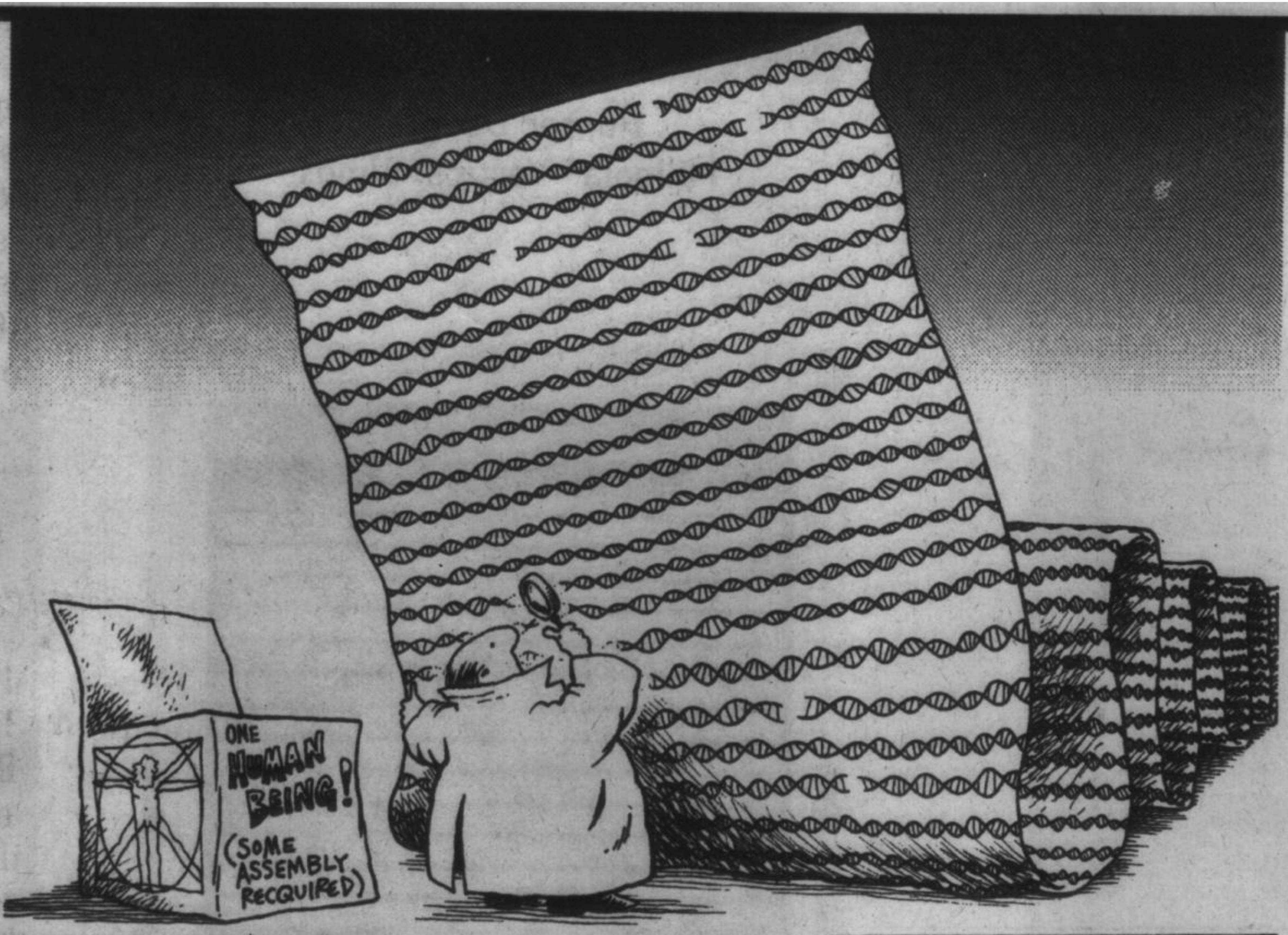


BY AUTH FOR THE PHILADELPHIA INQUIRER

UNIVERSITY OF EXETER

Modern Approaches to Sequencing

Dr Konrad Paszkiewicz, Head, Exeter Sequencing Service,

Wellcome Trust Biomedical Informatics Hub,

January 2013

Contents

- Review of Sanger Sequencing
- Timeline and impact of human genome project
- Second generation sequencing technologies
- Third generation sequencing technologies
- Sequencing – back on the benchtop

