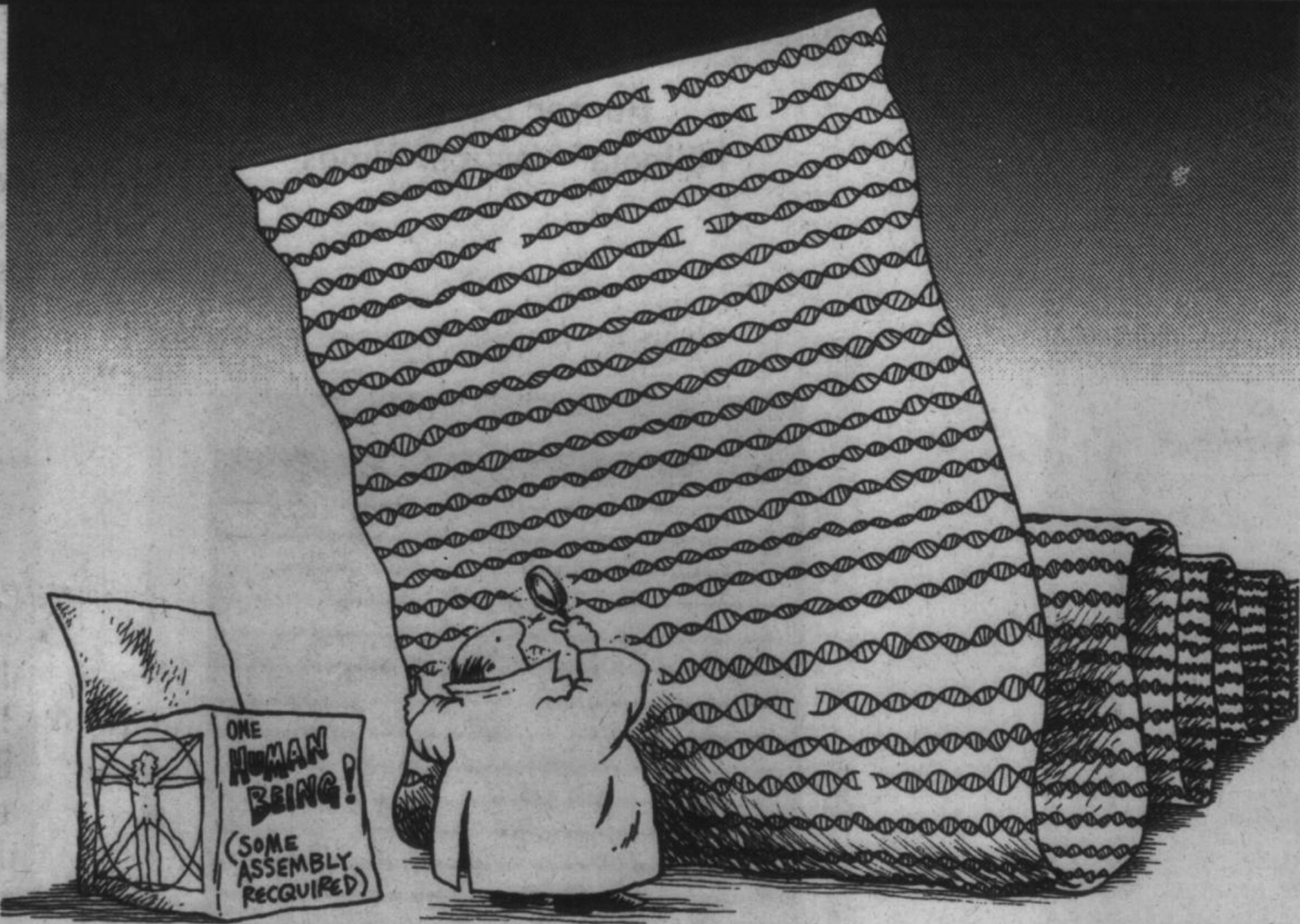




BY AUTH FOR THE PHILADELPHIA INQUIRER



# Modern Approaches to Sequencing

Dr Konrad Paszkiewicz, Head, Exeter Sequencing Service,  
Director Wellcome Trust Biomedical Informatics Hub,  
January 2014



# Contents

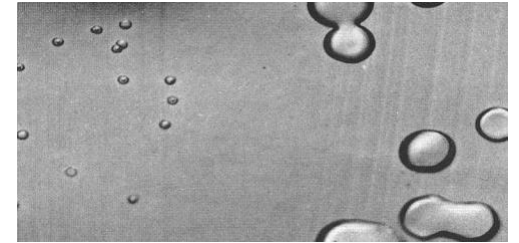
- Review of Sanger Sequencing
- Timeline and impact of human genome project
- Second generation sequencing technologies
- Third generation sequencing technologies
- Nanopore sequencing technologies

# USB Nanopore sequencer

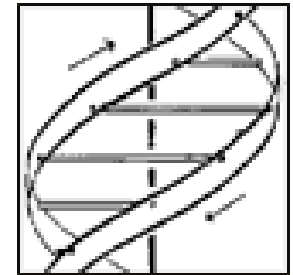


# Timeline

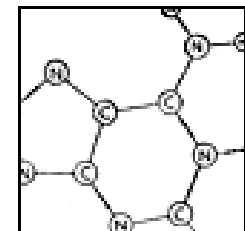
**1944:** Avery, O.T., et al “Studies on the chemical nature of the substance inducing transformation of Pneumococcal types”



**1953:** Watson J.D. and Crick F.H.C. “A Structure for Deoxyribose Nucleic Acid”



**1953:** Watson J.D. and Crick F.H.C. “Genetical Implications of the structure of Deoxyribonucleic Acid”



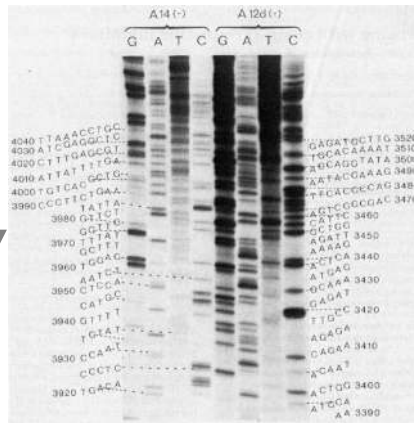
**1959** – First homeogenous DNA purified

# Timeline

**1970** – First discovery of type II restriction enzymes

**1972:** sequencing of the first gene from RNA by Walter Fiers

**1976:** sequencing of the first complete genome by Fiers (Bacteriophage MS2 which infects *E.coli*)



**1977:** Maxam AM, Gilbert W. "A new method for sequencing DNA".

**1977:** Sanger F, Nicklen S, Coulson AR. "DNA sequencing with chain-terminating inhibitors"

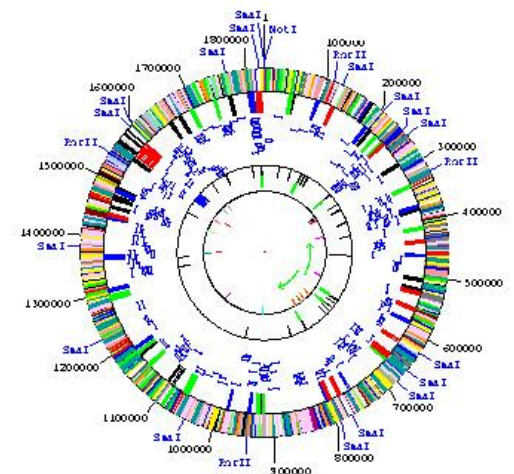
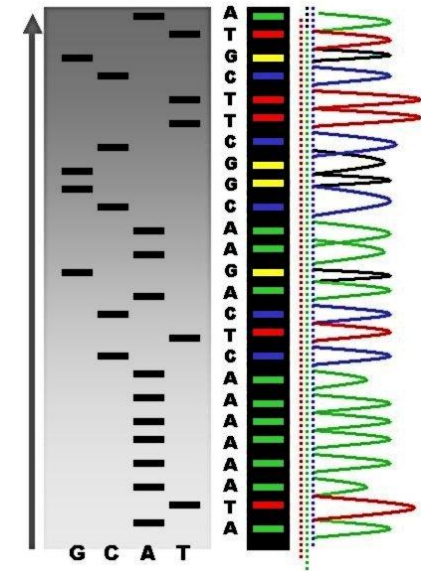
# Timeline

**1985-86:** Leroy Hood use fluorescently labeled ddNTPs, set the stage for automated sequencing

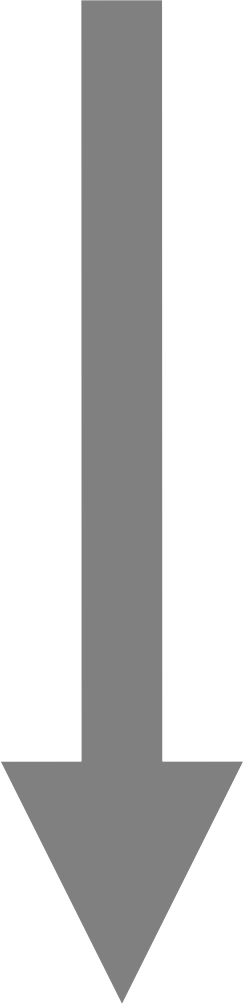
**1987:** Applied Biosystems markets first automated sequencing machine (ABI 370)

**1990:** National Institutes of Health (NIH) begins large-scale sequencing trials (\$0.75/base) Human Genome Project (HGP) begins, \$3-billion and 15 years

**1995:** Craig Venter at TIGR published the *Haemophilus influenzae* genome. First use of whole-genome shotgun sequencing



# Timeline



**1998:** Green & Ewing publish “phred” base caller/scorer

**2000:** Sydney Brenner and Lynx Therapeutics publishes “MPSS”, parallelized bead-base sequencing tech, launches “Next-Gen”

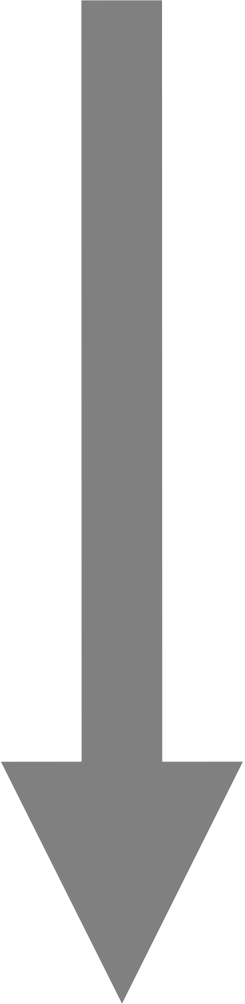
**2001:** HGP/Celera draft assembly published in Nature/Science

**2003:** HGP “complete” genome released. ENCODE project launched

**2004:** 454 releases pyrosequencer, costs 6-fold less than automated Sanger sequencing



# Illumina-era Timeline



**1998:** Shankar Balasubramanian and David Klenerman patent "A method for reproducing molecular arrays" and found Solexa



**2001:** £12 million series A funding



**2003:** Mercier, J.-F., et al. "Solid phase DNA amplification"

**2004:** Solexa acquires Solid phase DNA amplification method



# Illumina-era Timeline

**2006:** Solexa release Genome Analyser I

**2007:** Illumina acquires Solexa

**2008:** Illumina releases GAI

**2008:** Human microbiome project launched

**2010:** Illumina HiSeq 2000 released

**2011:** MiSeq launched

**2012:** ENCODE publications

**2012:** Illumina HiSeq 2500 released

**2013-2015:** RIP 454



# Review of Sanger Sequencing





## Fred Sanger 1918-2013

Double Nobel laureate and developer of the dideoxy sequencing method, first published in December 1977. [Credit: Wellcome Images]

"Fred Sanger is a quiet giant, whose discoveries and inventions transformed our research world." (A.Bradley, WTSI.)

# The challenge of DNA sequencing

- 1953 – Double helix discovered
- 1971 - First DNA sequence determined (12bp!)
- 1977 – Sanger sequencing method published
- Why did it take so long?

# Some possible reasons

- The chemical properties of different DNA molecules were so similar that it appeared difficult to separate them
- The chain length of naturally occurring DNA molecules is much greater than for proteins and made complete sequencing seem unapproachable.
- The 20 amino acid residues found in proteins have widely varying properties that had proven useful in the separation of peptides.
  - Only four bases in DNA made sequencing a more difficult problem for DNA than for protein.
- No base-specific DNAases were known.
  - Protein sequencing had depended upon proteases that cleave adjacent to certain amino acids
- DNA was considered boring compared to proteins

# Steps on the road to sequencing

- 1959 – First homeogenous DNA purified
- 1970 – First discovery of type II restriction enzymes
- 1972 – First RNA gene sequence published (lac operon)
- 1975 – Sanger first publishes his plus/minus method of sequencing (unable to distinguish homopolymers)
- 1977 – Maxam & Gilbert publish their method (could distinguish homopolymers)
- 1977 – Sanger publishes Dideoxy sequencing method

# Maxam-Gilbert Sequencing

*Proc. Natl. Acad. Sci. USA*  
Vol. 74, No. 2, pp. 560-564, February 1977  
Biochemistry

## A new method for sequencing DNA

(DNA chemistry/dimethyl sulfate cleavage/hydrazine/piperidine)

ALLAN M. MAXAM AND WALTER GILBERT

Department of Biochemistry and Molecular Biology, Harvard University, Cambridge, Massachusetts 02138

*Contributed by Walter Gilbert, December 9, 1976*

**ABSTRACT** DNA can be sequenced by a chemical procedure that breaks a terminally labeled DNA molecule partially at each repetition of a base. The lengths of the labeled fragments then identify the positions of that base. We describe reactions that cleave DNA preferentially at guanines, at adenines, at cytosines and thymines equally, and at cytosines alone. When the products of these four reactions are resolved by size, by electrophoresis on a polyacrylamide gel, the DNA sequence can be read from the pattern of radioactive bands. The technique will permit sequencing of at least 100 bases from the point of labeling.

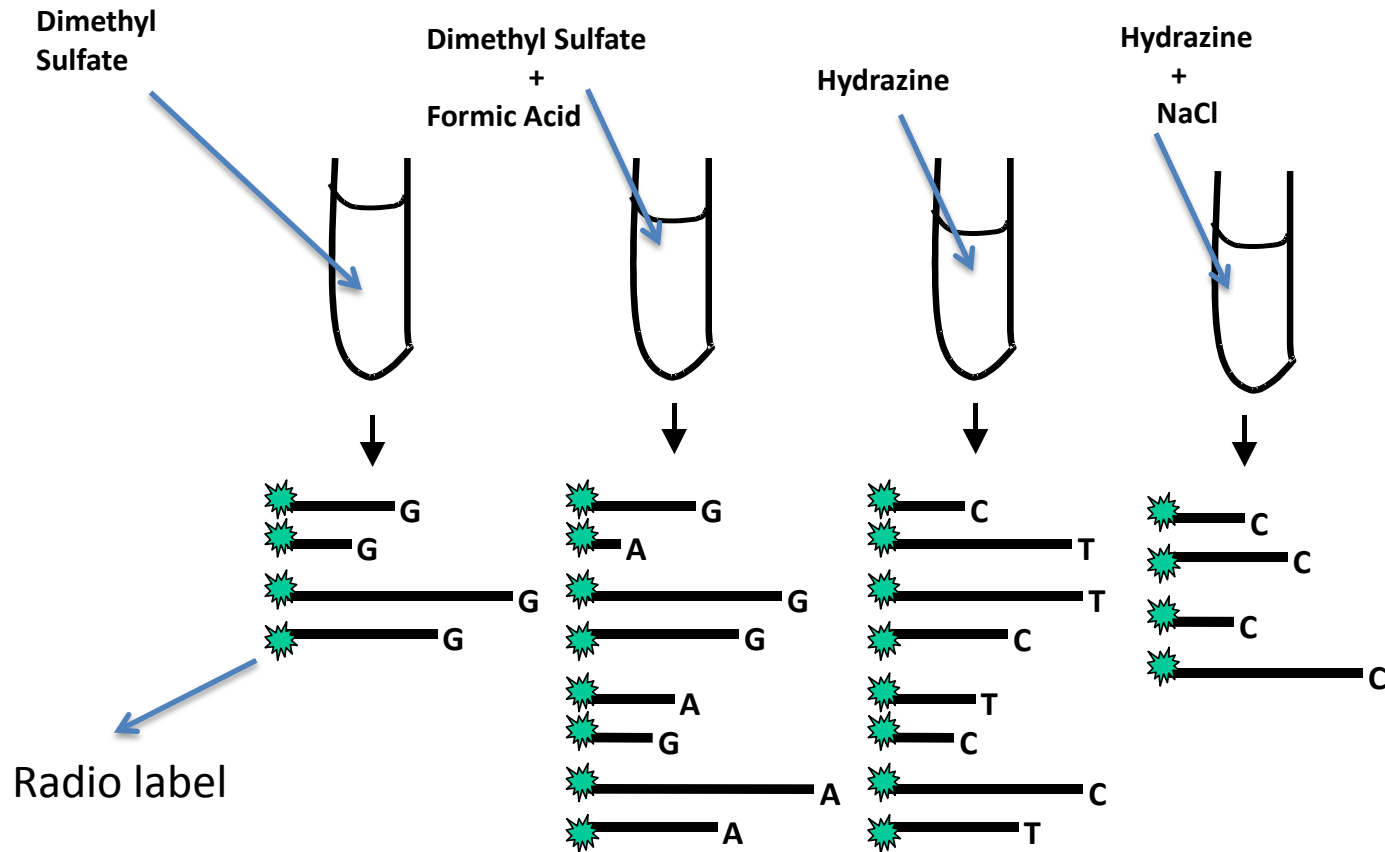
We have developed a new technique for sequencing DNA molecules. The procedure determines the nucleotide sequence of a terminally labeled DNA molecule by breaking it at adenine, guanine, cytosine, or thymine with chemical agents. Partial cleavage at each base produces a nested set of radioactive

## THE SPECIFIC CHEMISTRY

**A Guanine/Adenine Cleavage (2).** Dimethyl sulfate methylates the guanines in DNA at the N7 position and the adenines at the N3 (3). The glycosidic bond of a methylated purine is unstable (3, 4) and breaks easily on heating at neutral pH, leaving the sugar free. Treatment with 0.1 M alkali at 90° then will cleave the sugar from the neighboring phosphate groups. When the resulting end-labeled fragments are resolved on a polyacrylamide gel, the autoradiograph contains a pattern of dark and light bands. The dark bands arise from breakage at guanines, which methylate 5-fold faster than adenines (3).

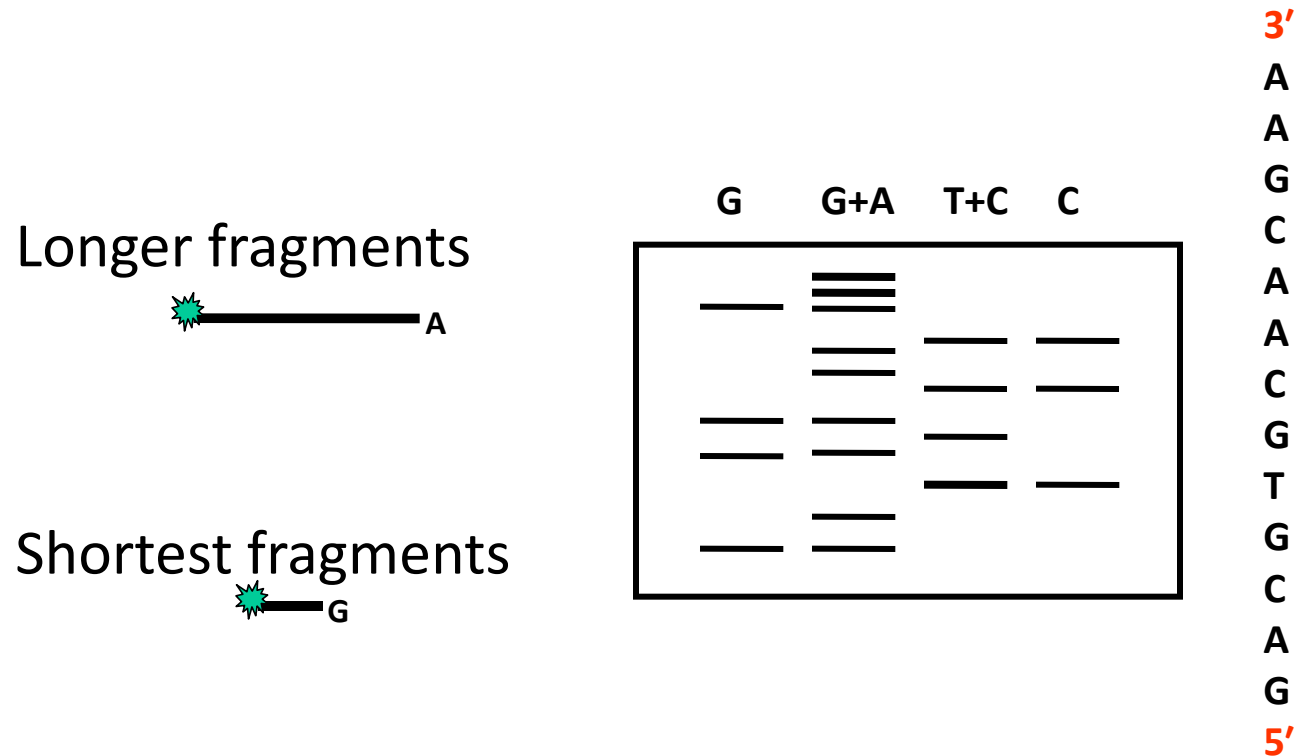
This strong guanine/weak adenine pattern contains almost half the information necessary for sequencing; however, ambiguities can arise in the interpretation of this pattern because the intensity of isolated bands is not easy to assess. To determine

# Maxam-Gilbert Sequencing



Maxam-Gilbert sequencing is performed by **chain breakage** at specific nucleotides.

# Maxam-Gilbert Sequencing



Sequencing gels are read from **bottom to top** (5' to 3').

# Sanger di-deoxy sequencing method

*Proc. Natl. Acad. Sci. USA*  
Vol. 74, No. 12, pp. 5463-5467, December 1977  
Biochemistry

## DNA sequencing with chain-terminating inhibitors

(DNA polymerase/nucleotide sequences/bacteriophage  $\phi$ X174)

F. SANGER, S. NICKLEN, AND A. R. COULSON

Medical Research Council Laboratory of Molecular Biology, Cambridge CB2 2QH, England

*Contributed by F. Sanger, October 3, 1977*

**ABSTRACT** A new method for determining nucleotide sequences in DNA is described. It is similar to the "plus and minus" method [Sanger, F. & Coulson, A. R. (1975) *J. Mol. Biol.* 94, 441-448] but makes use of the 2',3'-dideoxy and arabinonucleoside analogues of the normal deoxynucleoside triphosphates, which act as specific chain-terminating inhibitors of DNA polymerase. The technique has been applied to the DNA of bacteriophage  $\phi$ X174 and is more rapid and more accurate than either the plus or the minus method.

The "plus and minus" method (1) is a relatively rapid and simple technique that has made possible the determination of the sequence of the genome of bacteriophage  $\phi$ X174 (2). It depends on the use of DNA polymerase to transcribe specific regions of the DNA under controlled conditions. Although the method is considerably more rapid and simple than other

a stereoisomer of ribose in which the 3'-hydroxyl group is oriented in *trans* position with respect to the 2'-hydroxyl group. The arabinosyl (ara) nucleotides act as chain terminating inhibitors of *Escherichia coli* DNA polymerase I in a manner comparable to ddT (4), although synthesized chains ending in 3' araC can be further extended by some mammalian DNA polymerases (5). In order to obtain a suitable pattern of bands from which an extensive sequence can be read it is necessary to have a ratio of terminating triphosphate to normal triphosphate such that only partial incorporation of the terminator occurs. For the dideoxy derivatives this ratio is about 100, and for the arabinosyl derivatives about 5000.

## METHODS

# Sanger sequencing

AGCTGCCCCG



**ddATP** +  
four dNTPs

**ddA**  
dAdGdCdTdGdCdCdCdG



**ddCTP** +  
four dNTPs

dAdG**ddC**  
dAdGdCdTdG**ddC**  
dAdGdCdTdGdC**ddC**  
dAdGdCdTdGdCdC**ddC**



**ddGTP** +  
four dNTPs

dA**ddG**  
dAdGdCdT**ddG**  
dAdGdCdTdGdCdCdC**ddG**



**ddTTP** +  
four dNTPs

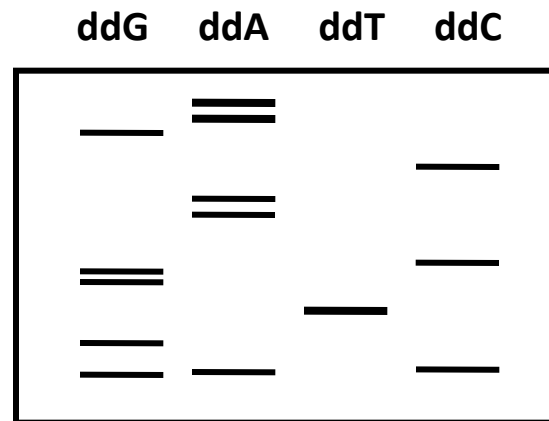
dAdGdC**ddT**  
dAdGdCdTdGdCdCdCdG

# Sanger di-deoxy method

Longer fragments



Shortest fragments



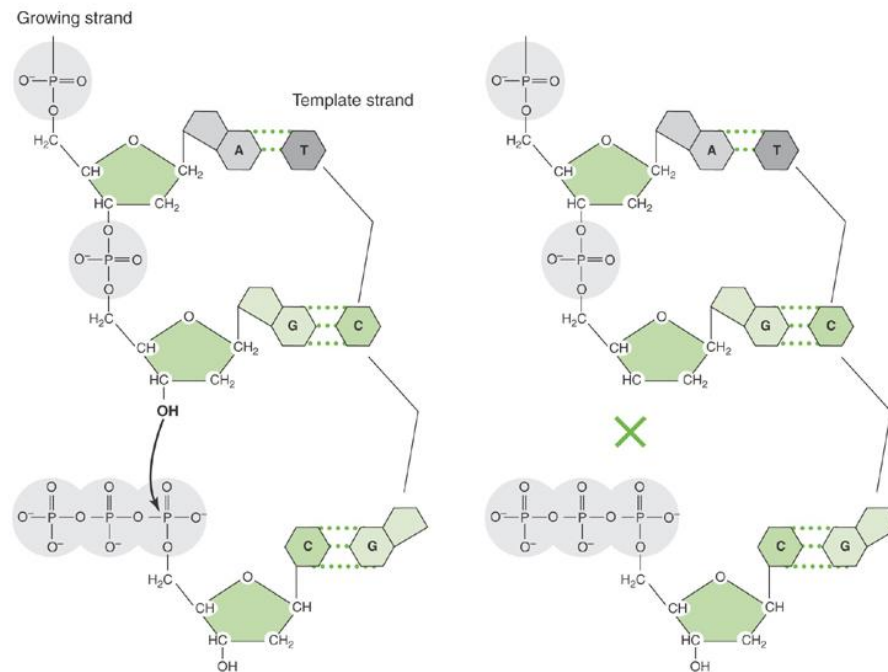
3'  
A  
A  
G  
C  
A  
A  
C  
G  
T  
G  
C  
A  
G  
5'

# Sanger Sequencing

- With addition of enzyme (DNA polymerase), the primer is extended until a ddNTP is encountered.
- The chain will end with the incorporation of the ddNTP.
- With the proper dNTP:ddNTP ratio, the chain will terminate throughout the length of the template.
- All terminated chains will end in the ddNTP added to that reaction.

# How is sequencing terminated at each of the 4 bases?

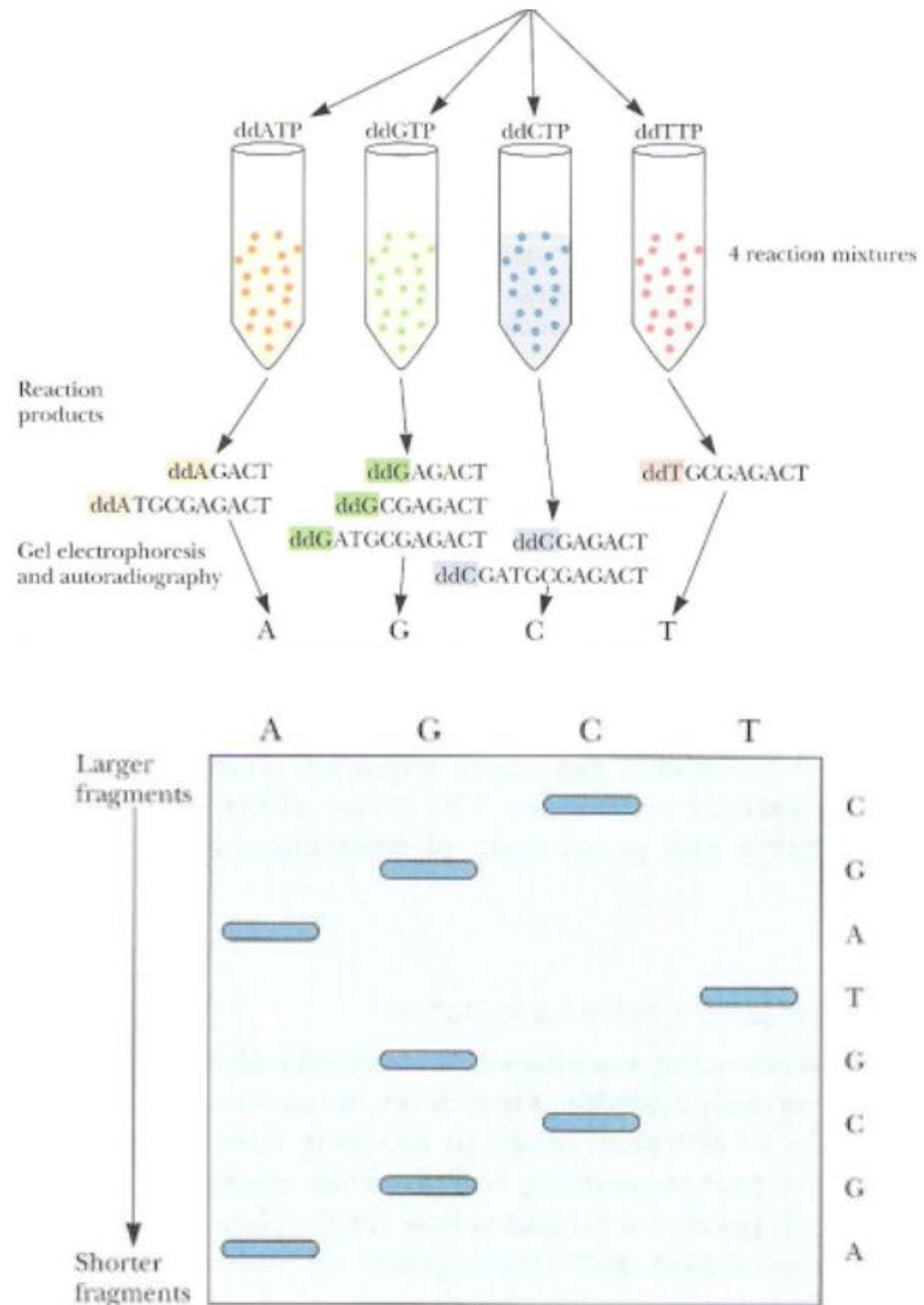
The 3'-OH group necessary for formation of the phosphodiester bond is missing in ddNTPs



Chain terminates  
at ddG

# Dideoxy Method

- Run four separate reactions each with different ddNTPs
- Run on a gel in four separate lanes
- Read the gel from the bottom up



# Improvements to Sanger's original method

- Cycle sequencing is chain termination sequencing performed in a thermal cycler.
  - Requires a heat-stable DNA polymerase.
- Fluorescent dyes are multicyclic molecules that absorb and emit fluorescent light at specific wavelengths.
  - Examples are fluorescein and rhodamine derivatives.
  - For sequencing applications, these molecules can be covalently attached to nucleotides.

# Dye Terminator Sequencing

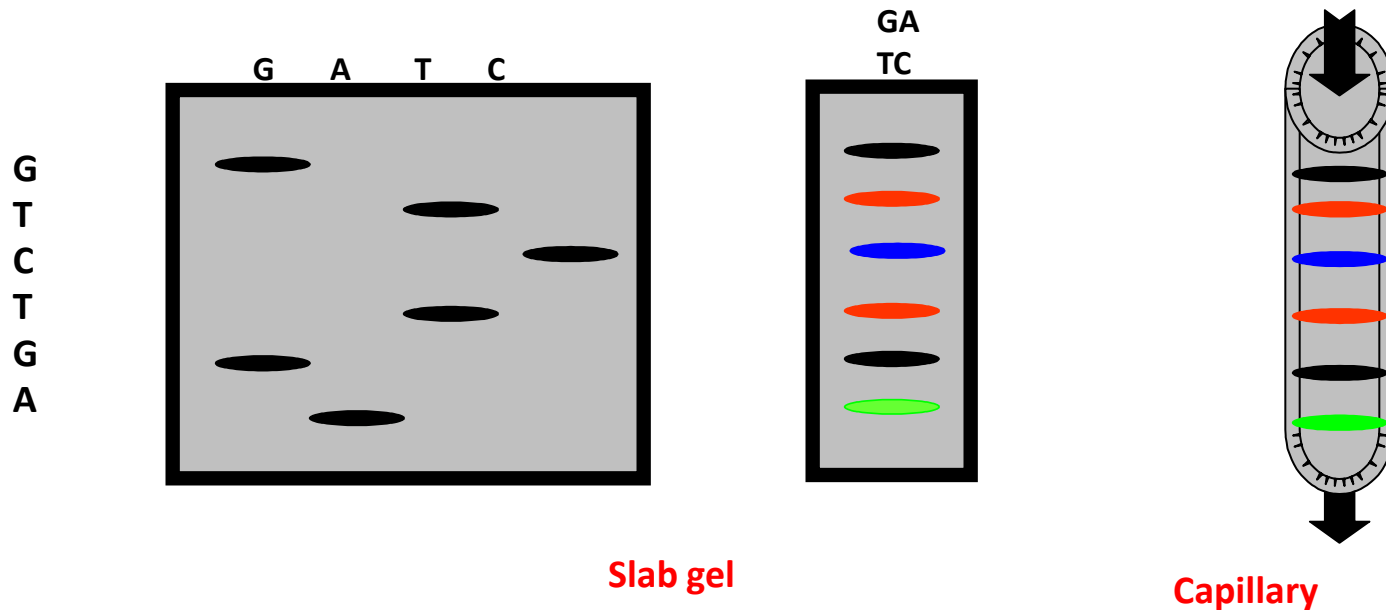
- A distinct dye or “color” is used for each of the four ddNTP.
- Since the terminating nucleotides can be distinguished by color, all four reactions can be performed in a single tube.



The fragments are distinguished by size and “color.”

