Oral Presentation

- **Find an research article for yourself.**
- Make slides and present this article in class.
- You need to decide which article you want to talk before Oct. 22nd.
- This article must be a research article, which uses some computational methods for systems biology.
- This article will be used for the oral presentation and final exam.

Homework Assignment (1)

- Your Name
- Do you have a computer? Or do you have chances to use a computer?
- If yes, what operating system does this computer have? (MS Windows, Mac OS, UNIX/Linux)
- Do you have permission to install a software on this computer?
- Do you have experience in programming? If yes, what programming language did you use? (C/C++, Fortran, Perl, Python, Ruby, VB etc.)
- Do you know how to use R? If no, how difficult is it for you to master this software? (easy; need time and help; very difficult; impossible for me to master it)

(due by 11:59PM, August 31, 2015)

Next-generation Sequencing

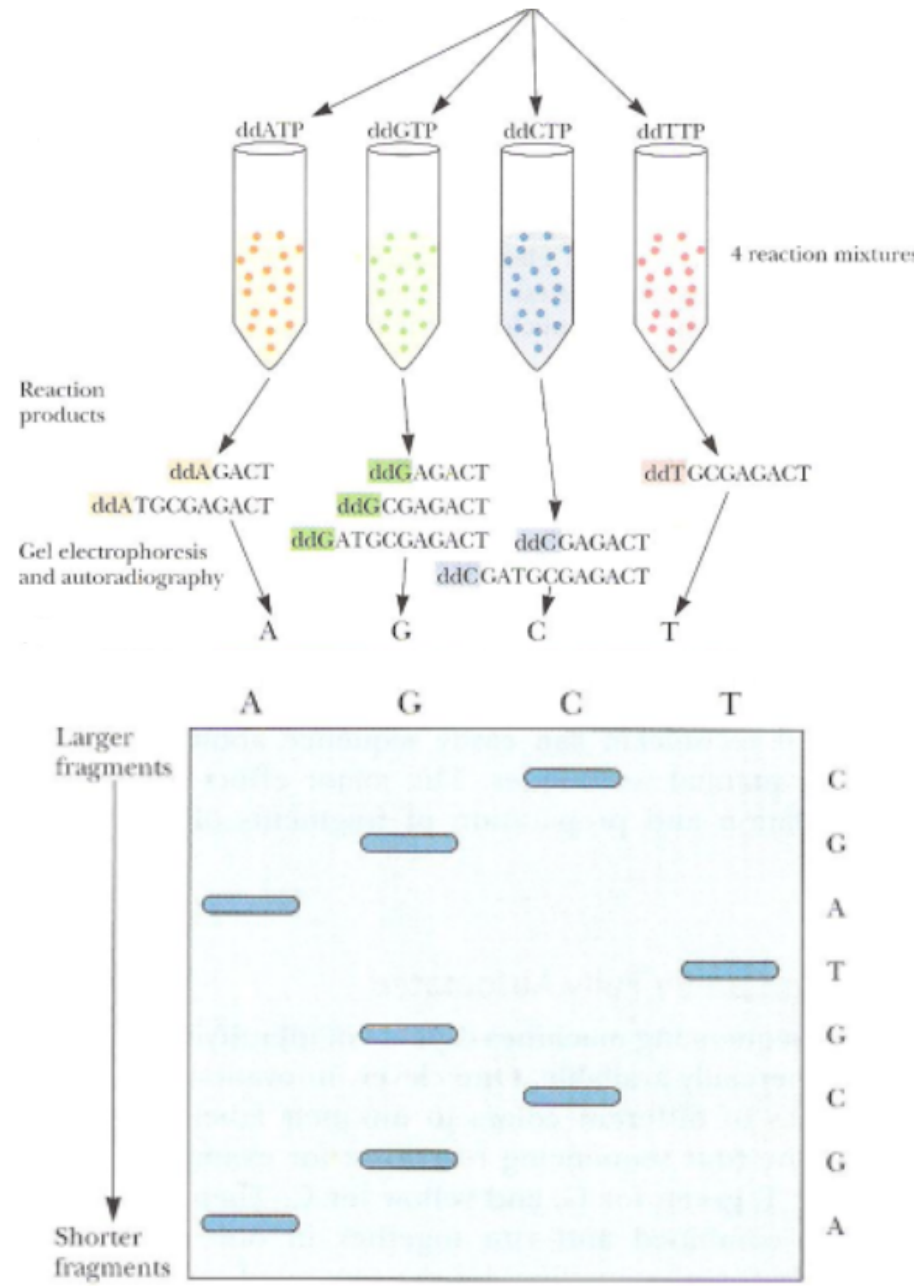
Lecture 2

NGS

- Introduction to the background
- NGS workflow and accuracy
- Data format and quality control
- Assembly
- RNA-seq
  - Aligner
  - Analysis tools
  - Applications, such as MiRNA
- Chip-seq
  - Applications

Sanger Method

- Primer attachment and extension of bases
- Run four separate reactions each with different ddNTPs. All terminated chains will end in the ddNTP added to that reaction.
- Run on a gel in four separate lanes
- Read the gel from the bottom up



Human Genome Project

- The human genome is about 3 billion bp
- Began in 1990. The Human Genome Project was declared complete in April 2003. An initial rough draft of the human genome was available in June 2000 and by February 2001.
- Cost = $3 billion
- fostered development of faster, less expensive sequencing techniques.

Next-Generation Sequencing

- Sequencing by synthesis (SBS), pyrosequencing, sequencing by ligation
- Advantages:
  - Accurate
  - Parallel processing
  - Easily automated
  - Eliminates the need for labeled primers and nucleotides
  - No need for gel electrophoresis

Platform

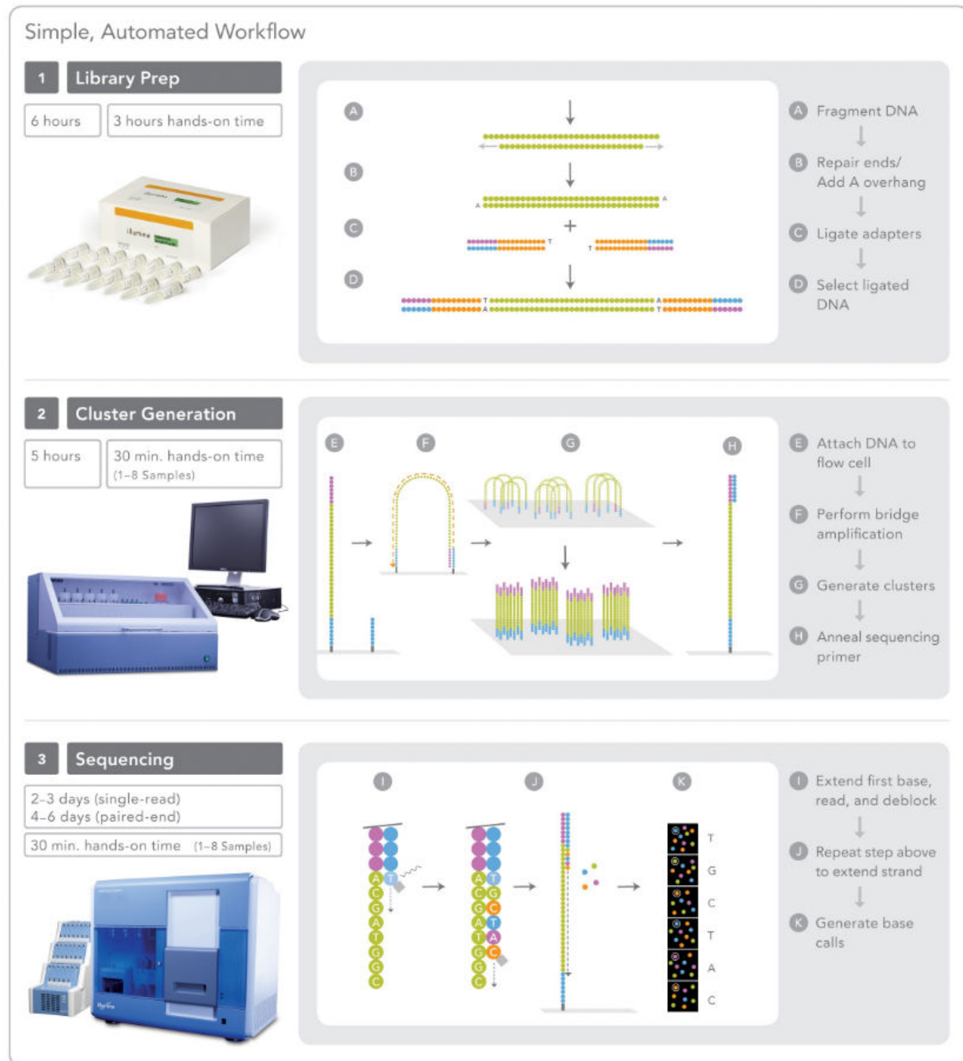
- Illumina
- SOLiD (Life Technology)
- 454 (Roche)
- Helicos
- Pacific Biosciences
- Ion Torrent (Life Technology)

NGS

- Introduction to the background
- NGS workflow and accuracy
- Data format
- Assembly
- RNA-seq
  - Aligner
  - Analysis tools
  - Applications, such as MiRNA
- Chip-seq
  - Applications

Work flow

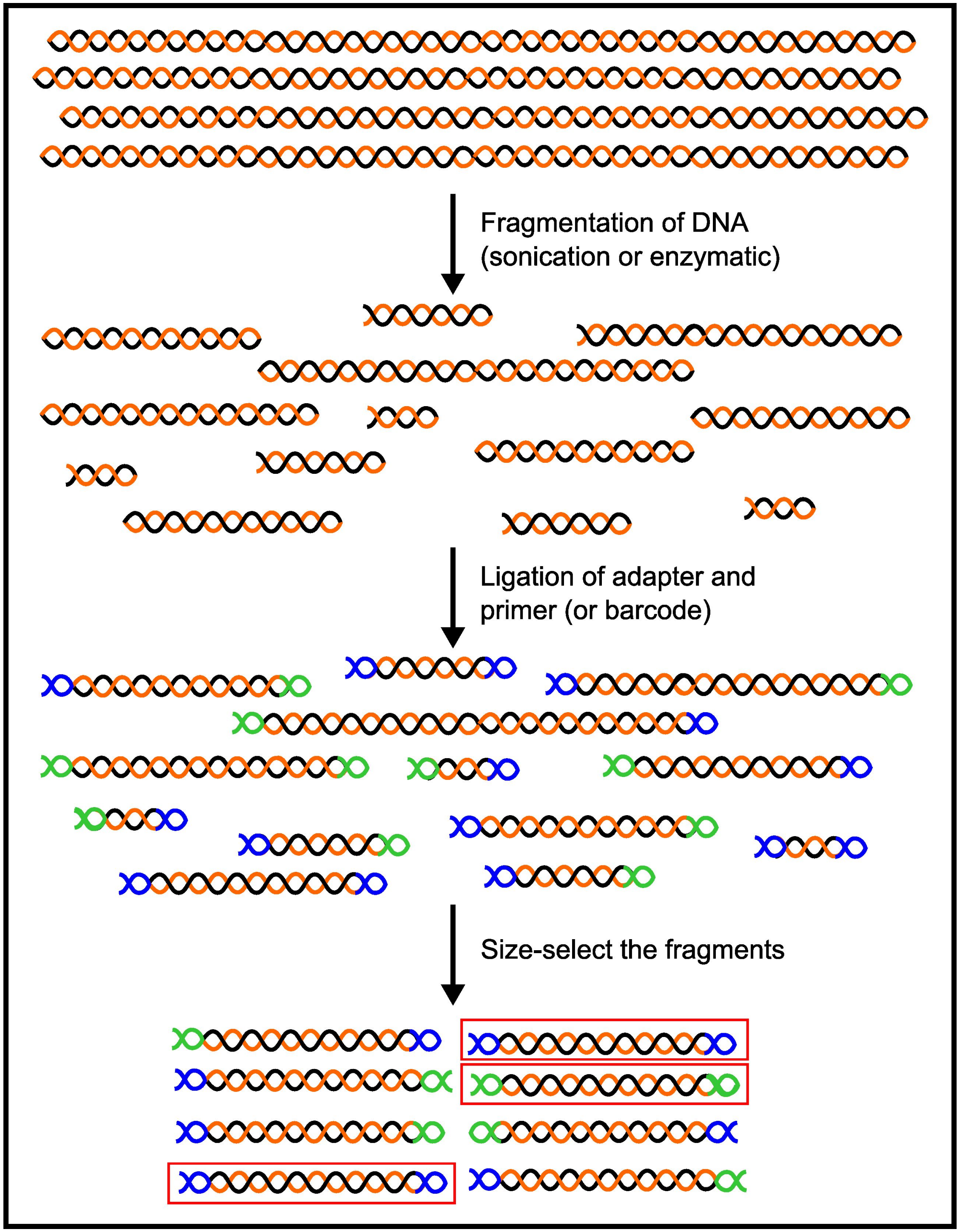
1 Library preparation
2 Amplification
3 Sequencing and imaging
4 Data analysis



Library preparation

- Most platforms adhere to a common library preparation procedure with minor modifications, before a 'run' on the instrument.
- Procedure includes DNA fragmenting, DNA repair, adaptor ligation, and size-selection.
- This process typically results in considerable sample loss with limited throughput.

Nature Methods 6, (2009)

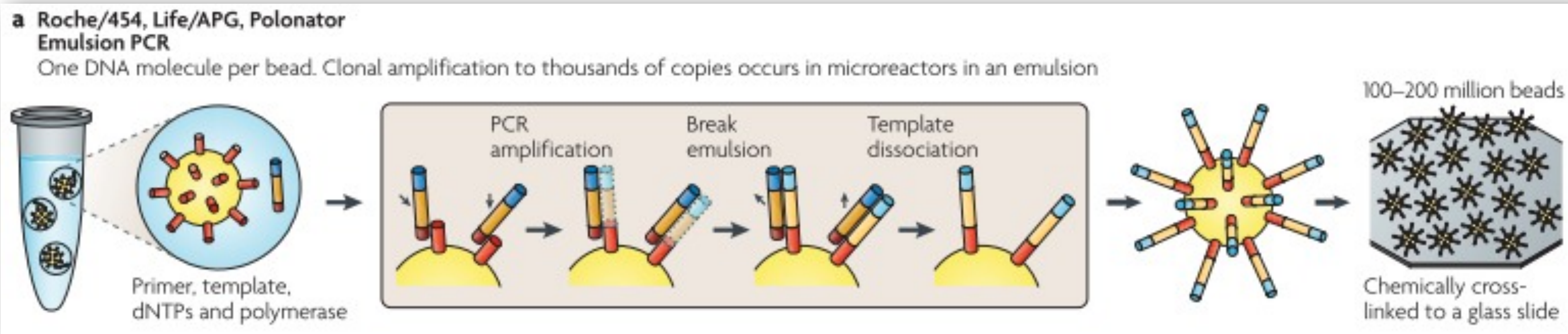


1. fragmenting the DNA (sonication, nebulization, or shearing)
2. DNA repair and end polishing (blunt end, phosphorylated end that is ready for ligation)
3. platform-specific adaptor ligation.
4. Size-selection

Amplification

- Problem: most imaging systems do not designed to detect single fluorescent event.
- Solution: need amplified templates.
- Issue: need a robust method that is capable to produce a representative, non-biased source of nucleic acid material from the genome under investigation.
- Options: emulsion PCR (emPCR) or solid phase amplification

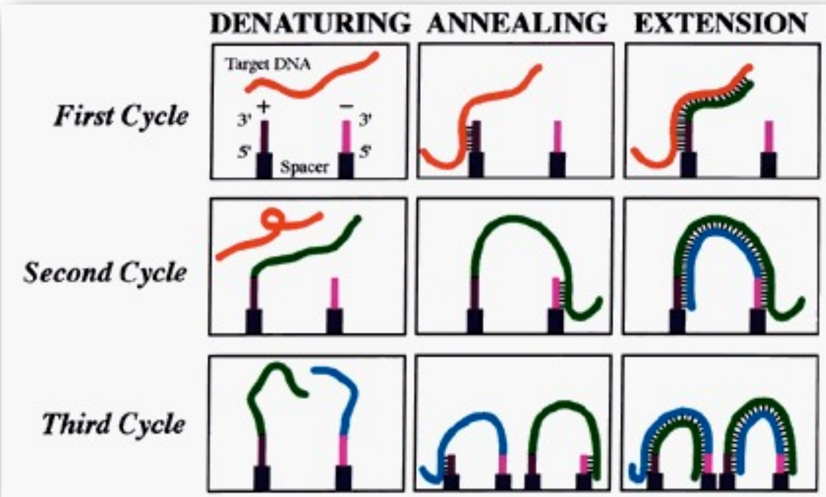
Emulsion PCR



Metzker et al, 2010

- EmPCR: preparing sequencing templates in a cell-free system, and advantage of avoiding the arbitrary loss of genomic sequences.
- A library of fragment is created, and adaptors containing universal priming sites are ligated to the target ends, allowing complex genomes to be amplified with common PCR primers.
- After ligation, the DNA is separated into single strands and captured onto beads under conditions that favor one DNA molecule per bead.
- Millions beads can be immobilized in a individual PicoTiterPlate (PTP) wells in which the NGS chemistry can be performed.

