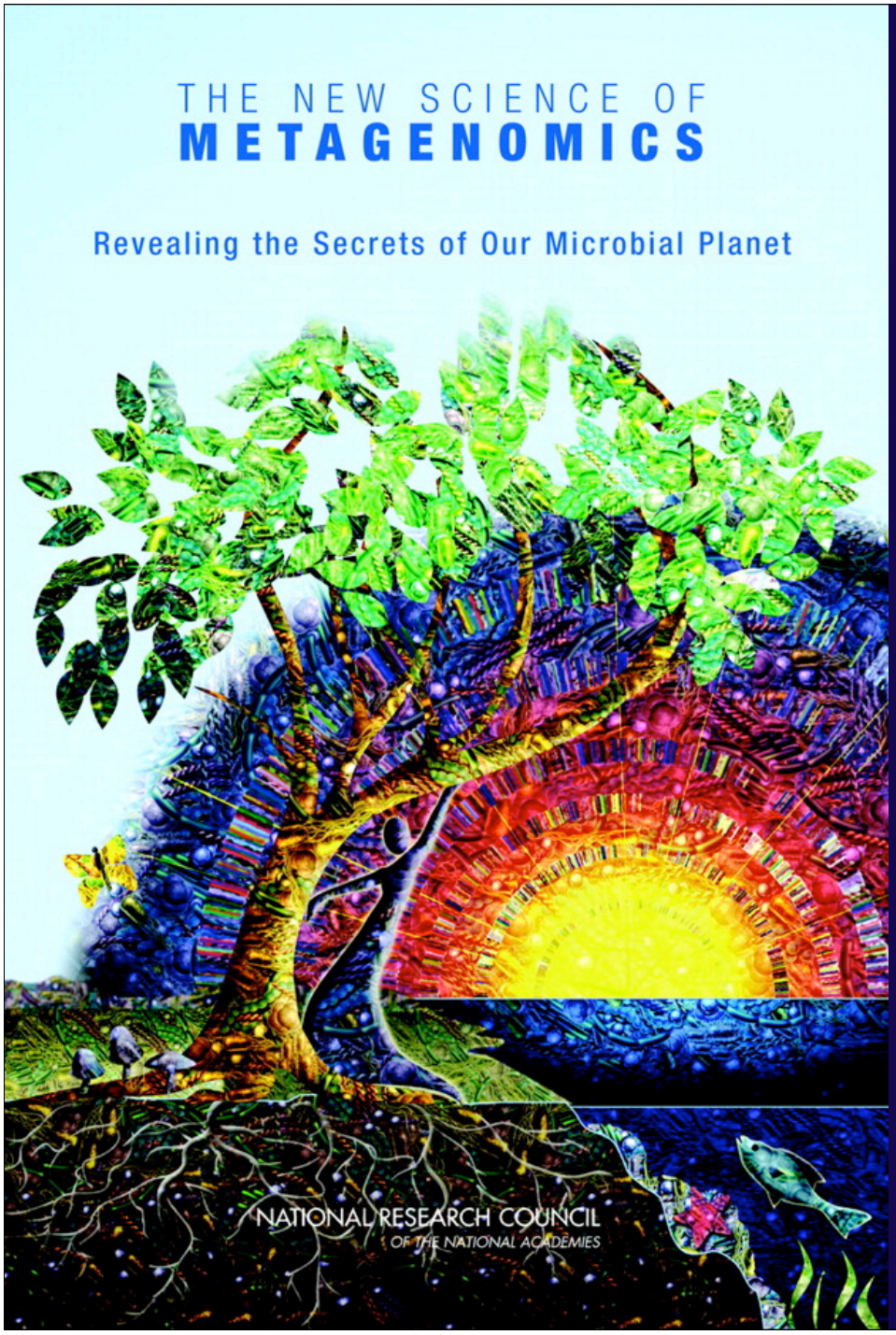
- **Nanopore sequencing**
  Real time single molecular sequencing
  By transit through pores in a membrane

- **EM sequencing**
  Sequence determination using electron microscopy

- AFM sequencing
  Sequence determination using various types of STEM and AFM

Applications of NexGen Sequencing

- Whole genome sequencing – human variation, disease risk
- Genome scanning – cancer SNPs
- Population wide genomic sequencing – HIV variants
- Ancient DNA sequencing - Neanderthal DNA sequencing
- Epigenetic analysis – CpG methylation
- Protein-DNA binding sites - ChIPs
- Transcription analysis – human cell expression profiling
- sIRNA analysis – expression control
- Metagenomics – population profiling life's diversity
- Synthetic genomics – engineering advantageous organisms



Using DNA sequencing to assess all the genes in a sample to learn -

- **What genes are there?**
- **What functions are there?**
- **What organisms are there?**
- **What populations are there?**
- **How do they interact?**
- **How do they interdepend?**
- **How do they change?**

J. Craig Venter's The Sorcerer's Path

Sampling the Sea's Diversity

METAGENOMICS / SARGASSO SEA

Done with “old” technology - capillary sequencers - circa 2004.
Results with NexGen sequencing likely to be 1,000 fold more revealing

- Whole-genome shotgun sequencing of microbial populations in samples
- 1.045 billion base pairs (non-redundant)
- At least 1800 genomic species (relatedness criteria)
- 1 48 new bacterial phylotypes.
- Over 1.2 million previously unknown genes
- More than 782 new rhodopsin-like photoreceptors.

Evidently we have a lot to learn about diversity in the oceans!!!

- *Commercial impact? --- ExxonMobil committed $600 Million to Craig Venter ‘s Synthetic Genomics to develop engineered organisms for biofuel production - based partly on biodiversity / DNA sequencing results.*

Questions

- Is an ensemble sometimes better than one?
- Can single molecule SBS be used as the read out for computational devices based on DNA?
- Can single molecule SBS be used as the basis for programmable molecular assemblers?

Questions?