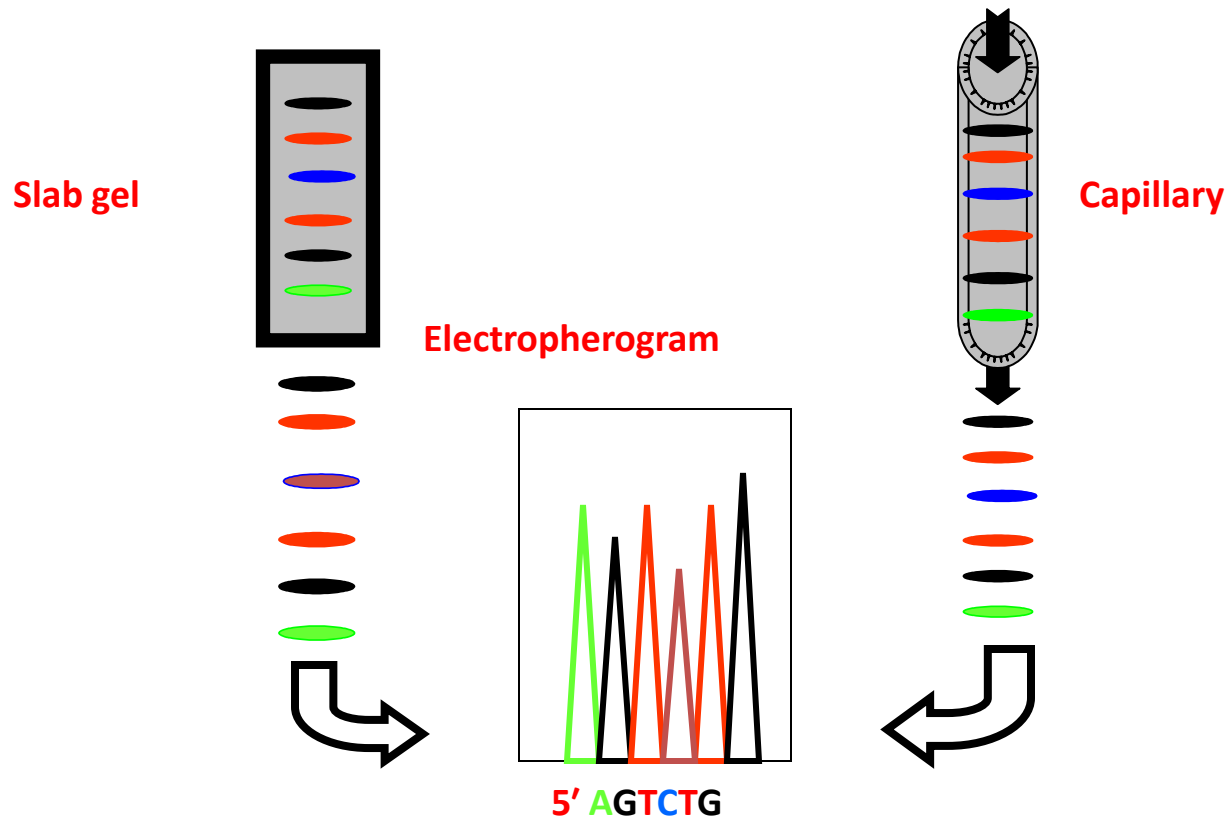
# Dye Terminator Sequencing

The DNA ladder is resolved in one gel lane or in a capillary.

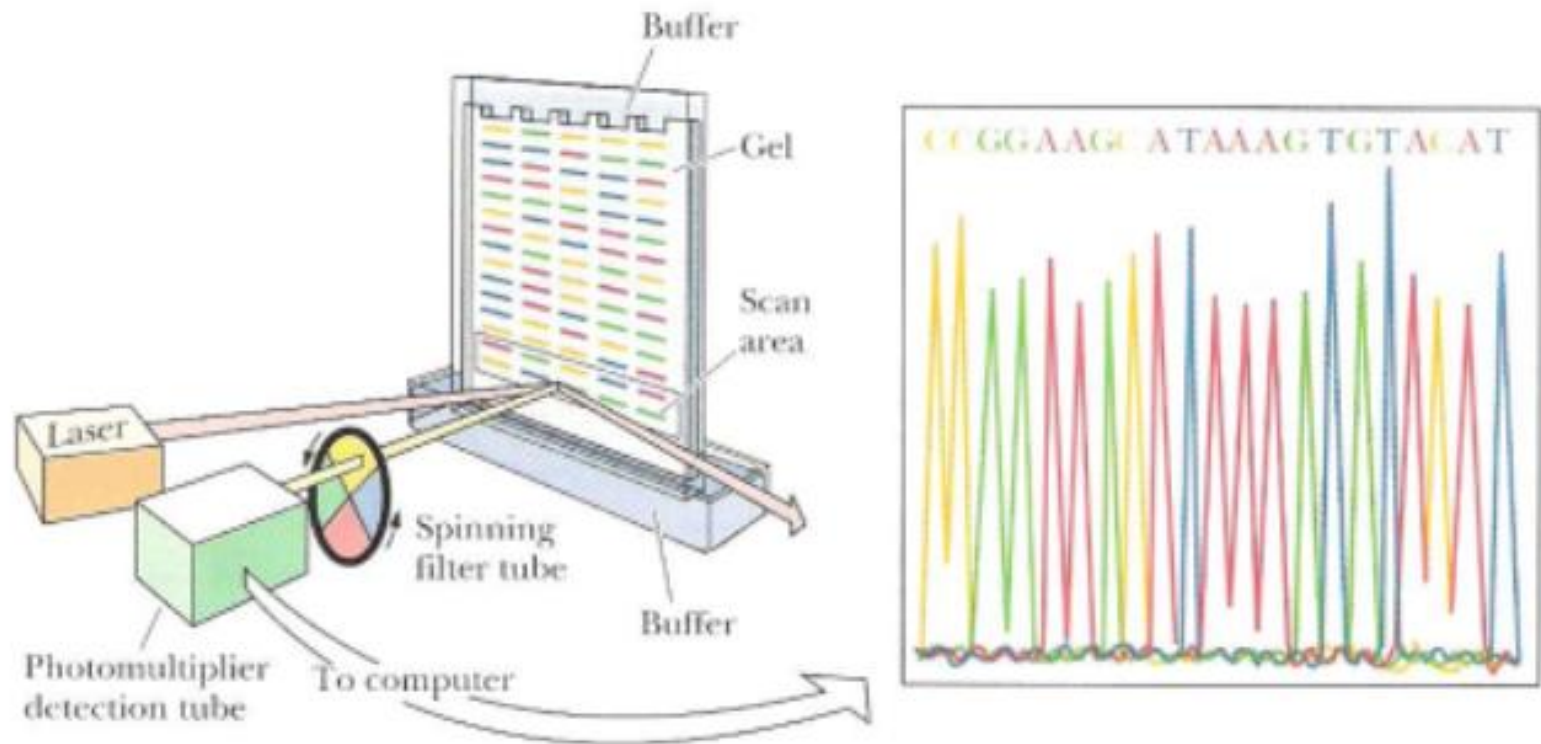


# Dye Terminator Sequencing

- The DNA ladder is read on an **electropherogram**.

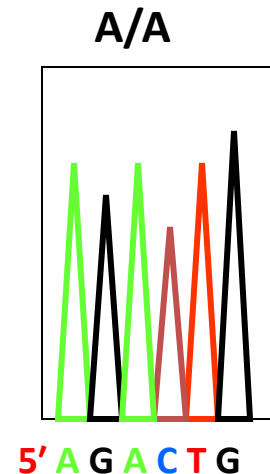
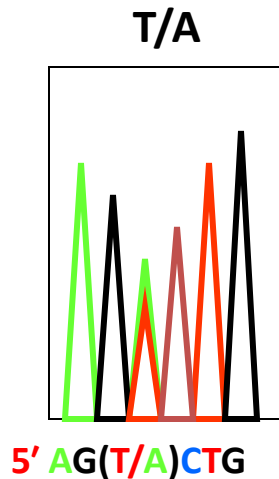
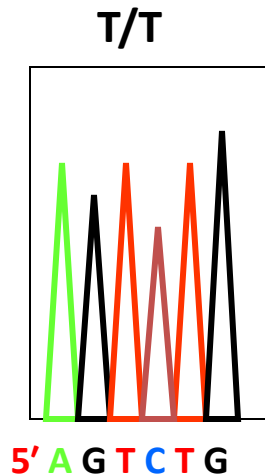


# Automated Version of the Dideoxy Method



# Automated Sequencing

- Dye primer or dye terminator sequencing on capillary instruments.
- Sequence analysis software provides analyzed sequence in text and electropherogram form.
- Peak patterns reflect mutations or sequence changes.



## First generation (Sanger) sequencing

|                    |                                                                                                      |
|--------------------|------------------------------------------------------------------------------------------------------|
| Throughput         | 50-100kb, 96 sequences per run                                                                       |
| Read length        | 0.5-2kbp                                                                                             |
| Accuracy           | high quality bases - 99%: ~900bp<br>very high quality bases - 99.9%:<br>~600bp<br>99.999%: 400-500bp |
| Price per raw base | ~\$200,000/Gb                                                                                        |

# Sanger Sequencing

## Useful videos

- <http://www.youtube.com/watch?v=91294ZAG2hg&feature=related>
- <http://www.youtube.com/watch?v=bEFLBf5WEtc&feature=fvwrel>

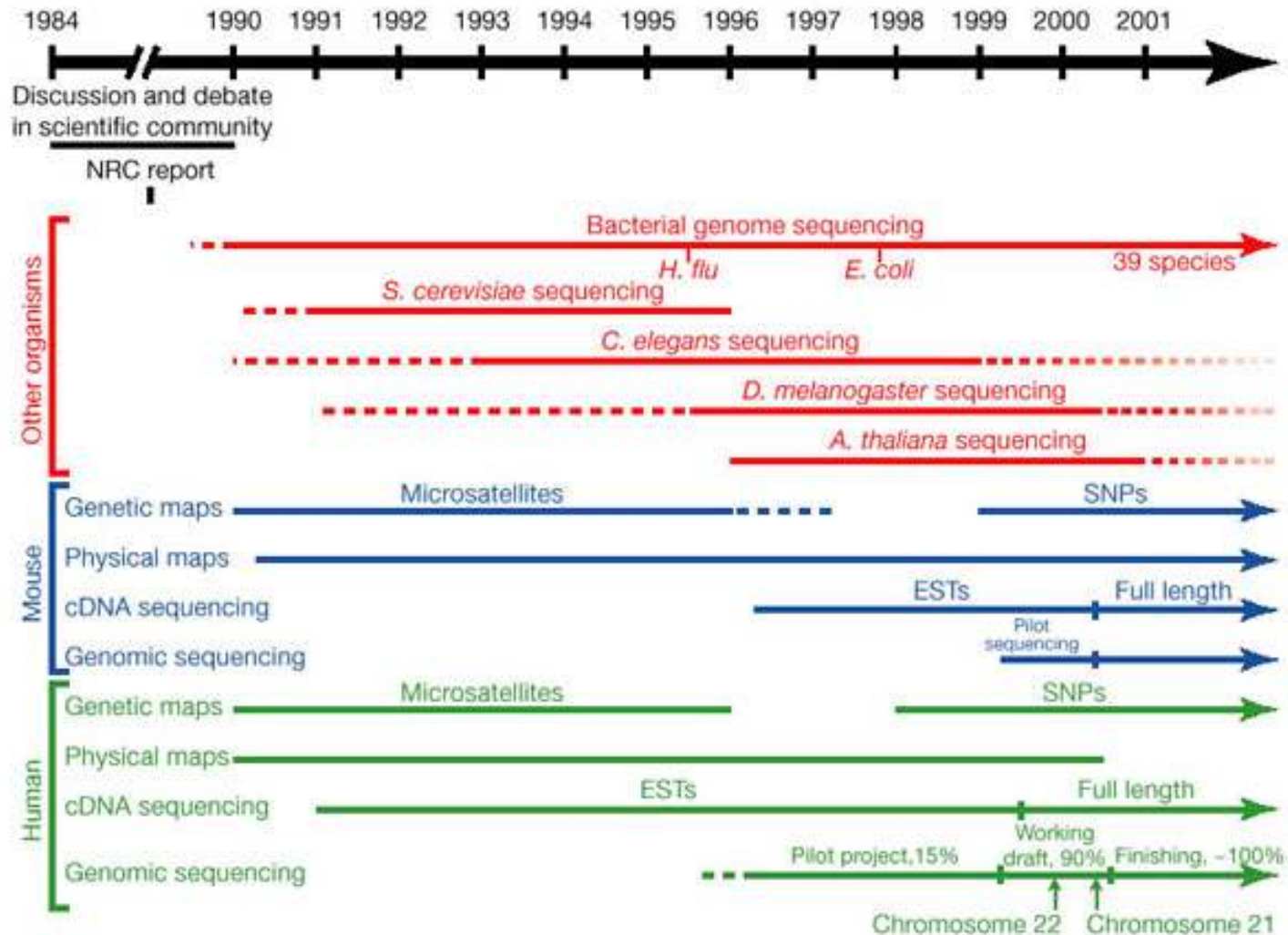
# Human genome project



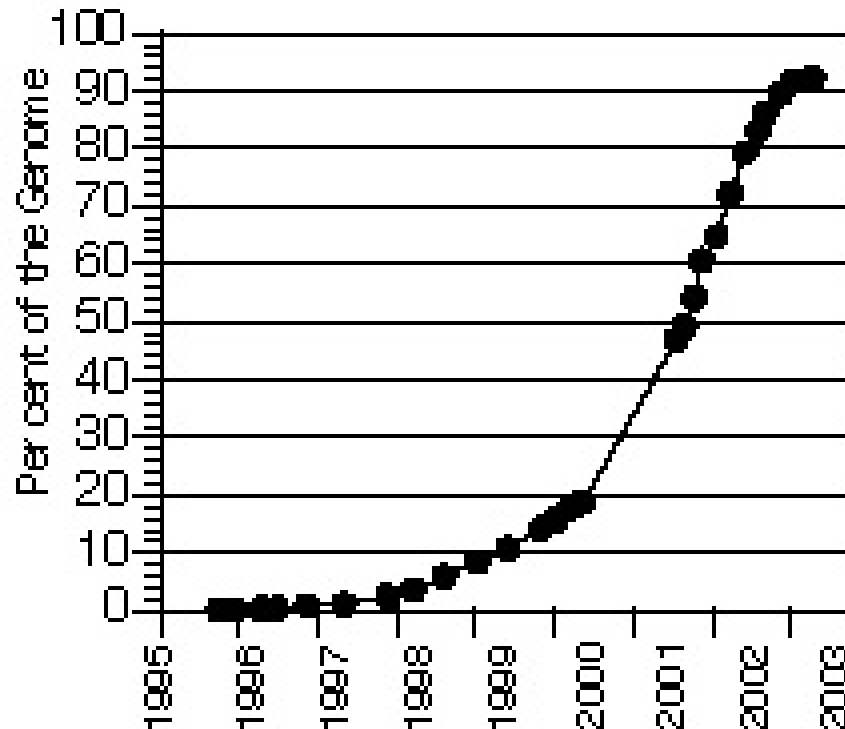
# Human Genome Project

- One of the largest scientific endeavors
  - Target accuracy 1:10,000 bases
  - Started in 1990 by DoE and NIH
  - \$3Billion and 15 years
  - Goal was to identify 25K genes and 3 billion bases
- Used the Sanger sequencing method
- Draft assembly done in 2000, complete genome by 2003, last chromosome published in 2006

# Human Genome Project



# Human Genome Project



This blog post indicates ~2.86Gbase/3.1Gbase of the non-redundant genome has been sequenced in hg18 or ~**92%** centromeres, telomeres, and highly repetitive regions left

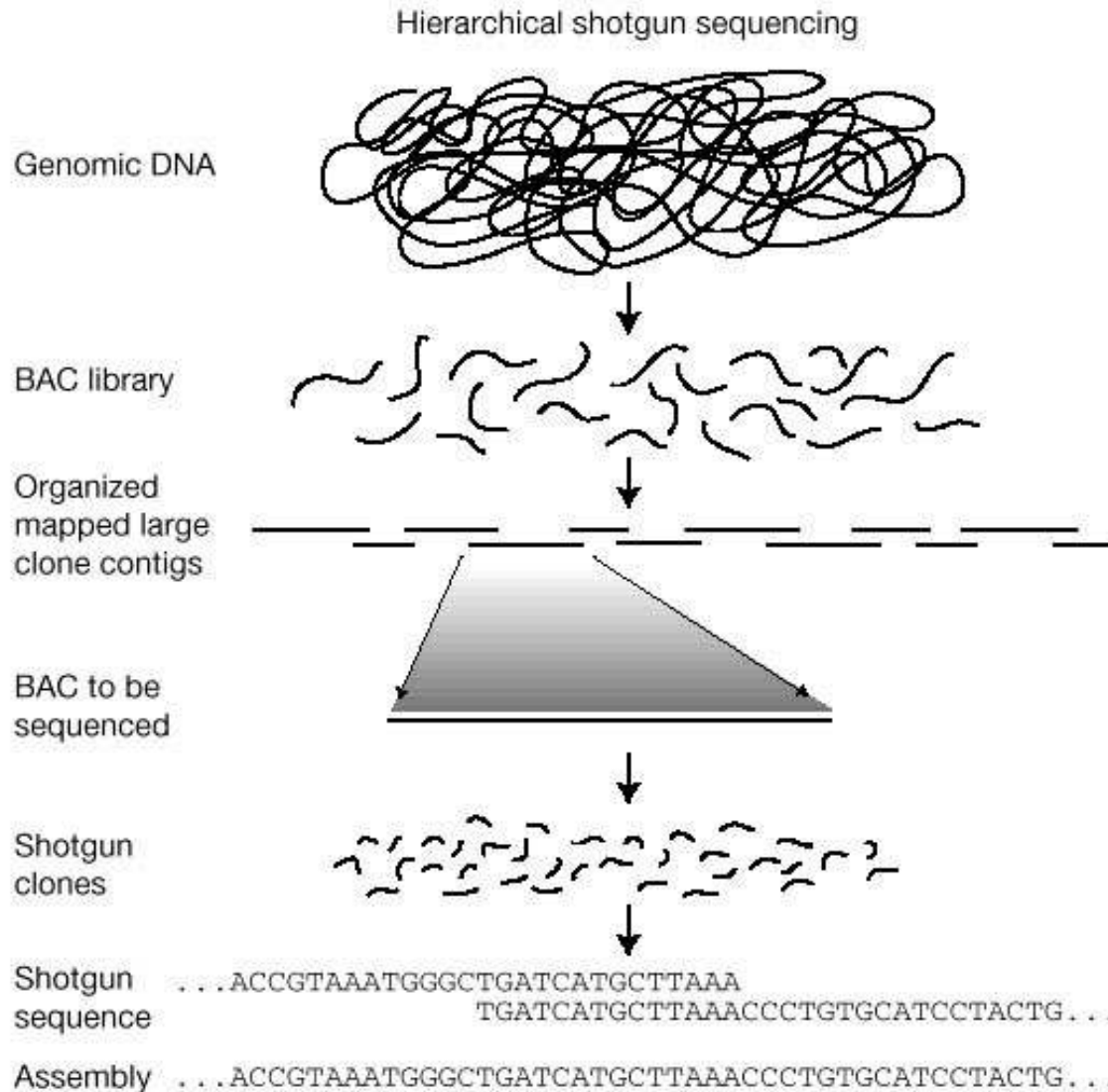
# How it was Accomplished

- Public Project
  - Hierarchical shotgun approach
  - Large segments of DNA were cloned via BACs and located along the chromosome
  - These BACs were shotgun sequenced
- Celera
  - Pure shotgun sequencing
  - Used public data (released daily) to help with assembly

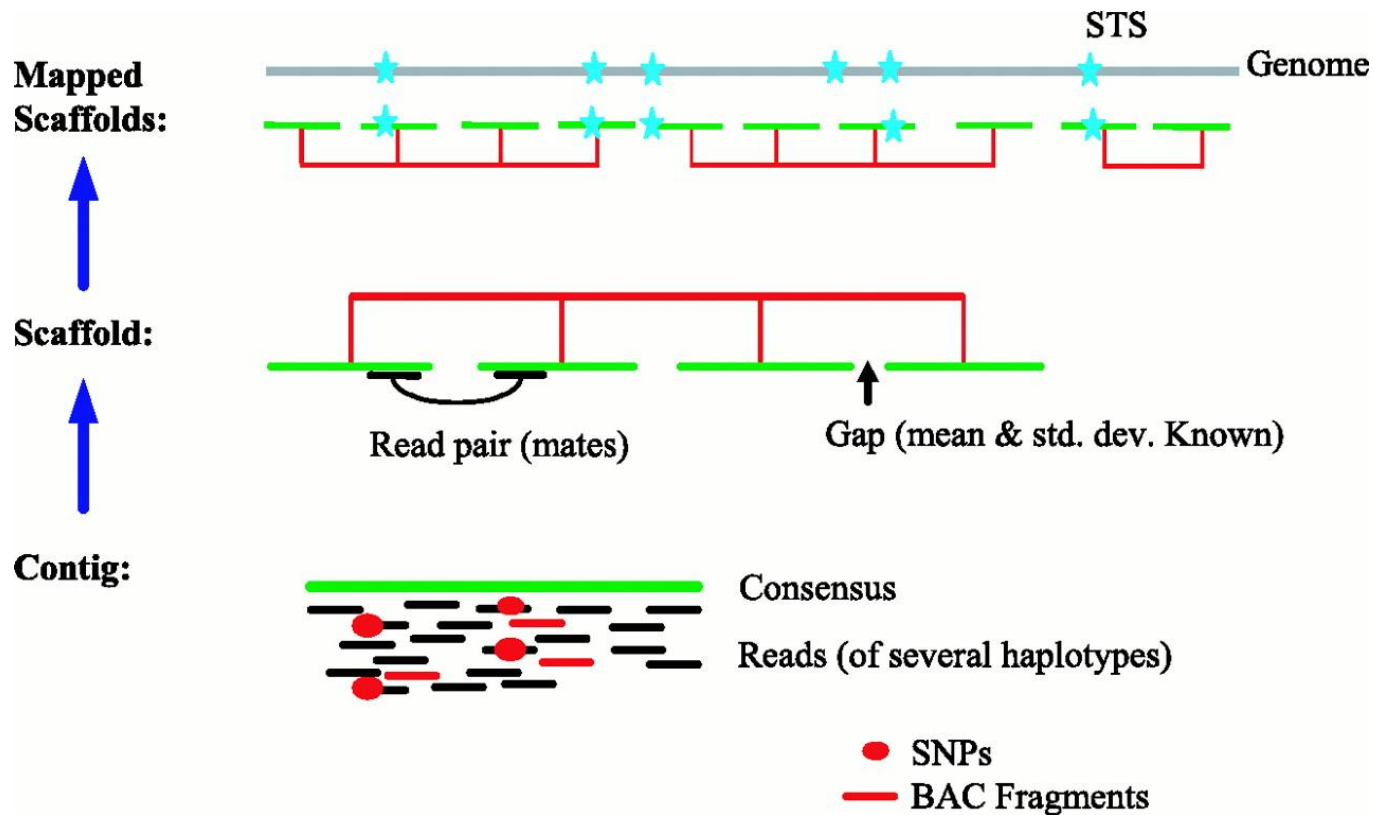
# Shotgun Sequencing

- Celera
  - Started in Sept 1999, goal was to do in \$300M and 3 years what the public project was doing for \$3B and 15 years!
  - Whole-genome shotgun sequencing
  - Used both whole-genome assembly and regional chromosome assembly
  - Incorporated data from the public project
  - Raised ethical concerns about the ownership of the human genome and patentability of genes

# Hierarchical Sequencing

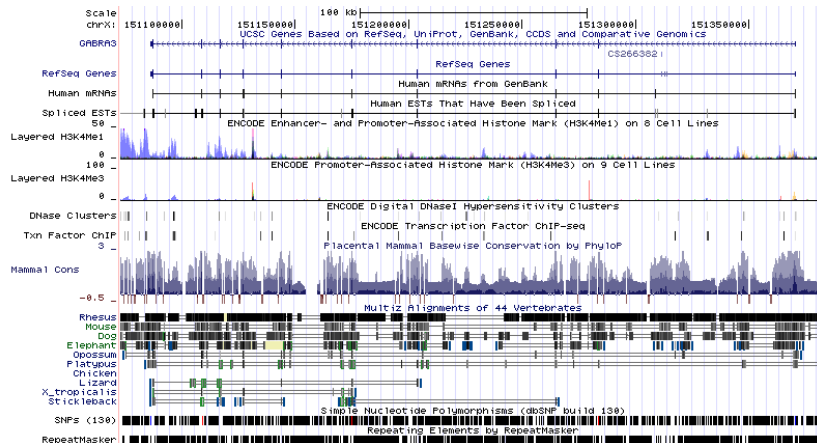


# Celera Shotgun Sequencing



- Used paired-end strategy with variable insert size: 2, 10, and 50kbp

# HGP Data Access

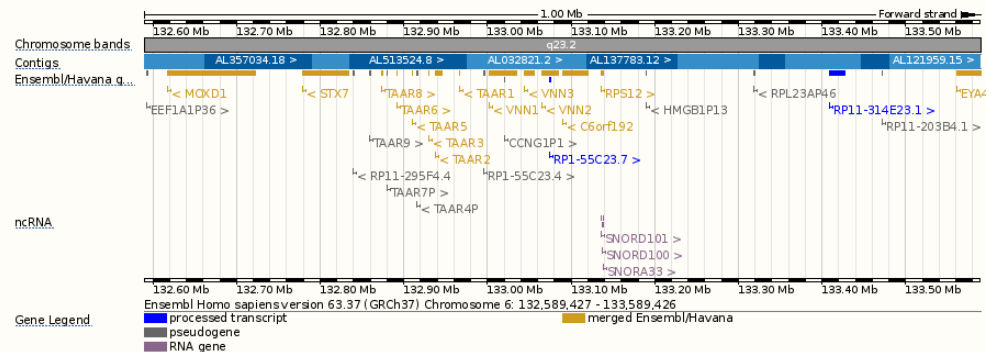


ORIGIN

```

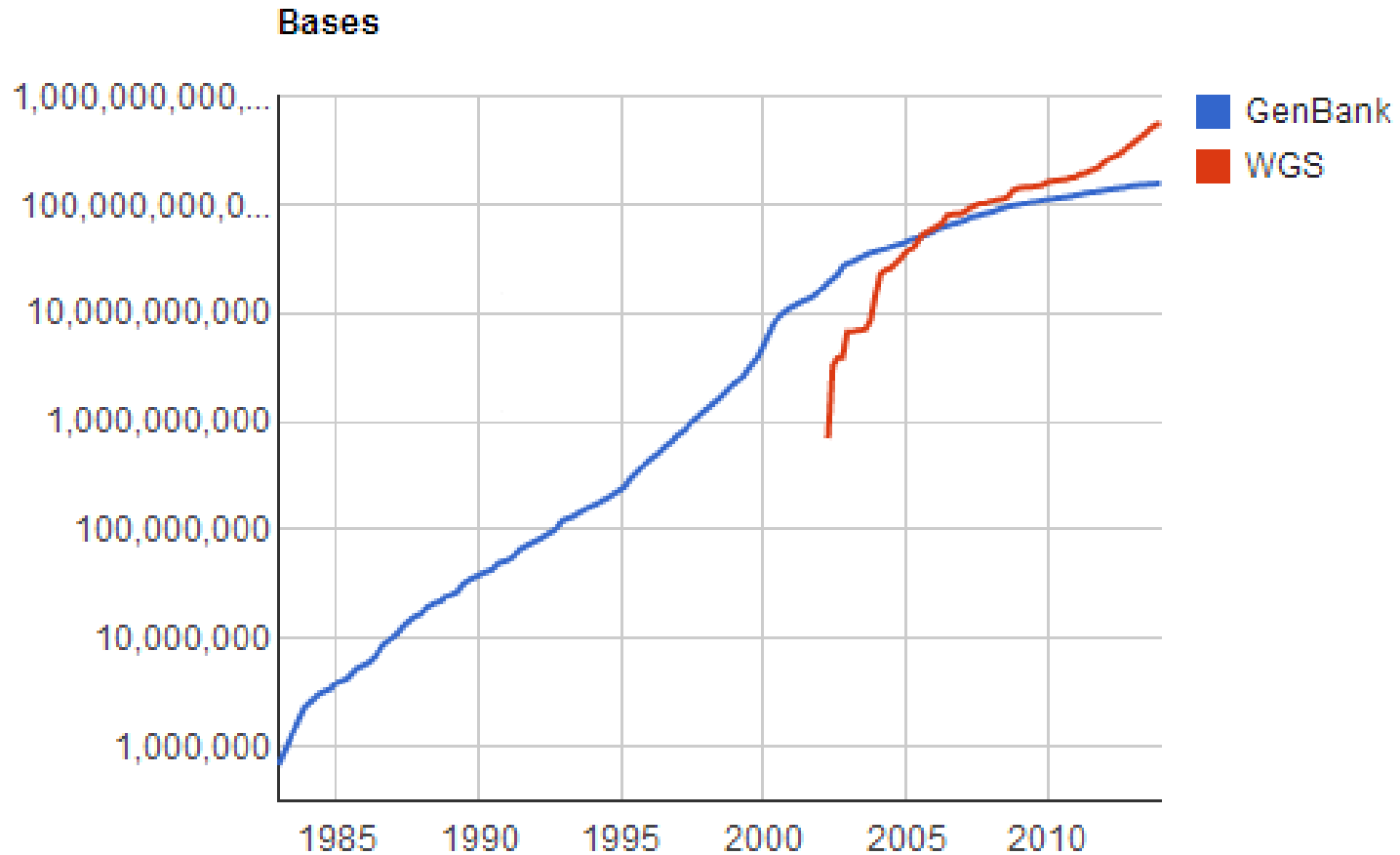
1  actttccgctc tttgttagga tgactggaac ttgtaccact tatctggaag gcagcccggt
61  ttgtctctatc aaaaatgtaa atgtgagcgg gcacaatggt ccaacgcctg taatcccgag
121 actttcggag gccgaggcgg gtggatcacc tgaggtcagg agttggagag cagctcggcc
181 aacatggtga aaccccatct ctactaaaaa tacaaaaaatt agccggcgct ggtggttgt
241 gccgttaatc ccagctattc gggaggctga gccaggagaa tcgcttgaa cccaggaggcg
301 gaggtgttag tgagacgaga ttgcgccatt gcactccagc caggttgaca agagcaaaac
361 tccgtctcaa aaaaaaaaaa agtaaaagtaa aatgttcttt aatctagcaa ttttacttct
421 agaagctaaa cctacagatg tacaccacat gtaagccaga atcgtttaca aagagatata
481 ttcaacttg aaaccccgct tctactaaaa atacaaaaaa ttgctggcg atggtggcag
541 gcgcctatag tcccagctac tcgggaggct gaggcaggag aatggcgtga acccgcgagg
601 cagagcttgc agtgagccga gatcgcgcca ctgcactcca gccctgggcta cagagcaaga
661 ctccatctta aaaaaaaaaa aaaaagggaa tagcaagac ttggaataaa cgtatatgct
721 cattgaaaag tgaggagtta aataaattat gtcacatcta agcaagagaa tactacacag
781 ctttcaaaa gaactaggct catctaaagc atctgataac agaaataaaa tacatattat
841 gaagttaaaa aatcaatata ctatagtagt aatatccttt ggaaaaaggt atttagtgt
901 gtgtgtctga aaagatacac aagaaataac taggtttctc aacaccgtaa cctgaatgt
961 acacatcatc ccgcccttgc cctgtacctt gttgactgct tgagcctgct gctaatcatt
1021 ctaatttata ctttatttta atatttttta actcatttat tttcttttta
1081 agactcttct tatttttgaa tggcactctt ccaaataaat ttttaaatca ttttatcaaa
1141 ttcttaaaag tatcctgttg gacatttgat tagaattata ctggataggc tgggtgtggt
1201 gggtcacac tgtaatccca gcaatttggg aggccaagga gggaggattg cttaggccca
1261 ggagtttgag actaatctgg gcaacatagc aagacccctc tctacaaaac ttttttaaaa
    
```

| Name                                   | Last modified     | Size | Description |
|----------------------------------------|-------------------|------|-------------|
| <a href="#">Parent Directory</a>       |                   | -    |             |
| <a href="#">chromApp.tar.gz</a>        | 20-Mar-2009 09:02 | 538K |             |
| <a href="#">chromFa.tar.gz</a>         | 20-Mar-2009 09:21 | 905M |             |
| <a href="#">chromFaMasked.tar.gz</a>   | 20-Mar-2009 09:30 | 477M |             |
| <a href="#">chromOut.tar.gz</a>        | 20-Mar-2009 09:03 | 163M |             |
| <a href="#">chromTrf.tar.gz</a>        | 20-Mar-2009 09:30 | 7.6M |             |
| <a href="#">est.fa.gz</a>              | 11-Aug-2011 10:57 | 1.4G |             |
| <a href="#">est.fa.gz.md5</a>          | 11-Aug-2011 10:57 | 44   |             |
| <a href="#">hg19.2bit</a>              | 08-Mar-2009 15:29 | 778M |             |
| <a href="#">md5sum.txt</a>             | 29-Jul-2009 10:04 | 457  |             |
| <a href="#">mrna.fa.gz</a>             | 11-Aug-2011 10:33 | 197M |             |
| <a href="#">mrna.fa.gz.md5</a>         | 11-Aug-2011 10:33 | 45   |             |
| <a href="#">refMrna.fa.gz</a>          | 11-Aug-2011 10:58 | 39M  |             |
| <a href="#">refMrna.fa.gz.md5</a>      | 11-Aug-2011 10:58 | 48   |             |
| <a href="#">upstream1000.fa.gz</a>     | 05-Aug-2011 16:32 | 7.5M |             |
| <a href="#">upstream1000.fa.gz.md5</a> | 05-Aug-2011 16:32 | 53   |             |
| <a href="#">upstream2000.fa.gz</a>     | 05-Aug-2011 16:34 | 14M  |             |
| <a href="#">upstream2000.fa.gz.md5</a> | 05-Aug-2011 16:34 | 53   |             |
| <a href="#">upstream5000.fa.gz</a>     | 05-Aug-2011 16:36 | 34M  |             |
| <a href="#">upstream5000.fa.gz.md5</a> | 05-Aug-2011 16:36 | 53   |             |
| <a href="#">xenoMrna.fa.gz</a>         | 11-Aug-2011 10:39 | 1.4G |             |
| <a href="#">xenoMrna.fa.gz.md5</a>     | 11-Aug-2011 10:39 | 49   |             |



Results in GenBank, UCSC, Ensembl & others

# Growth of Genbank

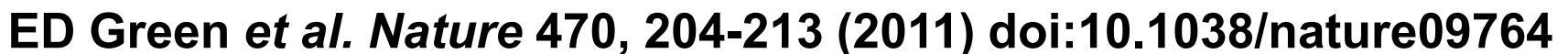


**December 2013 156,230,531,562 bases**

# Outcome of the HGP

- Spurred the sequencing of other organisms
  - 36 “complete” eukaryotes (~250 in various stages)
  - 1704 “complete” microbial genomes
  - 2685 “complete” viral genomes
- Enabled a multitude of related projects:
  - Encode, modEncode
  - HapMap, dbGAP, dbSNP, 1000 Genomes
  - Genome-Wide Association Studies, WTCCC
  - Medical testing, GeneTests, 23AndMe, personal genomes
  - Cancer sequencing, COSMIC, TCGA, ICGC
- Provided a context to organize diverse datasets