Solid-phase amplification



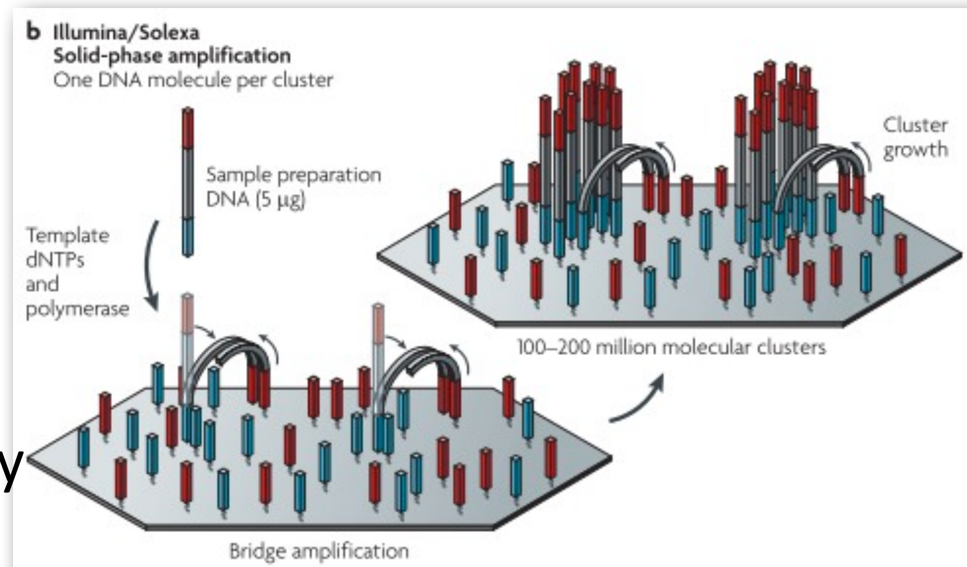
http://bit.ly/6JYlUz

Solid-phase amplification can be used to produce amplified clusters from fragment on a glass slide.

Solid-phase amplification can produce 100–200 million spatially separated template clusters.

High-density forward and reverse primers are covalently attached to the slide, and the ratio of the primers to the template on the support defines the surface density of the amplified clusters

http://www.youtube.com/watch?v=77r5p8lBwJk&NR=1



Metzker et al, 2010

Sequencing and imaging

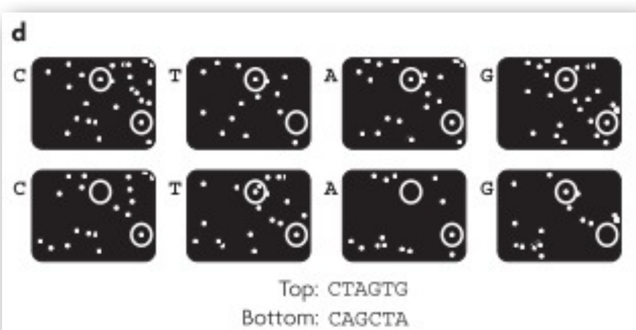
- Technologies:
  1. Cyclic reversible termination (Helicos BioSceinces, Illumina)
  2. Sequencing by ligation (SOLiD)
  3. Pyrosequencing (454)
  4. real-time sequencing (Pacific Biosciences)

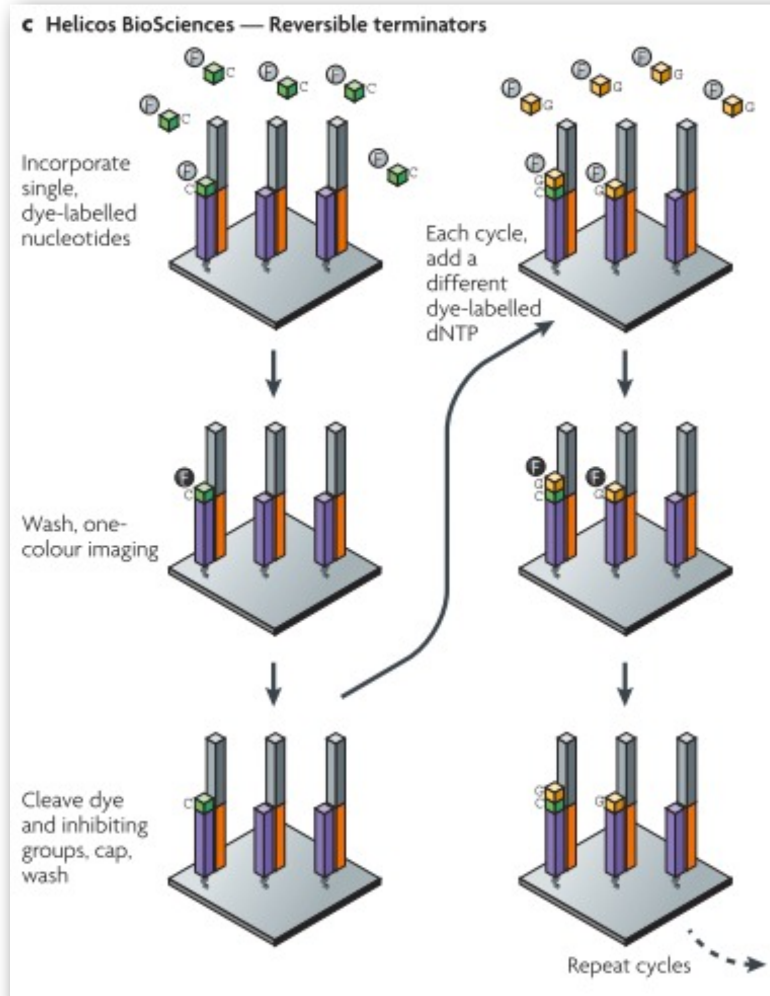
Cyclic reversible termination

Helicos: 1-colour

A cyclic method that comprises nucleotide incorporation, fluorescence imaging and cleavage