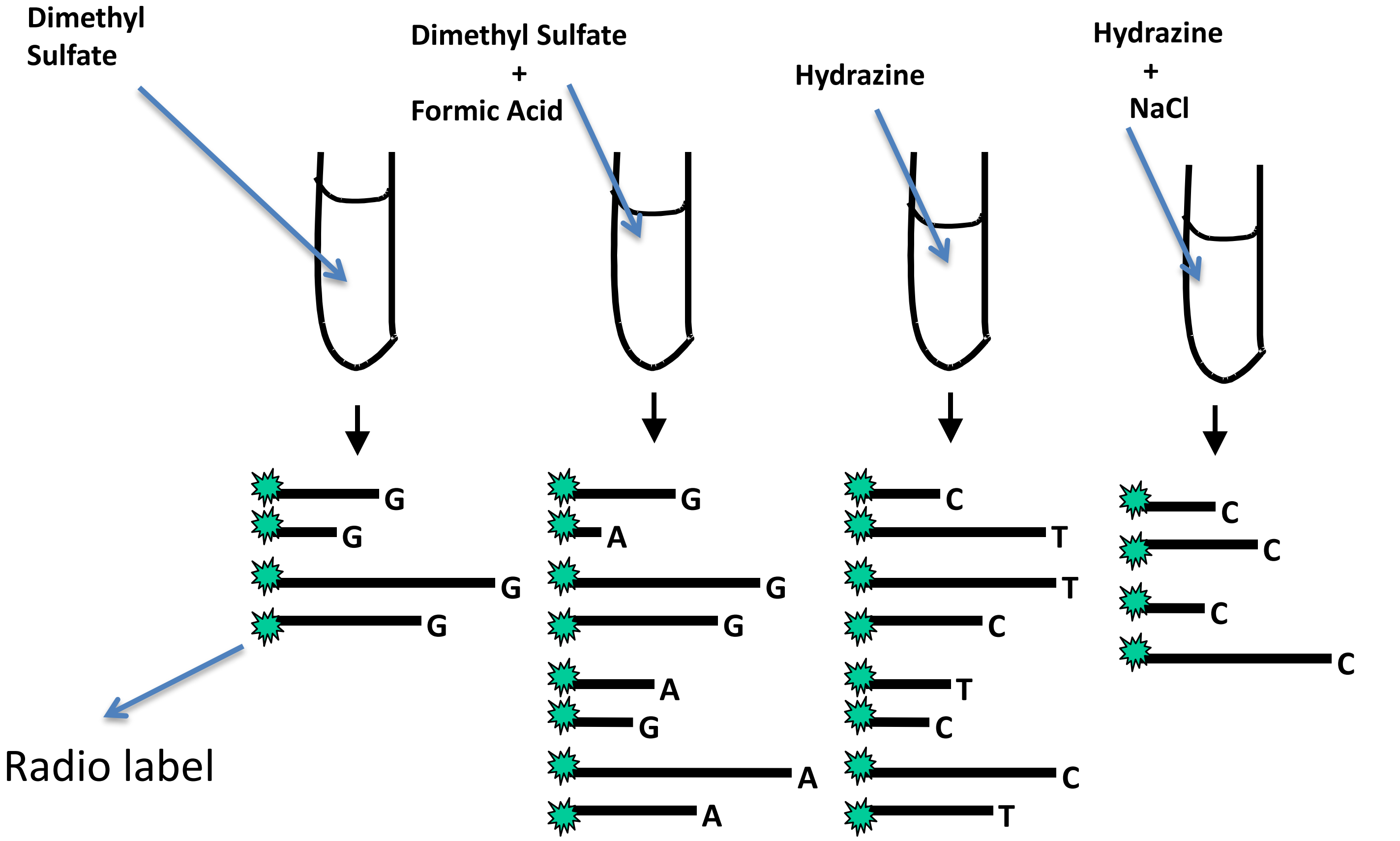
Review of Sanger Sequencing

Dr. Fred Sanger

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

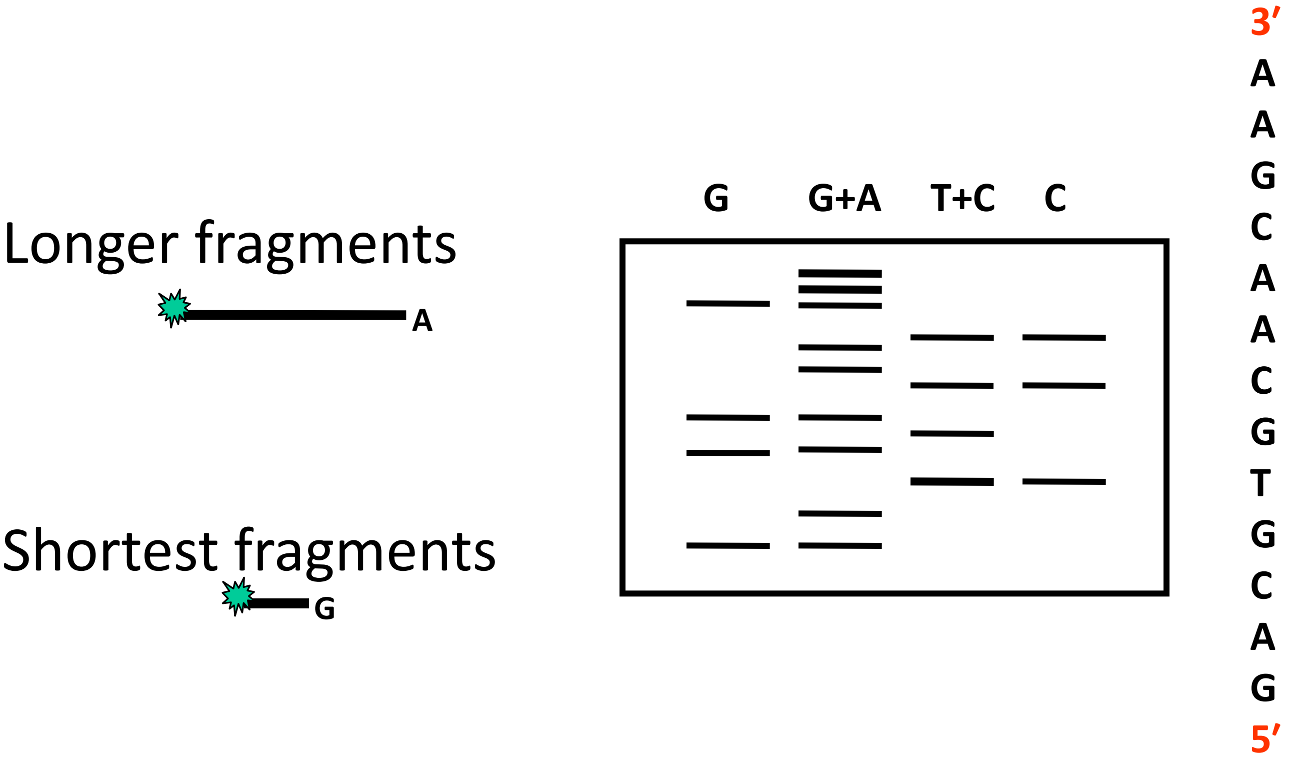
"Fred Sanger is a quiet giant, whose discoveries and inventions transformed our research world.” (A.Bradley, WTSI.)

Maxam-Gilbert Sequencing



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

Maxam-Gilbert Sequencing



Longer fragments

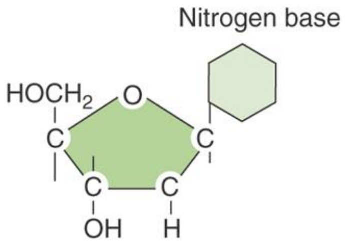
Shortest fragments

3′
A
A
G
C
A
A
C
G
T
G
C
A
G
5′

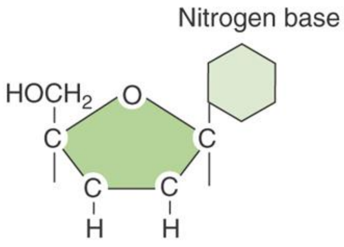
Sequencing gels are read from bottom to top (5′ to 3′).

Chain Termination (Sanger) Sequencing

- A modified DNA replication reaction.
- Growing chains are terminated by dideoxynucleotides.