## Genomic achievements since the Human Genome Project



# Economic Impact of the Project

- Battelle Technology Partnership Practice released a study in May 2011 that quantifies the economic impact of the HGP was **\$796 billion!**
- Genomics supports:
  - >51,000 jobs
  - Indirectly, 310,000 jobs
  - Adds at least \$67 billion to the US economy

# Second generation sequencing tech

# Second generation sequencing definition

“Synchronized reagent wash of nucleotide triphosphates followed by optical imaging” – *Niedringhaus, T. et al, Reviews Analytical Chemistry, 2011, 83 4327-4341*

# Illumina HiSeq



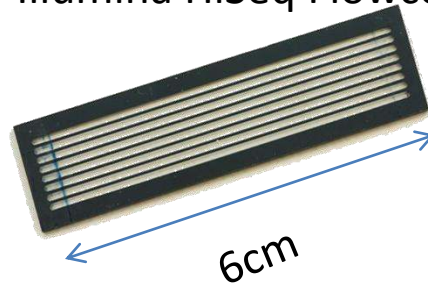
# Illumina HiSeq Key Features

- Advantages
  - Large volume of data (300Gb per run)
  - Short run time (< 1 day)
  - Straightforward sample prep
  - Well established open source software community
- Disadvantages
  - Requires pooling of large numbers of samples to achieve lowest costs
  - Short reads (36-150bp)

# Illumina Sequence By Synthesis

- Produces approximately 1.6 billion short reads (18bp-150bp) per flowcell
- Each run takes 2-9 days depending on the configuration
- Each flowcell is divided into either 2 or 8 separate lanes (channels)