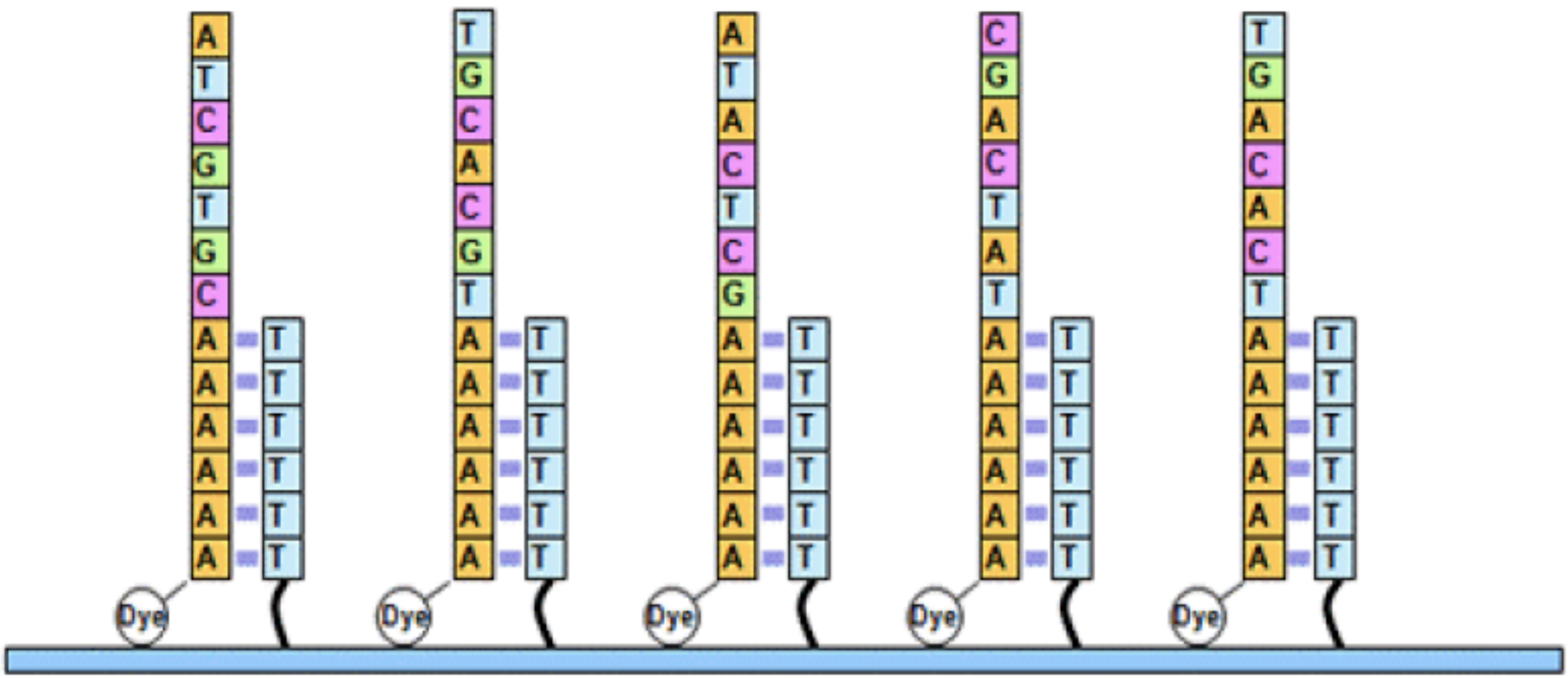
sequencing result



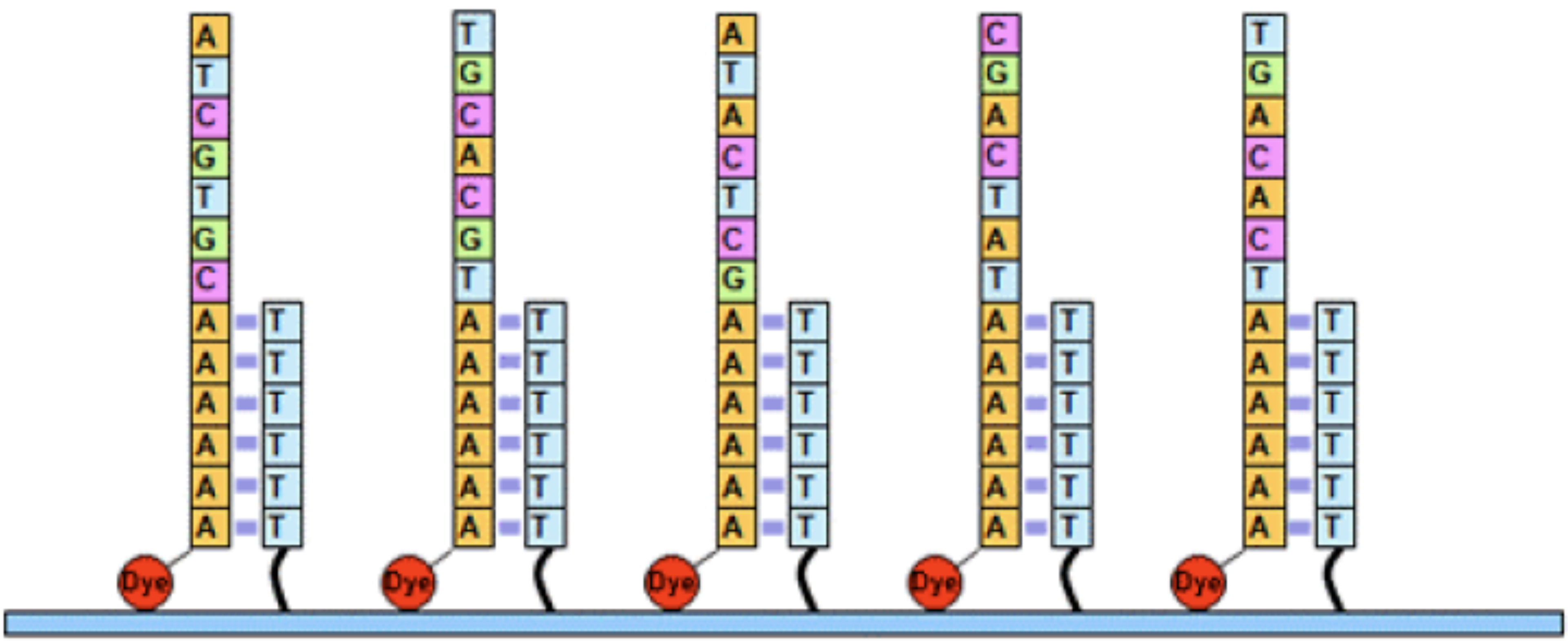


Metzker et al, 2010

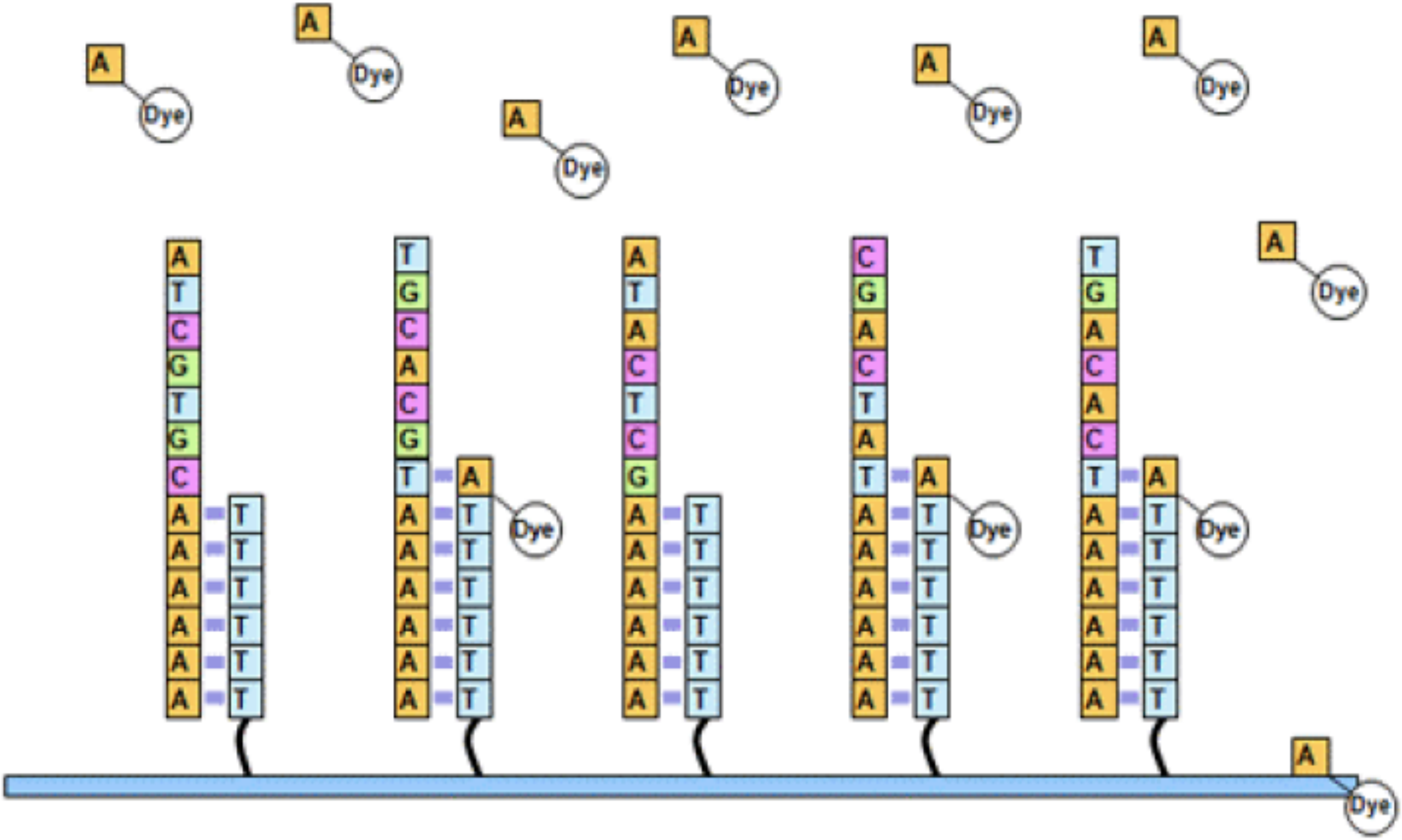
Amplified DAN templates and primers are immobilized on the flow cell



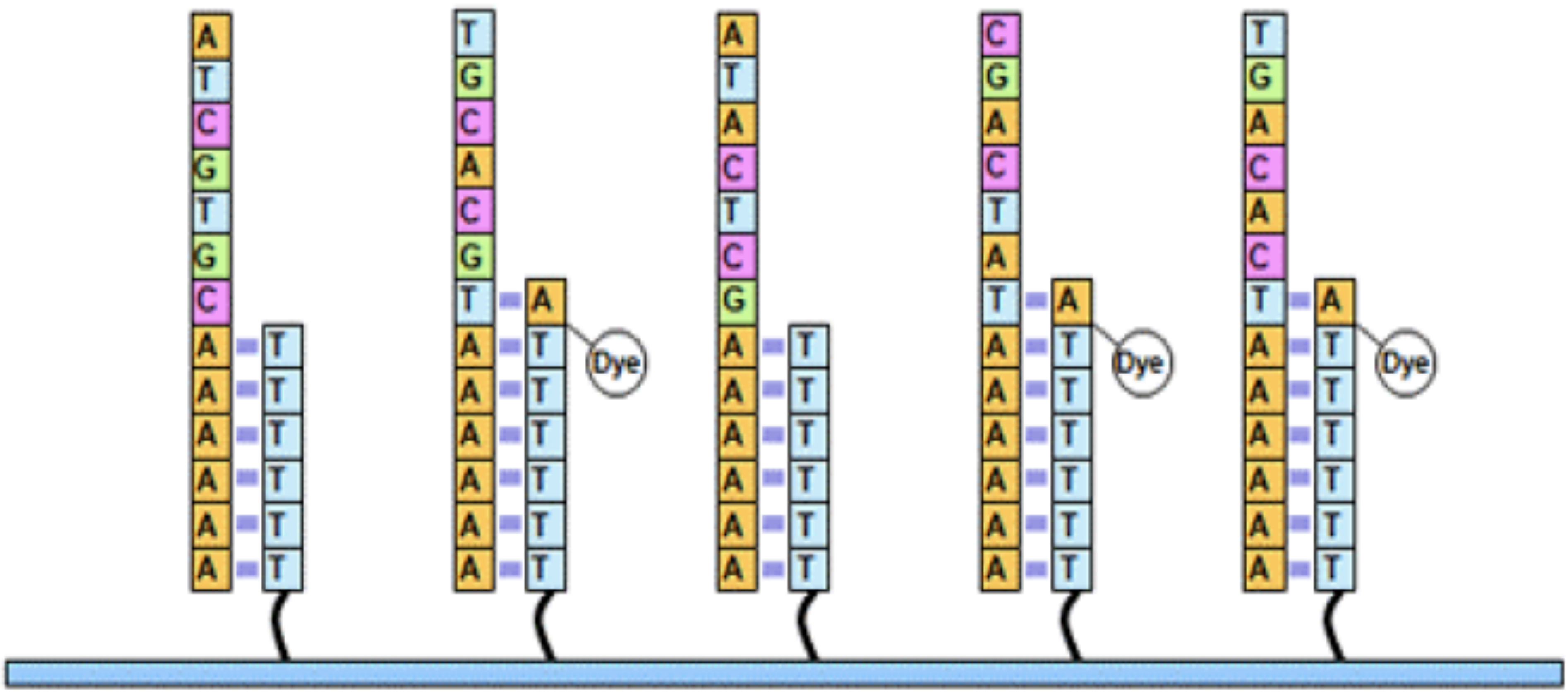
Visualize the template:primer duplexes by illuminating the surface with a laser and imaging with an electronic camera connected to a microscope. Record the positions of all the duplexes on the surface. After imaging, the dye molecules are cleaved and washed away.



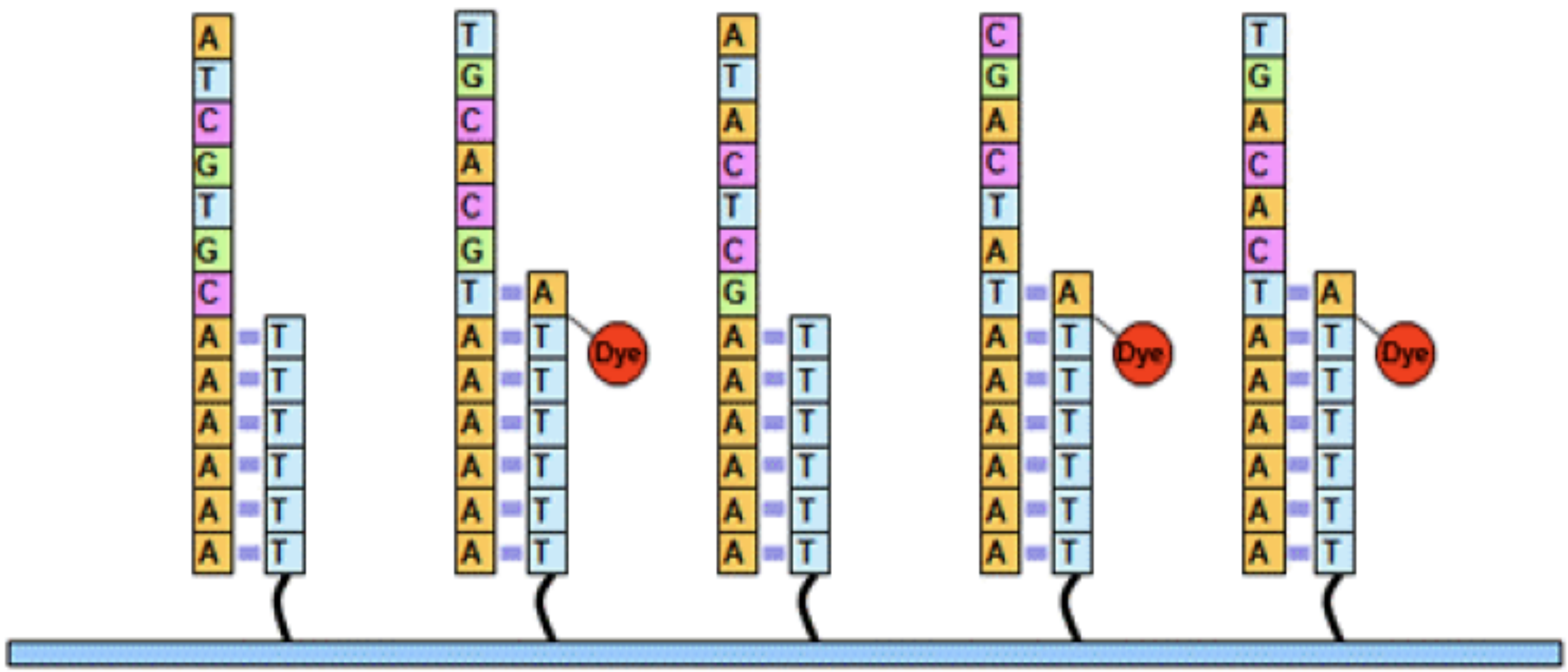
Flow in DNA polymerase and one type of fluorescently labeled nucleotide (for example A). The polymerase will catalyze the addition of labeled nucleotide to the appropriate primers.



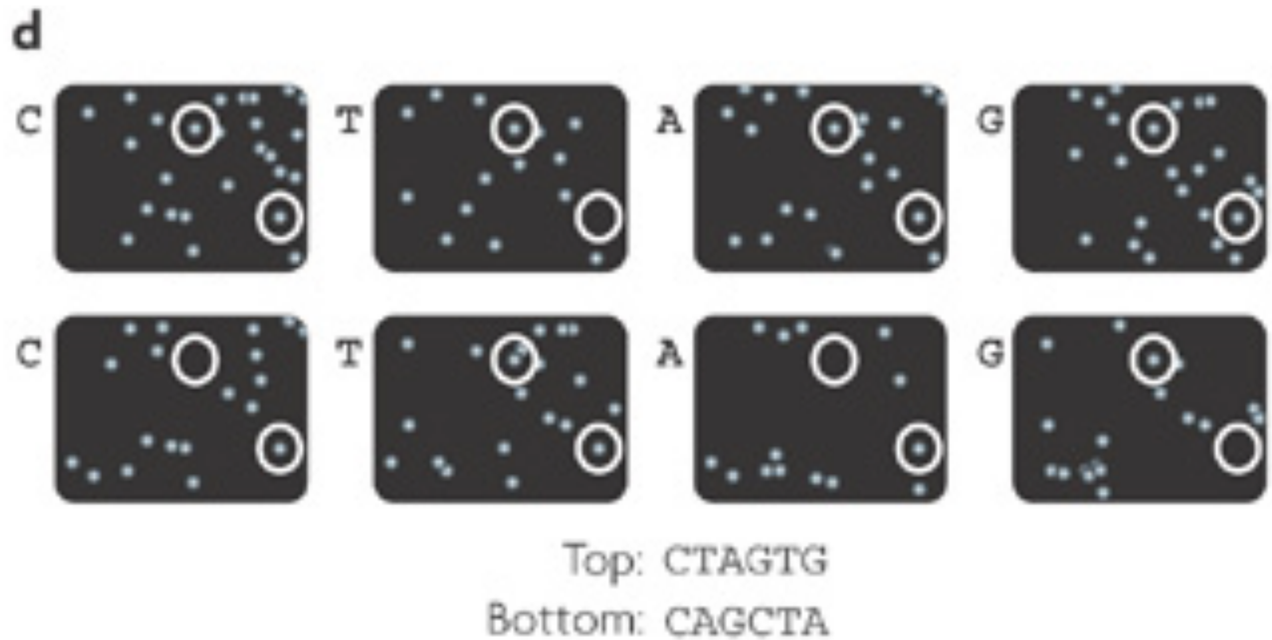
Wash out the polymerase and unincorporated nucleotides.



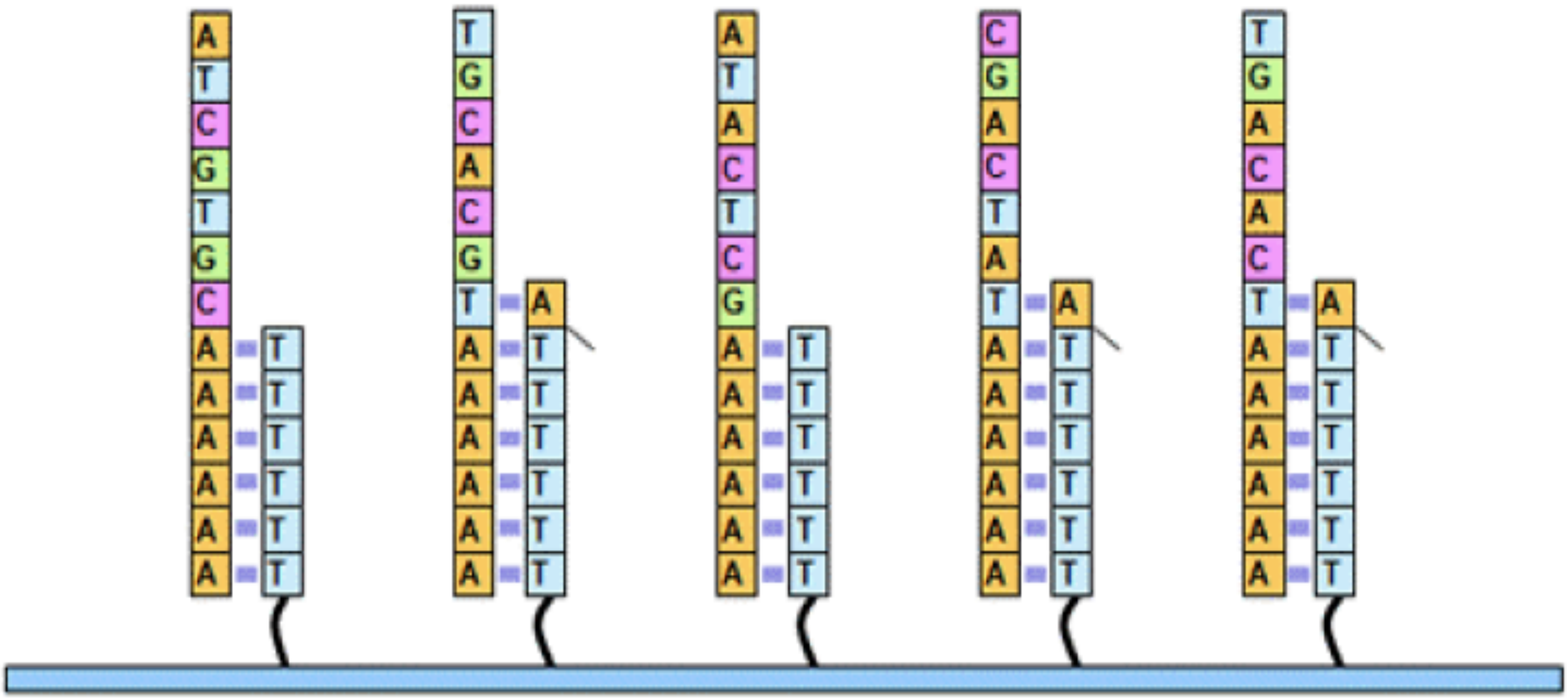
Visualize the incorporated labeled nucleotides by illuminating the surface with a laser and imaging with the camera. Record the positions of the incorporated nucleotides.