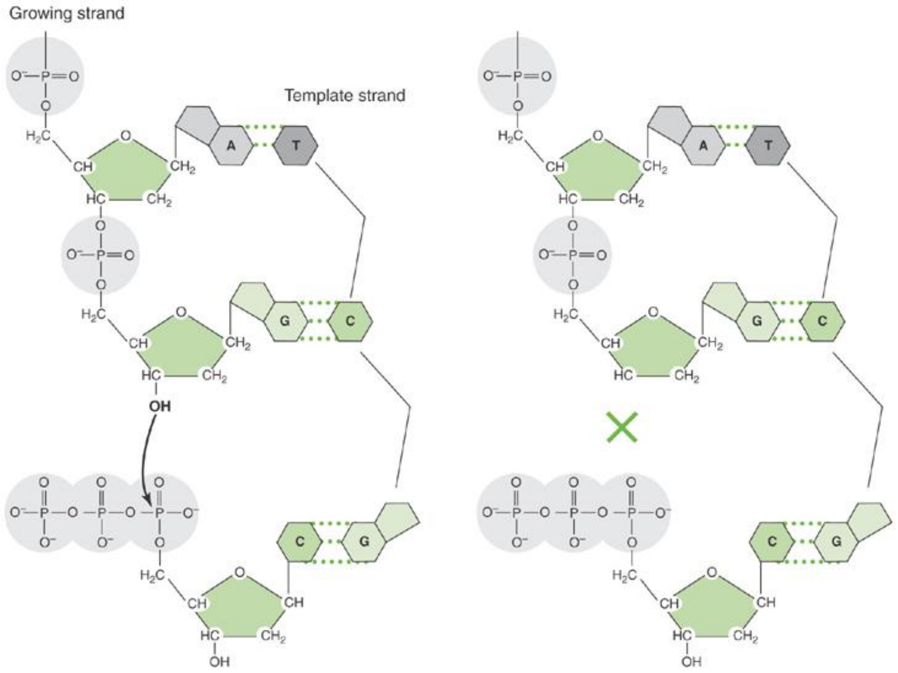
dNTP



ddNTP

Chain Termination (Sanger) Sequencing

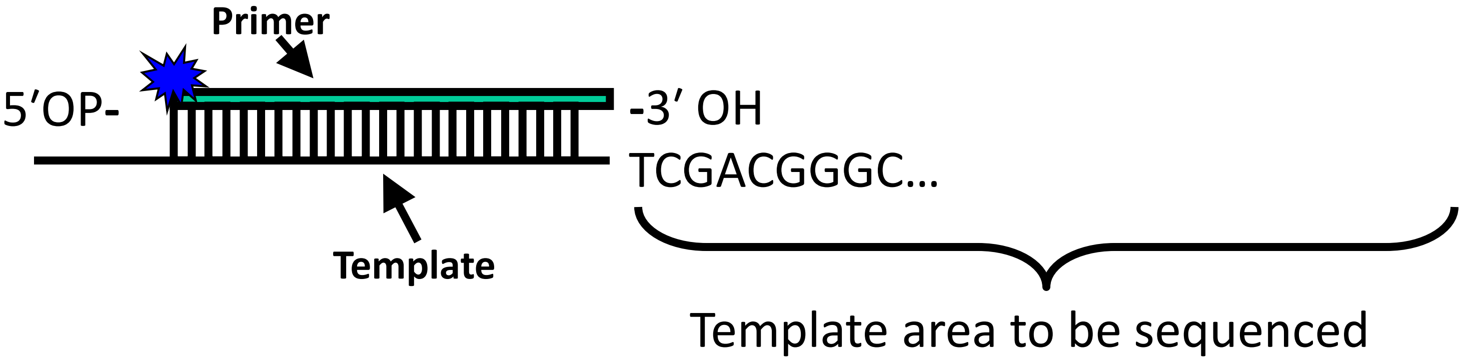
The 3′-OH group necessary for formation of the phosphodiester bond is missing in ddNTPs.



Chain terminates at ddG

Chain Termination (Sanger) Sequencing

- A sequencing reaction mix includes labeled primer and template.



- Dideoxynucleotides are added separately to each of the four tubes.

Chain Termination (Sanger) Sequencing

AGCTGCCCG

A	**ddATP +** **four dNTPs**	**ddA** **dAdGdCdTdGdCdCdCdG**
C	**ddCTP +** **four dNTPs**	**dAdGddC** **dAdGdCdTdGddC** **dAdGdCdTdGdCddC** **dAdGdCdTdGdCdCddC**