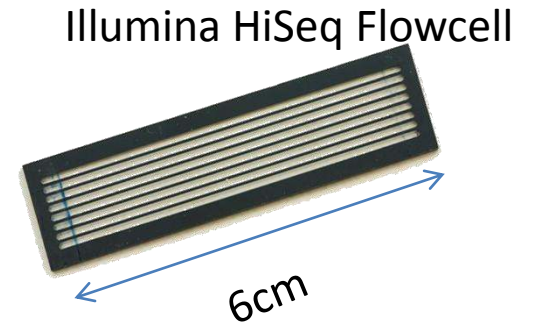
Illumina HiSeq Flowcell



# Illumina HiSeq setup



Automated sample preparation



Illumina HiSeq Flowcell

6cm

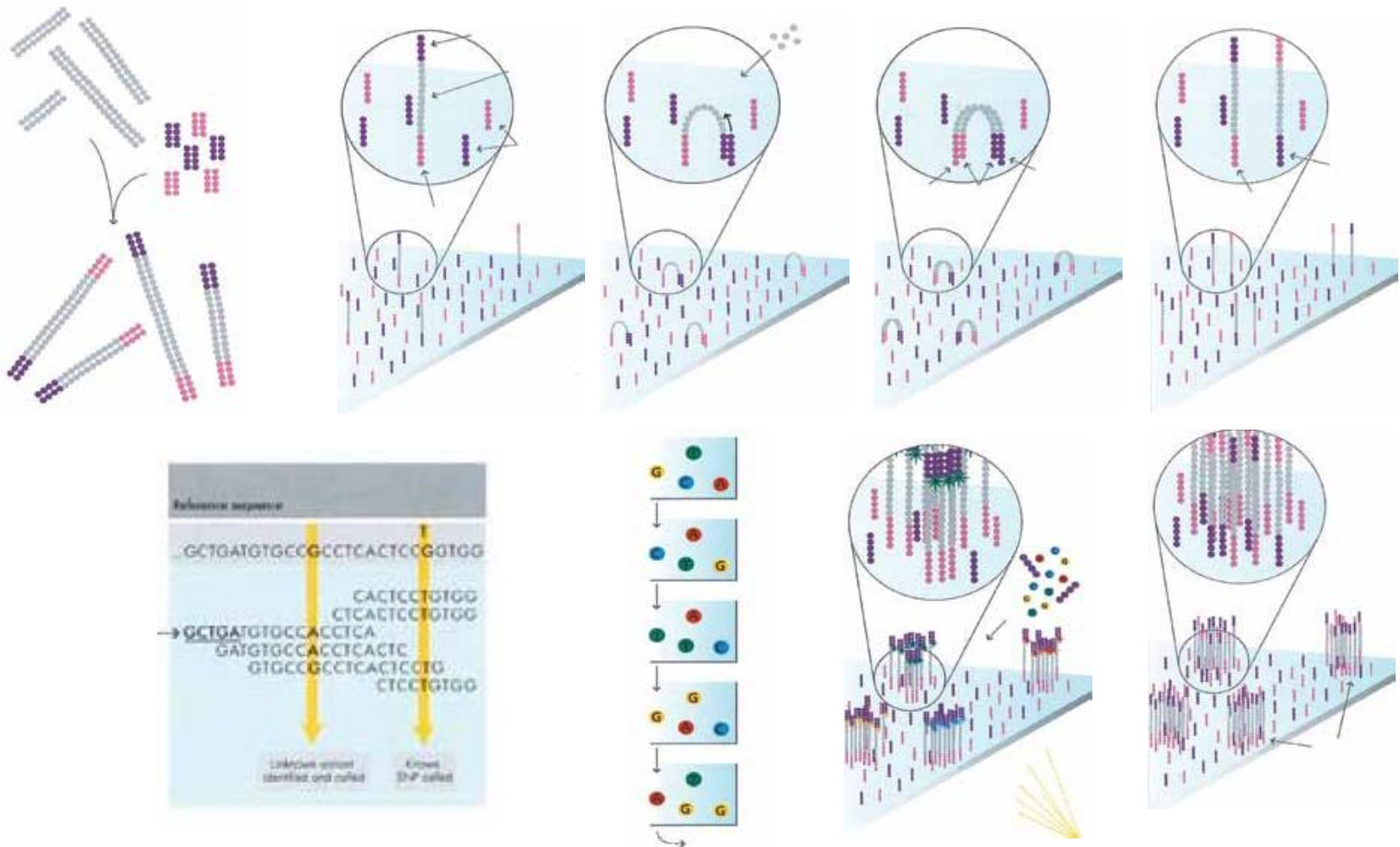


cBot Cluster generation



HiSeq 2500

# Illumina Sequencing

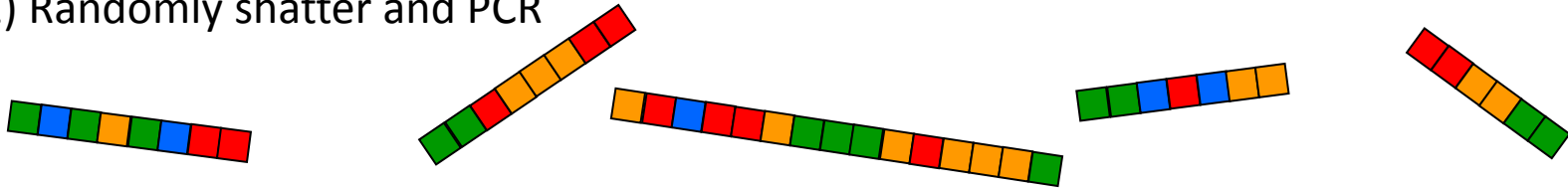


# DNA sample preparation (over-simplified)

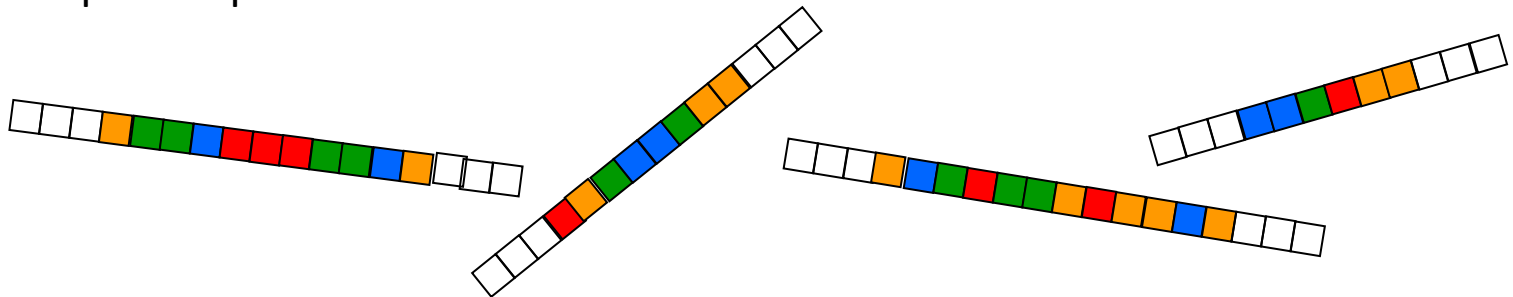
1) Extract DNA



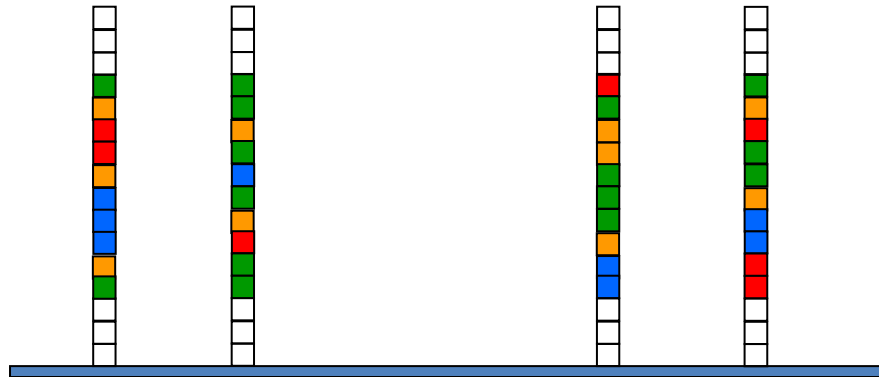
2) Randomly shatter and PCR



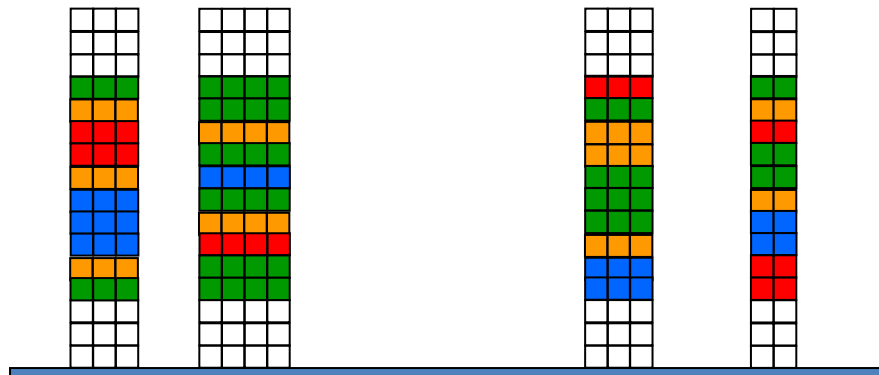
3) Attach adapter sequence 



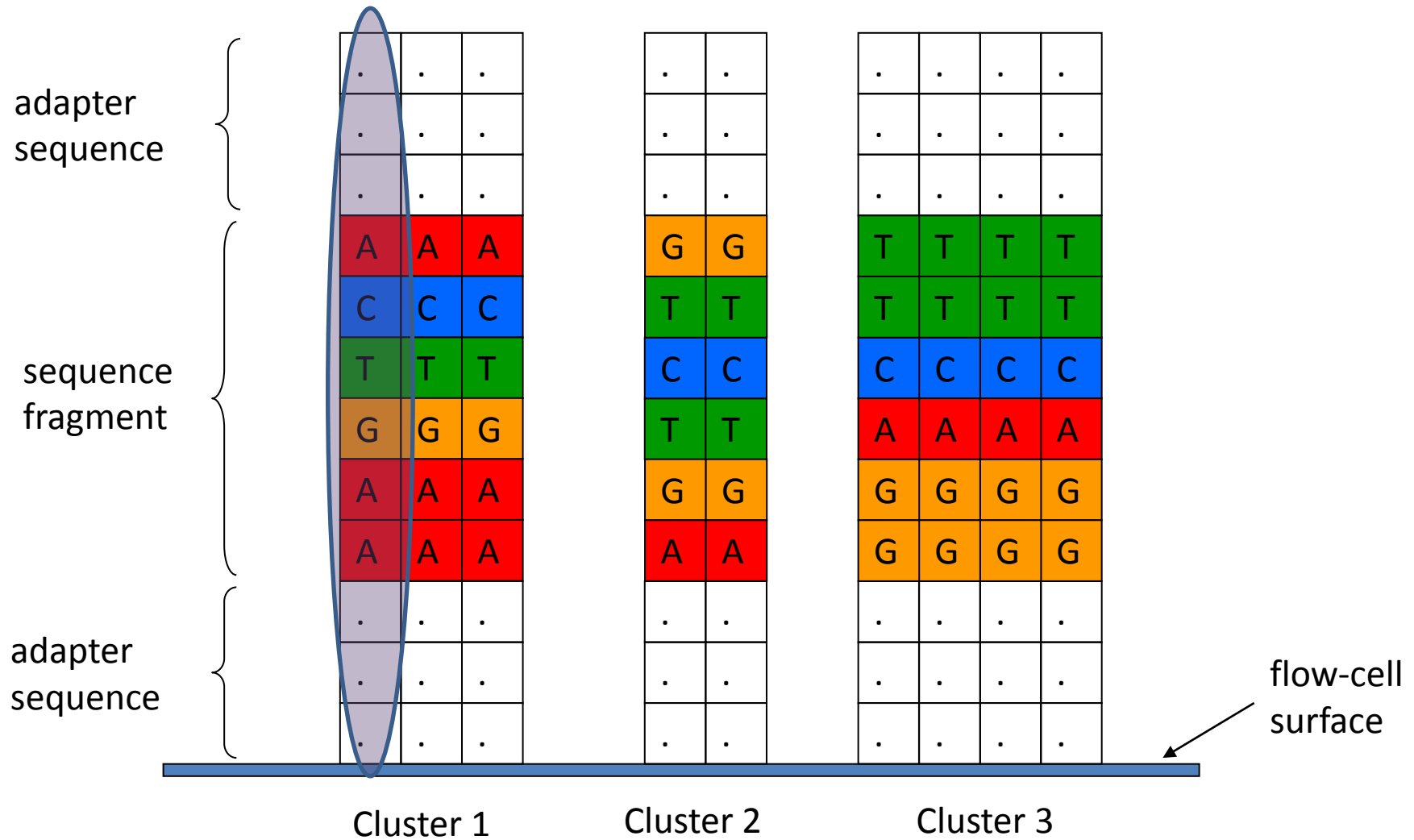
#### 4) Attach to flow-cell surface



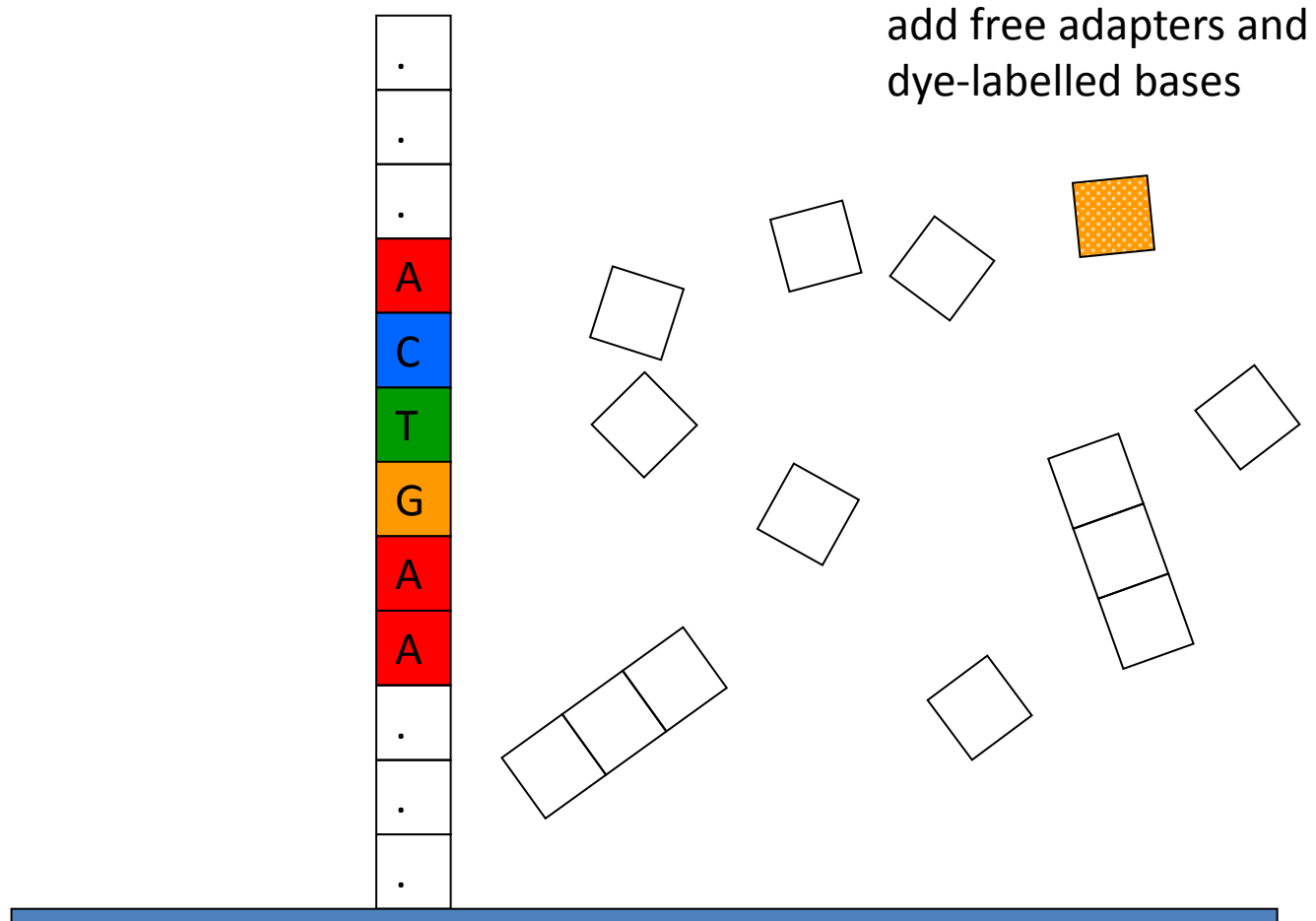
## 5) PCR-amplify into clusters



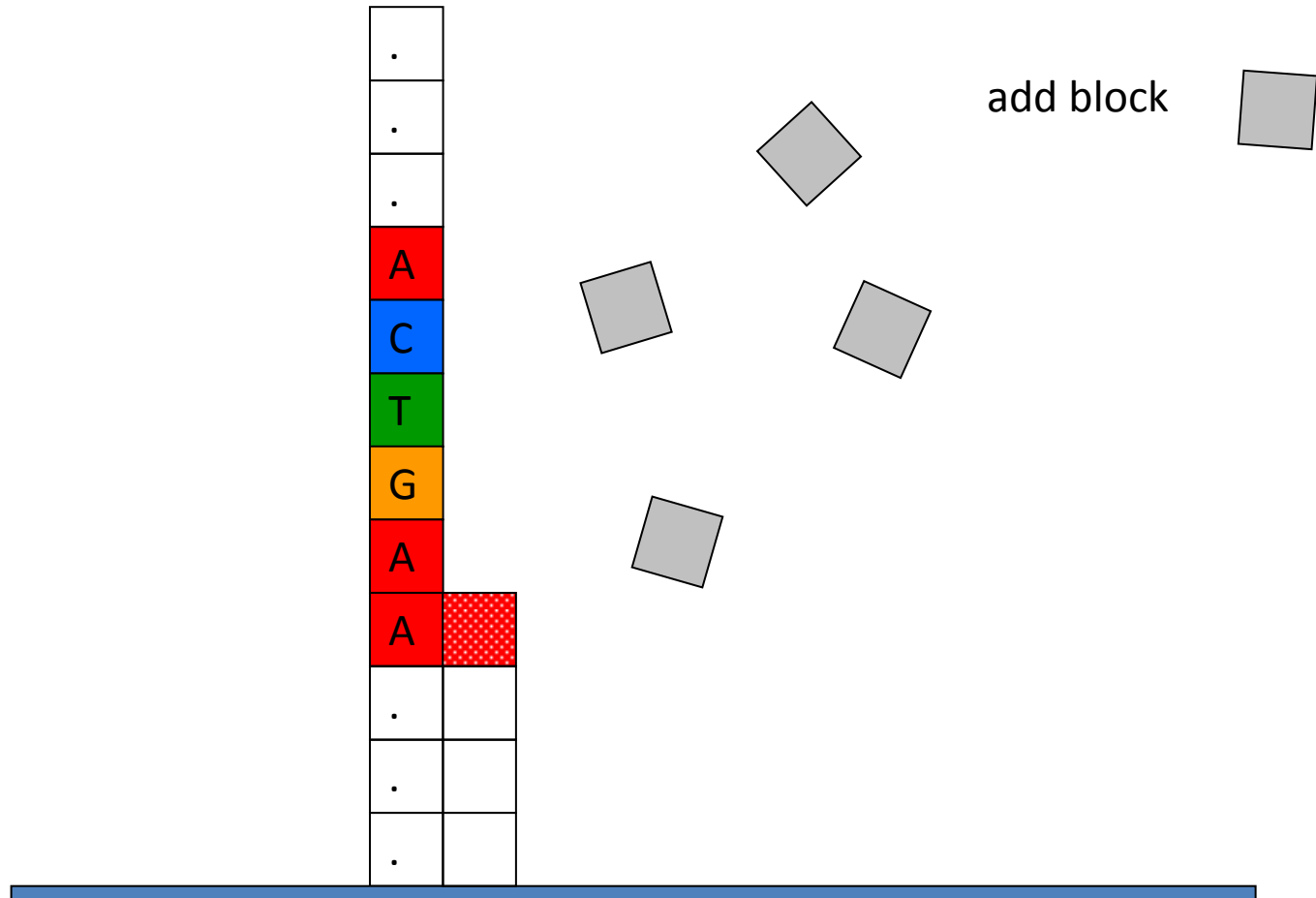
# Sequence clusters on the flow cell



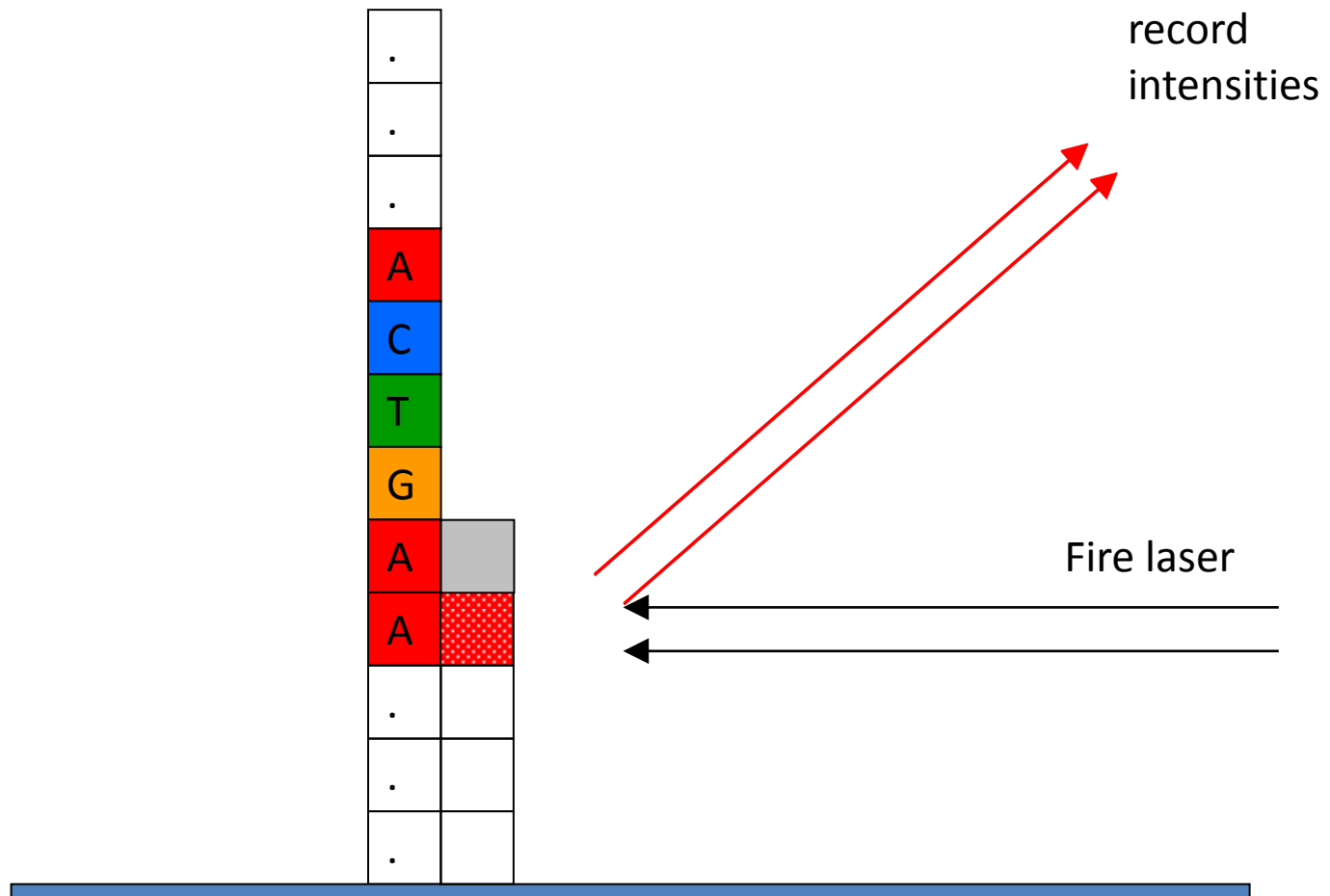
# Sequencing cycle 1



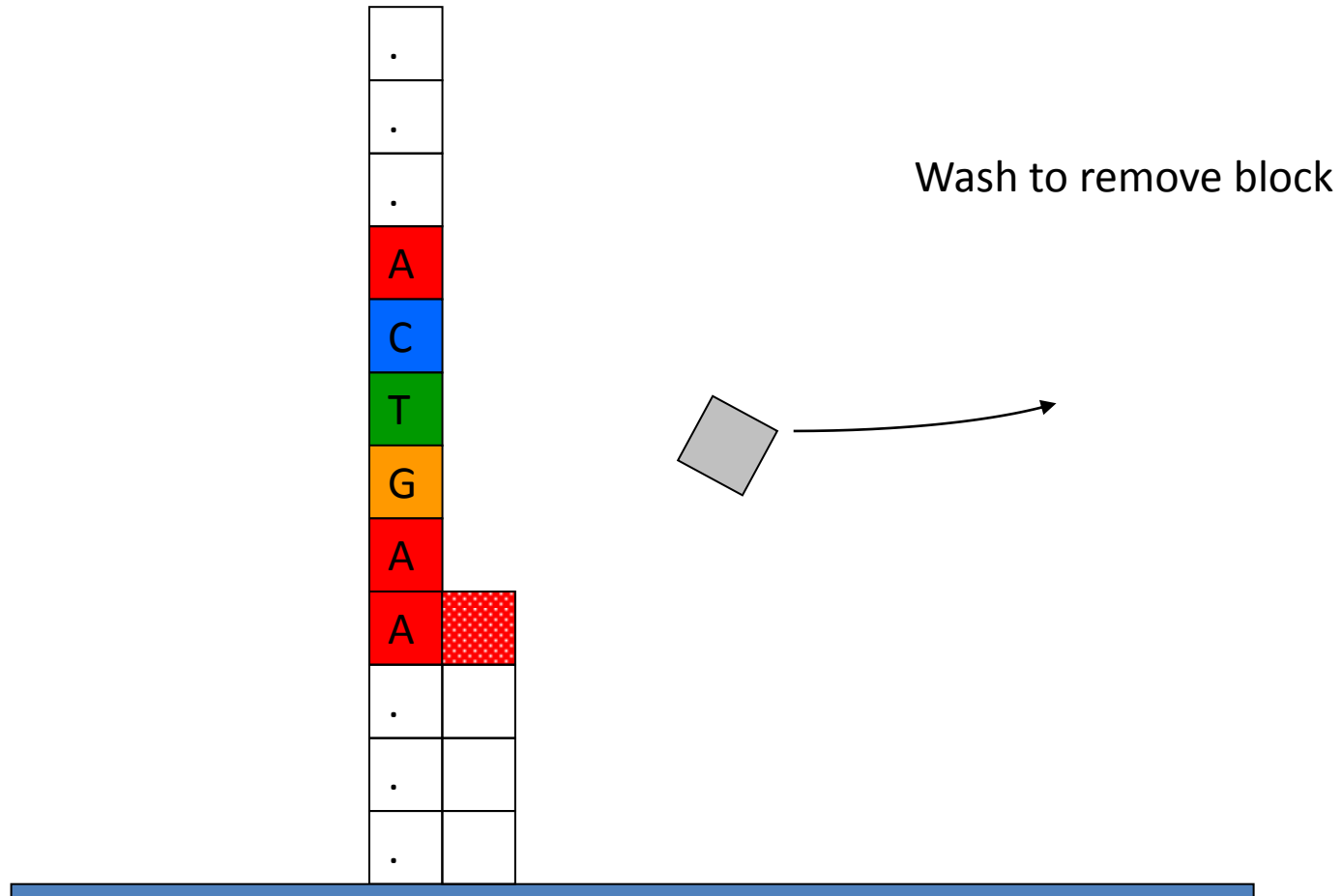
# Sequencing cycle 1



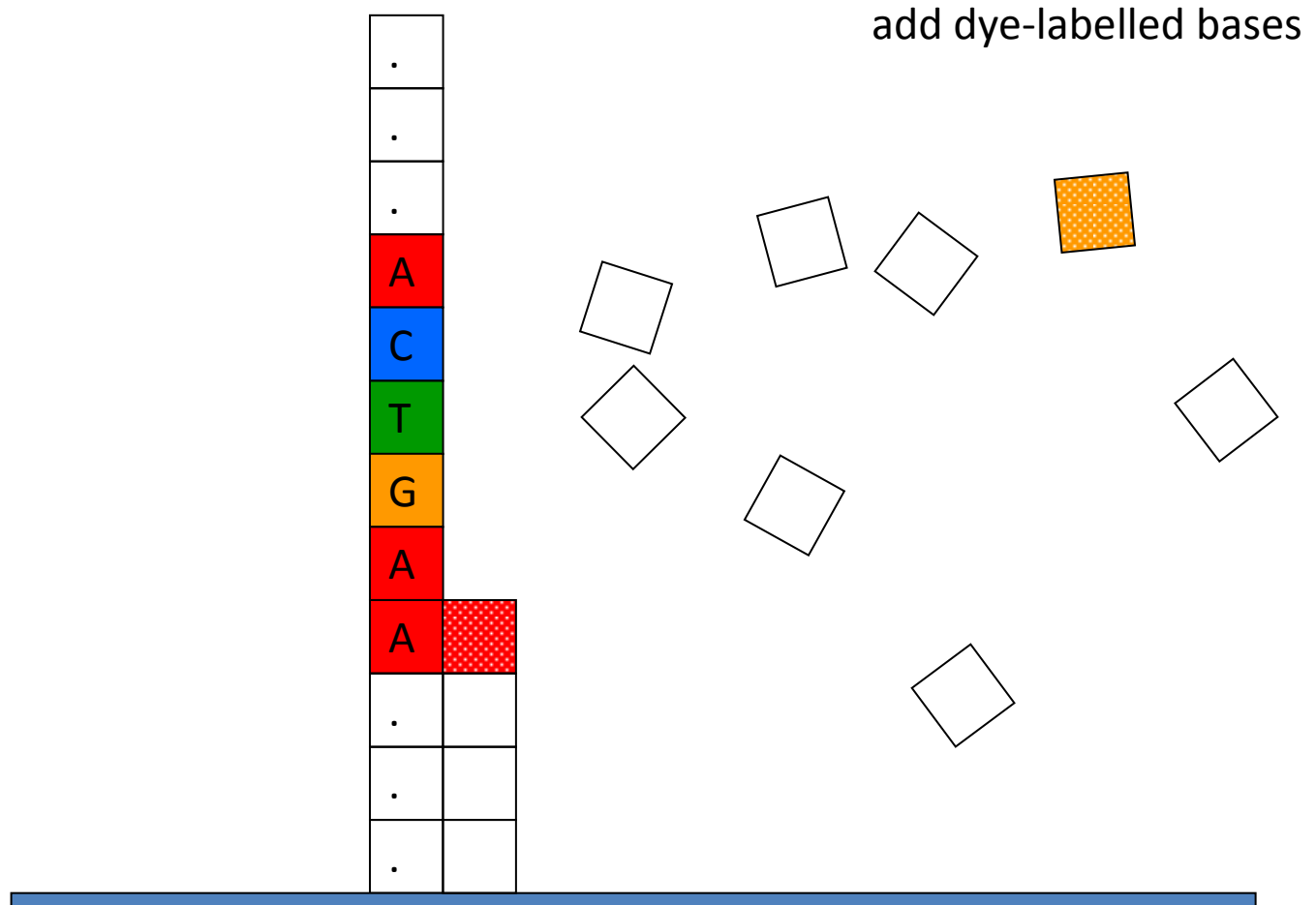
# Sequencing cycle 1



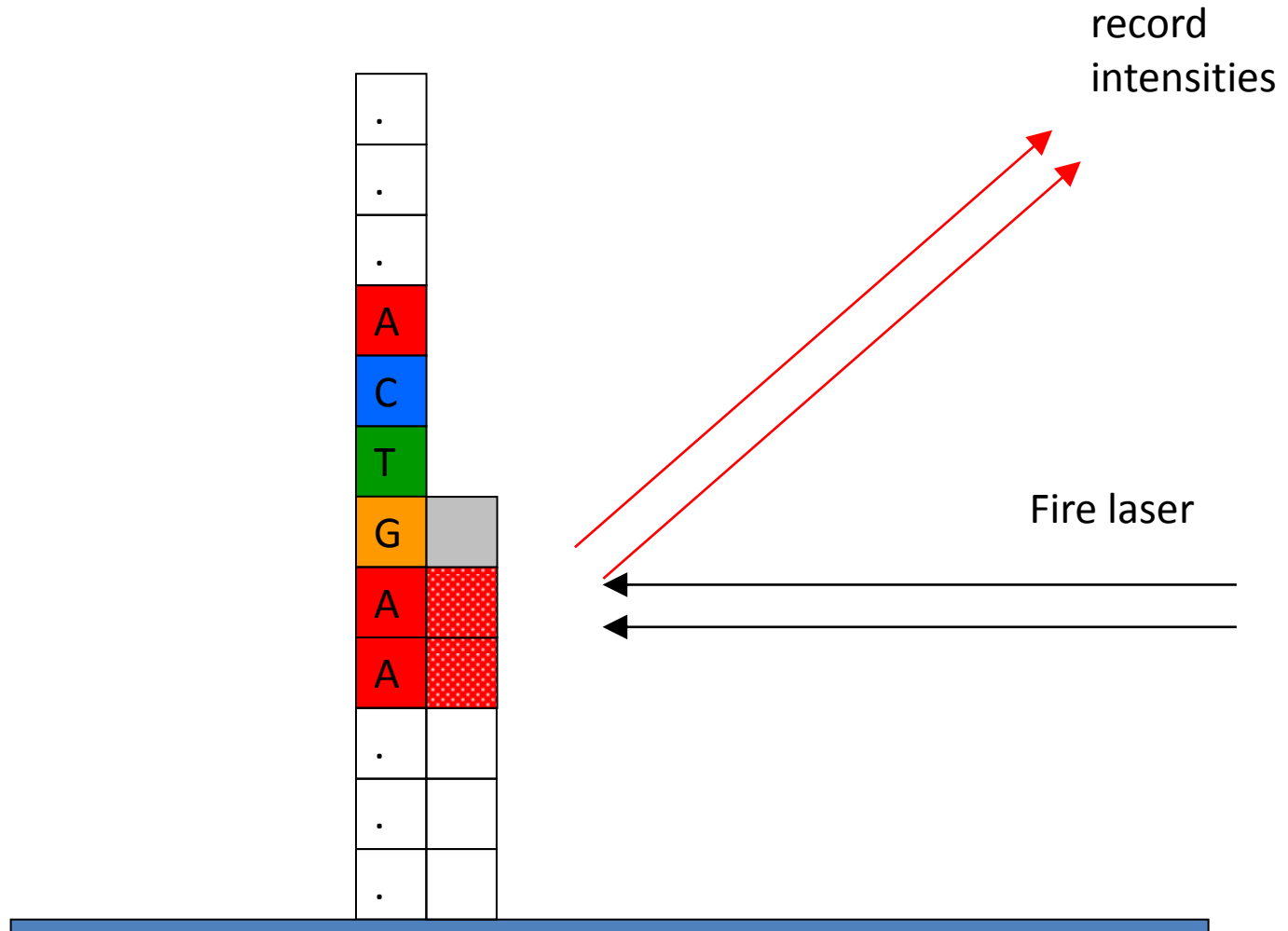
# Sequencing cycle 1



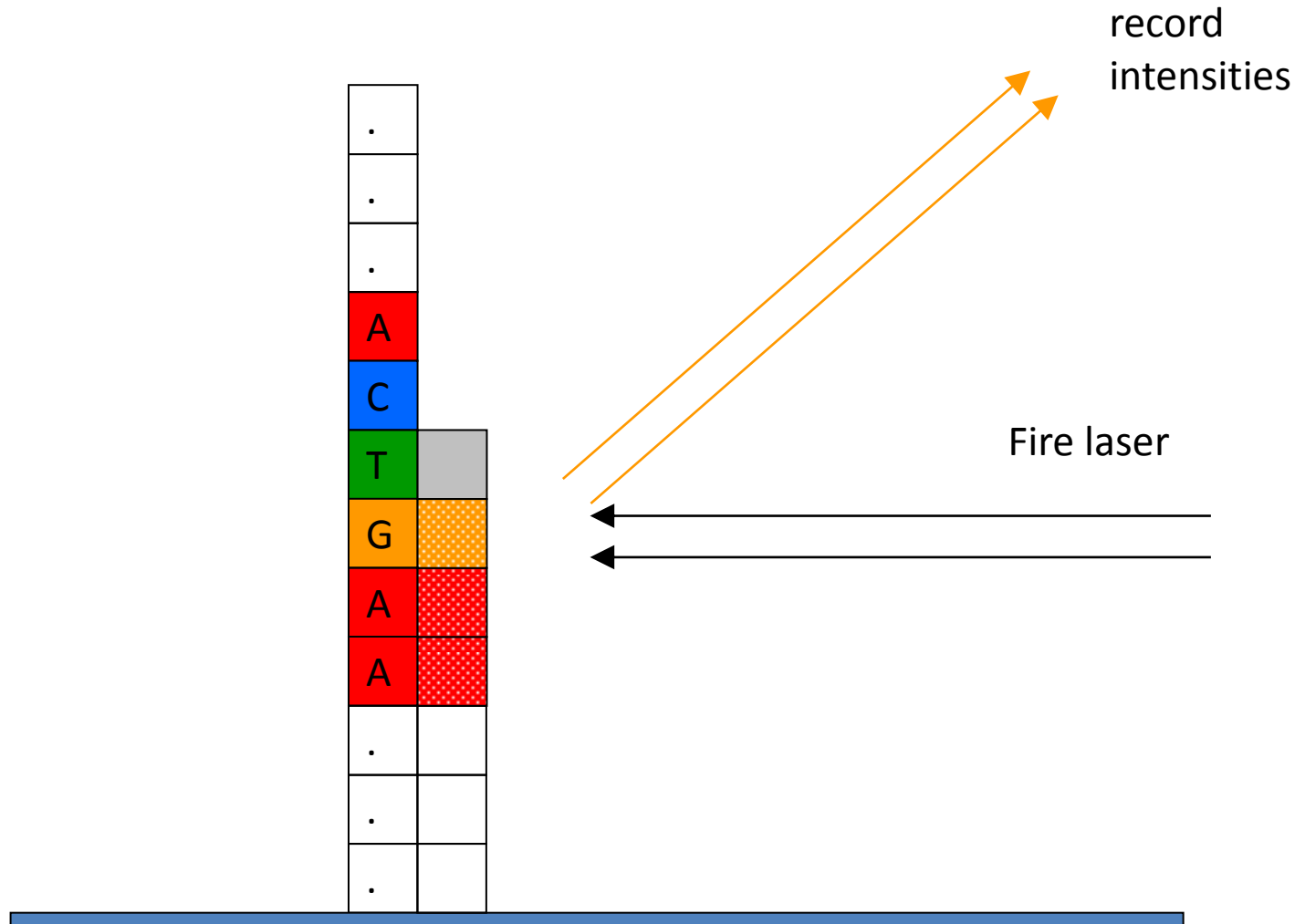
# Sequencing cycle 2



# Sequencing cycle 2



# Sequencing cycle 3



# Note

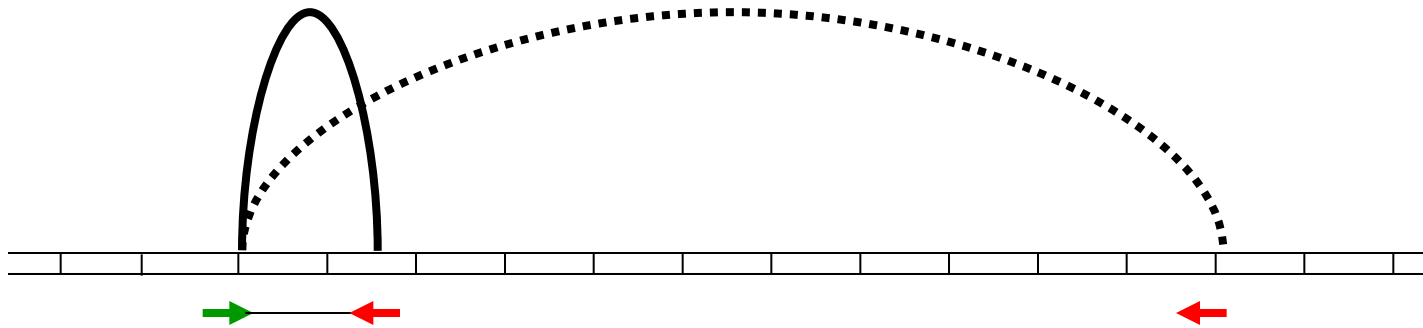
- The schema is over-simplified
- In reality
  - Sequencing is done from the top of the strand down towards the bottom of the flowcell
  - Reversible blocking agents are part of the modified dNTP fluorophores

# Paired-End Sequencing

- Provides distance relationship between two reads
- Important for many applications
  - Characterise insertions, deletions, copy number variants, rearrangements
  - Required for *De novo* assembly
  - Enables sequencing across repeats
  - Useful for Di-Tags, cDNA sequencing, etc.
- Enables sample multiplexing (identifier tags)
- Increases output per flowcell

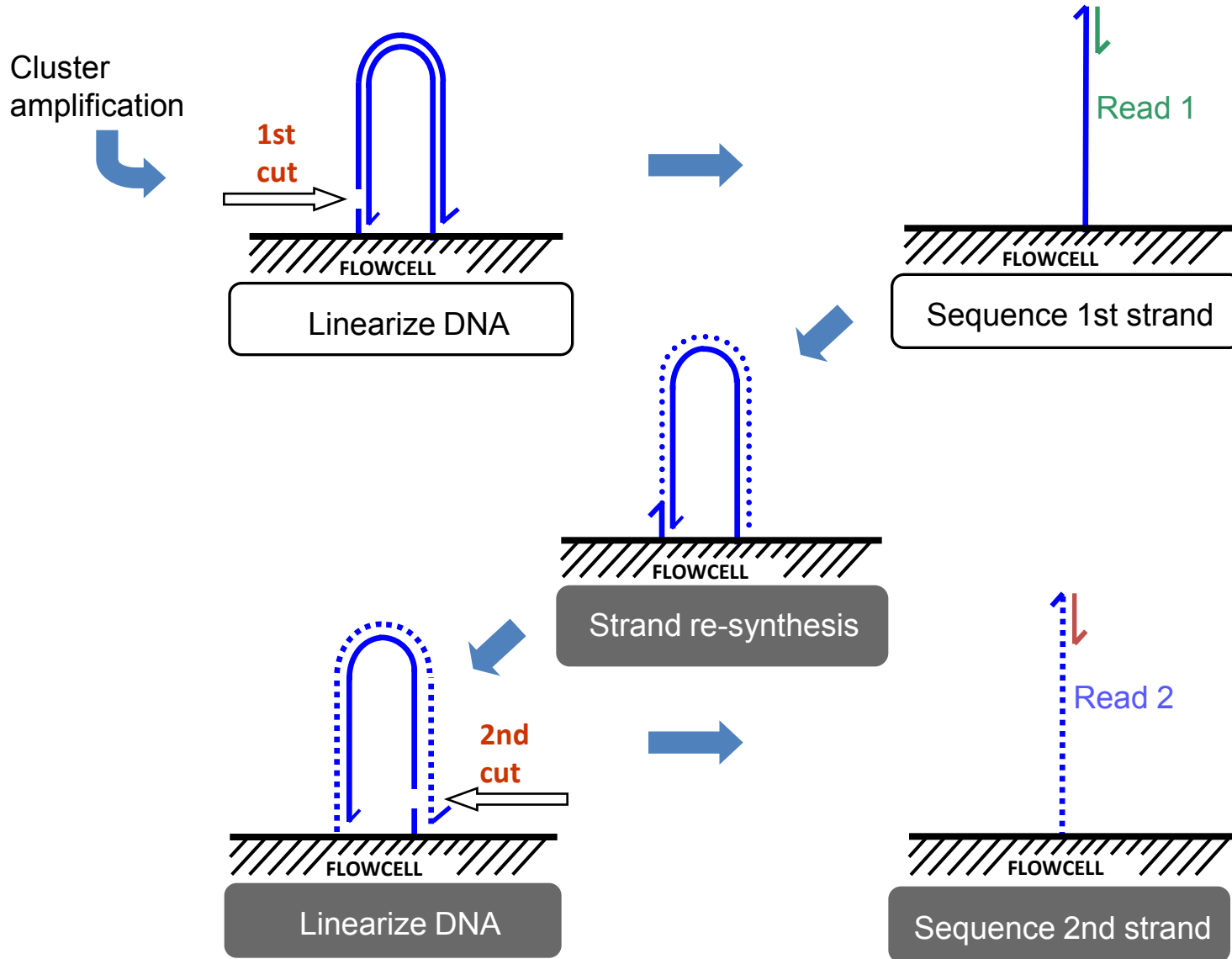
# Working with Paired Reads

- Applicable to different fragment size ranges
  - up to ~600 bp for standard libraries
  - 2 - 20kb mate-pair libraries

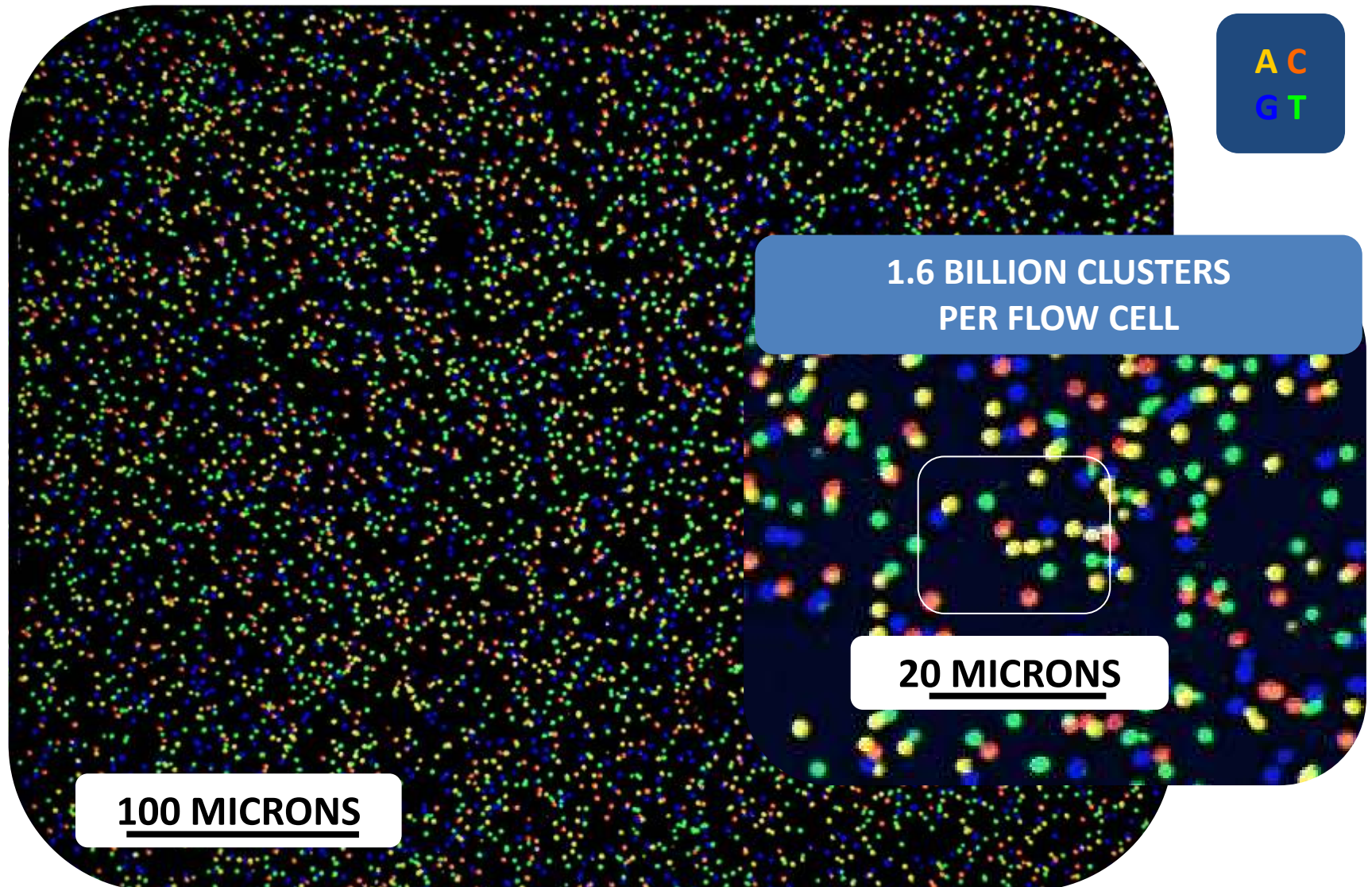


Enables alignment software to assign unique positions to previously *non-unique* reads

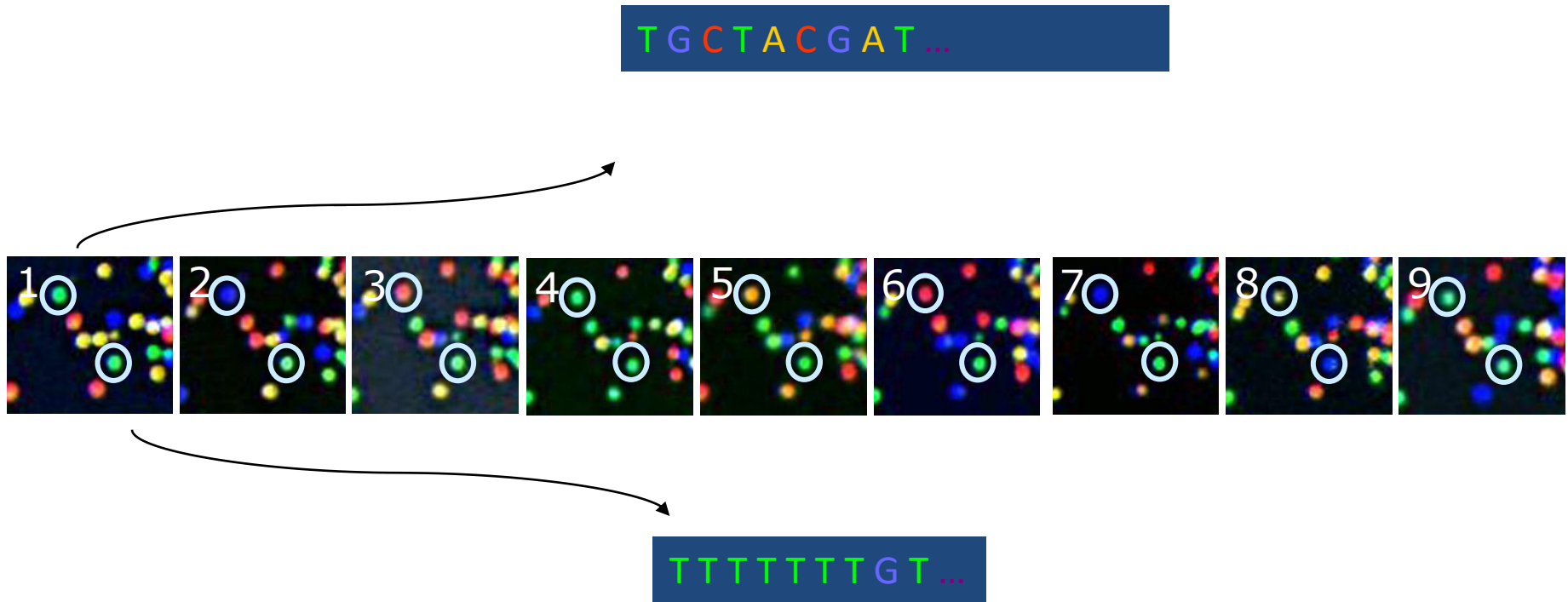
# Illumina Paired-End Sequencing



# Illumina Sequencing : How it looks



## Base calling from raw data



The identity of each base of a cluster is read off from sequential images.

**Current read lengths = 36-150 nt**

**Total sequence data for 1 paired-end run with 100bp = 300Gb!**

# HiSeq 2000 vs 2500 flowcells



HiSeq 2000  
8 lanes

12 day run time



HiSeq 2500  
2 lanes

2 day run time

# Comparison

| APPLICATION | RAPID RUN MODE | HIGH OUTPUT MODE |
|-------------|----------------|------------------|
|-------------|----------------|------------------|

|                                  |  |  |
|----------------------------------|--|--|
| ChIP-Seq<br>Transcription Factor |  |  |
|----------------------------------|--|--|

|  |                       |  |
|--|-----------------------|--|
|  | 40 Samples<br>7 Hours |  |
|--|-----------------------|--|

|  |  |                       |
|--|--|-----------------------|
|  |  | 200 Samples<br>2 Days |
|--|--|-----------------------|

|           |  |  |
|-----------|--|--|
| 1 x 36 bp |  |  |
|-----------|--|--|

|          |  |  |
|----------|--|--|
| mRNA-Seq |  |  |
|----------|--|--|

|  |                        |  |
|--|------------------------|--|
|  | 24 Samples<br>16 Hours |  |
|--|------------------------|--|

|  |  |                       |
|--|--|-----------------------|
|  |  | 120 Samples<br>5 Days |
|--|--|-----------------------|

|           |  |  |
|-----------|--|--|
| 2 x 50 bp |  |  |
|-----------|--|--|

|                                                                 |  |  |
|-----------------------------------------------------------------|--|--|
| TruSeq Exome Seq<br>62 MB Region<br>100x Coverage<br>2 x 100 bp |  |  |
|-----------------------------------------------------------------|--|--|

|  |                        |  |
|--|------------------------|--|
|  | 15 Samples<br>27 Hours |  |
|--|------------------------|--|

|  |  |                       |
|--|--|-----------------------|
|  |  | 85 Samples<br>12 Days |
|--|--|-----------------------|

|                                                   |  |  |
|---------------------------------------------------|--|--|
| Human Whole Genome<br>>30x Coverage<br>2 x 100 bp |  |  |
|---------------------------------------------------|--|--|

|  |                      |  |
|--|----------------------|--|
|  | 1 Sample<br>27 Hours |  |
|--|----------------------|--|

|  |  |                      |
|--|--|----------------------|
|  |  | 5 Samples<br>12 Days |
|--|--|----------------------|

# What does this mean?



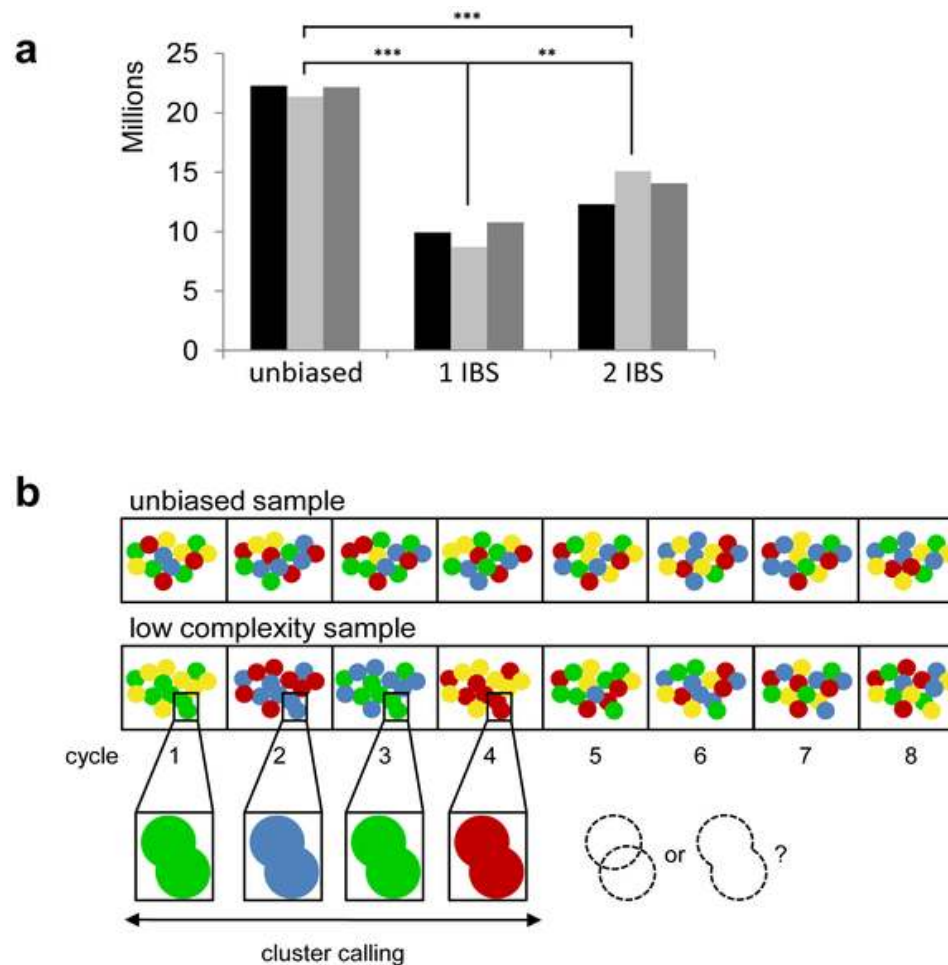
|  | Rapid run                    | Slow run                          |
|--|------------------------------|-----------------------------------|
|  | 48 genomes (£250 per sample) | 48 genomes/lane (£210 per sample) |
|  | 10 genomes (£510 per sample) | 10 genomes/lane (£350 per sample) |
|  | 8 genomes (£590 per sample)  | 8 genomes/lane (£400 per sample)  |
|  | 1 genome (£3400)             | 1 genome (£4000)                  |

# Potential issues with Illumina sequencing

- Low diversity sequences
  - 16S/amplicon sequences
  - Custom adaptors with barcodes at 5' end
- GC/AT bias
  - GC clusters are smaller than AT
  - (less of a problem post June 2011)
- Specific motifs which are difficult to sequence
  - GGC motif
  - Inverted repeats

*Nakamura, K., Oshima, T., Morimoto, T., Ikeda, S., Yoshikawa, H., Shiwa, Y., Ishikawa, S., et al. (2011). Sequence-specific error profile of Illumina sequencers. Nucleic acids research, gkr344–. Retrieved from <http://nar.oxfordjournals.org/cgi/content/abstract/gkr344v1>*

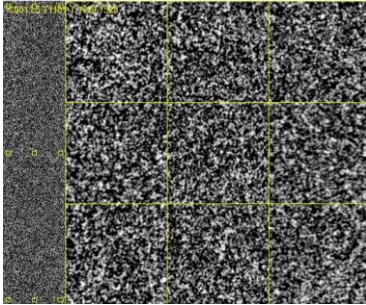
# Low Diversity samples



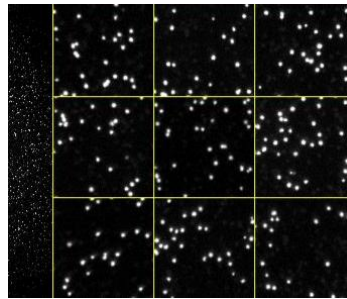
Krueger F, Andrews SR, Osborne CS (2011) Large Scale Loss of Data in Low-Diversity Illumina Sequencing Libraries Can Be Recovered by Deferred Cluster Calling. PLoS ONE 6(1): e16607. doi:10.1371/journal.pone.0016607  
<http://www.plosone.org/article/info:doi/10.1371/journal.pone.0016607>

# Low Diversity Example

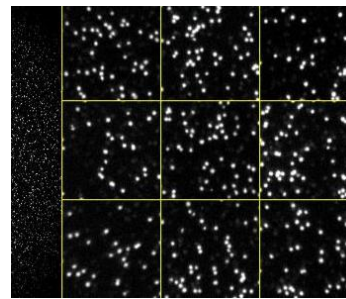
- Cycle 1



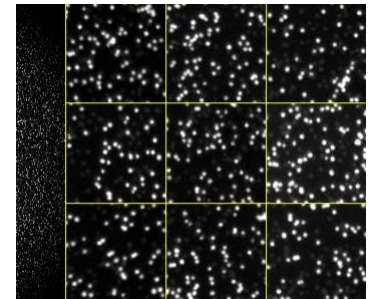
T



C



A



G

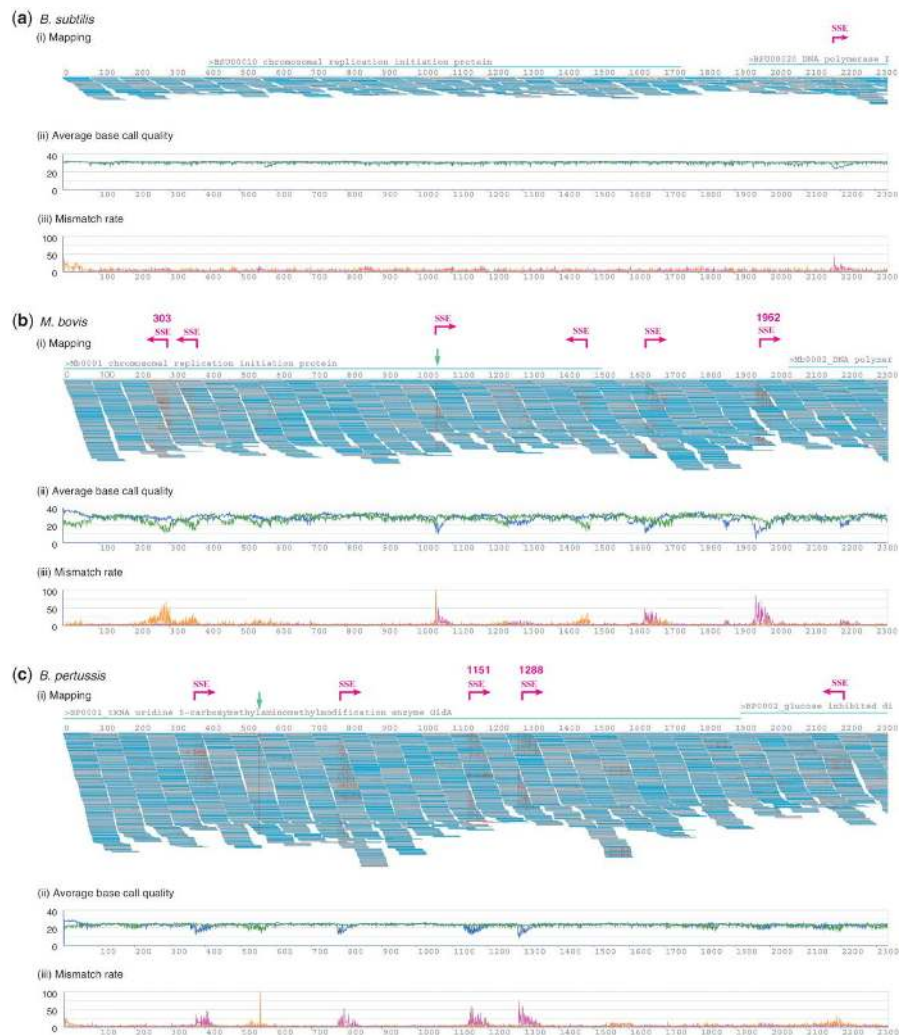
# Low diversity workarounds

- Multiplex with diverse samples and sequence across multiple lanes/runs
- Add PhiX control
- If dealing with amplicon, TraDIS or RAD-seq material, design multiple offset primers

# Low diversity issue 'fixed'

- Illumina are bringing out two improvements which are claimed to alleviate the problem
  - Software fix to prevent quality fall off due to incorrect phasing estimates. Available on MiSeq now.
    - Possibility of implementing those same changes on the HiSeq is uncertain
  - Ordered flowcells – base-caller will know apriori where clusters should be

# Sequence-specific error motifs (GCC and inverted repeats)



Nakamura K et al. Nucl. Acids Res. 2011;nar.gkr344

# Illumina MiSeq

- Same technology and chemistry as HiSeq
- 2X300bp reads
- 15 Gbase/run
- Run 48-72 hours
- **\$800-\$1000 / run**
- **\$100K instrument**
- **\$50k for additional 2 year service contract**
- Capable of sequencing 48 5Mb bacterial genomes per run
- Libraries compatible with HiSeq



# Future Illumina developments

- 2x250bp reads (HiSeq fast run mode)
- Ordered flowcells
- 2x400bp reads (MiSeq)
- 10kb synthetic reads (approx. 5-6 million per lane)
  - Useful for phasing of haplotypes
  - Formed from short reads so repeat spanning is still problematic

# Other equipment (optional)



**Agilent Bravo liquid handling  
robot  
£85k**



**Agilent Tapestation  
£30k**

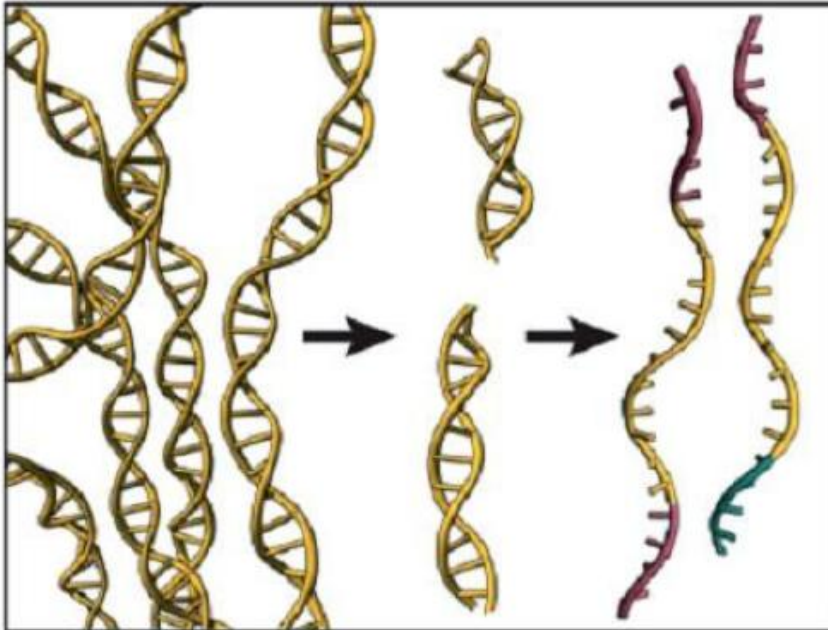


**Covaris 96-well sonicator  
£90k**

# Roche 454 Key Features

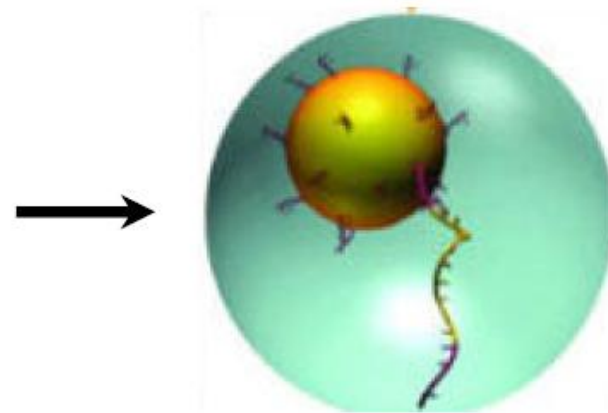
- Advantages
  - Long read lengths (200-1000bp)
  - Multiple samples possible
  - Short run time (< 1 day)
- Disadvantages
  - Relatively expensive (~£8k per run)
  - Low volume of sequence data (100Mb-1Gb)
  - Complex sample prep
  - Roche discontinuing support from 2015