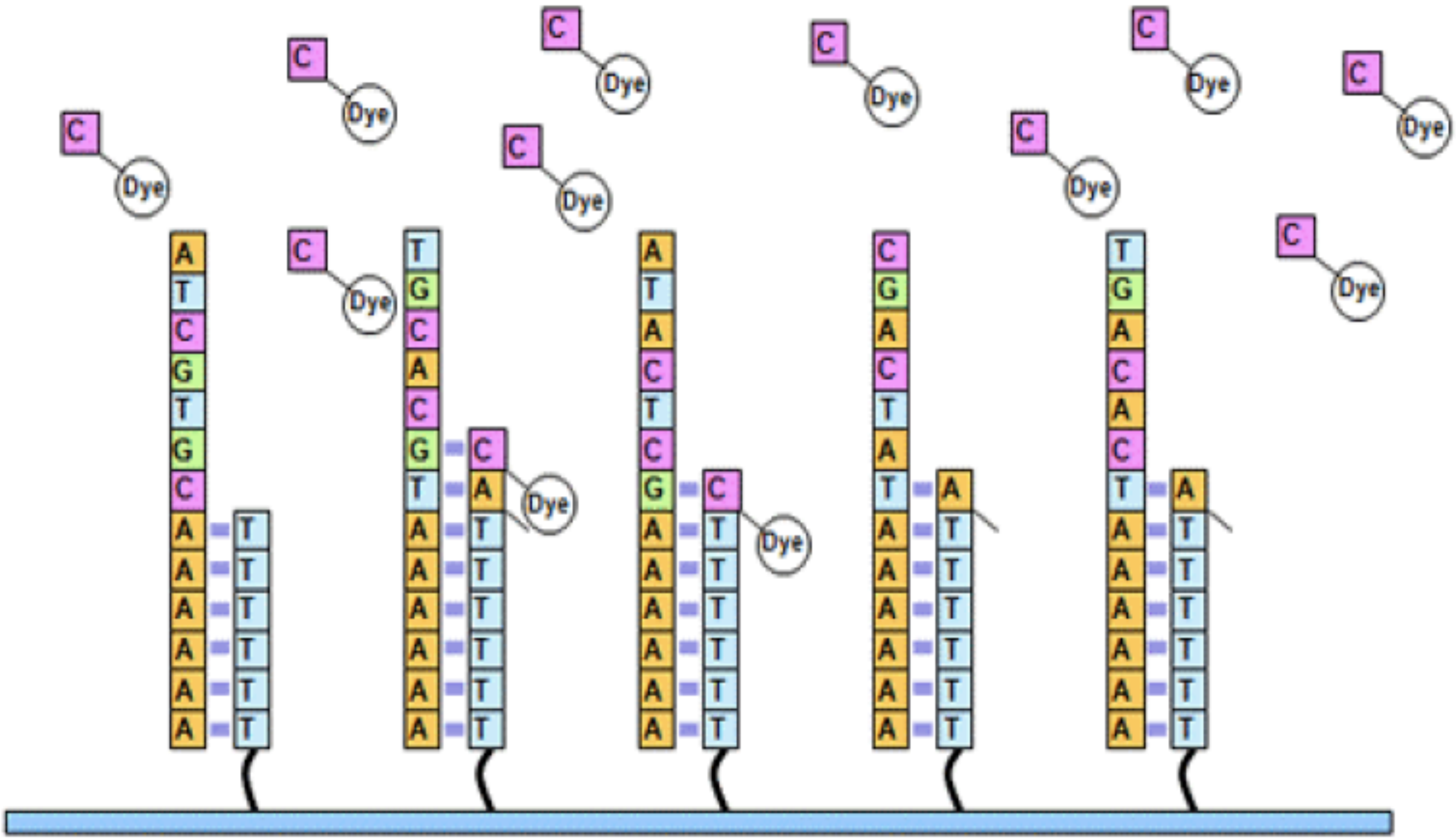
Cyclic reversible termination image of one cycle



Step 8: Remove the fluorescent label on each nucleotide.



Step 9: Repeat the process from step 5 with the next nucleotide (stepping through A, C, G and T), until the desired read-length is achieved.



Cyclic reversible termination

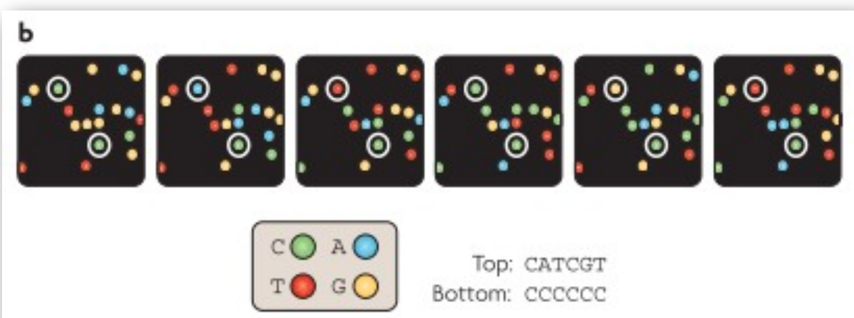
DNA synthesis is terminated after adding single nucleotide

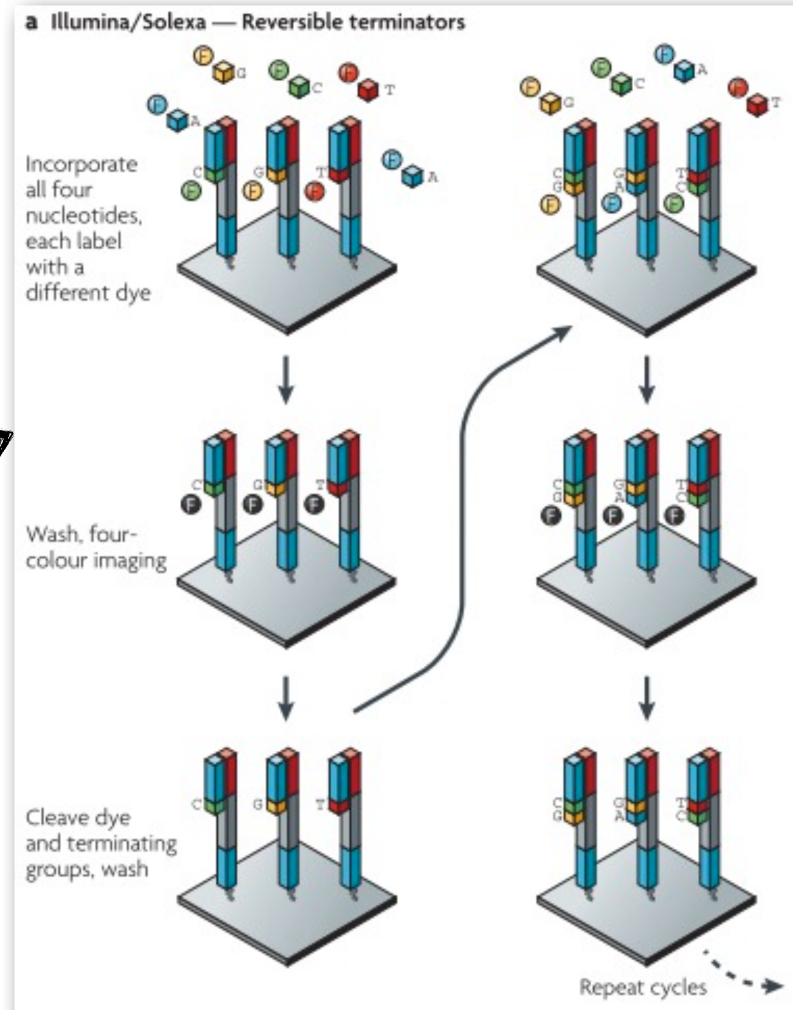
start/stop/start/stop/start/stop/...

Illumina: 4-colour

sequencing result

sequencing steps





Metzker et al, 2010

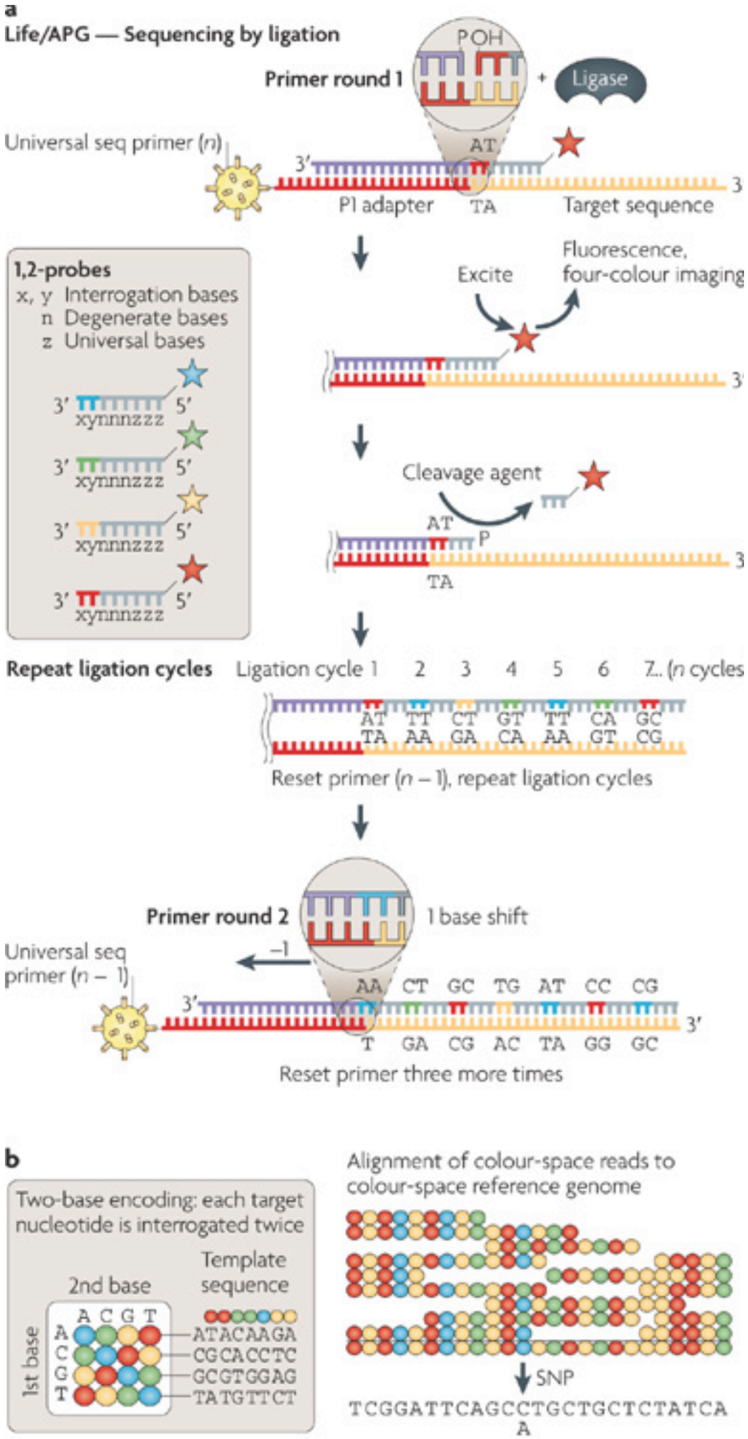
Video for illumina sequencing technology

https://www.youtube.com/watch?v=womKfikWlxM

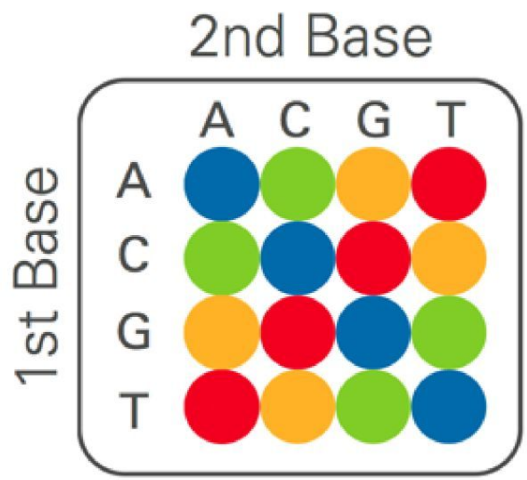
Sequencing and imaging

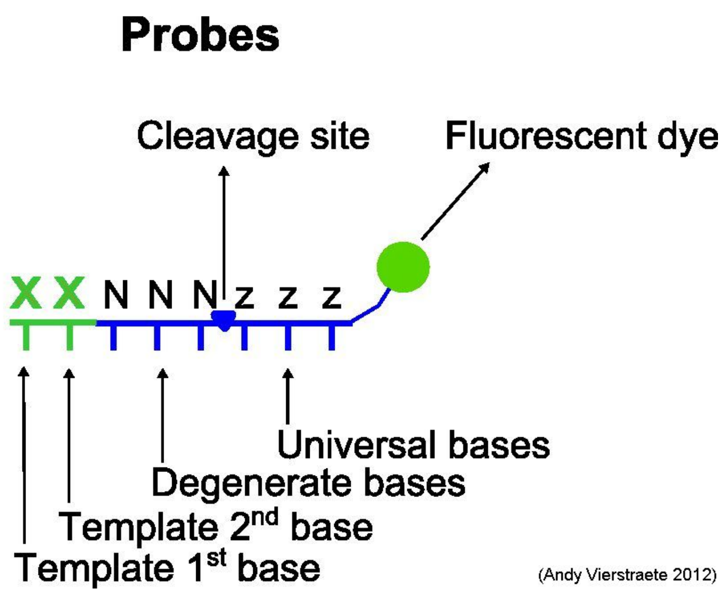
- Technologies:
  1. Cyclic reversible termination (Sequencing By Synthesis) (Helicos BioSceinces, Illumina)
  2. Sequencing by ligation (SOLiD)
  3. Pyrosequencing (454)
  4. real-time sequencing (Pacific Biosciences)