G	**ddGTP +** **four dNTPs**	**dAddG** **dAdGdCdTddG** **dAdGdCdTdGdCdCdCddG**
T	**ddTTP +** **four dNTPs**	**dAdGdCddT** **dAdGdCdTdGdCdCdCdG**

Chain Termination (Sanger) Sequencing

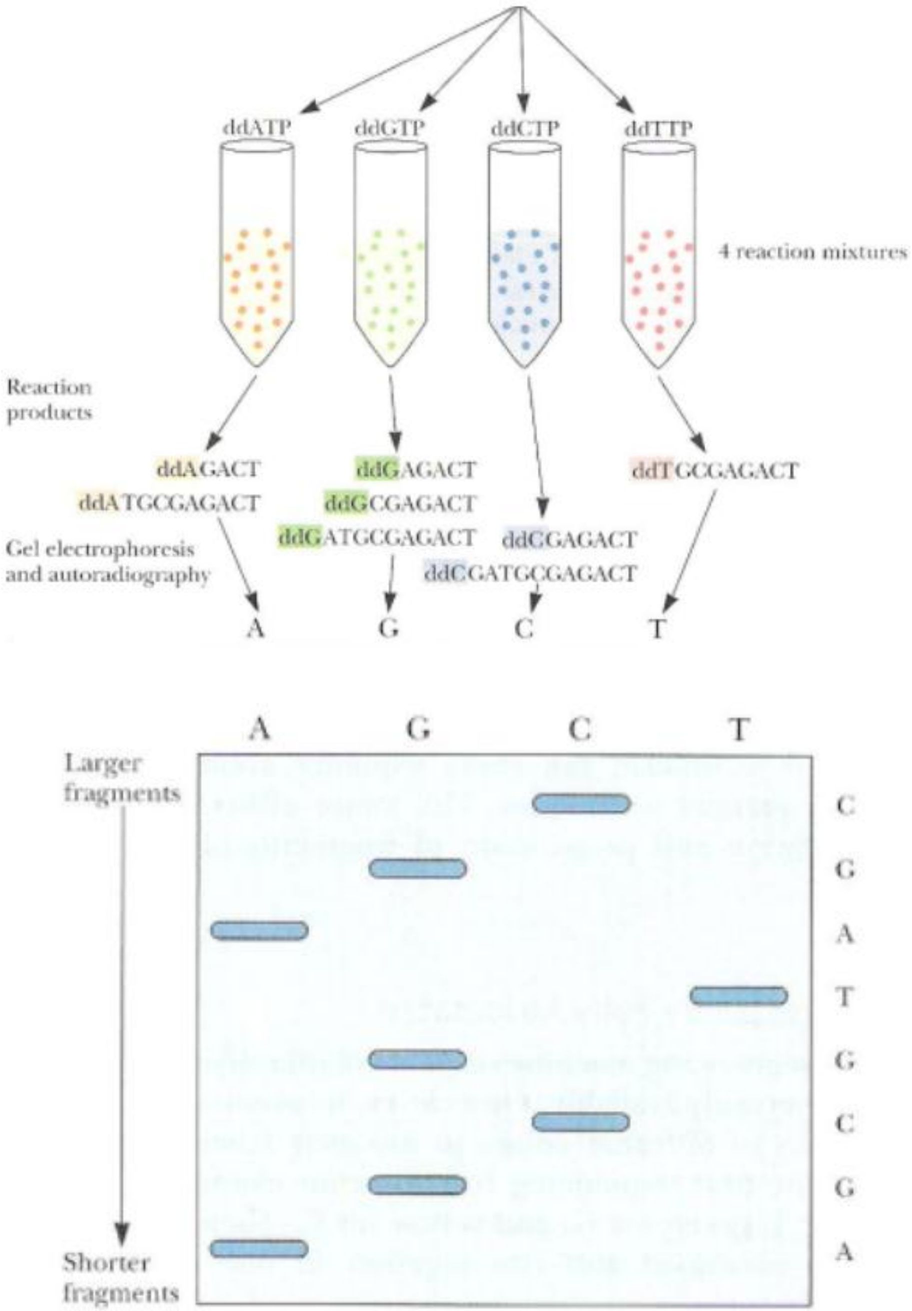
- With addition of enzyme (DNA polymerase), the primer is extended until a ddNTP is encountered.
- The chain will end with the incorporation of the ddNTP.
- With the proper dNTP:ddNTP ratio, the chain will terminate throughout the length of the template.
- All terminated chains will end in the ddNTP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