# 454 Step 1: Sample preparation



1. Genomic DNA is isolated and fragmented.
2. Adaptors are ligated to single stranded DNA
3. This forms a library

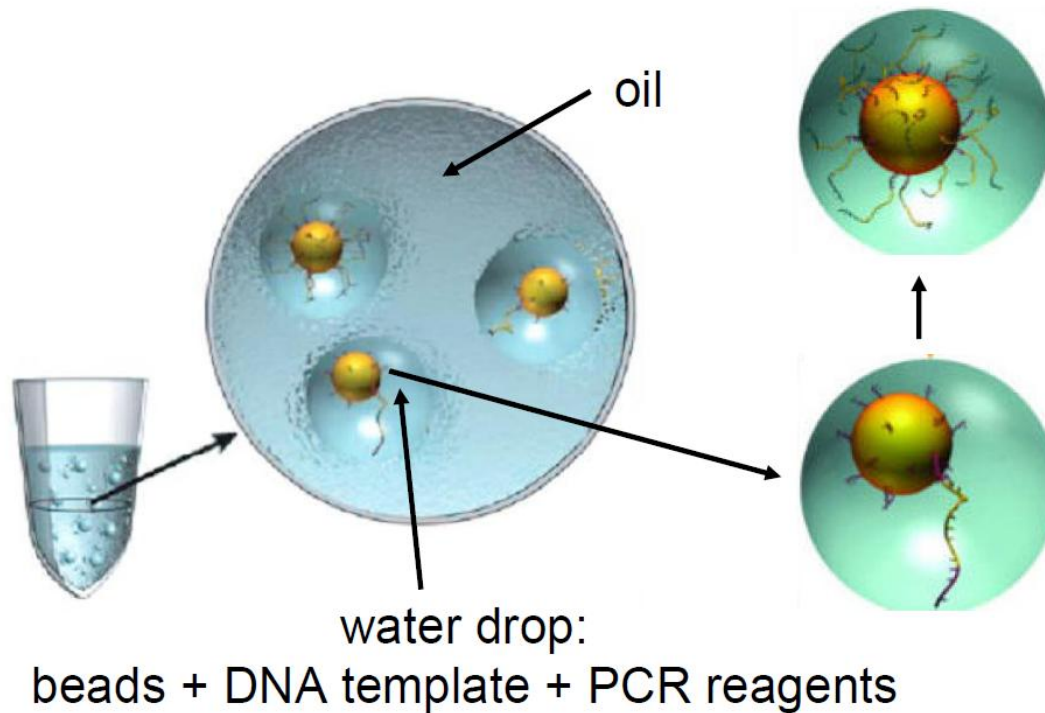
**One Fragment = One Bead**



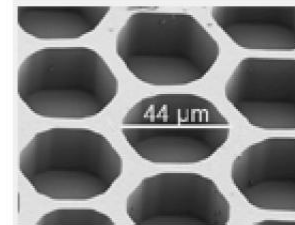
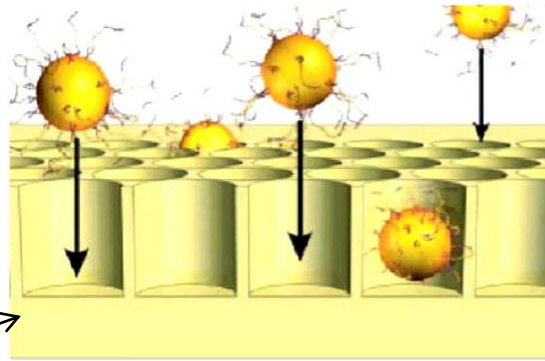
4. The single stranded DNA library is immobilised onto proprietary DNA capture beads

# 454 Step 2: Amplification

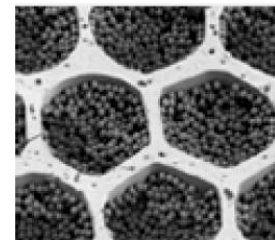
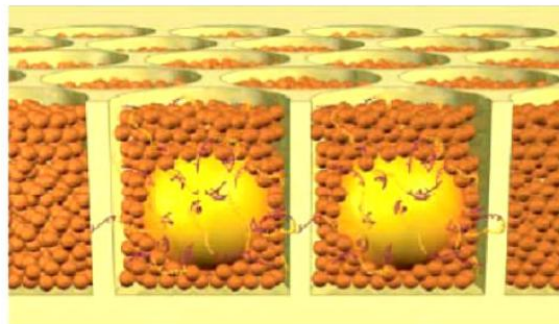
Water-based emulsion PCR



# 454 Step 3: Load emPCR products

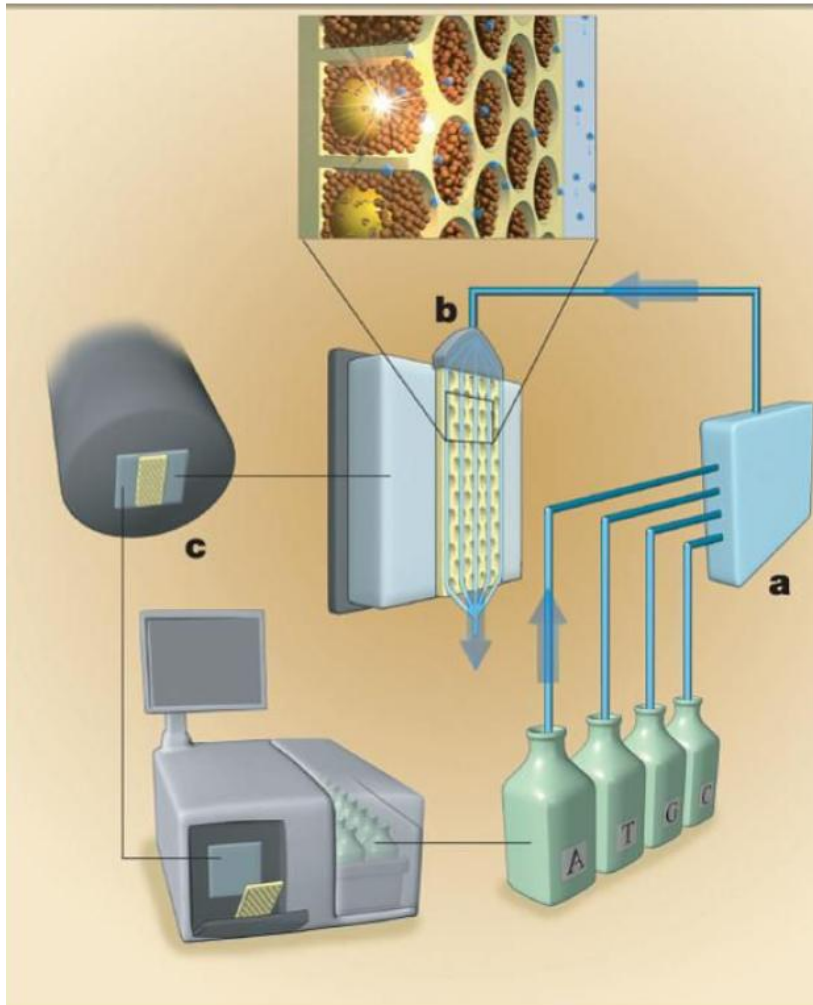


- enrich for DNA + beads
- diameter of the wells allows for only 1 bead/well



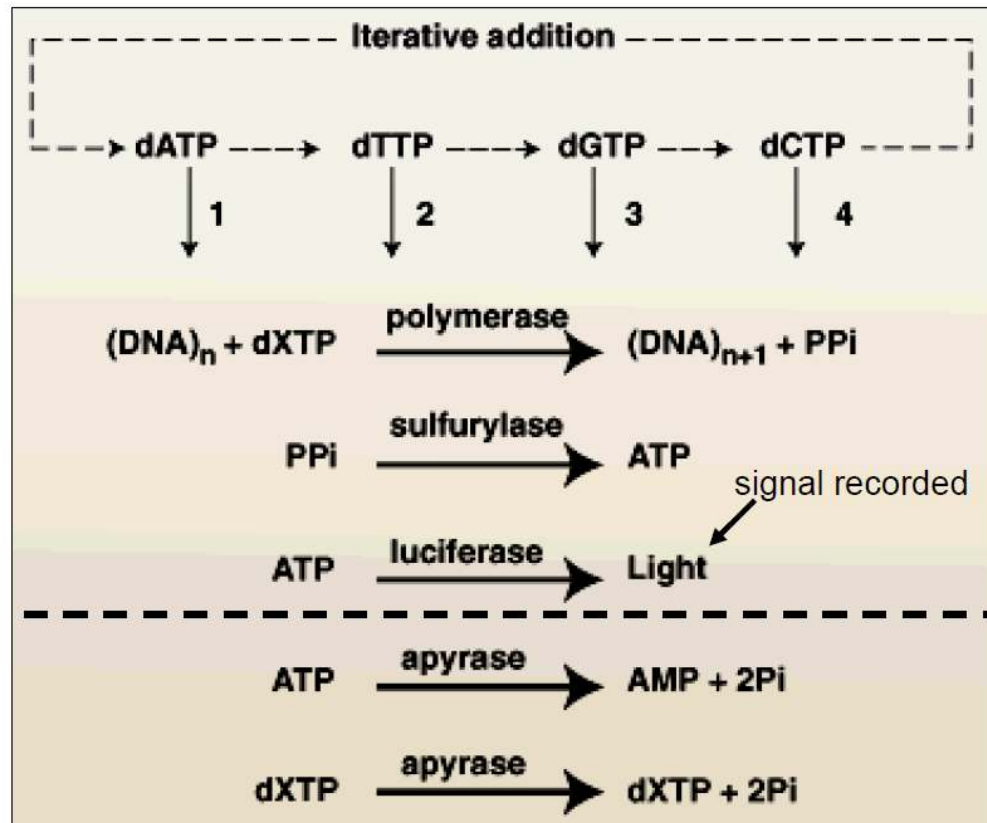
Smaller beads (red) carrying immobilized enzymes required for pyrophosphate sequencing are deposited into each well.

# 454 Step 4: Pyro-sequencing



1. Nucleotides are pumped sequentially across the plate
2. ~ 1 million reads obtained during 1 run
3. Addition of nucleotides to DNA on a particular bead generates a light signal

# 454 Chemistry



# SOLiD

- Differs from Illumina and 454
  - No dXTP reagents are used
  - Oligonucleotide primer-based sequencing is used
  - Two bases are read at a time
  - High accuracy

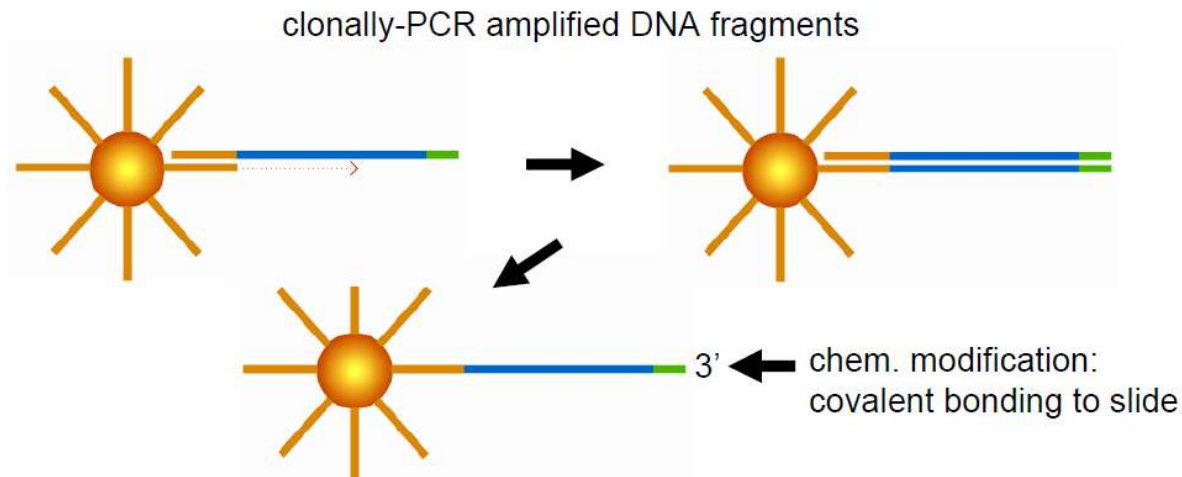
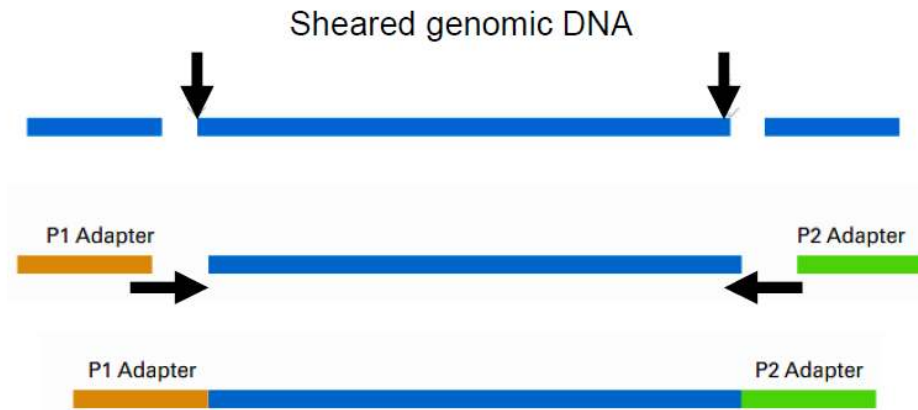
BUT – Only one colour is emitted

Need several sequencing steps to convert colour to a sequence

# Life Technologies SOLiD

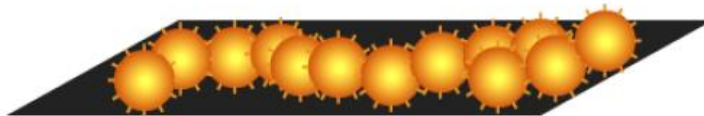
- Advantages
  - Two base encoding system
  - Every base read twice
  - Large volume of sequence data (270Gb per run possible)
- Disadvantages
  - Short read lengths (30-80bp)
  - Complex sample prep
  - Bioinformatics support less comprehensive
  - Paired-end reads more complex than Illumina or 454

# SOLiD: Step 1 Sample Prep

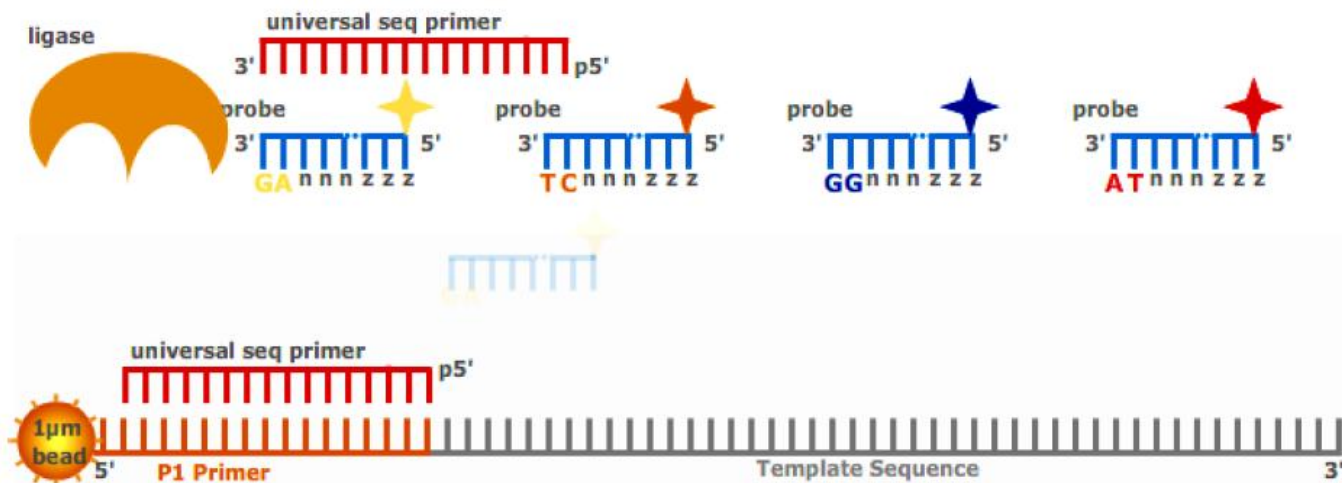


# SOLiD: Step 2 Attach beads

3'-modified beads  
deposited onto glass slide

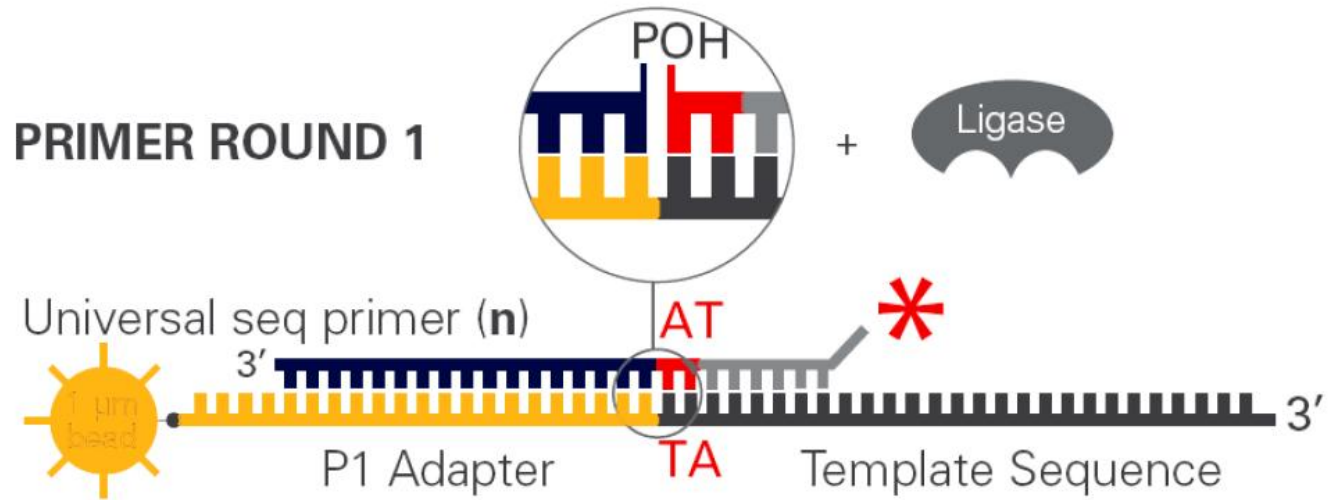


**Sequential ligation** with dye-labeled **oligonucleotides**



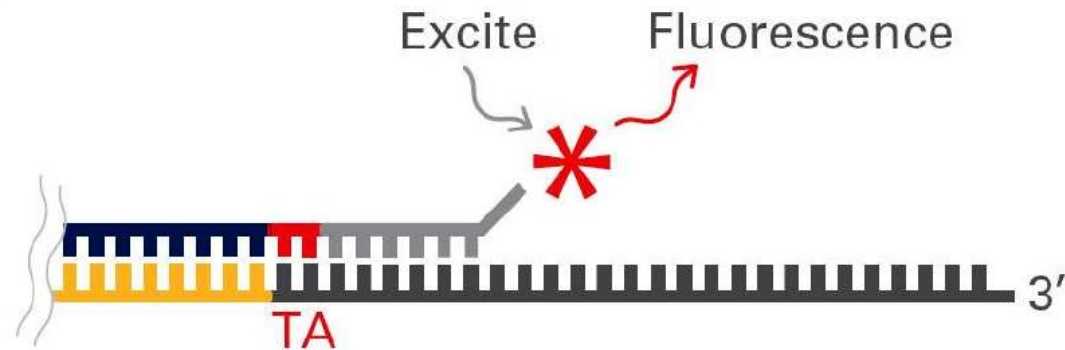
# SOLiD: Step 3 Sequencing 1

## 1. Prime and Ligate

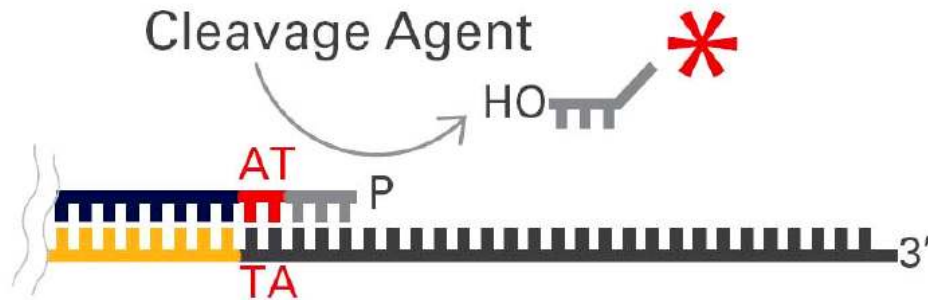


# SOLiD: Step 3 Sequencing 2

## 2. Image



## 3. Cleave off Fluor



# SOLiD Step 3 Sequencing 3

## 4. Repeat steps 1-4 to Extend Sequence

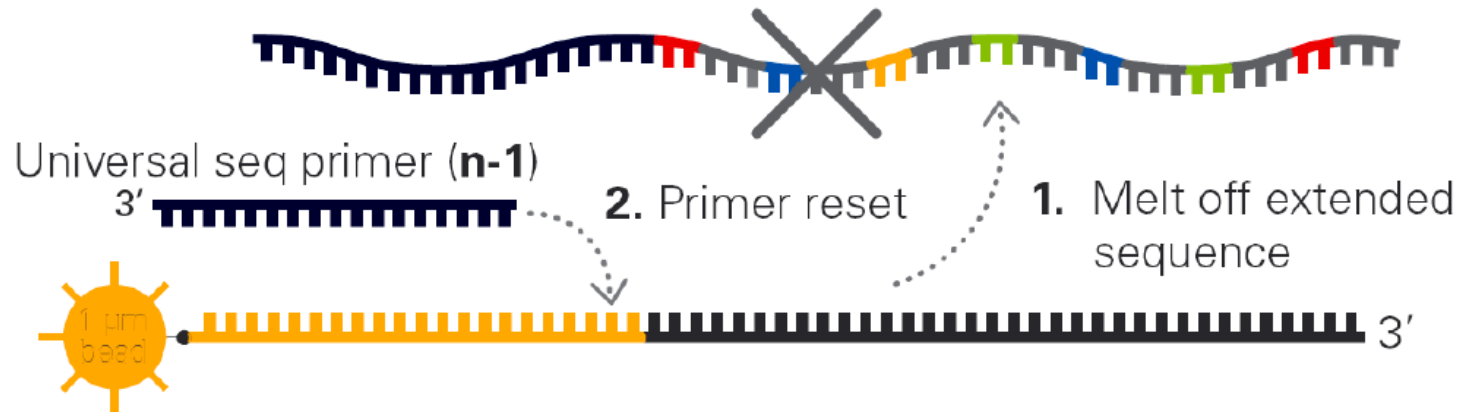
Ligation cycle      1      2      3      4      5      6      7 ... (n cycles)



A random primer is ligated to the template only when the labeled nucleotide complements the fifth nucleotide on the template, counting from the end of the previously ligated primer.

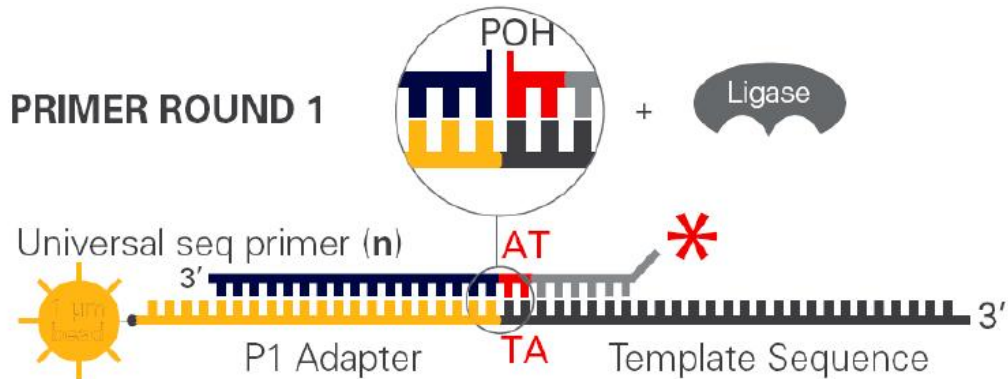
# SOLiD Step 3 Sequencing 4

## 5. Primer Reset

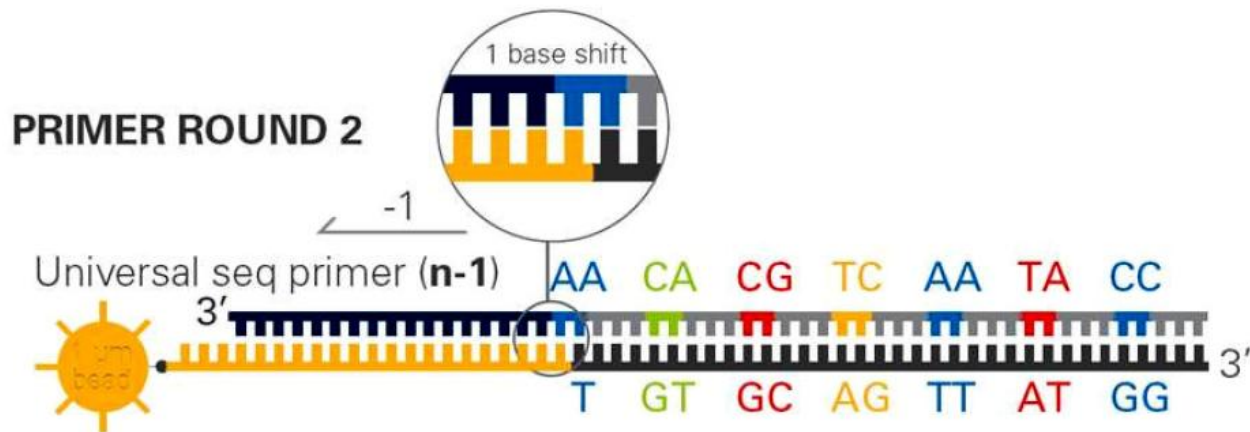


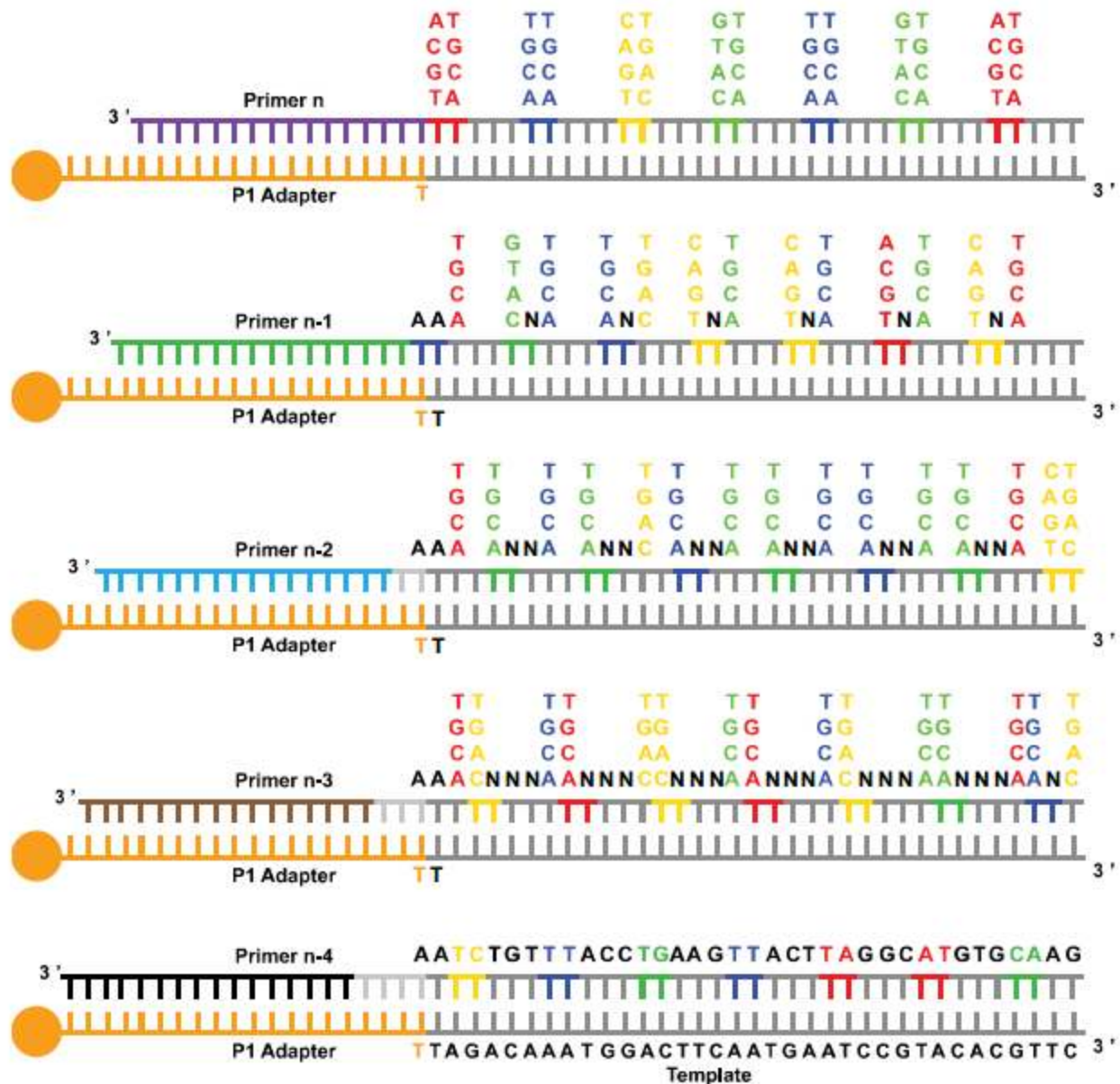
# SOLiD Step 3 Sequencing 5

## 1. Prime and Ligate



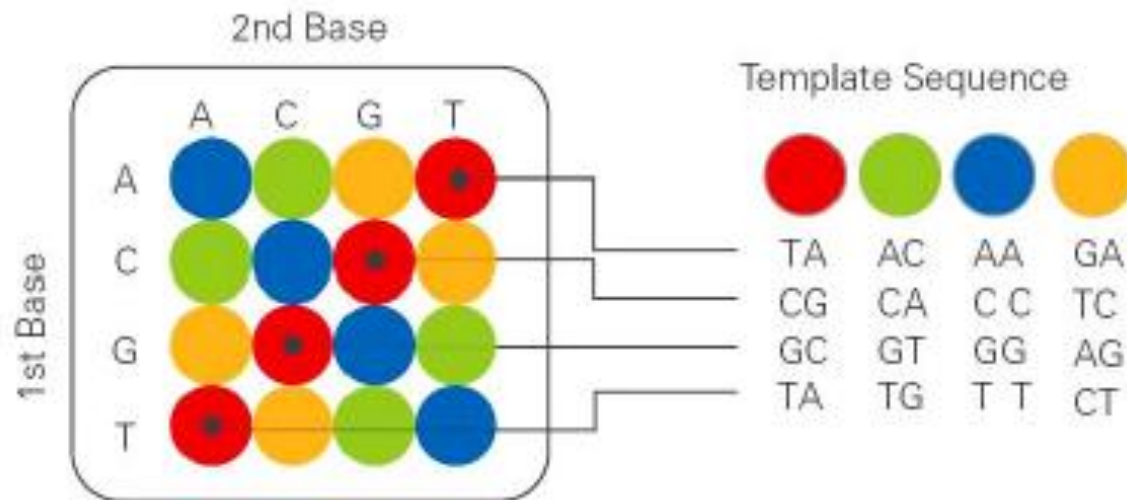
## 6. Repeat steps 1-5 with new primer





# SoLID Colour space

Possible dinucleotides encoded by each color



# Common features

- Most 2<sup>nd</sup> generation platforms share the following:
  - Adaptor sequences to fix probes to a surface/bead
  - Amplification
  - Use of fluorescent probes/CCD devices or pH sensors
  - Capable of paired-end reads
  - Post-processing software to determine image quality
  - Shorter read lengths compared to traditional capillary based sequencers
  - Much higher data volumes (~Gb)
  - Sequence a human genome in a matter of days

# Common features

Images

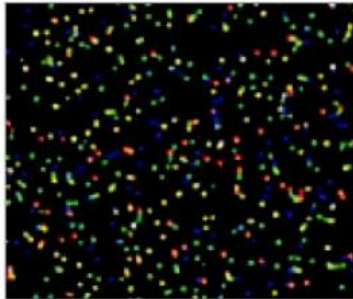


Image Analysis

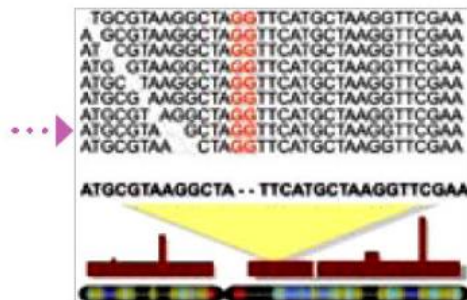
| Line No. | X  | Y    | Cycle 1 - A C G T | Cycle 2 - A C G T       |
|----------|----|------|-------------------|-------------------------|
| 0        | 12 | 604  | 1000              | 403.0 300.0 300.0 200.0 |
| 0        | 12 | 779  | 300               | 403.0 113.0 230.0 110.0 |
| 0        | 12 | 100  | 1000              | 100.0 200.0 403.0 403.0 |
| 0        | 12 | 300  | 1000              | 100.0 200.0 403.0 403.0 |
| 0        | 12 | 1107 | 1207              | 10.0 40.0 40.0 10.0     |
| 0        | 12 | 1074 | 400               | 20.0 40.0 40.0 40.0     |
| 0        | 12 | 807  | 300               | 100.0 400.0 40.0 300.0  |
| 0        | 12 | 807  | 1700              | 40.0 40.0 40.0 40.0     |
| 0        | 12 | 300  | 300               | 400.0 400.0 40.0 40.0   |
| 0        | 12 | 807  | 1700              | 40.0 40.0 40.0 40.0     |
| 0        | 12 | 300  | 300               | 400.0 400.0 40.0 40.0   |
| 0        | 12 | 807  | 1700              | 40.0 40.0 40.0 40.0     |
| 0        | 12 | 300  | 300               | 400.0 400.0 40.0 40.0   |
| 0        | 12 | 807  | 1700              | 40.0 40.0 40.0 40.0     |
| 0        | 12 | 300  | 300               | 400.0 400.0 40.0 40.0   |
| 0        | 12 | 807  | 1700              | 40.0 40.0 40.0 40.0     |

Base Calling

```

ATGGCCTGGGCTAGTTTCGATTACGA
CCTGGGCTAGTTTCGATTACGATCGA
GCTAGTTTCGATTACGATCGATCGTT
ATCGATCGTTGCATGCTGGGGTAGTG
TTTCGATTACGATCGATCGTTGCATGC
TCGATTACGATCGATCGTTGCATGCTG
CTAGTTTCGATTACGATCGATCGTTGC
TCGATTACGATCGATCGTTGCATGCTG
TACGATCGATCGTTGCATGCTGGGGTA
TCGATCGTTGCATGCTGGGGTAGTGC
TCGATTACGATCGATCGTTGCATGCTG
CGATTACGATCGATCGTTGCATGCTGC
TAGTTTCGATTACGATCGATCGTTGC
CGATTACGATCGATCGTTGCATGCTGC
ACGATCGATCGTTGCATGCTGGGGTAG
    
```

Aligned Reads



# Phred Score

- Phred program:  
[http://en.wikipedia.org/wiki/Phred\\_base\\_calling](http://en.wikipedia.org/wiki/Phred_base_calling)
- $Q = -10 \log_{10}(P)$
- $P = 10^{(-Q/10)}$

**Phred quality scores are logarithmically linked to error probabilities**

| Phred Quality Score | Probability of incorrect base call | Base call accuracy |
|---------------------|------------------------------------|--------------------|
| 10                  | 1 in 10                            | 90 %               |
| 20                  | 1 in 100                           | 99 %               |
| 30                  | 1 in 1000                          | 99.9 %             |
| 40                  | 1 in 10000                         | 99.99 %            |
| 50                  | 1 in 100000                        | 99.999 %           |

# Bioinformatics implications

- 100-10,000 fold increase in data volumes
- Tool development
- Data quality is poorer
- Less bioinformatics manpower available per sequencing project
- Finished genomes are usually of poorer quality than Sanger 'gold-standard' genomes
- Due to data volume, other applications have become feasible
- E.g. RNA-seq, ChIP-seq, Meth-Seq.