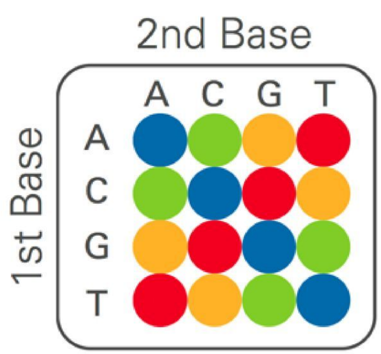
Sequencing by ligation

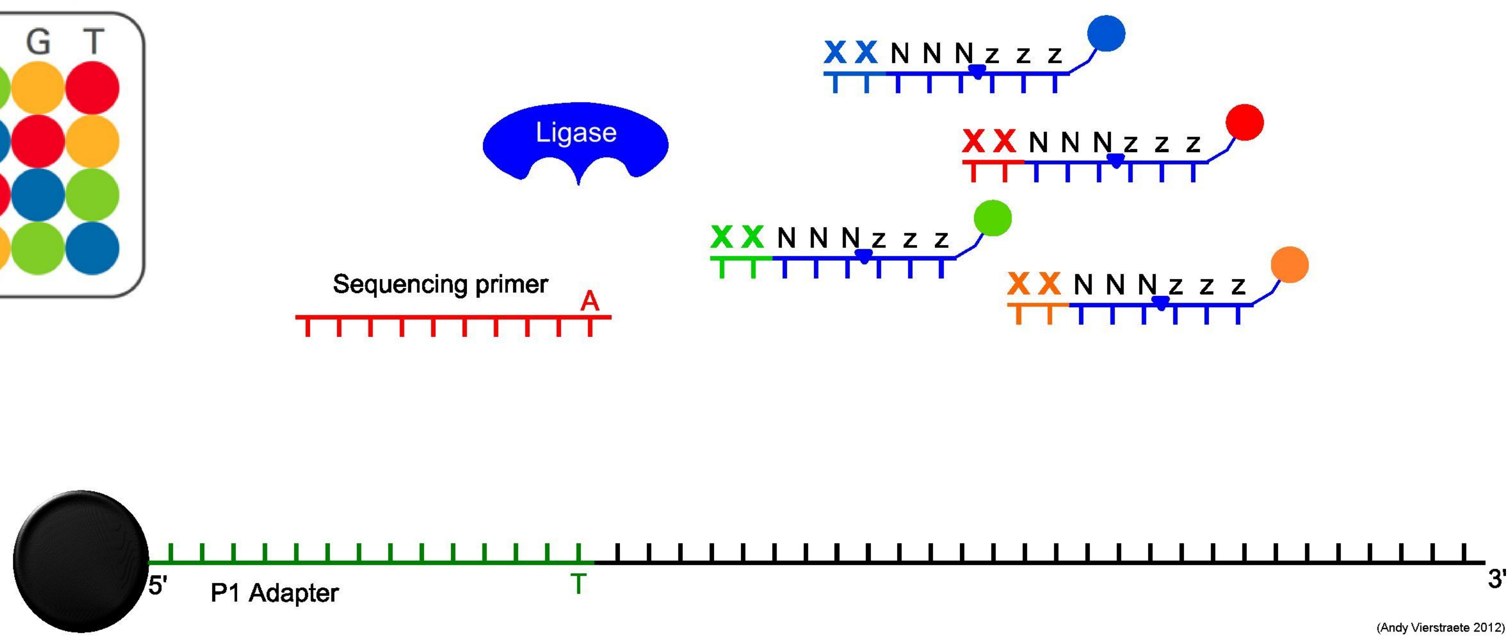


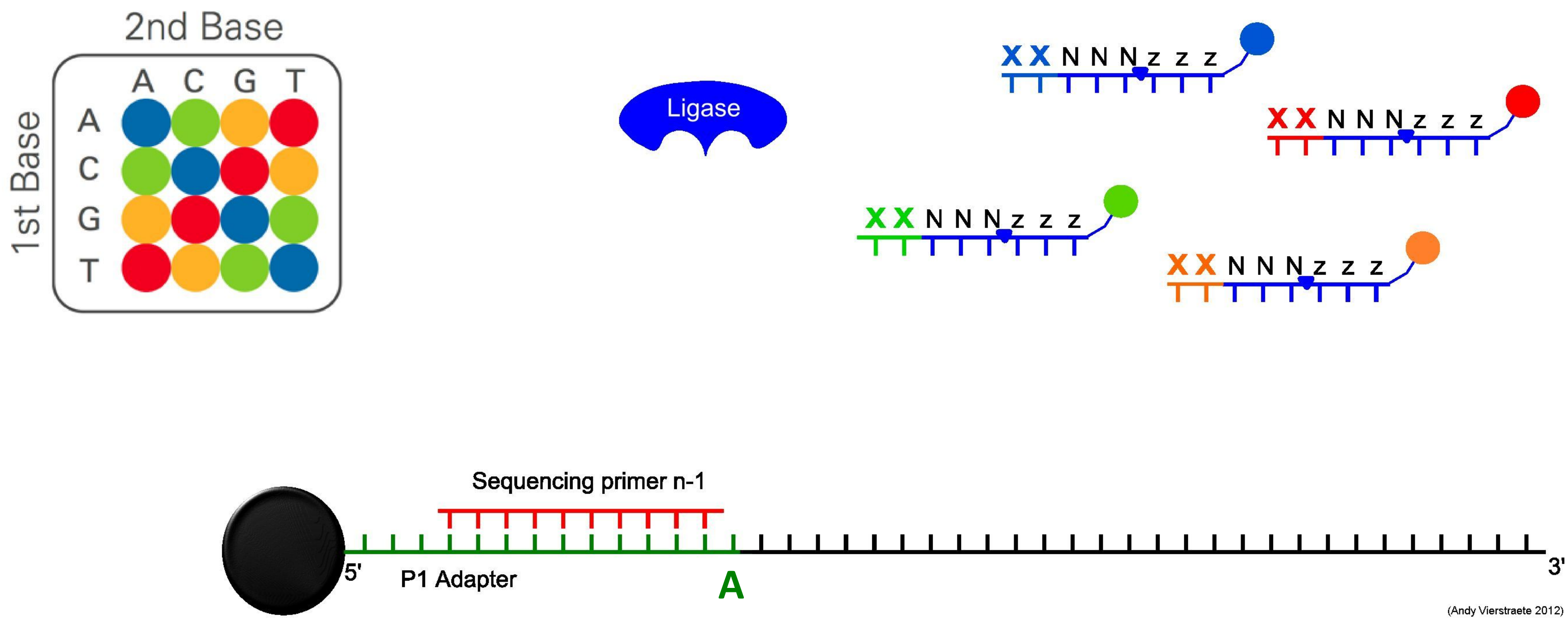
Use of DNA ligase and two-base-encoded probes. In its simplest form, a fluorescently labelled probe hybridizes to its complementary sequence adjacent to the primed template. DNA ligase is then added to join the dye-labelled probe to the primer. Non-ligated probes are washed away, followed by fluorescence imaging to determine the identity of the ligated probe. The cycle can be repeated by using cleavable probes to remove the fluorescent dye.

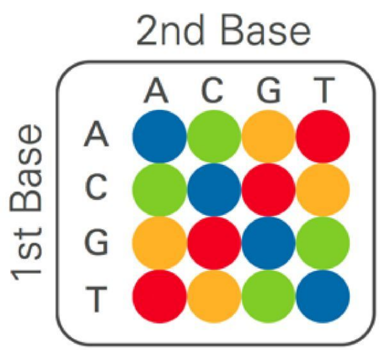

2nd Base
A C G T
1st Base
A
C
G
T


Probes
Cleavage site
Fluorescent dye
X X N N N z z z
Universal bases
Degenerate bases
Template 2nd base
Template 1st base
(Andy Vierstraete 2012)

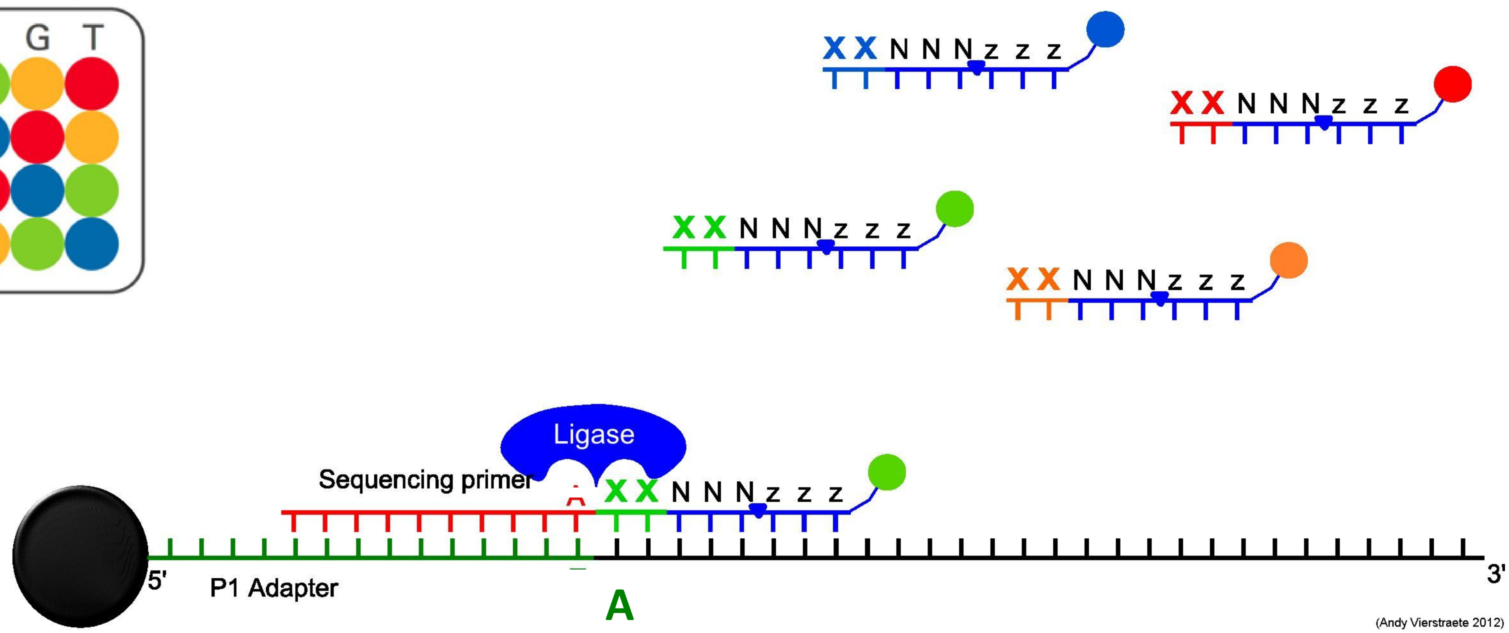

2nd Base
A C G T
1st Base
A
C
G
T


X X N N N z z z
Ligase
Sequencing primer
A
5'
P1 Adapter
T
3'
(Andy Vierstraete 2012)

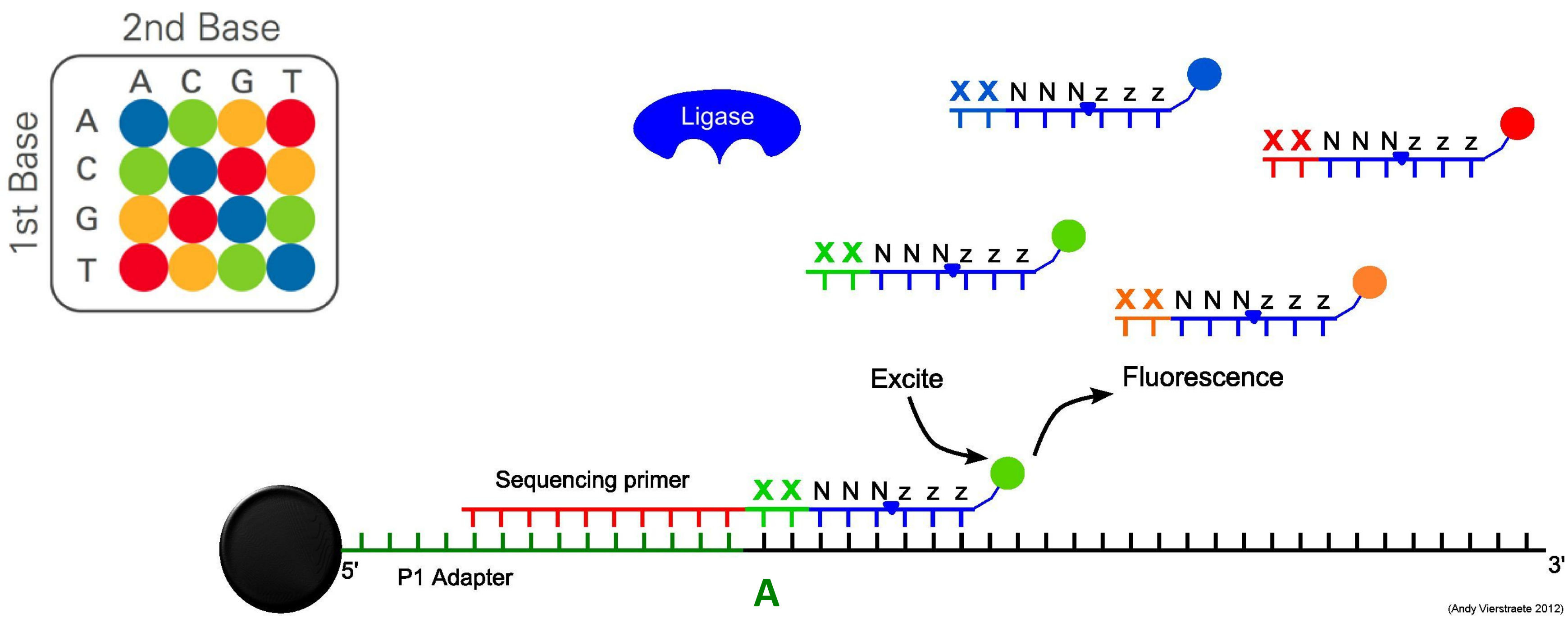

2nd Base
A C G T
1st Base
A
C
G
T
Ligase
X X N N N z z z
X X N N N z z z
X X N N N z z z
X X N N N z z z
Sequencing primer n-1
5'
P1 Adapter
A
3'
(Andy Vierstraete 2012)



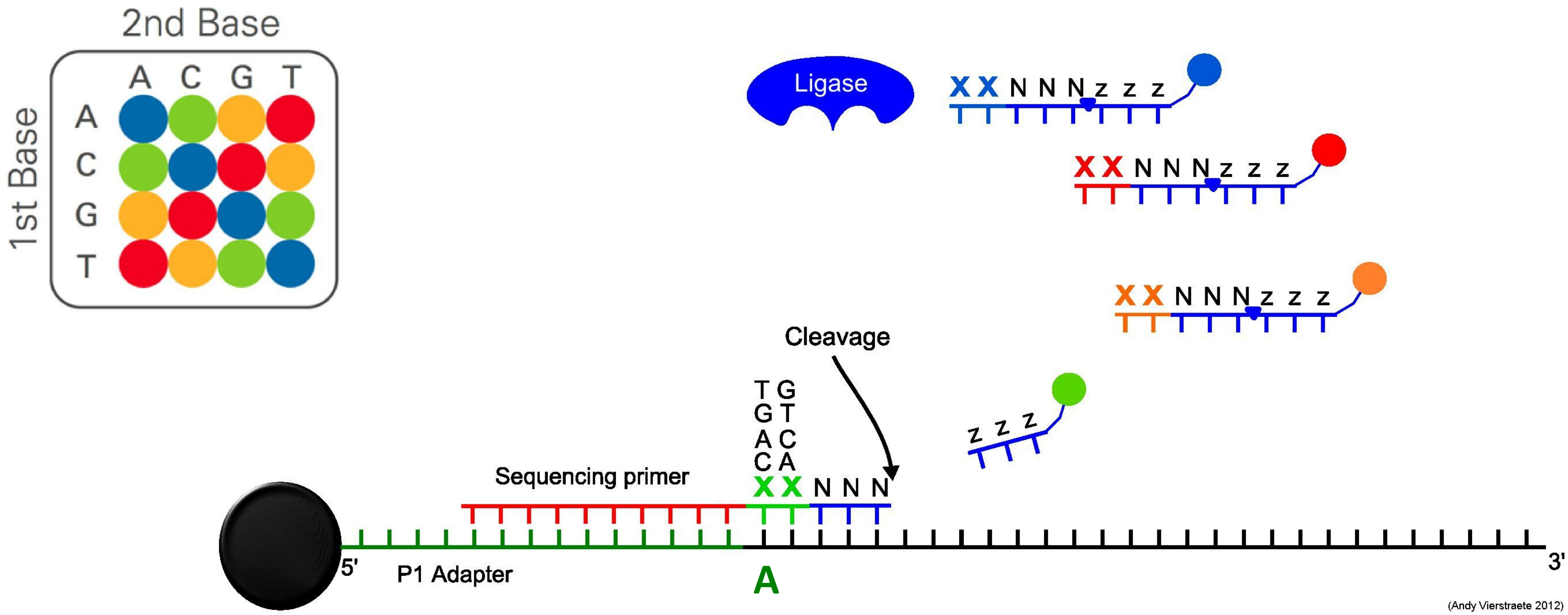
2nd Base
A C G T
1st Base
A
C
G
T



X X N N N z z z
Ligase
Sequencing primer
5'
P1 Adapter
A
3'
(Andy Vierstraete 2012)