Chain Termination (Sanger) Sequencing

- The collection of fragments is a sequencing ladder.
- The resulting terminated chains are resolved by electrophoresis.
- Fragments from each of the four tubes are placed in four separate gel lanes.

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



Cycle Sequencing

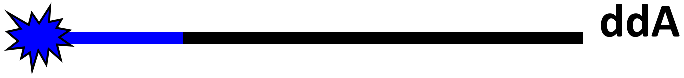
- Cycle sequencing is chain termination sequencing performed in a thermal cycler.
- Cycle sequencing requires a heat-stable DNA polymerase.

Fluorescent Dyes

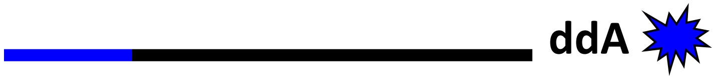
- Fluorescent dyes are multicyclic molecules that absorb and emit fluorescent light at specific wavelengths.
- Examples are fluorescein and rhodamine derivatives.
- For sequencing applications, these molecules can be covalently attached to nucleotides.

Fluorescent Dyes

- In dye primer sequencing, the primer contains fluorescent dye–conjugated nucleotides, labeling the sequencing ladder at the 5′ ends of the chains.

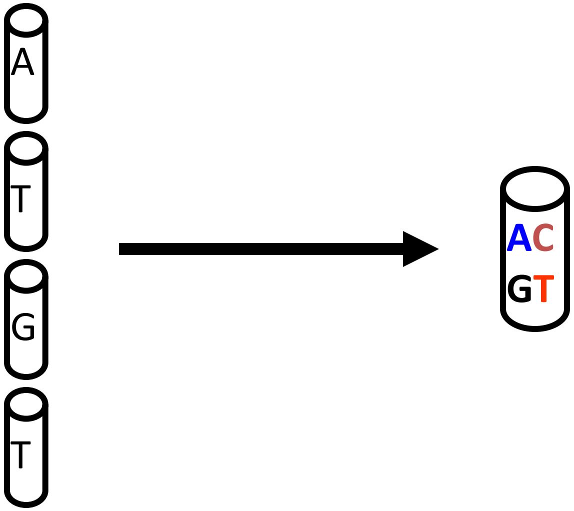


- In dye terminator sequencing, the fluorescent dye molecules are covalently attached to the dideoxynucleotides, labeling the sequencing ladder at the 3′ ends of the chains.