# Benchtop sequencers

# The NGS Market

- Currently dominated by Illumina (70% instruments)
- Market split into:
  - Low throughput but fast: clinical applications and sequencer for individual labs
  - Very high throughput: genome centers and large-scale projects
- E.g Illumina HiSeq 2000 vs. MiSeq
  - 300Gbase per 10 day run vs 7 Gbase in 48 hours

Niedringhaus, T. P., Milanova, D., Kerby, M. B., Snyder, M. P., & Barron, A. E. (2011). Landscape of next-generation sequencing technologies. *Analytical chemistry*, 83(12), 4327–41. doi:10.1021/ac2010857

# Benchtop sequencers

- Roche 454 Junior, Illumina Miseq are essentially miniature versions of the 454 and HiSeq
- Life Technologies Ion Torrent and Ion Proton are benchtop sequencers derived from 454 pyrosequencing
- Designed for individual groups
- Typical instrument cost is \$150k (inc 3 year service contract)
- Typical run cost in consumables: \$1000/run (at maximum output)

# Illumina MiSeq

- Same technology and chemistry as HiSeq
- 2X250bp
- 7.5 Gbase/run
- Run 48 hours
- **\$800 / run**
- **\$100K instrument**
- **\$50k for additional 2 year service contract**
- No additional wet-lab equipment required
- Capable of sequencing 20-30 bacterial genomes per run
- RNA-seq of up to 6 samples
- Libraries compatible with HiSeq



# Roche 454 Junior

- Same chemistry
  - 100K reads, 700bp
  - 70 Mbases/run
  - Focus on clinical, 510K validated assays
  - **\$1000 per run**
  - **\$100K instrument**
- 
- Now uncompetitive – Roche reviewing future



# Life Technology Ion Torrent

454-like chemistry without dye-labelled nucleotides

- No optics, CMOS chip sensor
- Up to 400bp reads (single-end)
- 2 hour run-time (+5 hours on One Touch)
- Output is dependent on chip type (314, 316 or 318)
- 318 (11M wells) >1Gbase in 3 hours
- **\$700 per run**
- **\$50K for the instrument, plus \$75k for additional One Touch station and Server**
- **Libraries not compatible with Ion Proton**

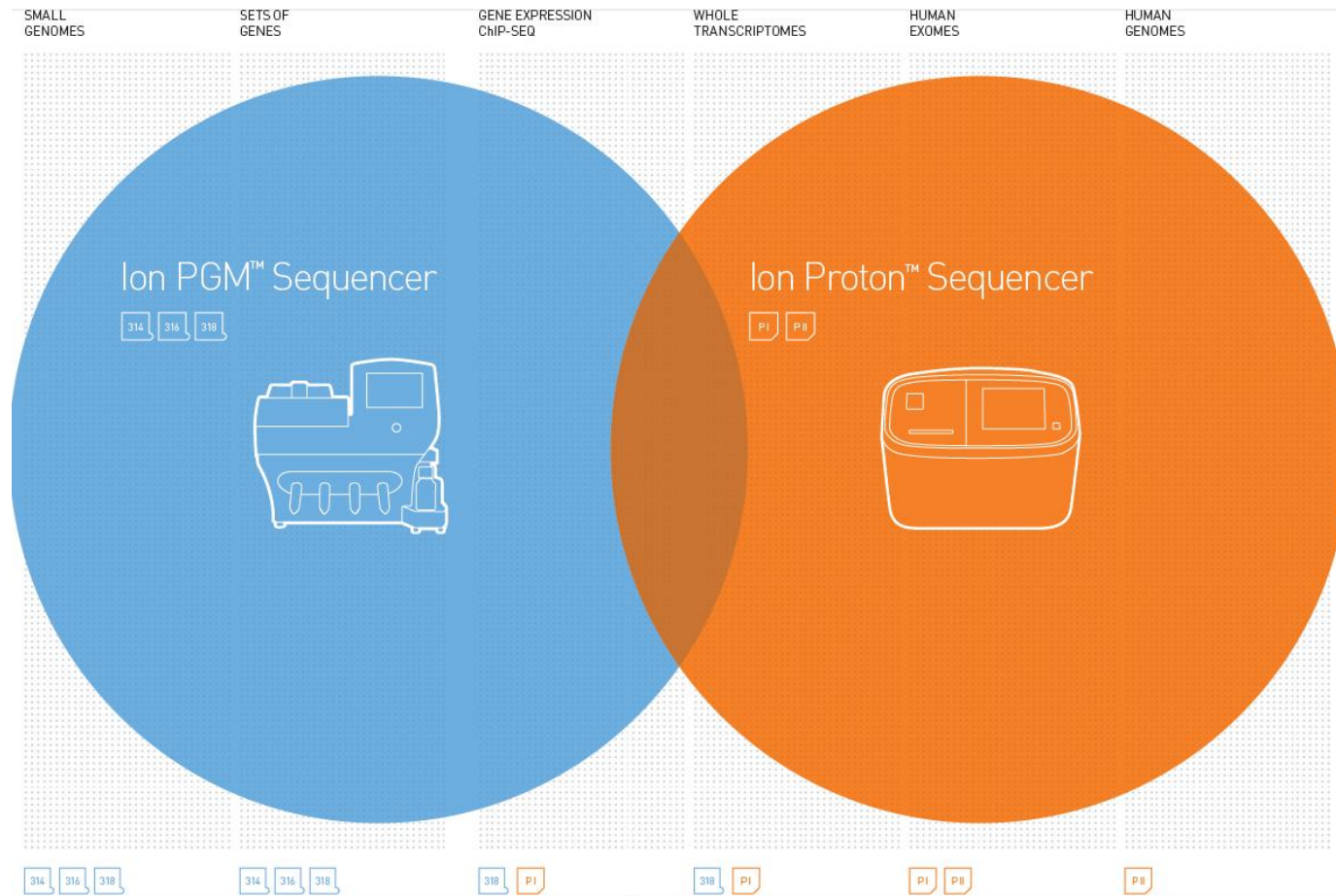


# Life Technology Ion Proton

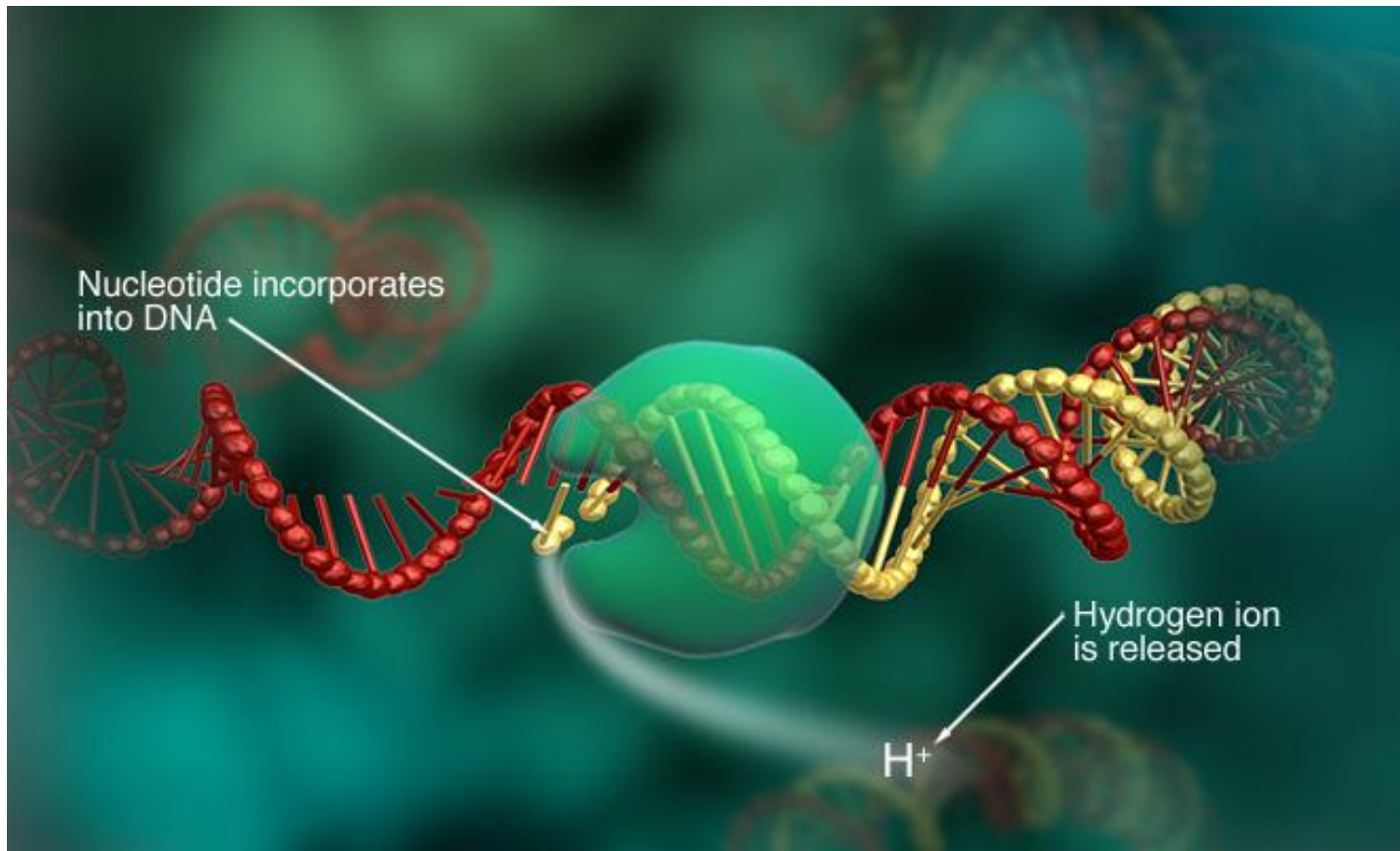
- 454-like chemistry without dye-labelled nucleotides
- No optics, CMOS chip sensor
- Up to 200bp reads (single-end)
- 2 hour run-time (+8 hours on One Touch)
- Output is dependent on chip type (P1 or P2 coming soon)
- 60-80 million reads (P1)
- **\$1500 per run**
- **\$150K for the instrument, plus \$75k for additional One Touch station and Server**
- **Libraries not compatible with Ion Torrent**



# Ion Torrent vs Ion Proton

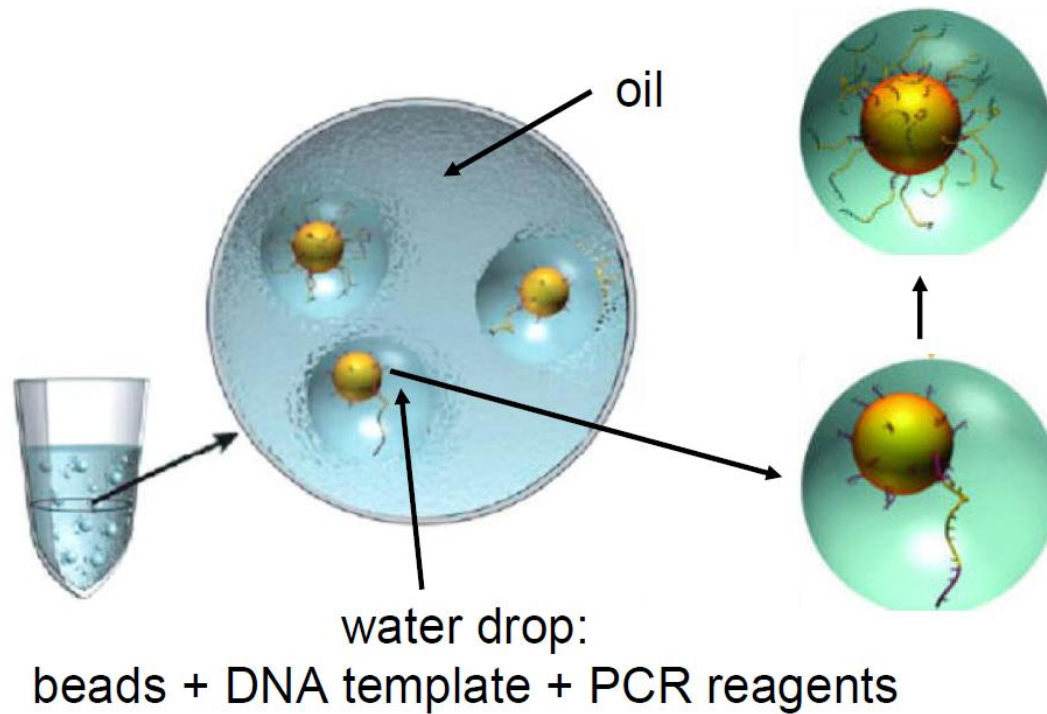


# Ion Torrent

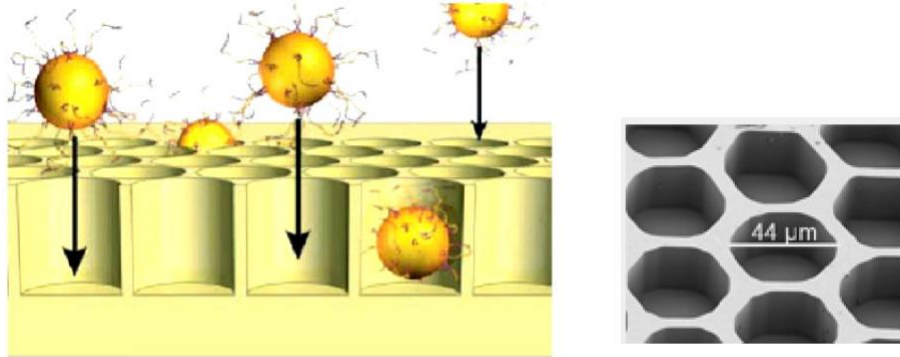


# Library prep

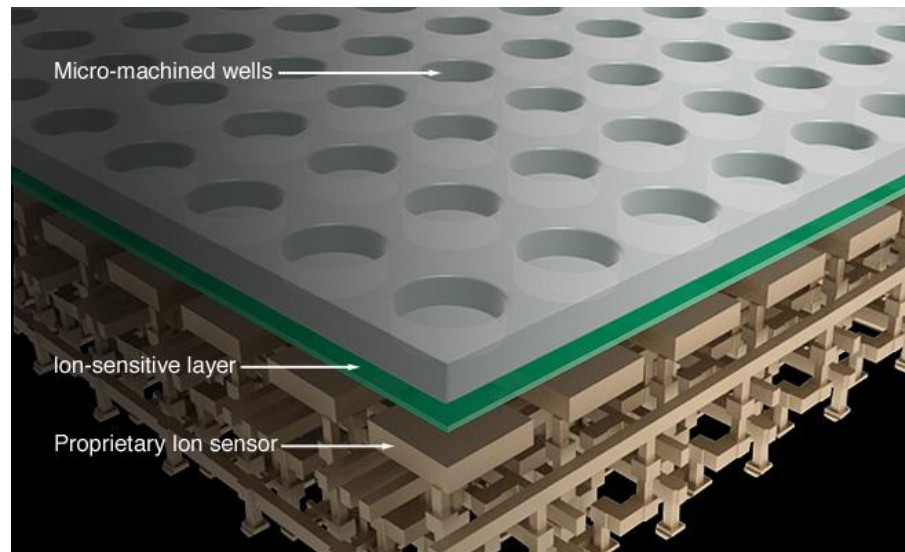
- 454 style library using emulsion PCR



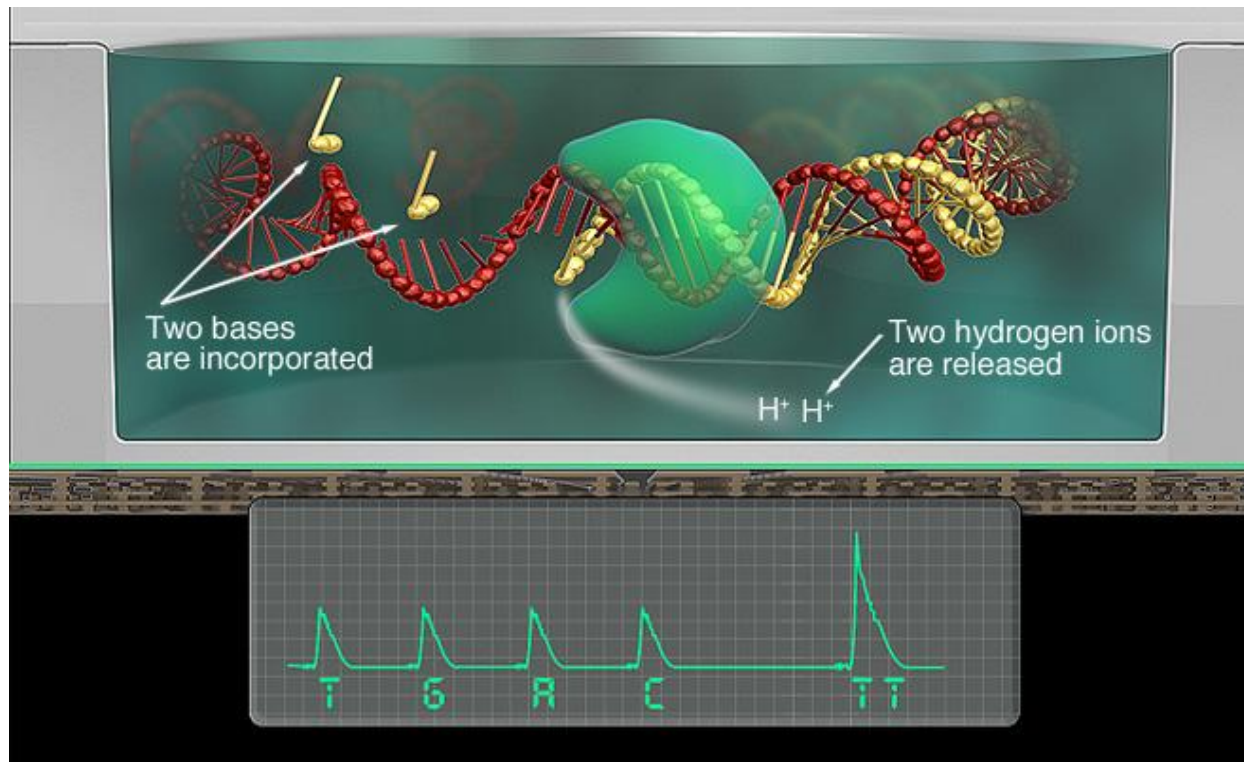
# Ion Torrent



- enrich for DNA + beads
- diameter of the wells allows for only 1 bead/well



# Ion System



# Benchtop sequencers



## Ion Proton (P1 chip)

- 60-80M reads
- Up to single-end 200 base pair runs
- 16Gb/run
- 4 hour run time
- **\$1500/run**
- **\$150K instrument**
- **One touch system required**



## Illumina MiSeq

- 30M reads
- 2X300bp
- 15 Gbase/run
- Run 48-72 hours
- **\$1000 / run**
- **\$100K instrument**
- **No additional equipment required**



## Roche 454 Junior

- Same chemistry
- 100K reads, 700bp
- 70 Mbases/run
- Focus on clinical, 510K validated assays
- **\$1000 per run**
- **\$100K instrument**

# Useful benchtop review paper

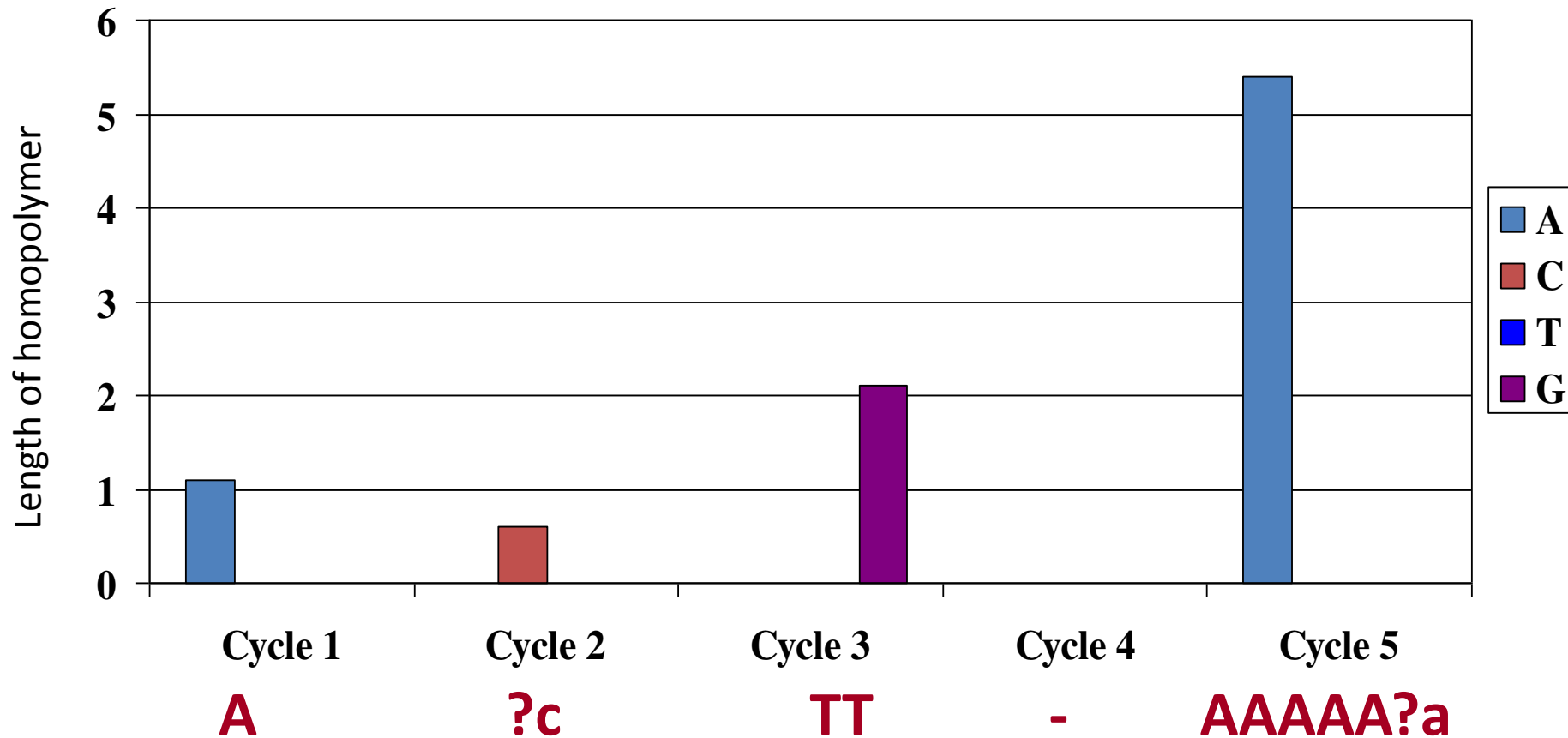
- Loman, N. J., Misra, R. V, Dallman, T. J., Constantinidou, C., Gharbia, S. E., Wain, J., & Pallen, M. J. (2012). Performance comparison of benchtop high-throughput sequencing platforms. *Nature biotechnology*, 30(5), 434–9. doi:10.1038/nbt.2198

# Possible problems

- These are common to all platforms
  - Biases introduced by sample preparation
  - Errors in base-calling
  - High GC/AT biases can cause difficulties
- 454 and Ion Torrent have difficulty sequencing homopolymeric tracts accurately
- Latest Ion Torrent reagent upgrades claim to reduce these
- Illumina also has specific motifs which are difficult to sequence

*Nakamura, K., Oshima, T., Morimoto, T., Ikeda, S., Yoshikawa, H., Shiwa, Y., Ishikawa, S., et al. (2011). Sequence-specific error profile of Illumina sequencers. Nucleic acids research, gkr344–. Retrieved from <http://nar.oxfordjournals.org/cgi/content/abstract/gkr344v1>*

# Homopolymer errors



- Different between signal of 1 and signal of 2 = **100%**.
- Different between signal of 5 and 6 is **20%**
- More difficult to decide if we have AAAAA or AAAAAA
- Is the final sequence: ACTTNAAAAA or ANTTAAAAAA or ATTAAAA.... Etc

# Third generation sequencers

# Third generation sequencers

- My definition: Single-molecule sequencing
- Currently only PacBio RS is commercially available

# Pacific Biosciences RS II

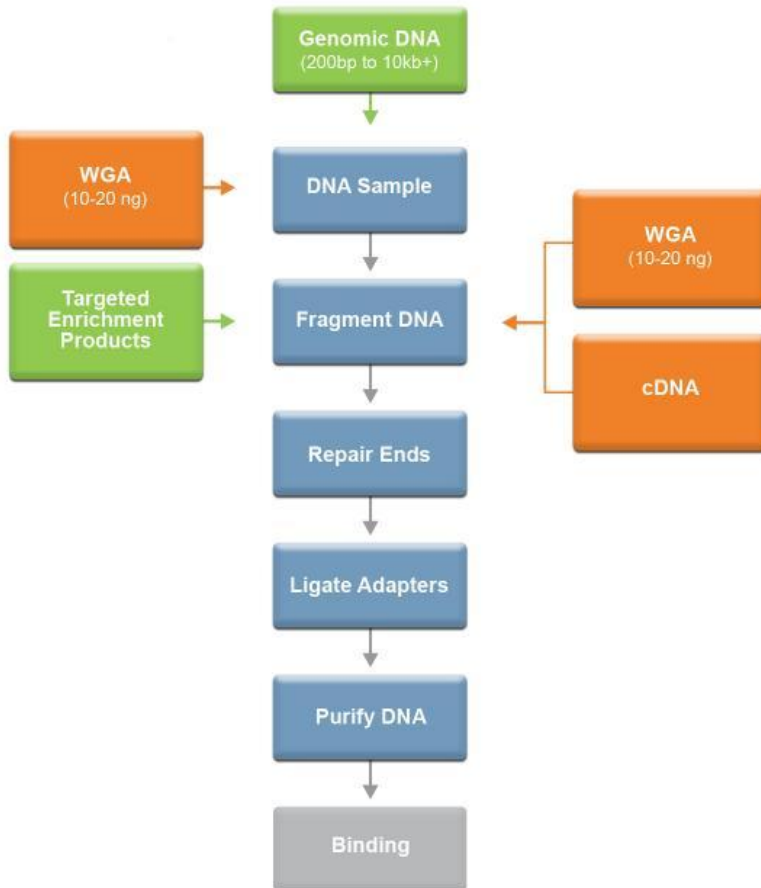


# Introduction

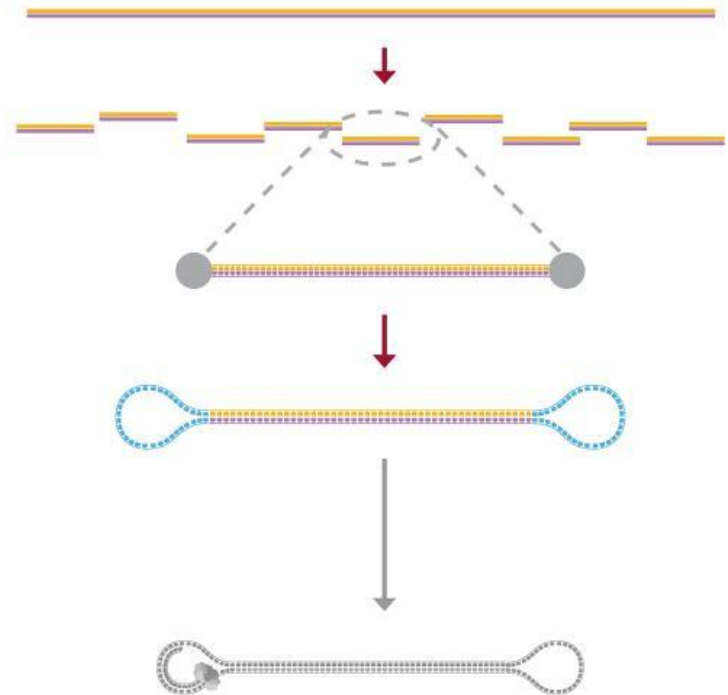
- Based on monitoring a single molecule of DNA polymerase within a zero mode waveguide (ZMW)
  - 150,000 ZMWs on a SMRT flowcell on PacBio RSII
- Nucleotides with fluorophore attached to phosphate (rather than base) diffuse in and out of ZMW (microseconds)
- As polymerase attaches complementary nucleotide, fluorescent label is cleaved off
- Incorporation excites fluorescent label for milliseconds -> nucleotide recorded