2nd Base
1st Base
A C G T
A
C
G
T
Ligase
XX N N N z z z
Excite
Fluorescence
Sequencing primer
5'
P1 Adapter
A
3'
(Andy Vierstraete 2012)



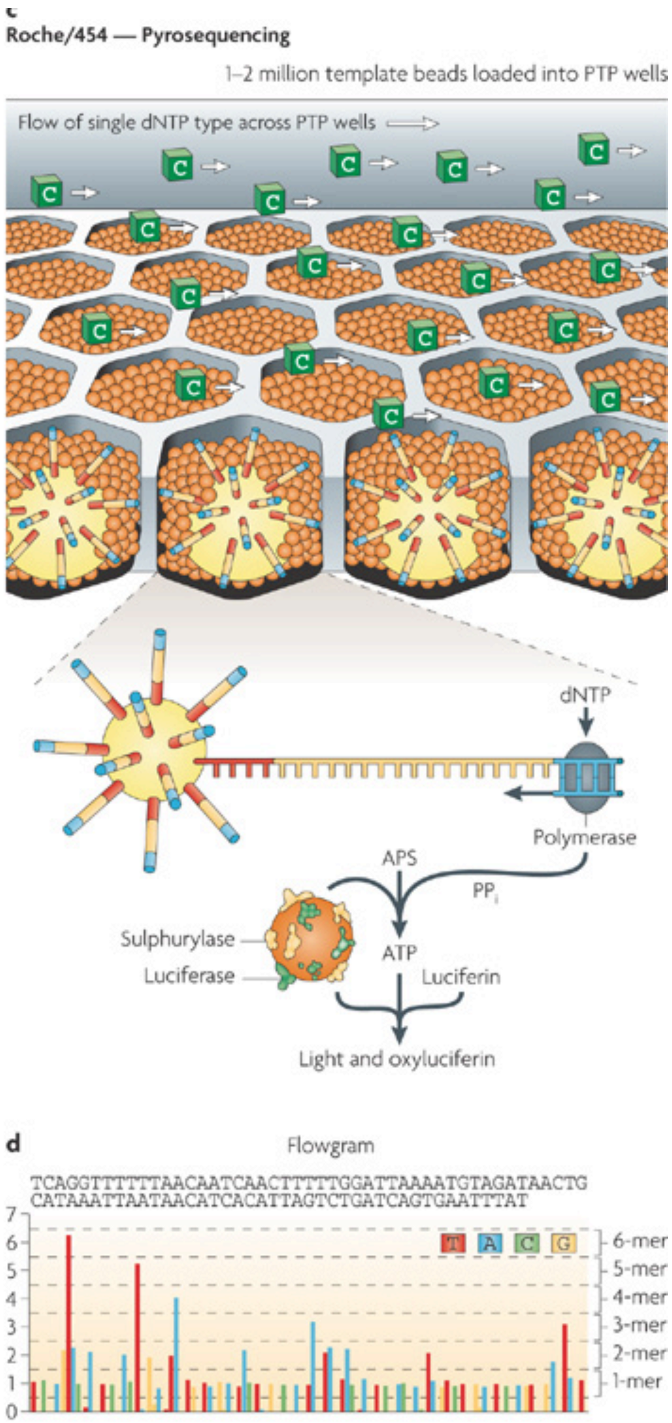
To sequence the skipped positions, the anchor and ligated oligonucleotides may be stripped off the target DNA sequence, and another round of sequencing by ligation started with an anchor one or more bases shorter.

Issues

- DNA ligase is sensitive to the structure of DNA.
- DNA ligase has very low efficiency when there are mismatches between the bases of the two strands.
- This sequencing by ligation method has been reported to have problem sequencing palindromic sequences.

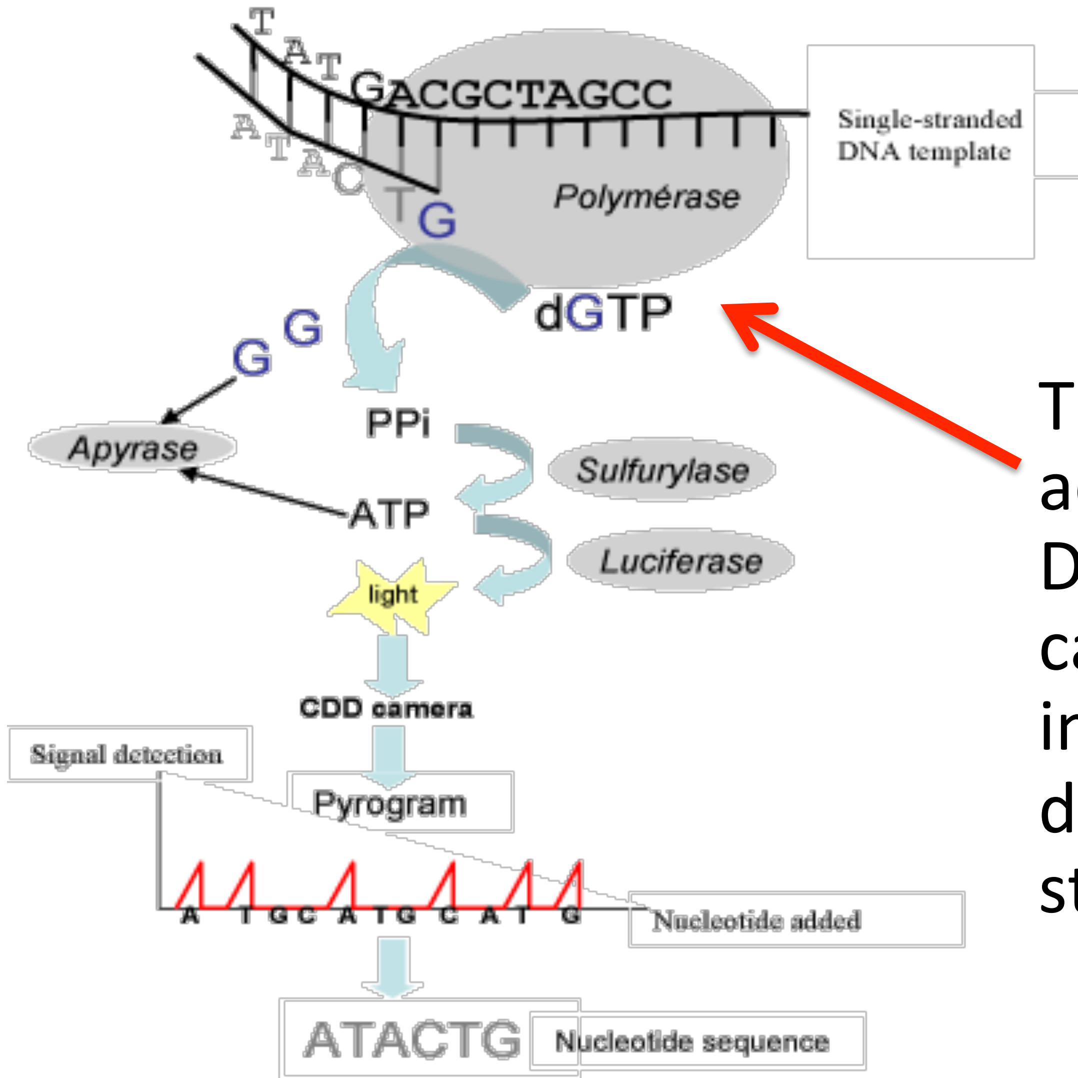
Sequencing and imaging

- Technologies:
  1. Cyclic reversible termination (Sequencing By Synthesis) (Helicos BioSceinces, Illumina)
  2. Sequencing by ligation (SOLiD)
  3. Pyrosequencing (454)
  4. Real-time sequencing (Pacific Biosciences)

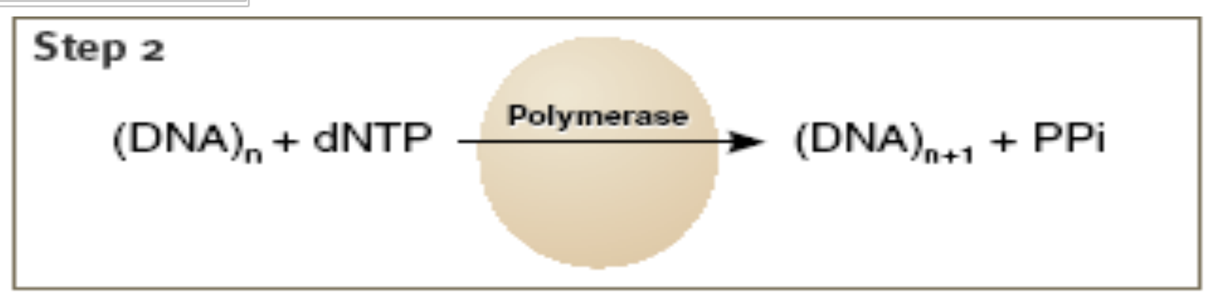


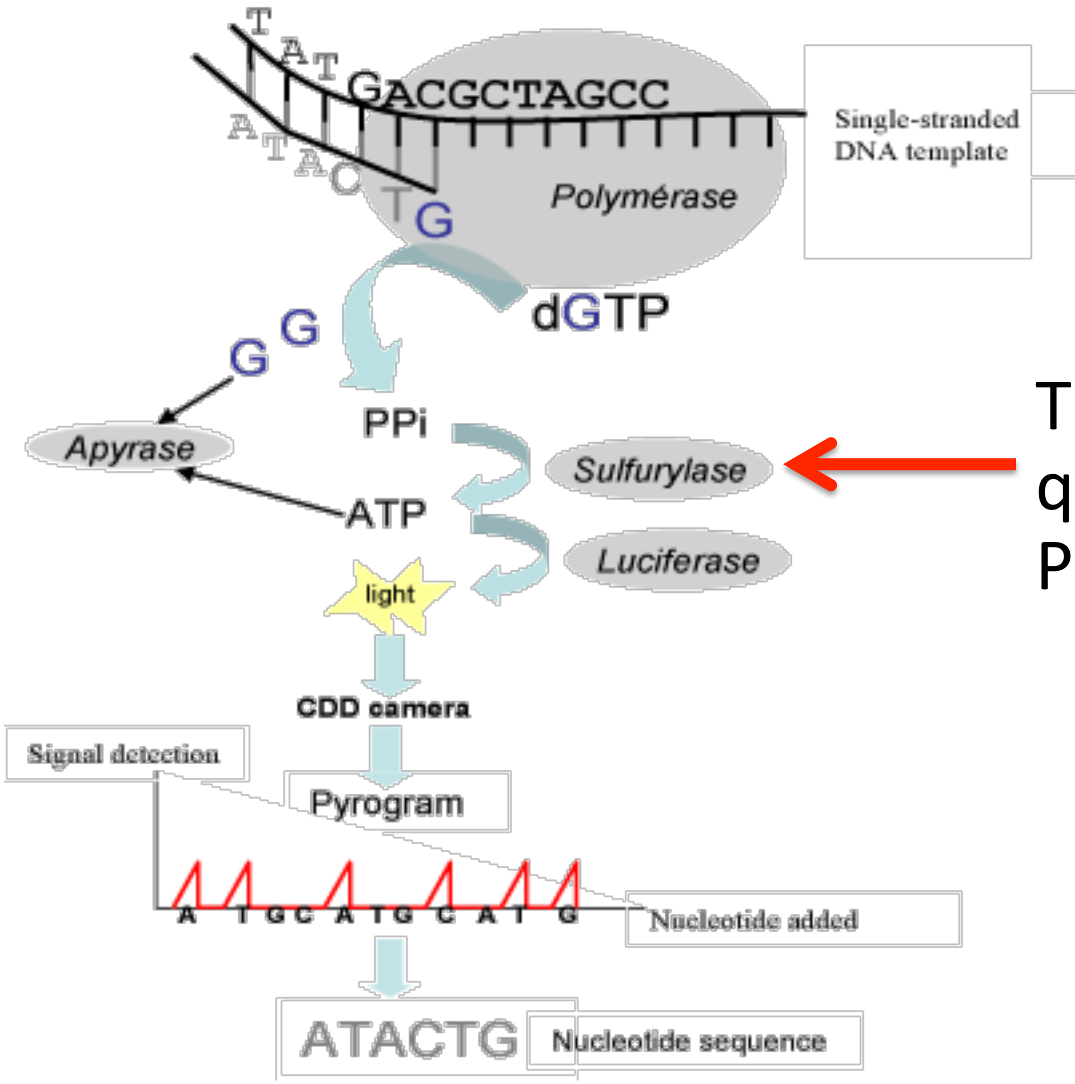
Pyrosequencing

Pyrosequencing measures the release of inorganic pyrophosphate by proportionally converting it into visible light using a series of enzymatic reactions.
The pyrosequencing method manipulates DNA polymerase by the single addition of a dNTP in limiting amounts. Upon incorporation of the complementary dNTP, DNA polymerase extends the primer and pauses. DNA synthesis is reinitiated following the addition of the next complementary dNTP in the dispensing cycle.
The order and intensity of the light peaks are recorded as flowgrams, which reveal the underlying DNA sequence