Dye Terminator Sequencing

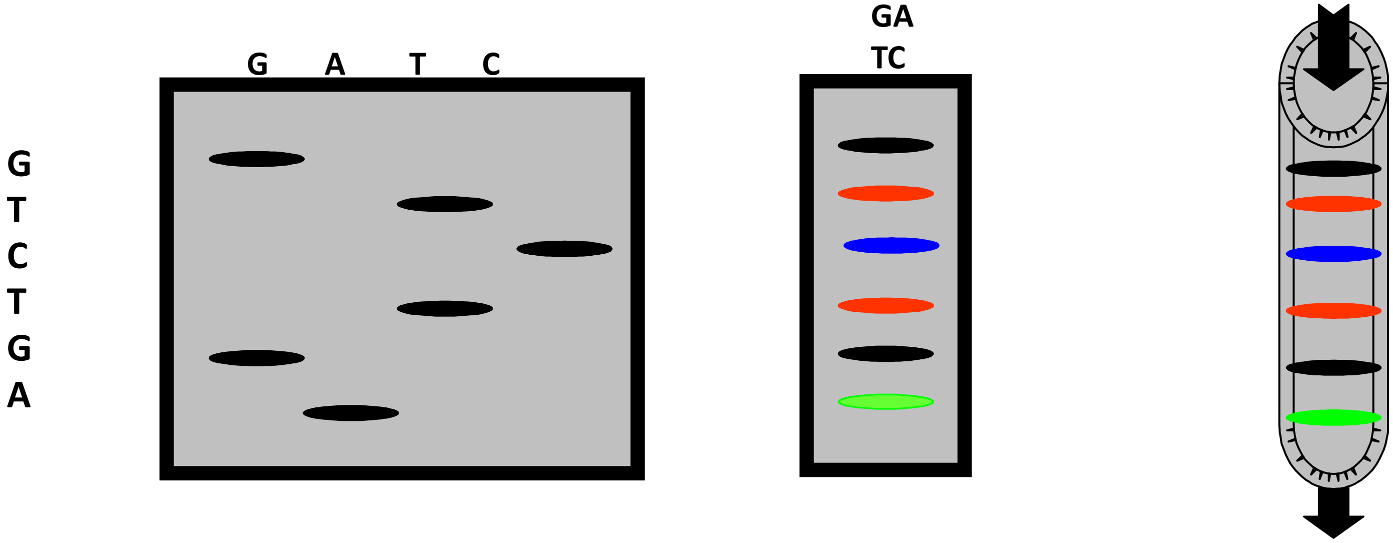
- A distinct dye or “color” is used for each of the four ddNTP.
- Since the terminating nucleotides can be distinguished by color, all four reactions can be performed in a single tube.



The fragments are distinguished by size and “color.”