# Library prep

## Sample Preparation



## Building of SMRTbell

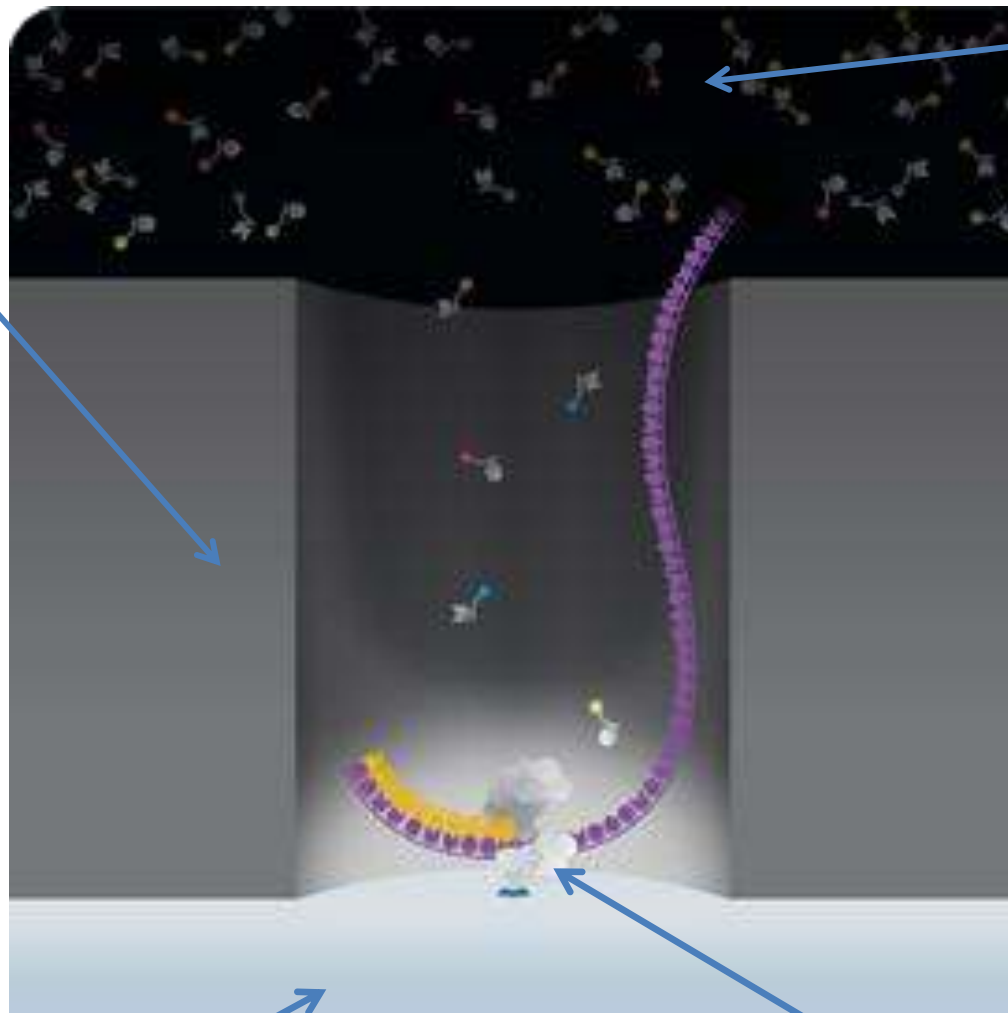


# SMRT Cell



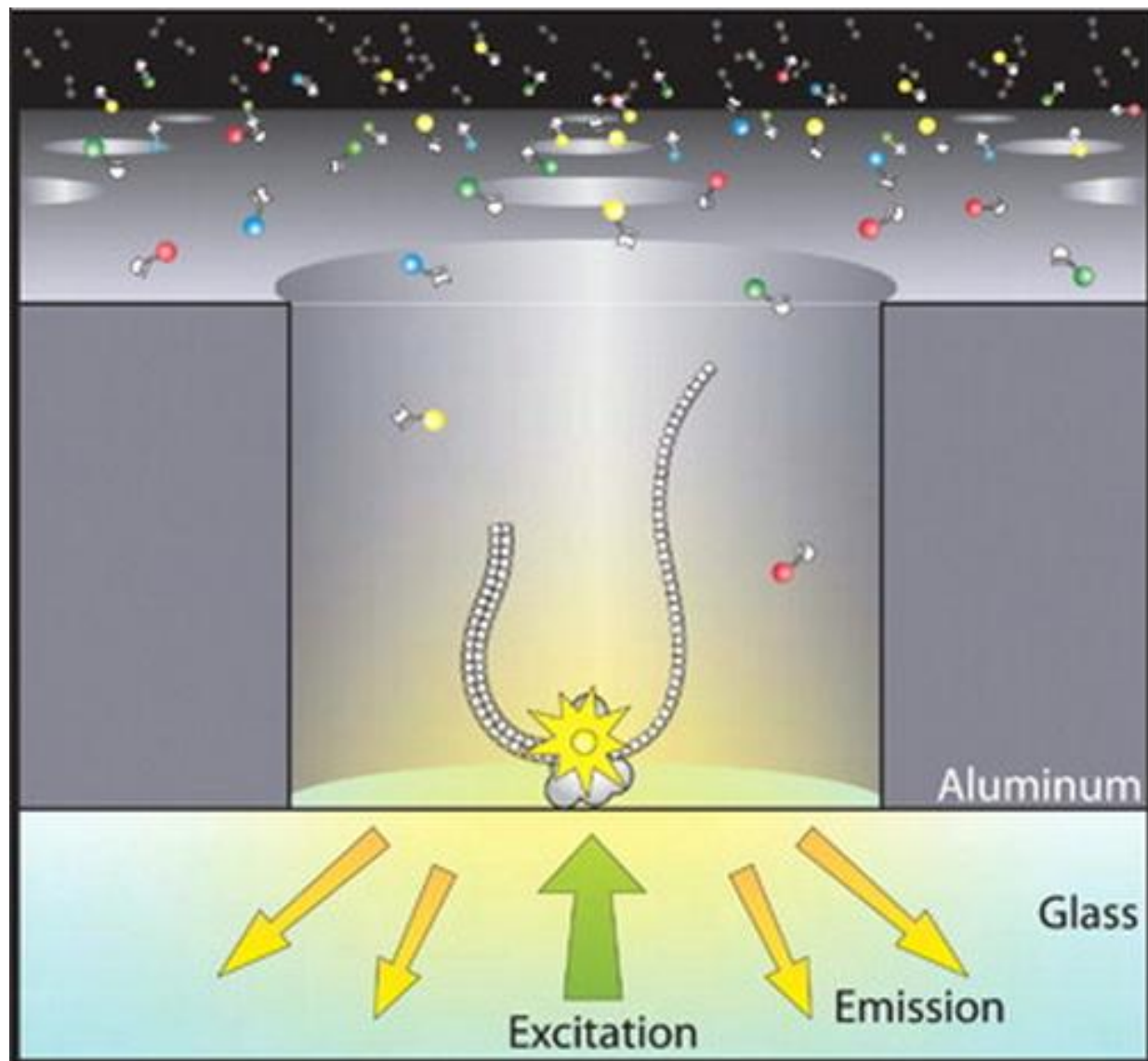
Zero mode waveguide

Free nucleotides

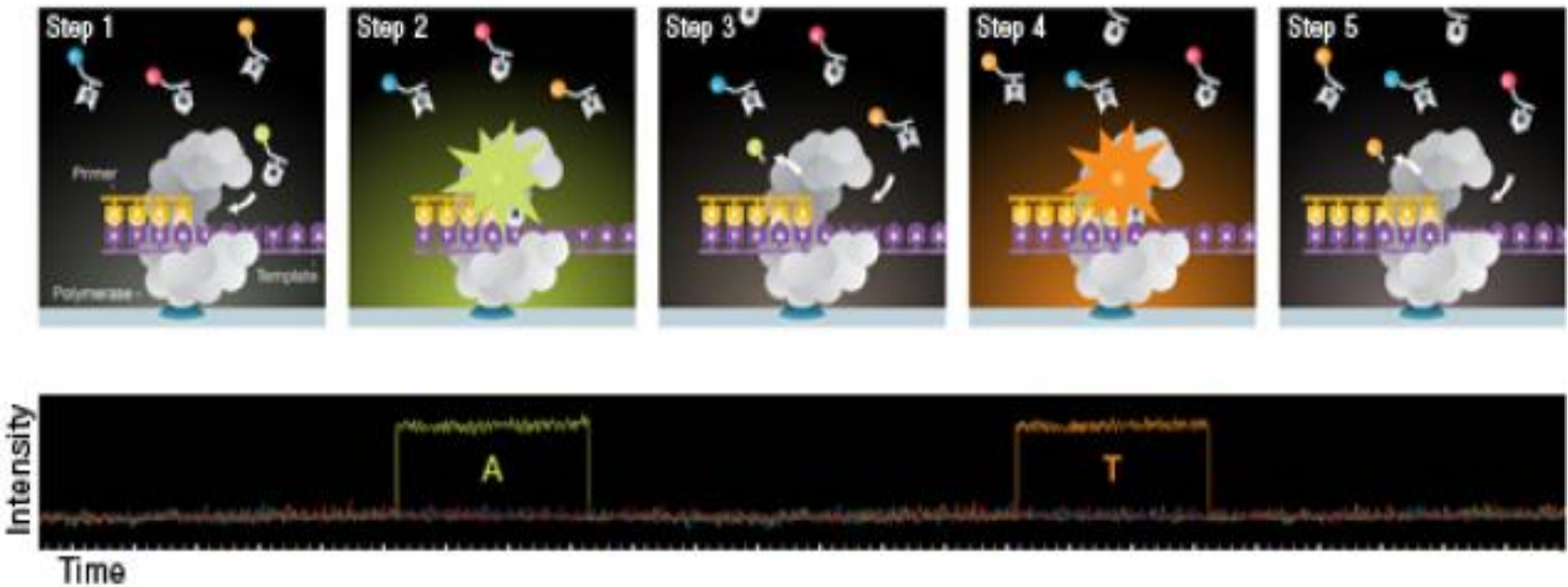


Laser and detector

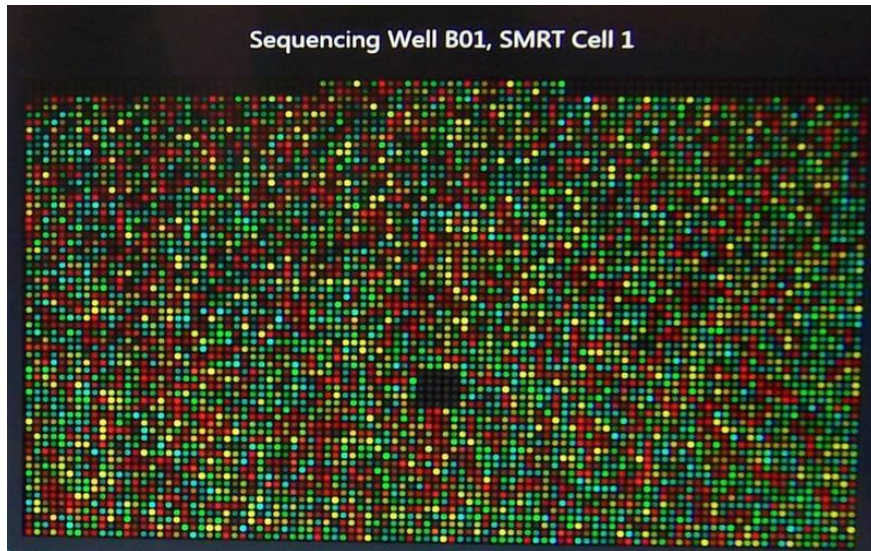
Immobilised DNA polymerase



# Observing a single polymerase

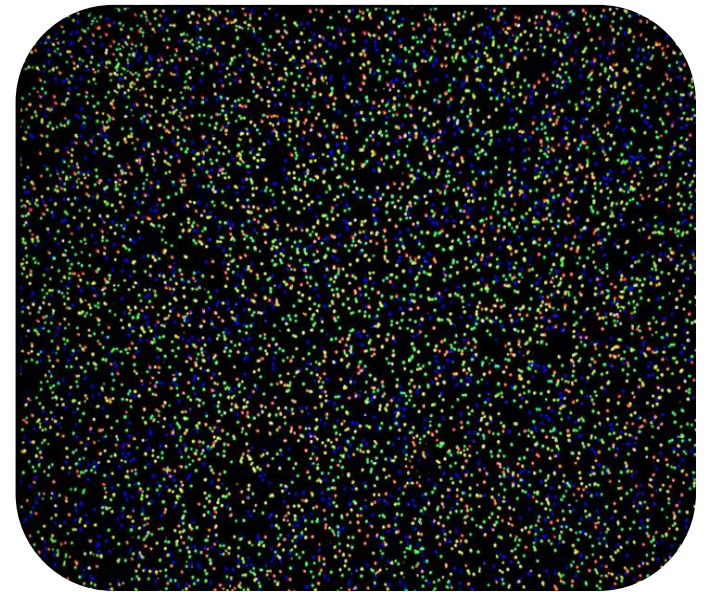


# What it looks like



PacBio ZMWs with single  
DNA strand

Ordered



Illumina DNA mono-colonial clusters

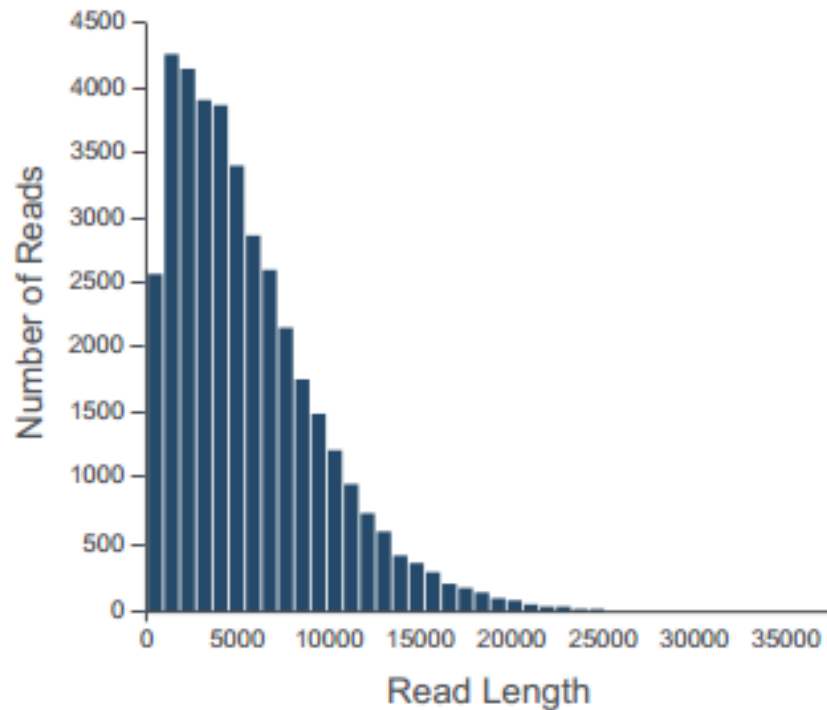
Unordered

# Output statistics

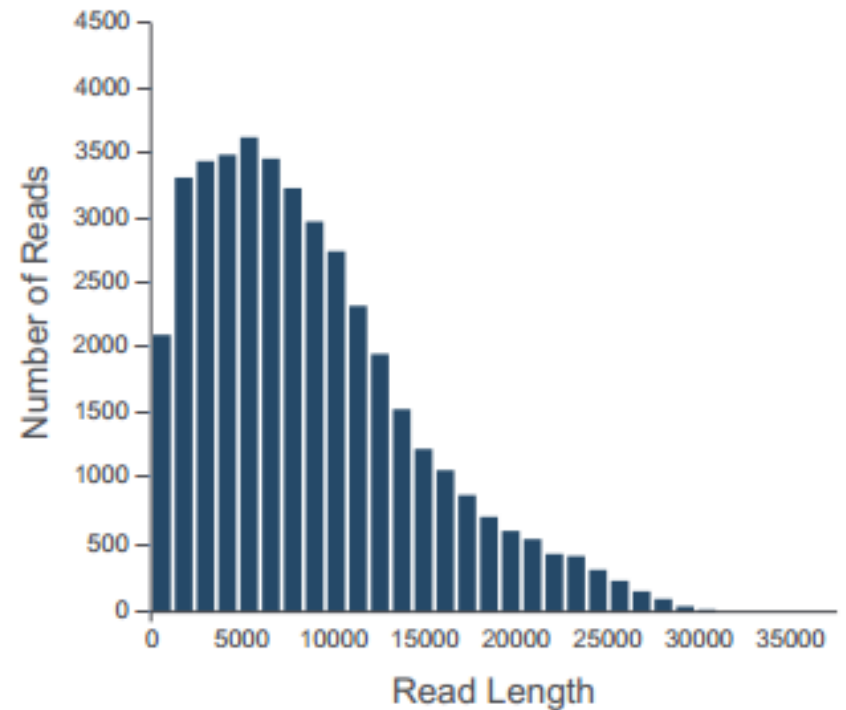
- Approximately 100,000-150,000 sequences per SMRT flowcell
- 300-500Mb output per SMRT flowcell
  - \$500 per run
- Library prep required
  - ~\$500 per sample
  - ~0.5ug per sample
- Size selection required to get the longest reads
- Read lengths
  - Distribution
  - Mean 8.5kb up to 20-25kb

# Read lengths

P4-C2 Chemistry



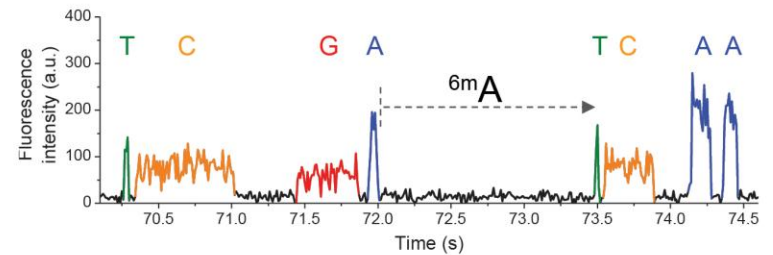
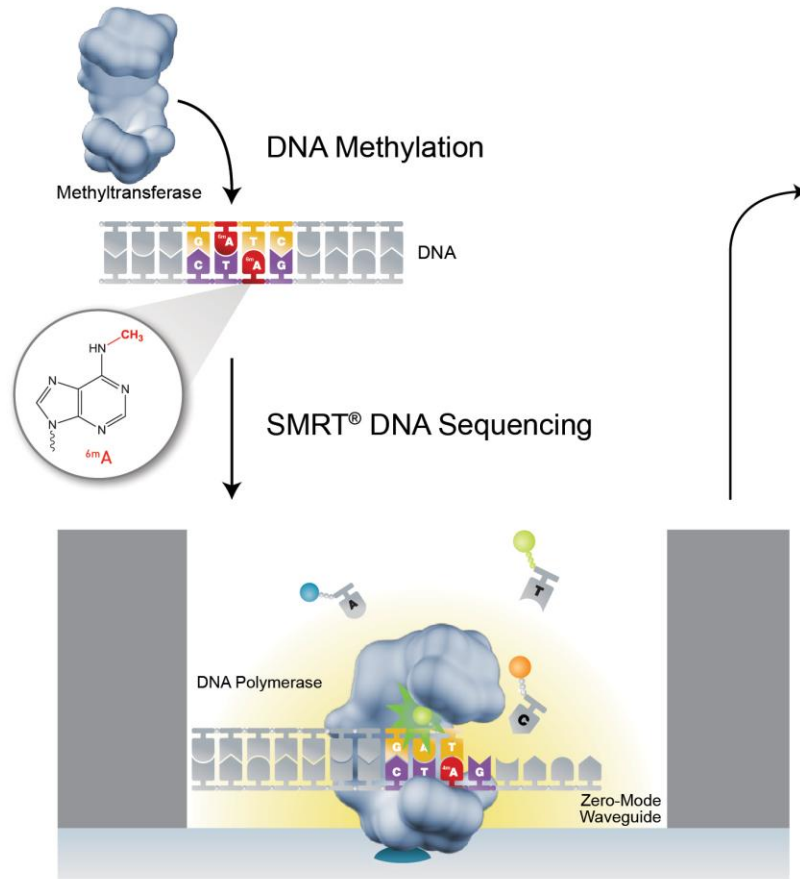
P5-C3 Chemistry



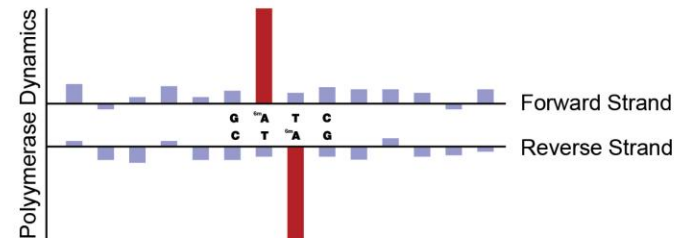
# Novel applications

- Epigenetic changes (e.g. Methylation) affect the amount of time a fluorophore is held by the polymerase
- Circularise each DNA fragment and sequence continuously

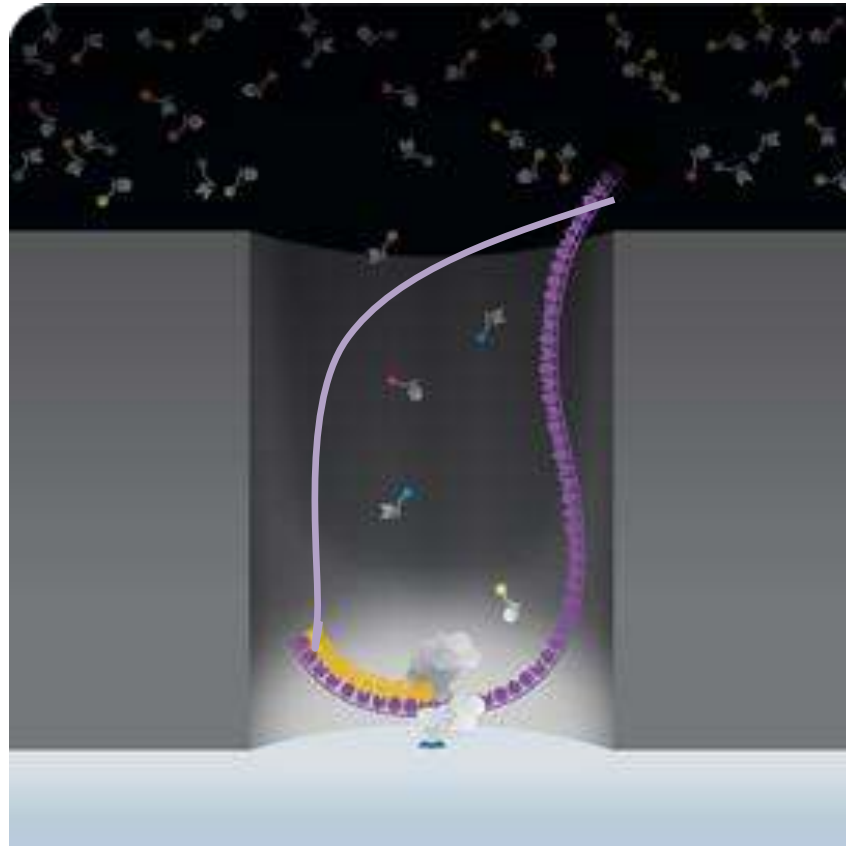
# Epigenetic changes



Analysis of Polymerase Kinetics



# Circular consensus sequencing



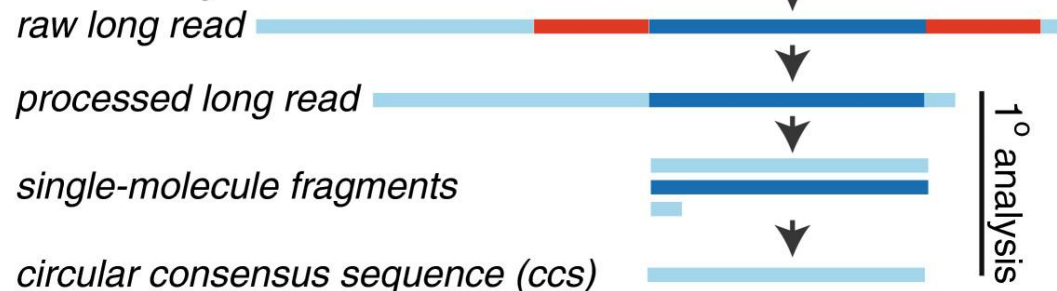
# Circular consensus sequencing

1. generate amplicon

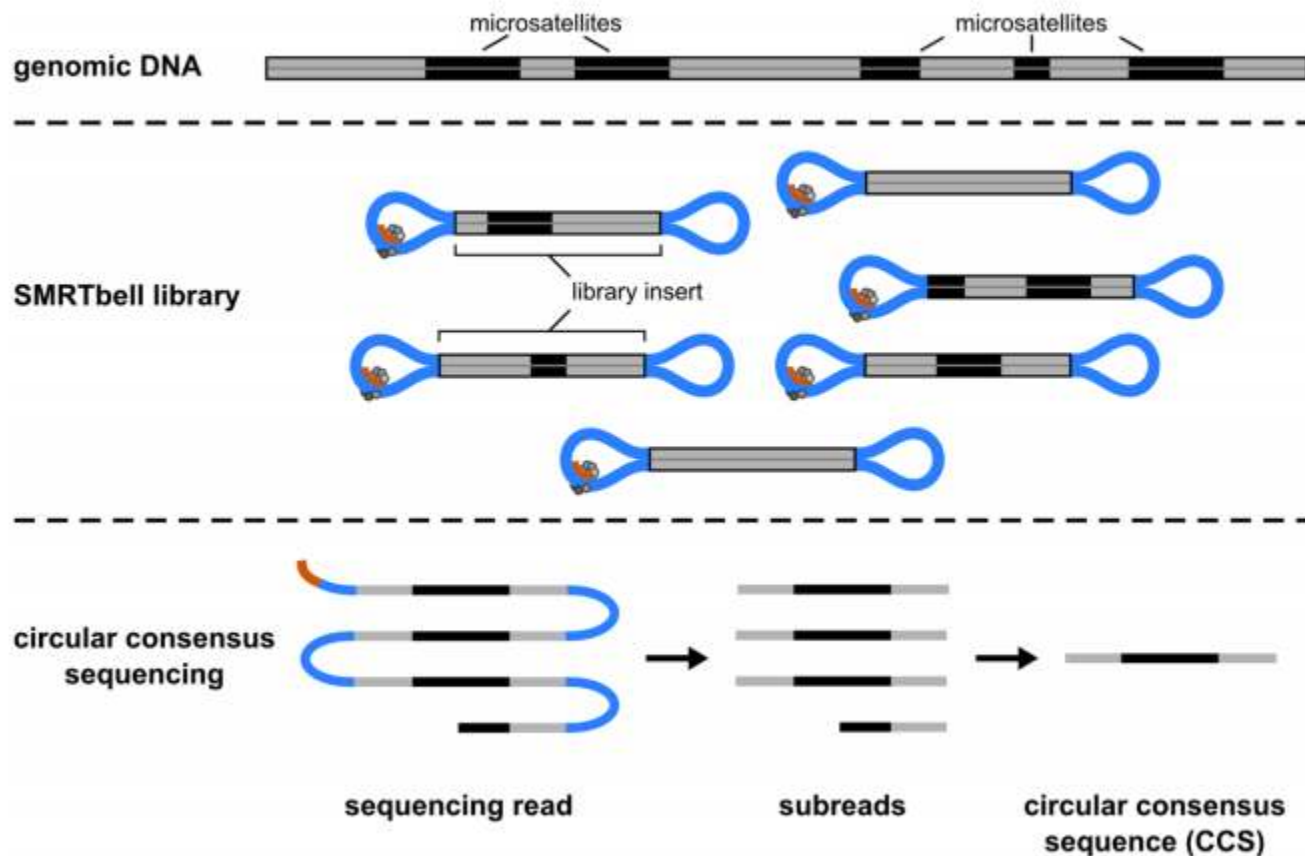
2. ligate adaptors

3. sequence

4. data analysis



# Circular consensus sequencing for rRNA or microsatellites



<http://www.sciencedirect.com/science/article/pii/S0167701213002728>

[http://www.biotechniques.com/multimedia/archive/00230/BTN\\_A\\_000114104\\_O\\_230651a.pdf](http://www.biotechniques.com/multimedia/archive/00230/BTN_A_000114104_O_230651a.pdf)

<http://www.microbiomejournal.com/content/1/1/10>

# Issues to be aware of

- PCR chimeras (affects all PCR amplicon methods)
- Chimeric sequences can be generated during library preparation
- Shorter sequences can be loaded preferentially
  - Uniform amplicon size reduces this
  - PacBio Magbead loading system

# Circular consensus sequencing (CCS)

- Raw error rates of a single pass read is high (10-15%)
- It is possible to read the same molecule repeatedly using CCS mode sequencing
- Can do this up to seven times to reduce error rates to around 0.1-0.5%
- Disadvantages
  - Reduction in read length proportional to number of passes (e.g 7 passes – max read length 3kb).
  - Reduction of total number of reads as some ZMW polymerases will fail

# Pacific Biosciences

- Advantages

- Longer reads lengths (median 8.5kb up to 25kb with P5-C3 chemistry)
- 40 minute run time
- Cost per run is low (\$400 per run plus \$400 per library prep)
- Same molecule can be sequenced repeatedly
- Epigenetic modifications can be detected
- Long reads enable haplotype resolution

- Disadvantages

- Library prep required (micrograms needed)
- If you use PCR based methods – it is NO LONGER single molecule
- Enzyme based
- Only 50,000 reads/run. 400-500Mb yield
- High (10-15%) error rate per run (but CCS can reduce this to  $<\sim 1\%$ )
- \$750k machine
- Lab requirements very stringent

# Bioinformatics Implications

- Relatively low data and high per base cost limits practical widespread use
- Can obtain useful 20-25kb fragments (C5 chemistry)
- Best used in conjunction with error correction algorithms utilising shorter PacBio reads or Illumina data
- Excellent to help scaffold genomes
- Able to generate complete bacterial genomes

 Koren, Sergey; Schatz, Michael C; Walenz, Brian P; Martin, Jeffrey; Howard, Jason T et al. (2012)

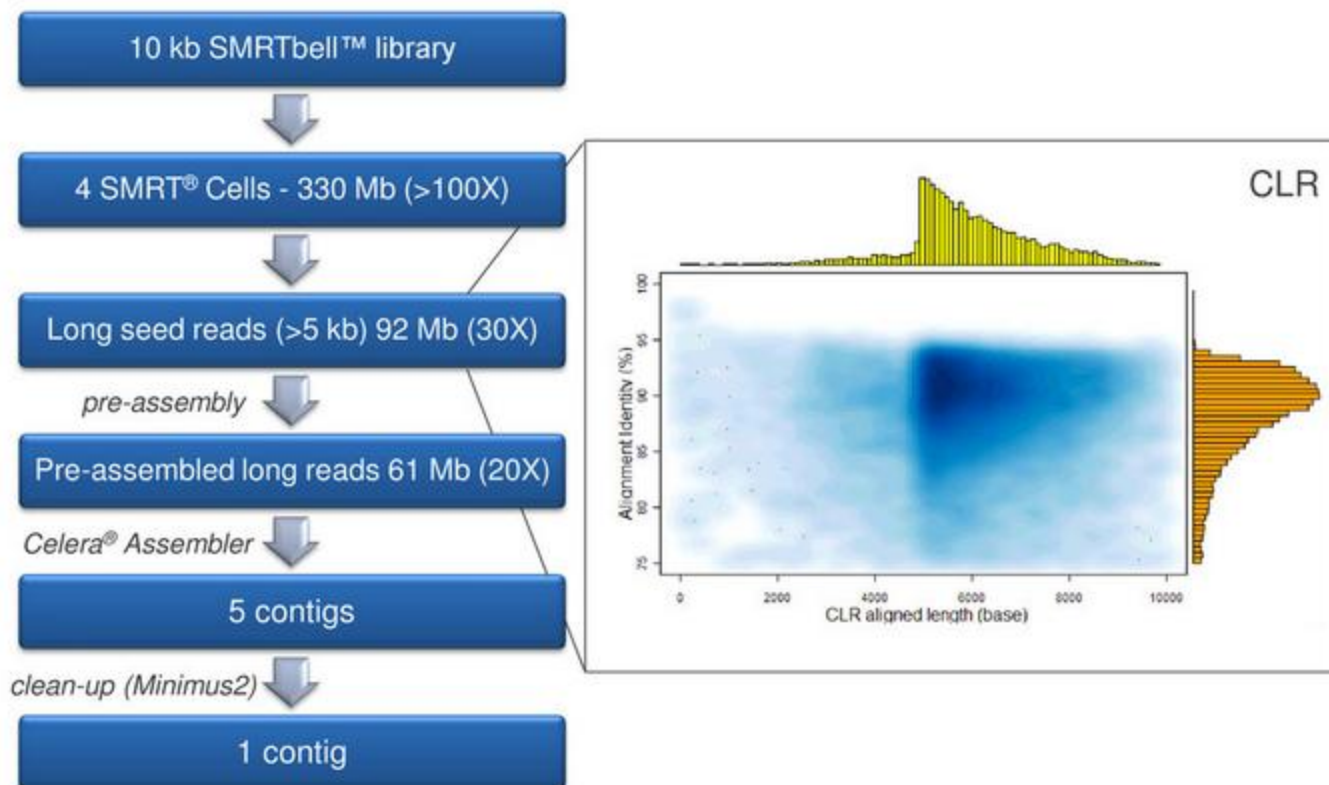
[Hybrid error correction and de novo assembly of single-molecule sequencing reads](#)

*Nature biotechnology* vol. 30 (7) p. 693-700

Chin, C.-S., Alexander, D. H., Marks, P., Klammer, A. A., Drake, J., Heiner, C., ... Korlach, J. (2013). Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data. *Nature methods*, 10(6), 563–9. doi:10.1038/nmeth.2474

# Hierarchical genome assembly

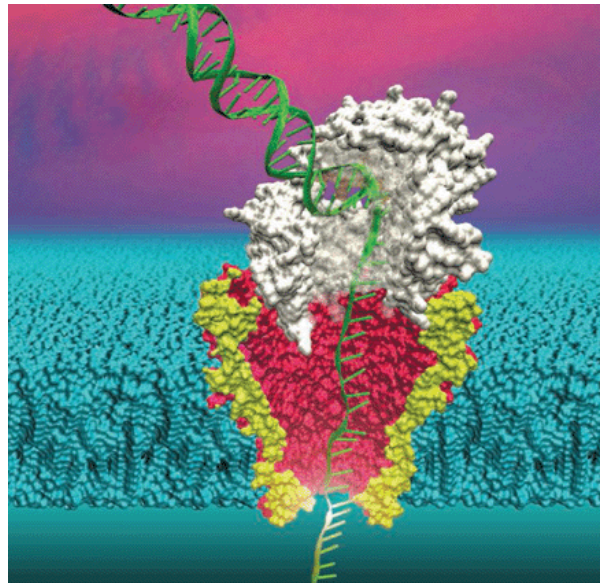
## HGAP Example - *Meiothermus ruber* (JGI)



# PacBio training resources

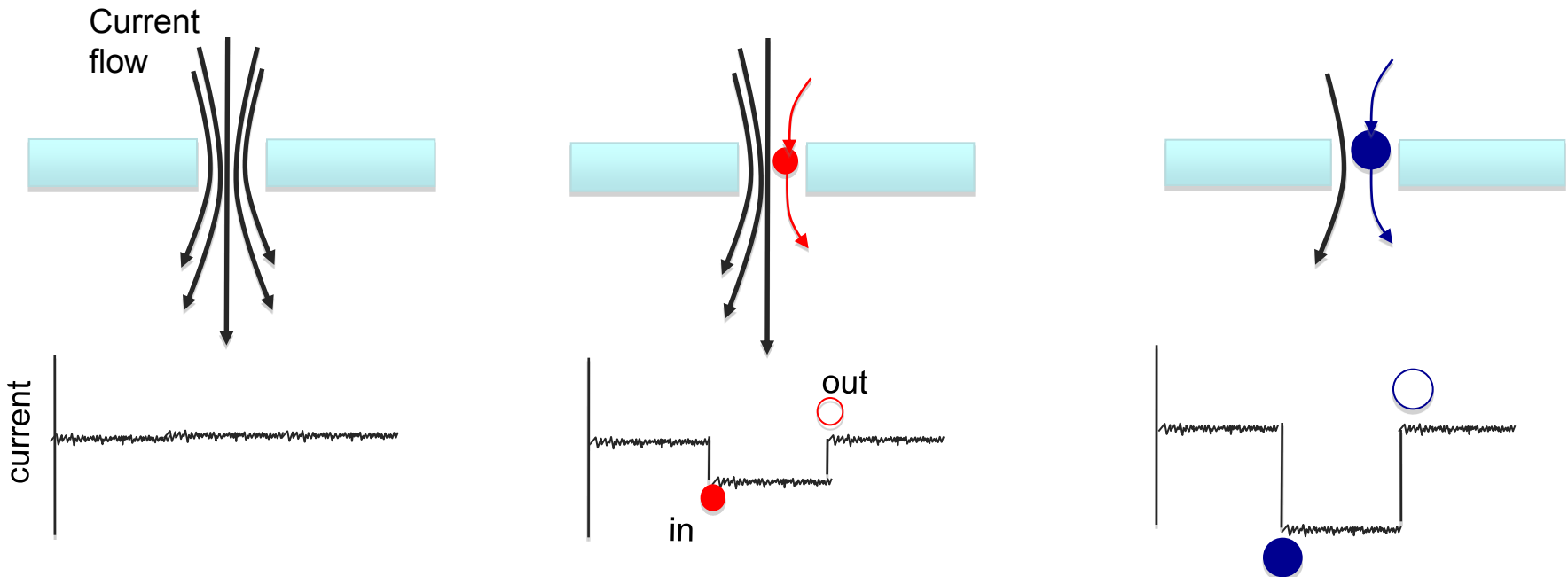
- <https://github.com/PacificBiosciences/Bioinformatics-Training/wiki>

# Nanopore sequencing



# What is a nanopore?

- Nanopore = ‘very small hole’
- Electrical current flows through the hole
- Introduce analyte of interest into the hole → identify “analyte” by the disruption or block to the electrical current

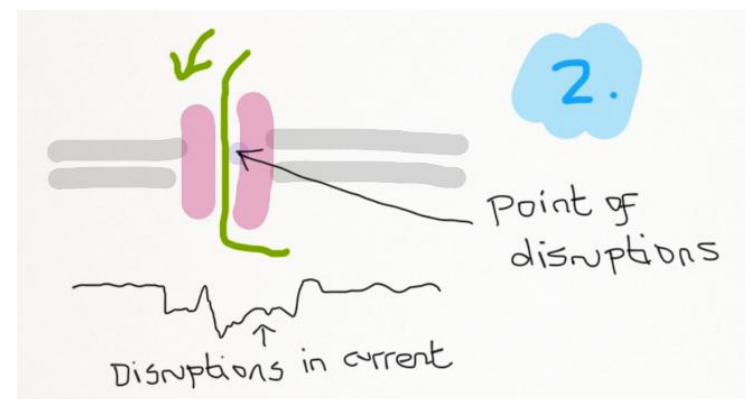
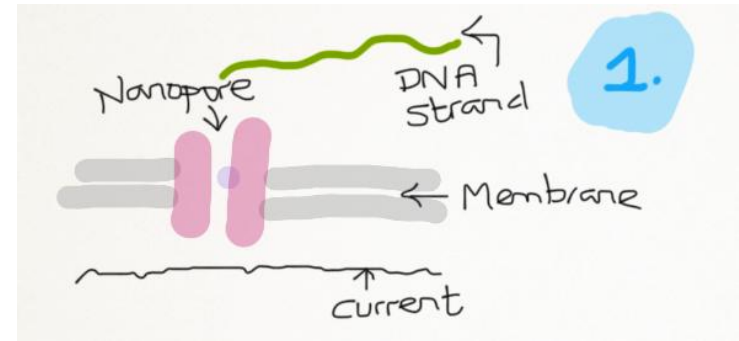


# What is a nanopore?

- Either biological or synthetic
- Biological
  - Lipid bilayers with alpha-haemolysin pores
  - Best developed
  - Pores are stable but bilayers are difficult to maintain
- Synthetic
  - Graphene, or titanium nitride layer with solid-state pores
  - Less developed
  - Theoretically much more robust

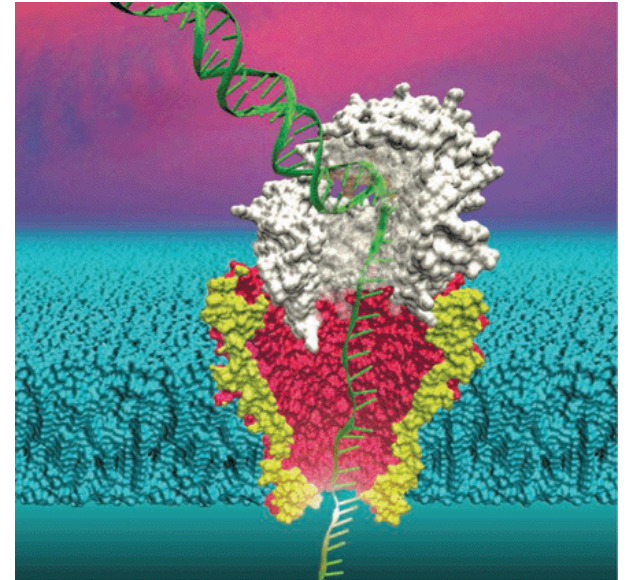
# Nanopore sequencing

- Theory is quite simple
- Feed a 4nm wide DNA molecule through a 5nm wide hole
- As DNA passes through the hole, measure some property to determine which base is present
- Holds the promise of no library prep and enormously parallel sequencing
- In practice this is not easy to achieve



# Nanopore sequencing

- In practice, it is much harder
- Problems:
  - DNA moves through the pore quickly
  - Holes are difficult/impossible to design to be thin enough so that only one base is physically located within the hole
  - DNA bases are difficult to distinguish from each other without some form of labelling
  - Electrical noise and quantum effects make signal to noise ratios very low
  - Search space for DNA to find a pore is large