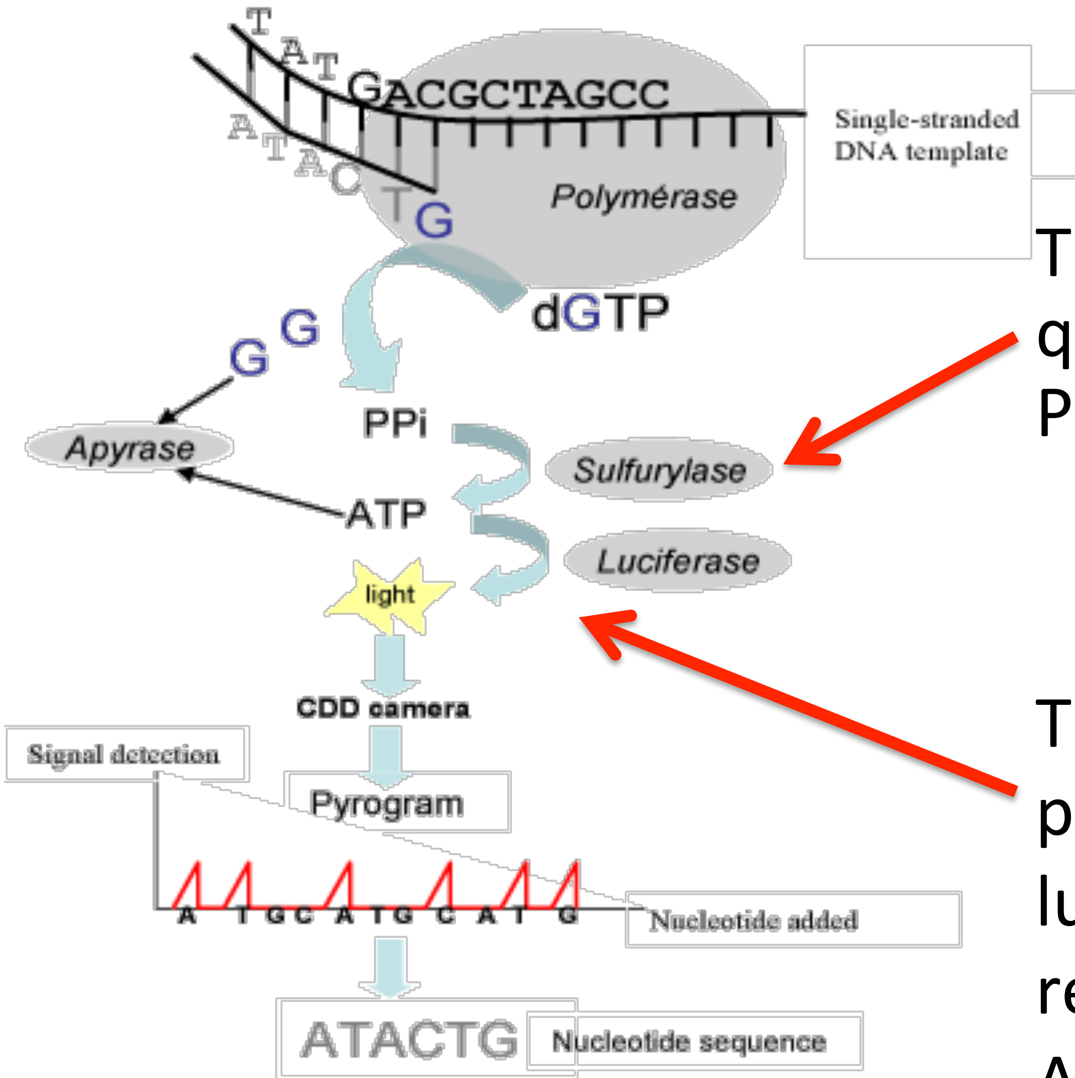


The first of four dNTPs is added to the reaction. DNA polymerase catalyzes the incorporation of the dNTP into the DNA strand.



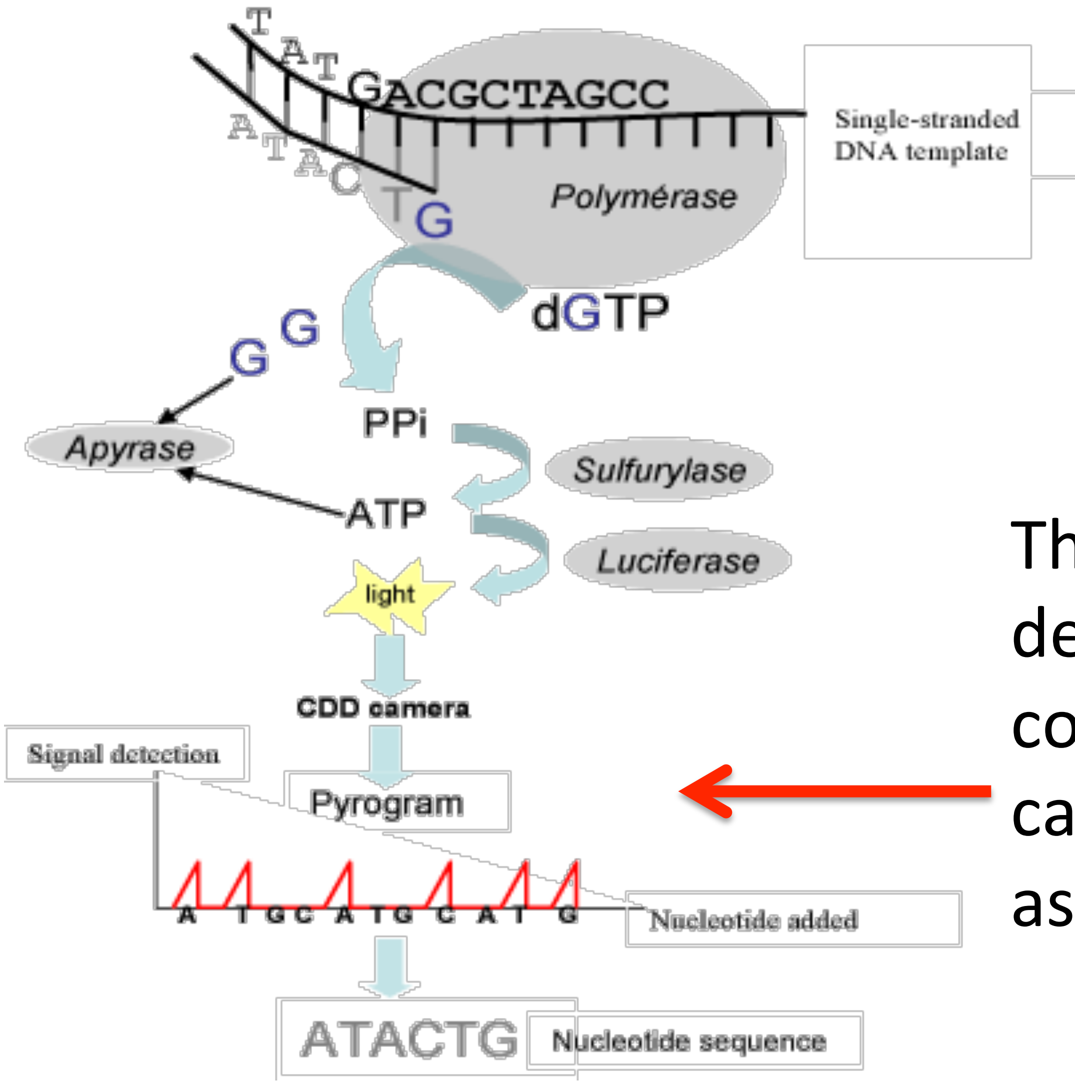


The ATP sulfurylase quantitatively converts PPi to ATP.

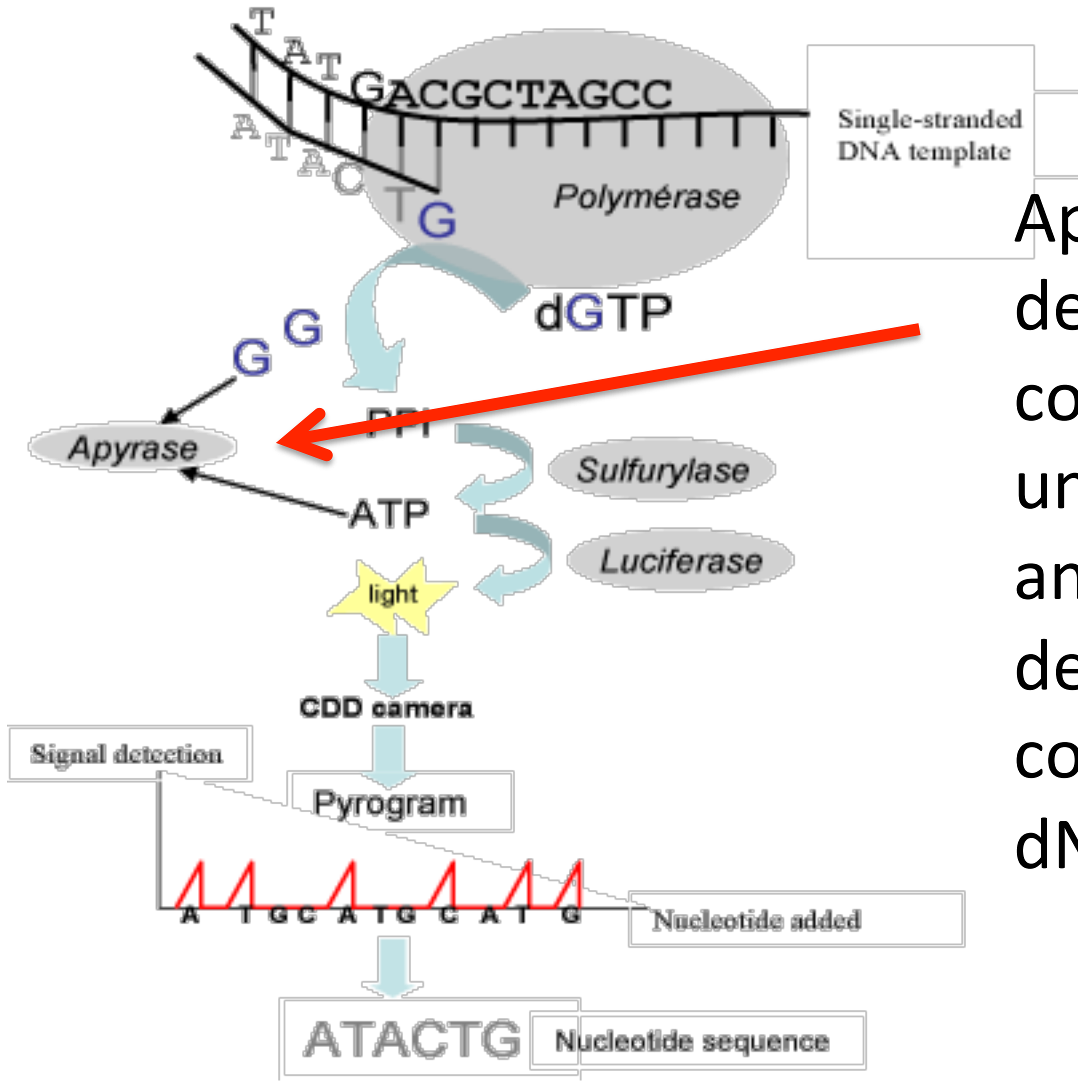


The ATP sulfurylase quantitatively converts PPi to ATP.

The signal light is produced by the luciferase-catalyzed reaction in presence of ATP.

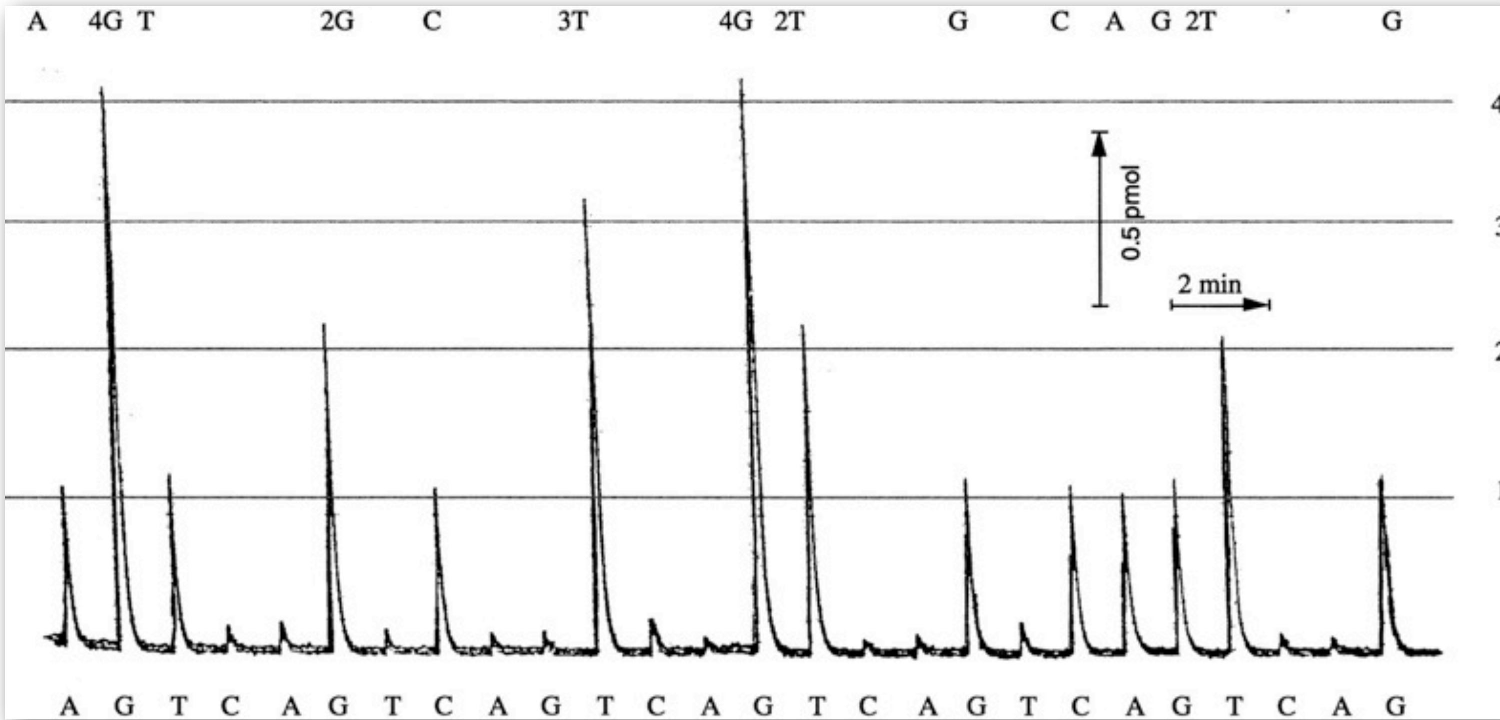


The signal light is detected by a charge coupled device (CCD) camera and integrated as a peak in a Pyrogram.



Apyrase, a nucleotide degrading enzyme, continuously degrades unincorporated dNTPs and excess ATP. When degradation is complete, another dNTP is added.