Dye Terminator Sequencing

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

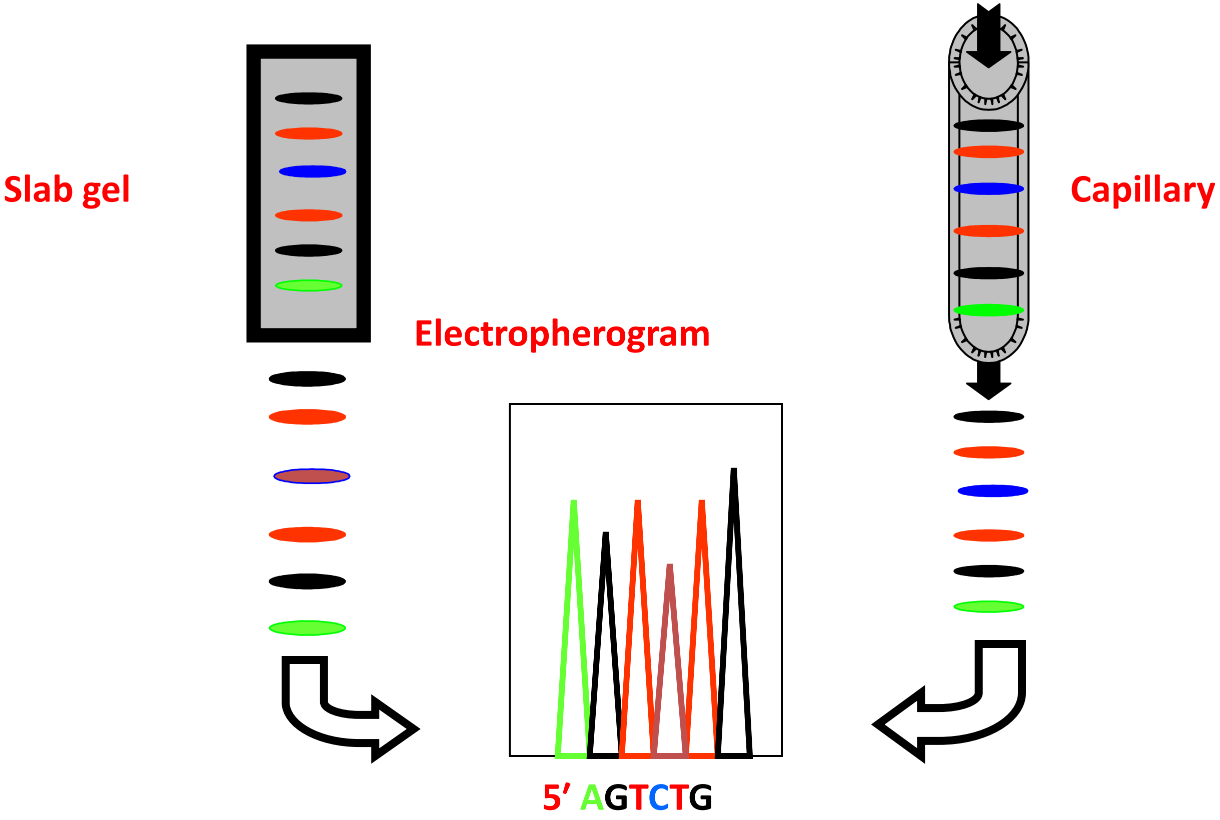


Slab gel

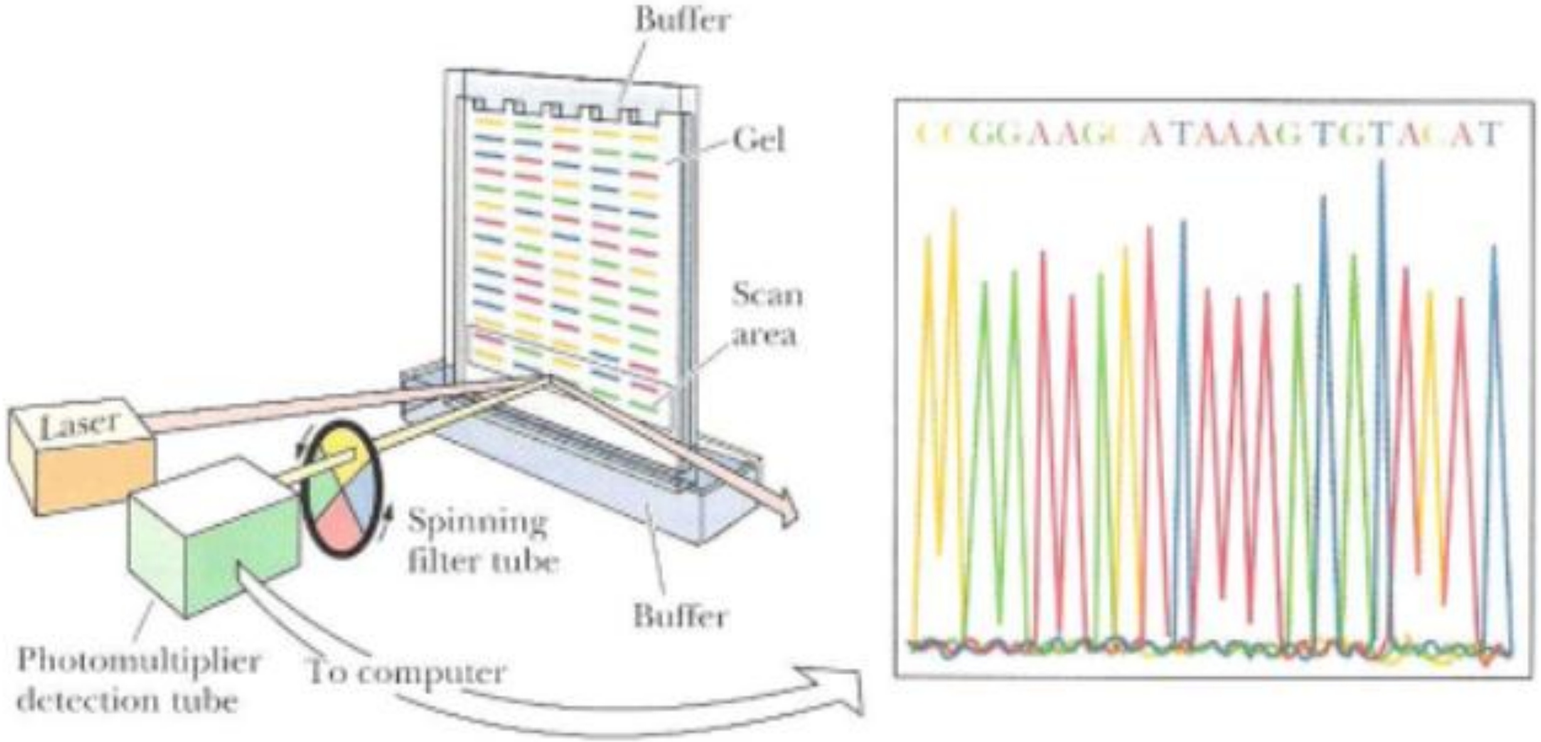
Capillary

Dye Terminator Sequencing

- The DNA ladder is read on an electropherogram.