# Approaches to simplify nanopore sequencing

- Slow down movement of bases through nanopore
  - Use an enzyme to chop DNA up and sequence individual bases as they pass through a pore
  - And/or use an enzyme to slow the progress of DNA through a pore
  - Monitor capacitative changes in the bilayer
- Hybridize labels to single stranded DNA
  - Force the labels to disassociate as they pass through the pore
  - Detect the labels

# Companies involved



- Company which appears closest to commercialisation
- Two approaches to sequencing
  - Exo-nuclease sequencing (originally part of a co-marketing agreement with Illumina)
  - Strand sequencing
- Both use synthetic membranes compatible with alpha-haemolysin derived pores
- Strand sequencing method is being commercialised

# Nucleotide Recognition

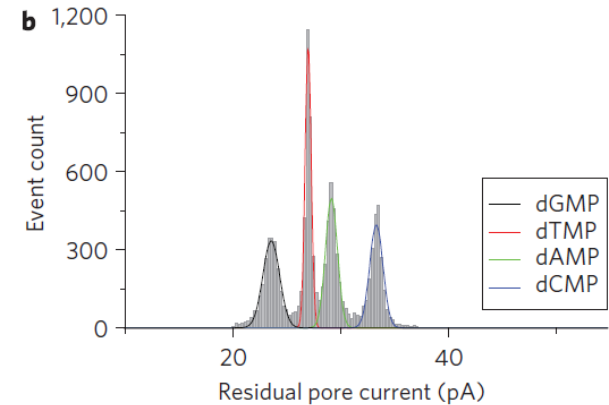
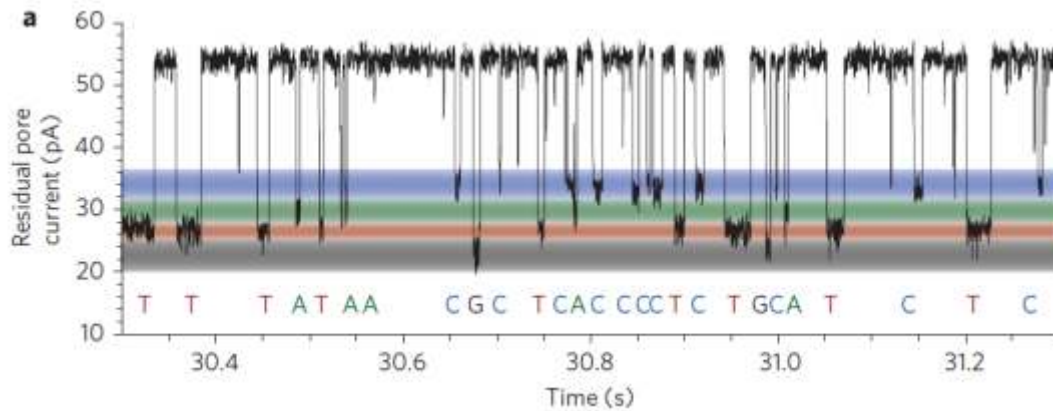
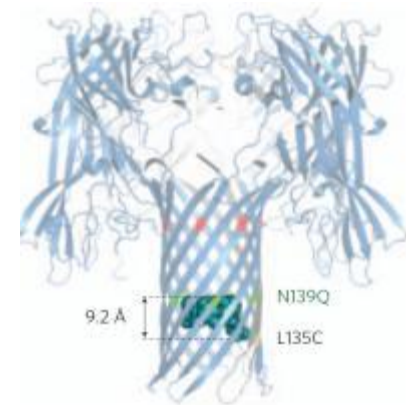
nature  
nanotechnology

ARTICLES

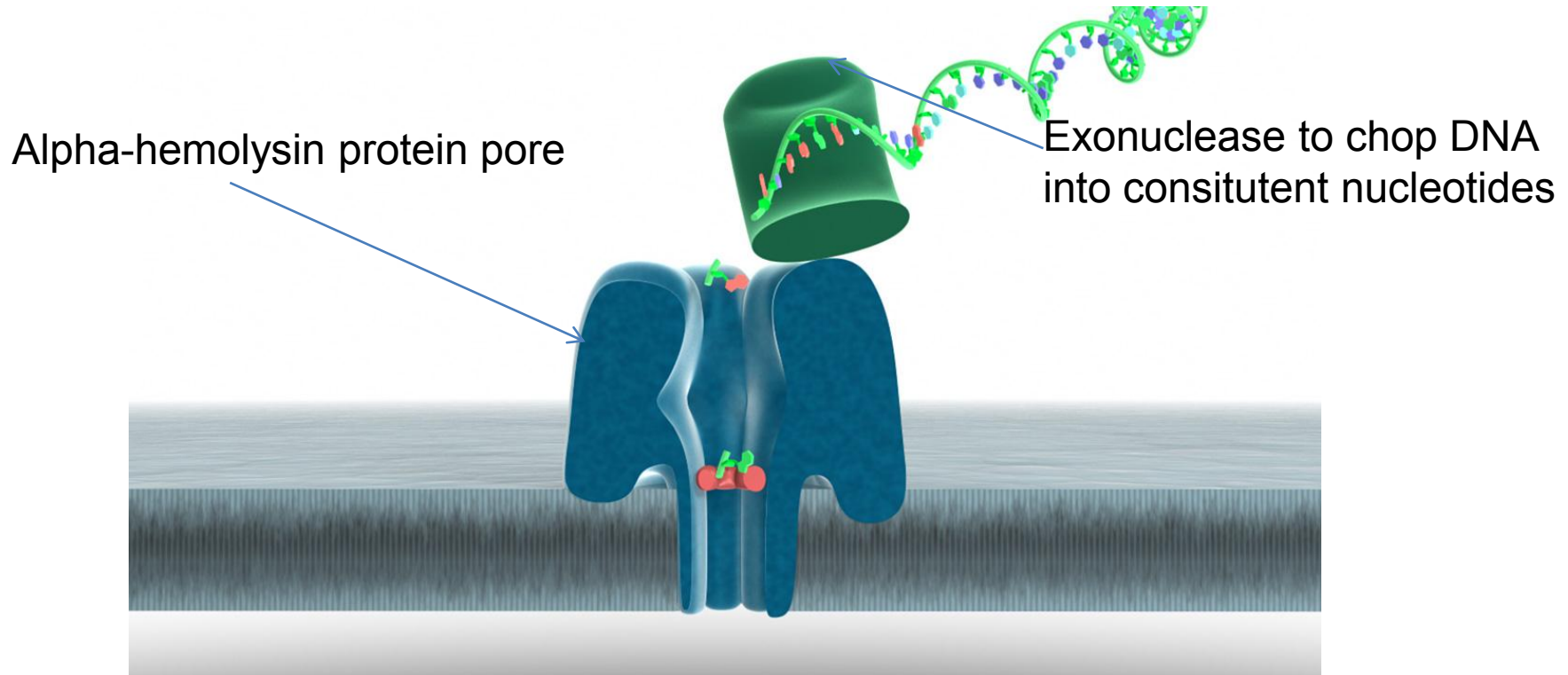
PUBLISHED ONLINE: XX XX 2009 | DOI: 10.1038/NNANO.2009.12

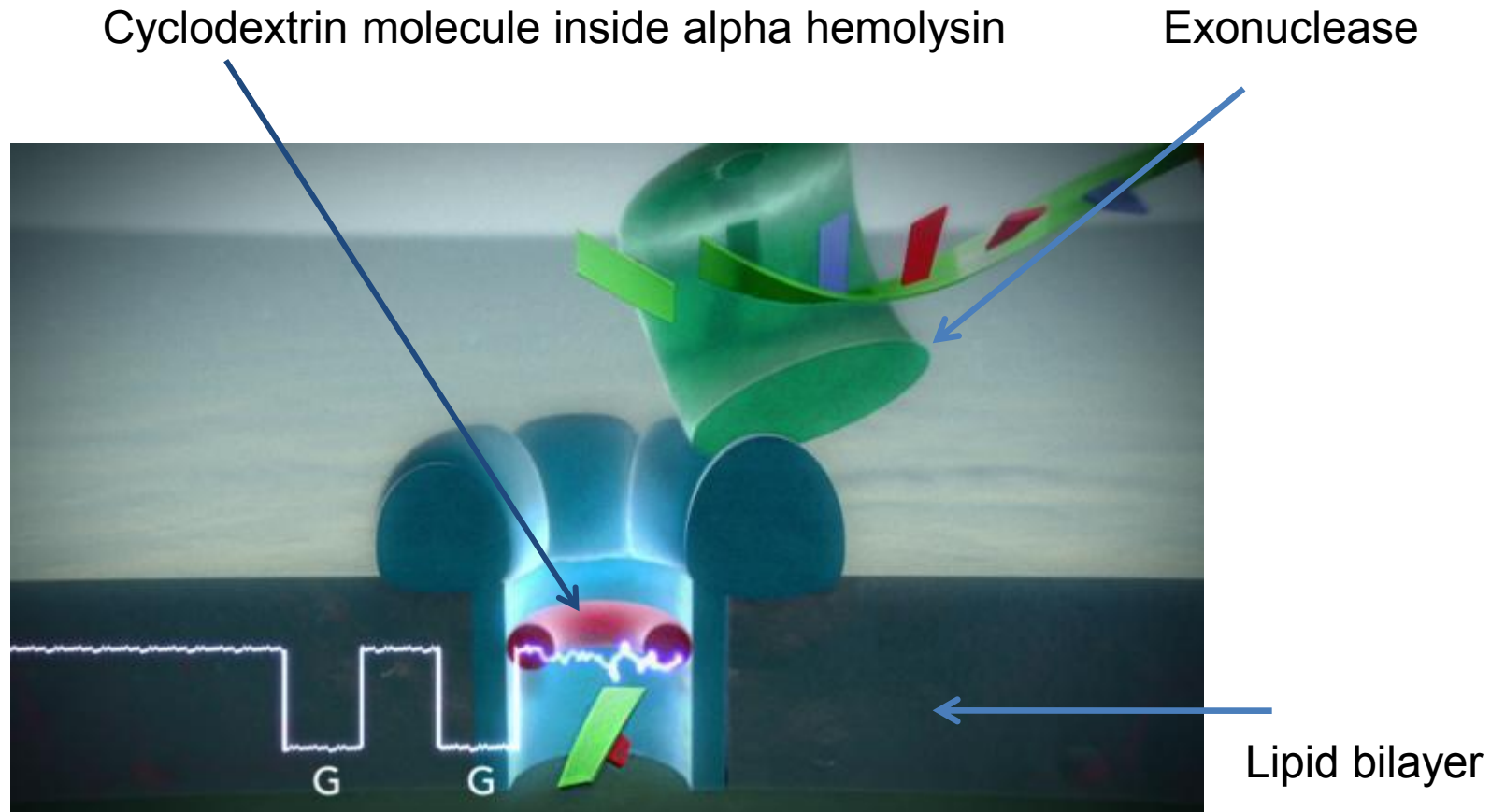
## Continuous base identification for single-molecule nanopore DNA sequencing

James Clarke<sup>1</sup>, Hai-Chen Wu<sup>2</sup>, Lakmal Jayasinghe<sup>1,2</sup>, Alpesh Patel<sup>1</sup>, Stuart Reid<sup>1</sup> and Hagan Bayley<sup>2\*</sup>



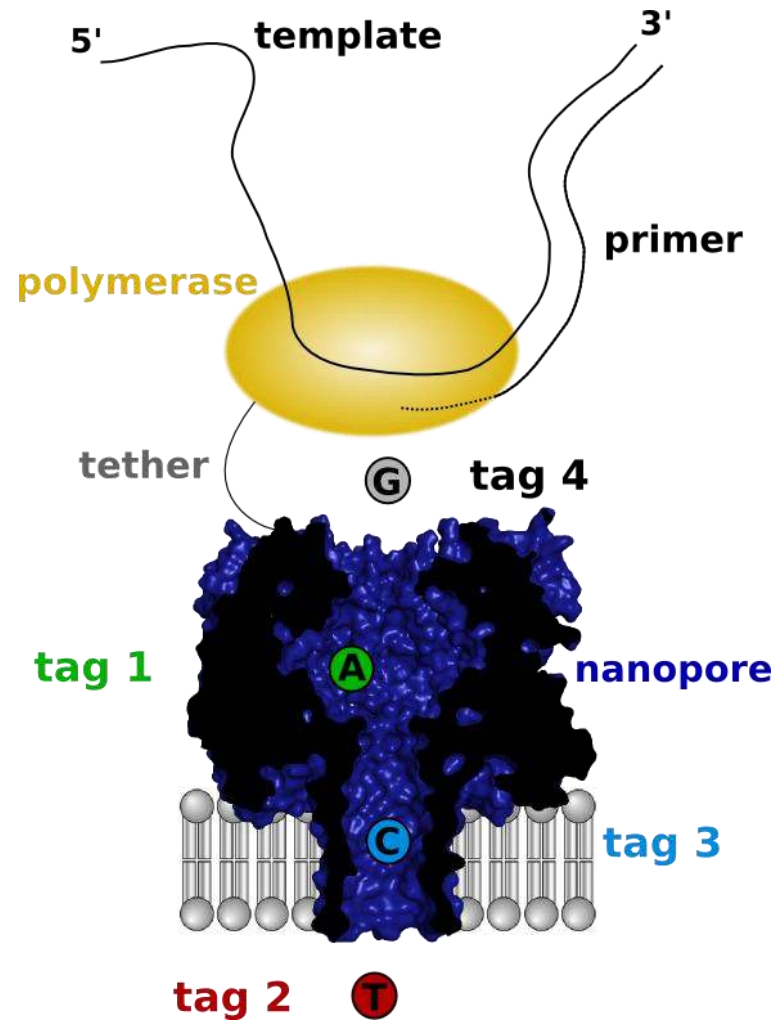
# Exonuclease sequencing





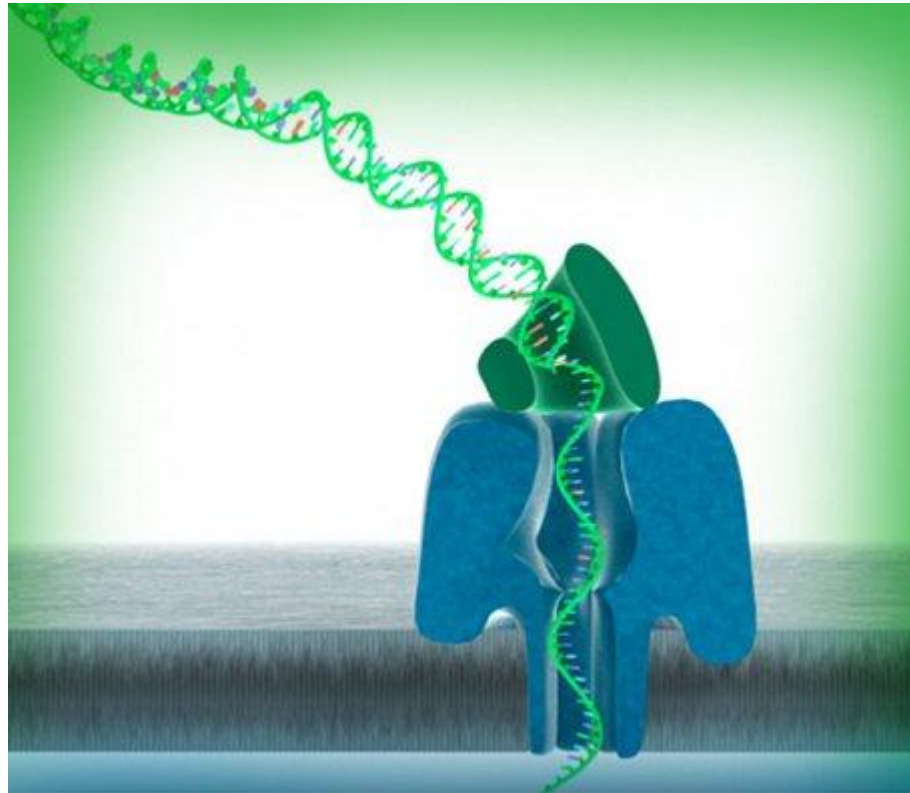
- Cyclodextrin inside alpha-hemolysin transiently binds to DNA base
- Interrupts the current through the pore
- Signal is indicative of base

# Similar to Genia approach



# Strand-sequencing

- Used in the recently advertised GRIDION and Minlon systems

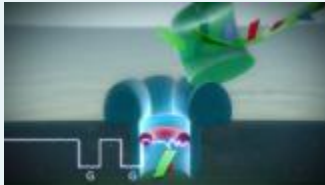


# Novel applications

Application  
Specific

**Adaptable protein nanopore:**

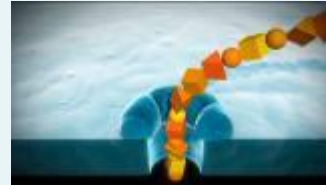
DNA Sequencing



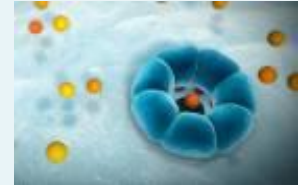
Proteins



Polymers



Small Molecules



Generic Platform

**Sensor array chip: many nanopores in parallel**



**Electronic read-out system**

# Oxford Nanopore Platforms

- GridION for sequencing centres
  - Promise of human genome in a few hours for around \$1000
  - 2000 and 5000 pore instruments
  - No estimated pricing of instrument



- MinION for individual researchers
  - \$900 for 512 pore chip
  - 100Mb-1Gb per MinION
  - Disposable after 6 hour run
  - 4% error rate in trials (mostly deletions)



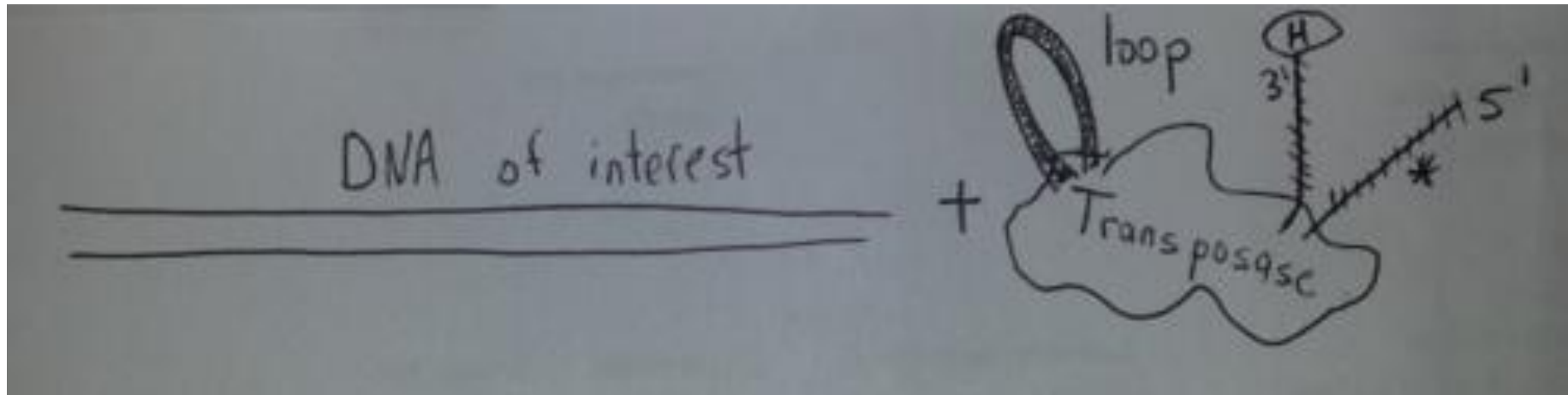
# Minlon Details

- Uses the strand sequencing technique
- Requires library preparation
- ‘Run –until’ technology
- More of a ‘sequencing sensor’ than a direct competitor to 2<sup>nd</sup> generation sequencing
- Very useful for detection
- Likely to become as ubiquitous as a PCR-machine

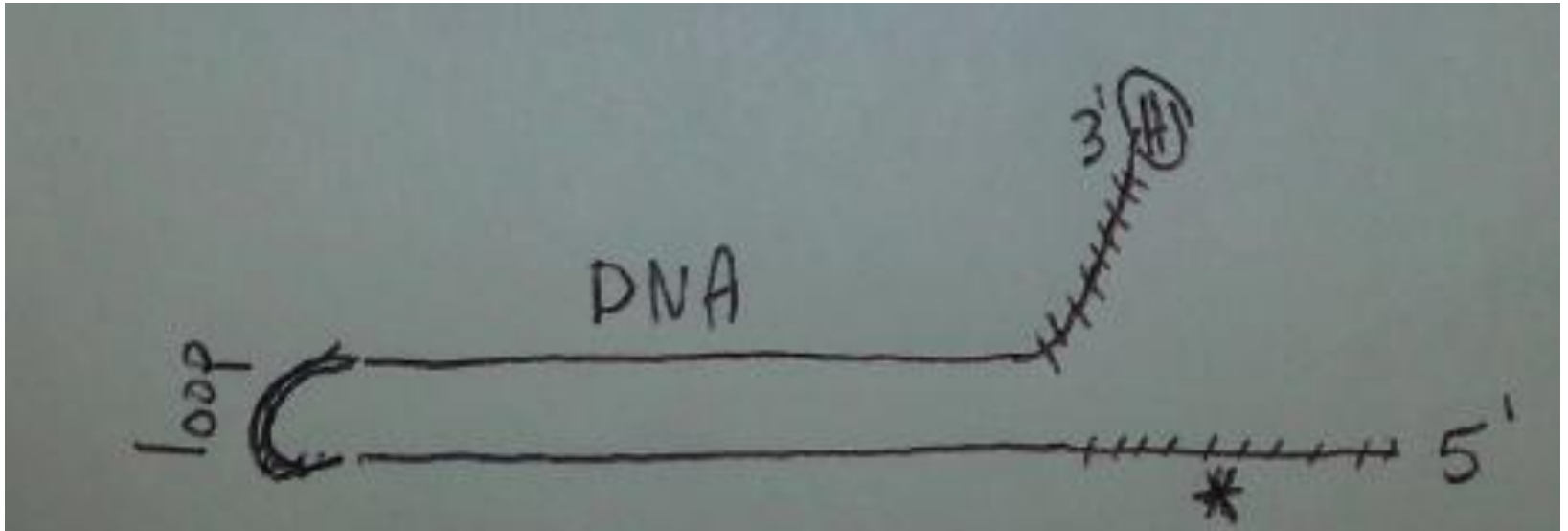
# Library preparation

- Transposase-based library preparation is still required
- Enzymatic biases will still be present *and may be more difficult to detect with lower number of reads*
- Efficiency of transposase may limit maximum read lengths
- Unclear whether system can be washed effectively part-way through a run to load different samples

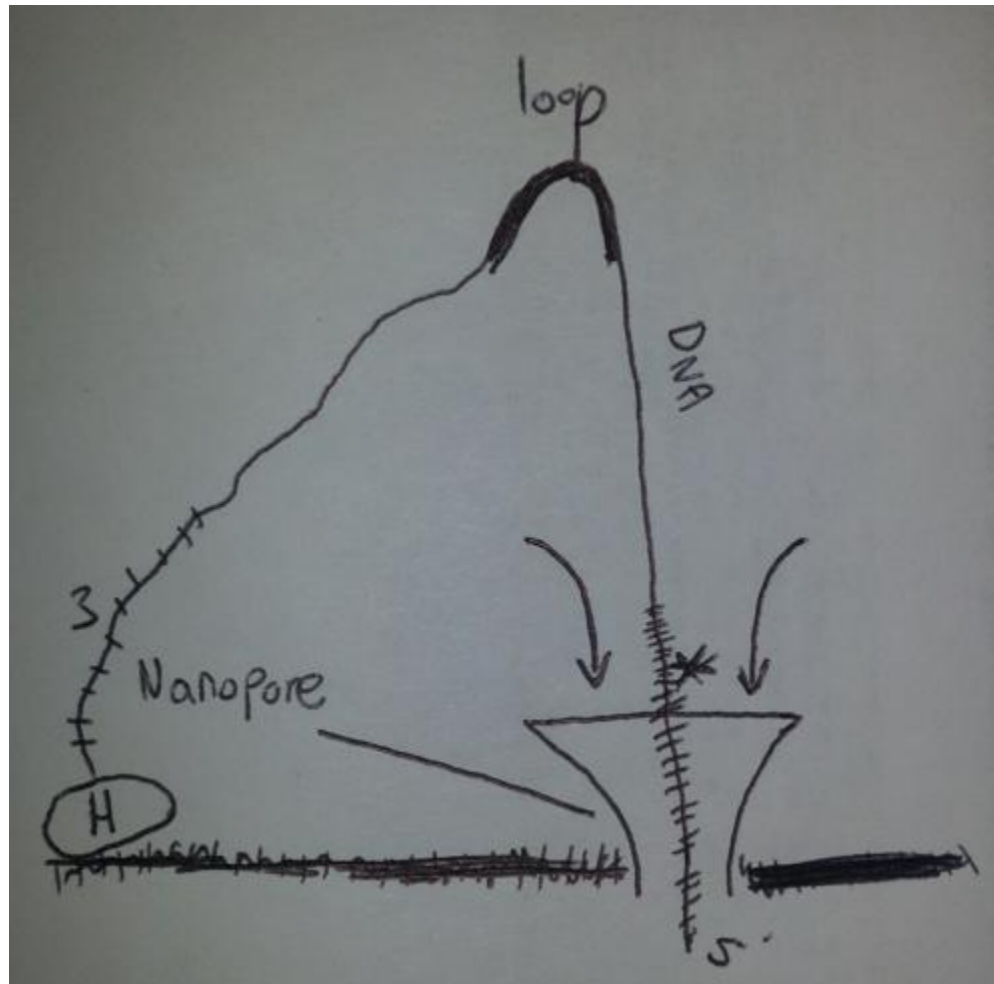
# Library preparation – Step 1



# Library preparation – Step 2

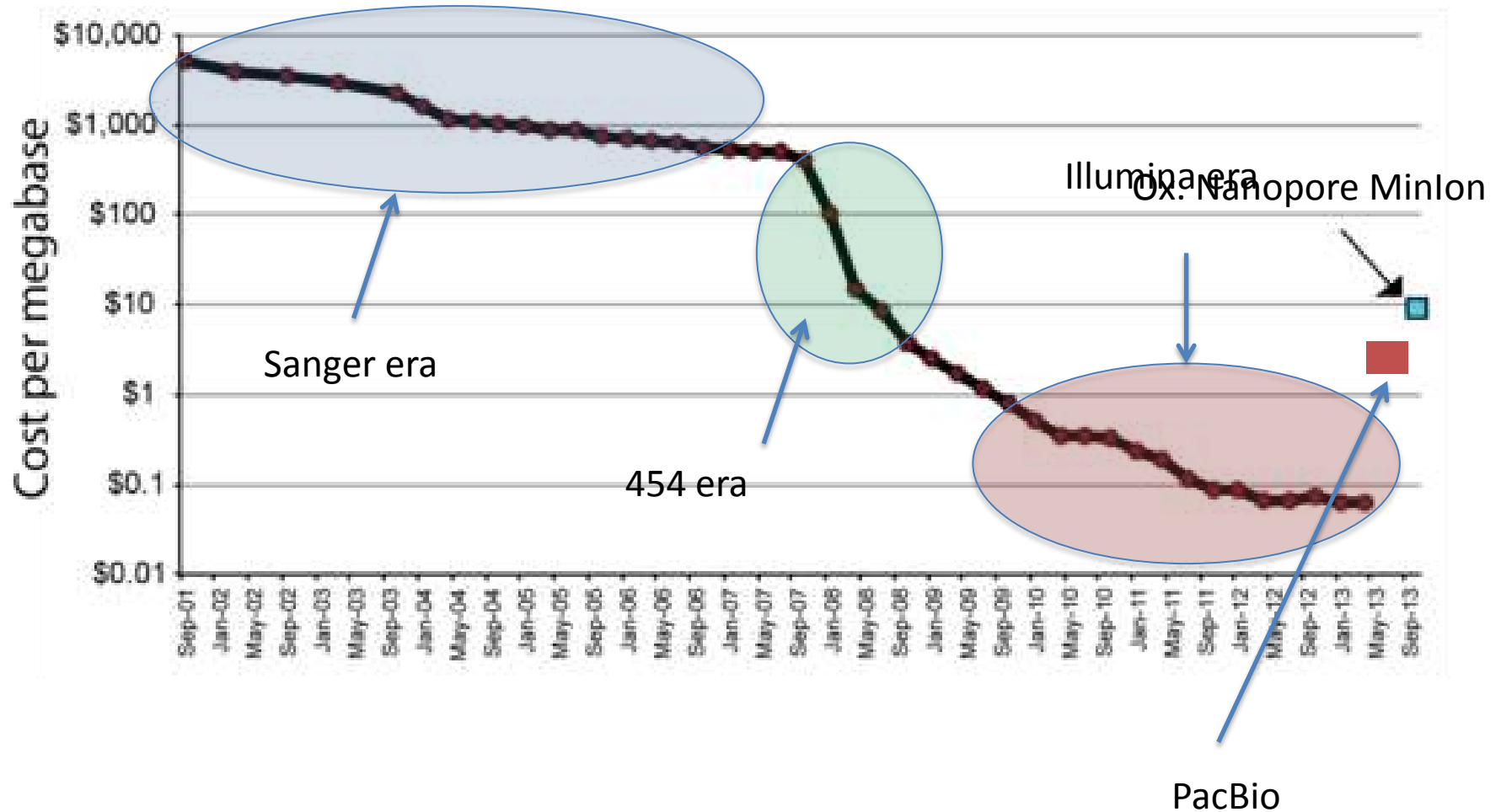


# Library preparation – Step 3



# Caveats

# Cost per megabase



# Oxford Nanopore is not single molecule

- The lipid bi-layers contain different types of nanopore
- Each has a different error profile
- It will still be necessary to over-sample and use sequences determined from complementary nanopores to reduce the overall error rate
- Will still likely need minimum of 5-10x coverage per genome (5-10 bacterial genomes per run)

# Oxford nanopore

- Potential Advantages

- Long reads lengths (10s – 100s kb)
- Protein → solid-state upgrades may eliminate reagent costs (3-5 years)
- Fast turn around
- Could measure epigenetic modifications and other molecules

- Potential Disadvantages

- Potentially non-stochastic errors (i.e. some sequences harder to sequence accurately)
- Library prep required
- Not single molecule
- Cost per base is ~\$10

# Bioinformatics Implications

- Will prove to be yet another step change as with 2<sup>nd</sup> generation sequencing
- Could obtain >100kb fragments
- Denovo assembly and phasing will be made easier
- Low number of reads per run and high per base cost may not make it useful for standard RNA-seq
- Burden will shift even further towards data management and downstream annotation
- ...it will lead to different bottlenecks

# Min Ion Access Programme

## MinION™ Access Programme

In late November 2013, Oxford Nanopore opened registration for a MinION Access Programme (MAP - product preview). This is a substantial but initially controlled programme designed to give life science researchers access to nanopore sequencing technology at no risk and minimal cost.

MAP participants will be at the forefront of applying a completely novel, long-read, real-time sequencing system to existing and new application areas. MAP participants will gain hands-on understanding of the MinION technology, its capabilities and features. They will also play an active role in assessing and developing the system over time. Oxford Nanopore believes that any life science researcher can and should be able to exploit MinION in their own work. Accordingly, Oxford Nanopore is accepting applications for MAP participation from all<sup>1, 2</sup>.

### About the programme

A substantial number of selected participants will receive a MinION Access programme package. This will include:

- At least one complete MinION system (device, flowcells and software tools).
- MAP participants will be asked to pay a refundable \$1,000 deposit on the MinION USB device, plus shipping.
- Oxford Nanopore will provide a regular baseline supply of flowcells sufficient to allow frequent usage of the system. MAP participants will ONLY pay shipping costs on these flowcells. Any additional flowcells required at the participants' discretion may be available for purchase at a MAP-only price of \$999 each plus shipping and taxes.
- Oxford Nanopore will provide Sequencing Preparation Kits. MAP participants may choose to develop their own sample preparation and analysis methods; however, at this stage on an unsupported basis.

<https://www.nanoporetech.com/technology/the-minion-device-a-miniaturised-sensing-system/map-application-form>

# Useful papers/videos

- <http://www.nanoporetech.com/technology/analytes-and-applications-dna-rna-proteins/dna-an-introduction-to-nanopore-sequencing>

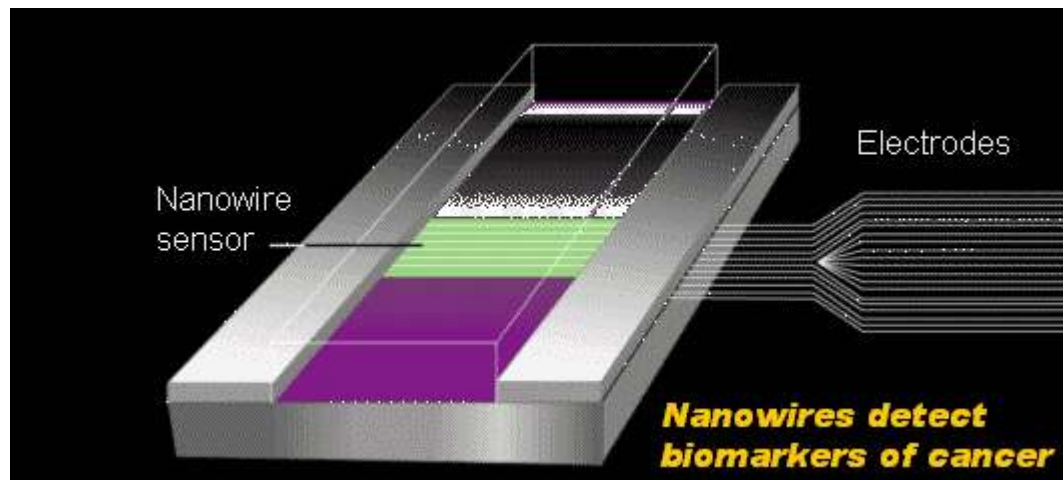
# Beyond nanopores

# General issues with nanopores

- Single base-pair resolution is not available
  - Typically 3-4 nucleotides fit into a nanopore
- Only one detector per DNA strand
- Fast translocation of DNA through pore
- Small signal and high noise
- Unstable lipid bilayers

# Nanowire alternatives

- QuantumDx QSEQ

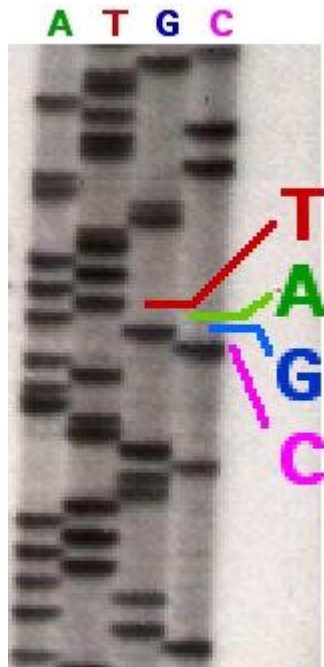


# Many others in development

- <http://www.allseq.com/knowledgebank/sequencing-platforms>

# Sequencing – back on the benchtop

1980



2000



2015?



Thanks to:

Audrey Farbos

Karen Moore

Wellcome Trust

**Contact me:**

[k.h.paszkievicz@exeter.ac.uk](mailto:k.h.paszkievicz@exeter.ac.uk)

Supported by  
**wellcome**trust