Pyrogram



Ronaghi, Genome Res 11:3-11 (2001)

AGGGGTGGCTTTGGGGTTGCAGTTG

Sequencing and imaging

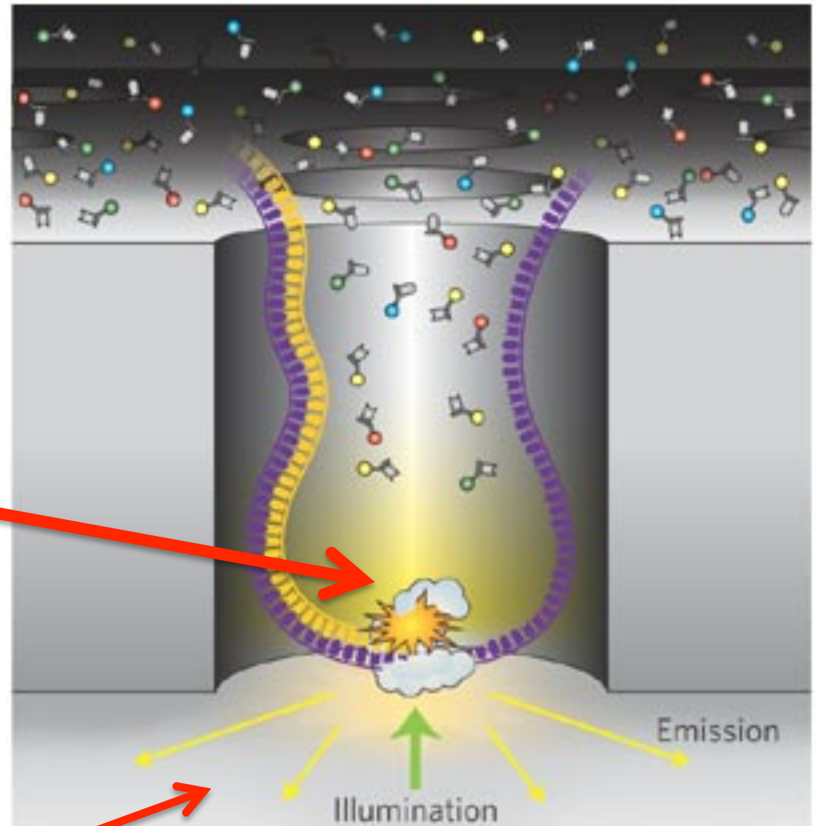
- Technologies:
  1. Cyclic reversible termination (Sequencing By Synthesis) (Helicos BioSceinces, Illumina)
  2. Sequencing by ligation (SOLiD)
  3. Pyrosequencing (454)
  4. Real-time sequencing (Pacific Biosciences)

Real-time sequencing

- Real-time sequencing is is a parallelized **single molecule DNA** sequencing by synthesis technology. Single molecule real time sequencing (also known as SMRT)
- Unlike reversible terminators, real-time nucleotides do not halt the process of DNA synthesis.
- The method of real-time sequencing involves imaging the continuous incorporation of dye-labelled nucleotides during DNA synthesis.

Real-time sequencing

A single DNA polymerase molecule is immobilized at the bottom of a waveguide.

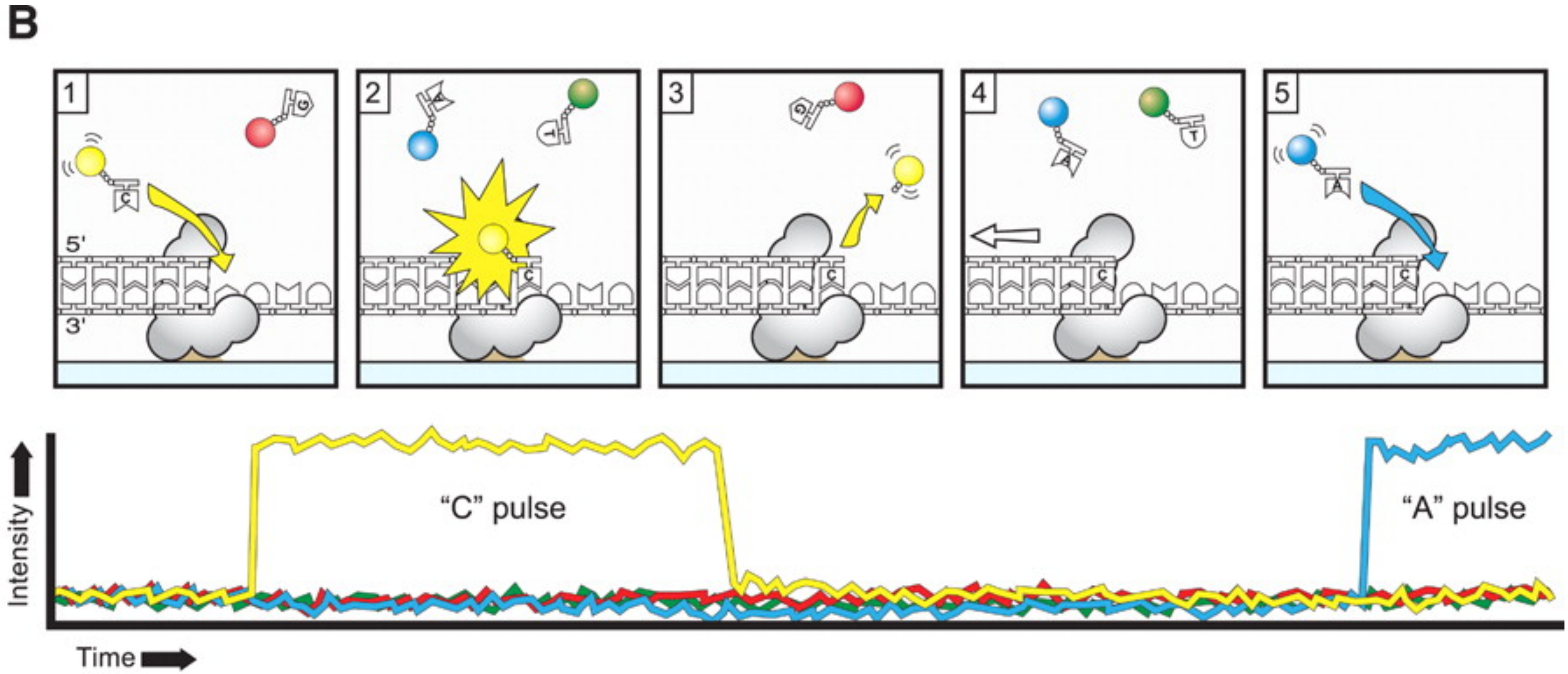


The detector can record the charge of H+

the bottom surface has an individual
zero-mode waveguide detectors

J Eid et al. Science 2009;323:133-138; David Pile Nature Photonics 3, 79 - 80 (2009)

Principle of single-molecule, real-time DNA



(1) A phospholinked nucleotide forms a cognate association with the template in the polymerase active site, (2) causing an elevation of the fluorescence output on the corresponding color channel. (3) Phosphodiester bond formation liberates the dye-linker-pyrophosphate product, which diffuses out of the ZMW, thus ending the fluorescence pulse. (4) The polymerase translocates to the next position, and (5) the next cognate nucleotide binds the active site beginning the subsequent pulse.

J Eid et al. Science 2009;323:133-138

Ion PGM™ System

https://www.youtube.com/watch?v=ZL7DXFPz8rU

The fastest and most affordable benchtop sequencer

Performance comparison

	Run time	Gb/run
Roche 454	8.5 hr	45
Illumina	9 days	35
SOLiD	14 days	50
Helicos	8 days	37
PGM	8 hrs	?

	Roche/454	SOLiD	Hi-Seq 2000	Pacific Biosci RS
Amplification	emPCR on bead surface	emPCR* on bead surface	Enzymatic amplification on glass surface	NA
Sequencing	Polymerase-mediated incorporation of unlabelled nucleotides	Ligase-mediated addition of 2-base encoded fluorescent oligonucleotides	Polymerase-mediated incorporation of end-blocked fluorescent nucleotides	Polymerase-mediated incorporation of terminal phosphate labelled fluorescent nucleotides
Detection	Light emitted from secondary reactions initiated by release of PPi	Fluorescent emission from ligated dye-labelled oligonucleotides	Fluorescent emission from incorporated dye-labelled nucleotides	Real time detection of fluorescent dye in polymerase active site during incorporation
Error model	Substitution errors rare, insertion/ deletion errors at homopolymers	End of read substitution errors	End of read substitution errors	Random insertion/ deletion errors
Read length	400 bp	75 bp	150 bp	>1,000 bp

Mardis, Nature 2011

Accuracy - base calling error

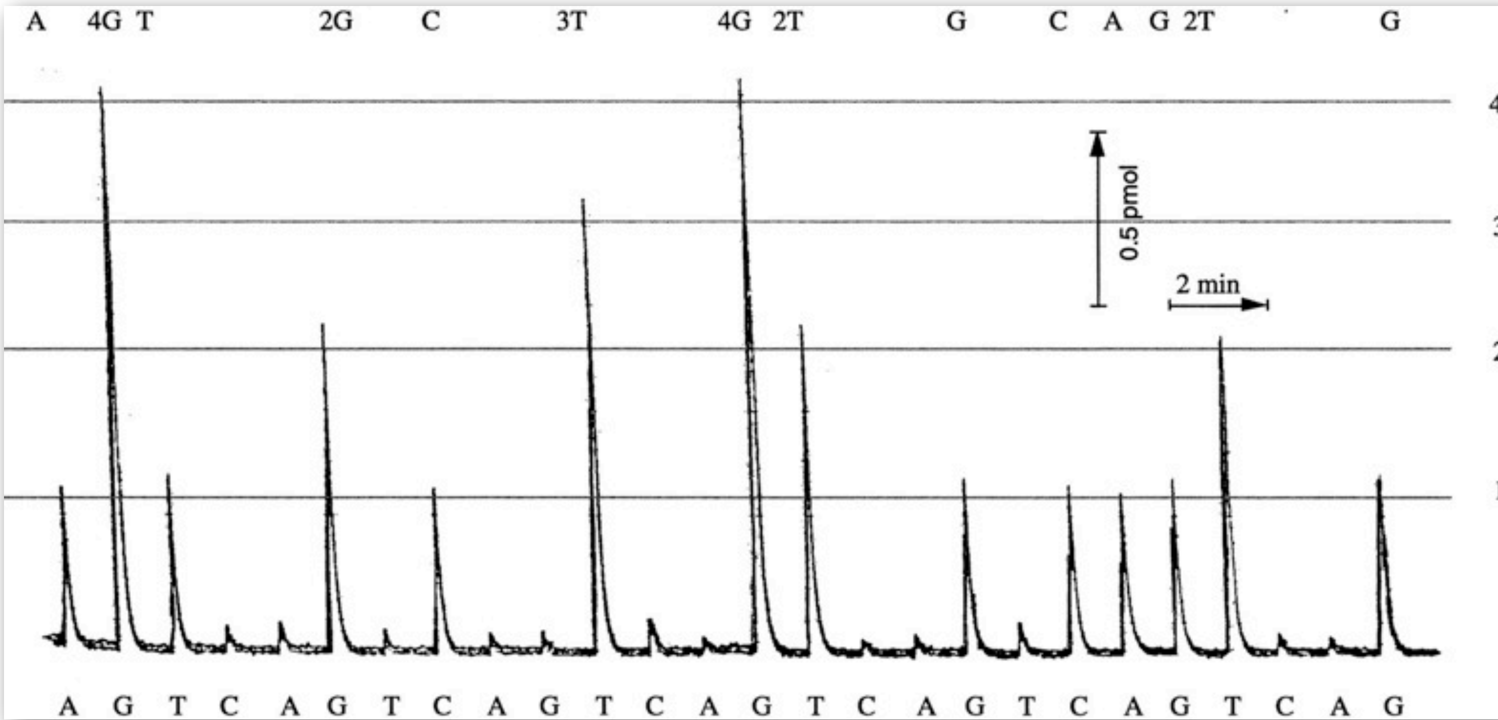
- base quality drops along read
  Sanger > SOLiD > Illumina > 454 > Helicos

Accuracy - homopolymer runs

- Issue for Roche 454:
  39% of errors are homopolymers
- A5 (e.g. TTTTT) motifs: 3.3% error rate
- A8 (e.g. TTTTTTTT)motifs: 50% error rate

- Reason: use signal intensity as a measure for homopolymer length

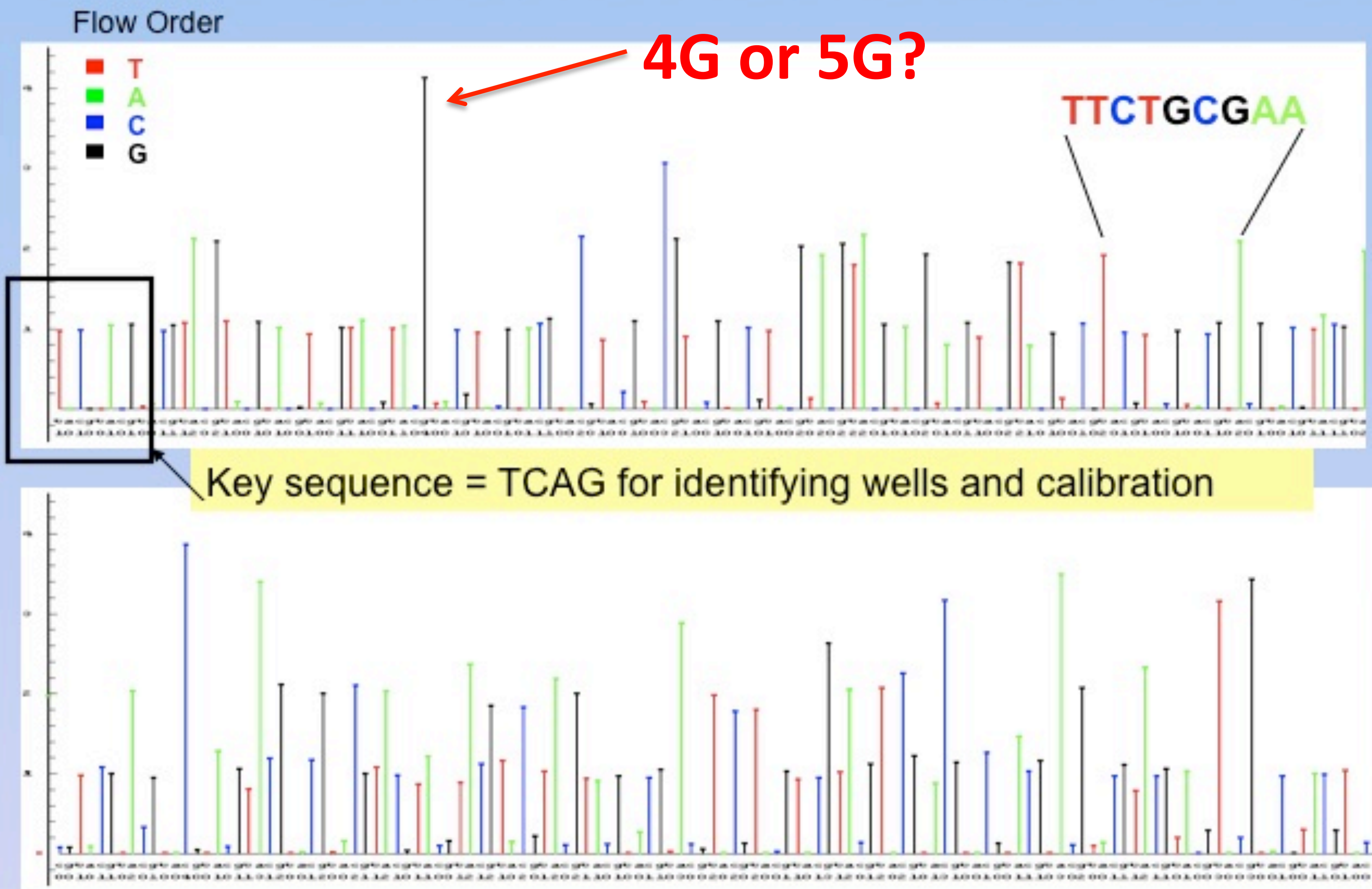
Accuracy - homopolymer runs



Ronaghi, Genome Res 11:3-11 (2001)

AGGGGTGGCTTTGGGGTTGCAGTTG

Example of a Flowgram



Accuracy - homopolymer runs

Real	**AGGGGTGGCTTTGGGGTTGCAGTTG**
As read by 454	**AGGGGGTGGCTTTGGGGTTGCAGTTG**

NGS

- Introduction to the background
- NGS workflow and accuracy
- Data format and quality control
- Assembly
- RNA-seq
  – Aligner
  – Analysis tools
  – Applications, such as MiRNA
- Chip-seq
  – Applications