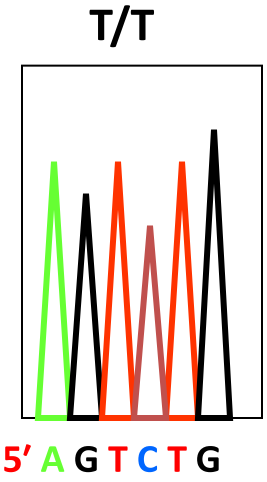


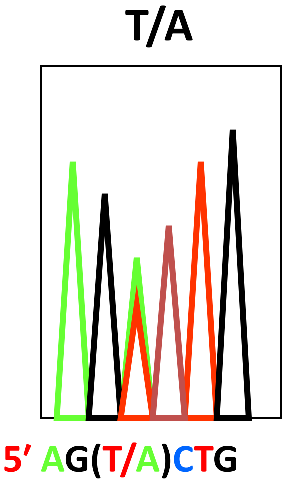
Automated Version of the Dideoxy Method

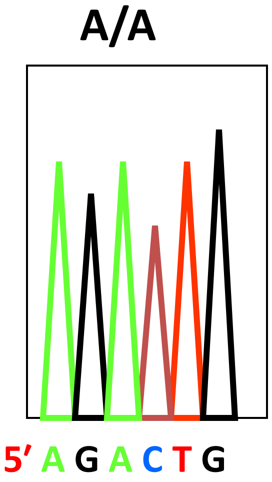


Automated Sequencing

- 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	0.5-1.1kbp
accuracy	high quality bases - 99%: ~900bp very high quality bases - 99.9%: ~600bp 99.999%: 400-500bp
price per raw base	~400k€/Gb

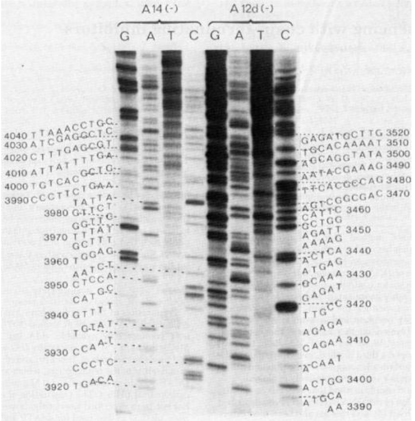
Sanger Sequencing
Useful videos

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

Timeline

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"

http://1.usa.gov/noyyPOhttp://bit.ly/nHn8Tw

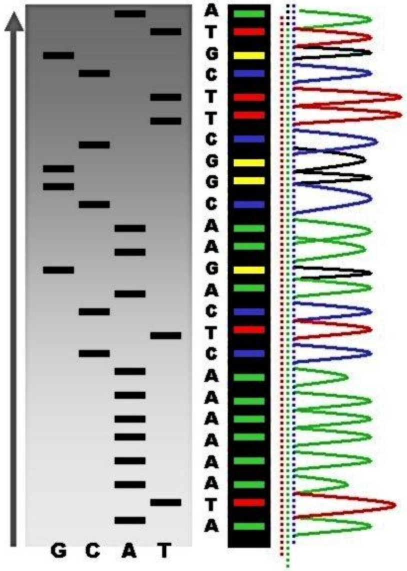
Timeline

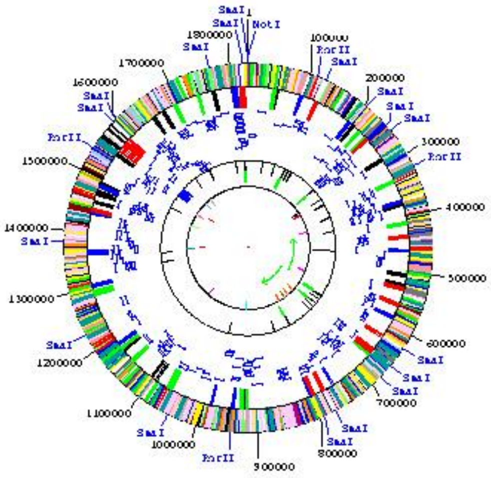
1985-86: Leroy Hood use fluorescently labeled ddNTPs, set the stage for automated sequencing

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

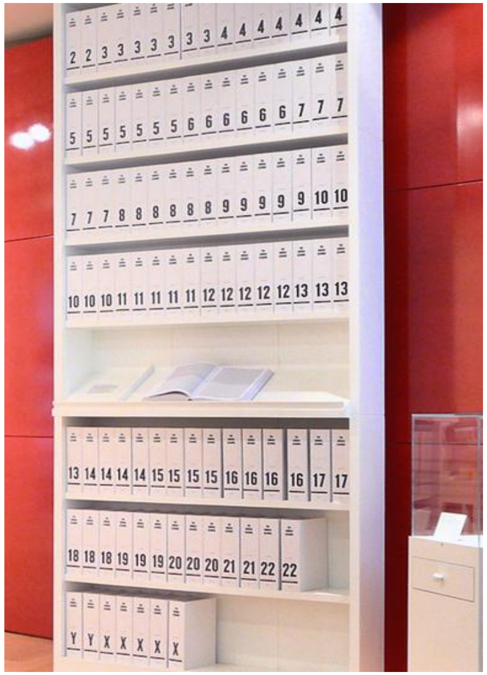




http://bit.ly/2KrFp0 http://bit.ly/qlQD18

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”

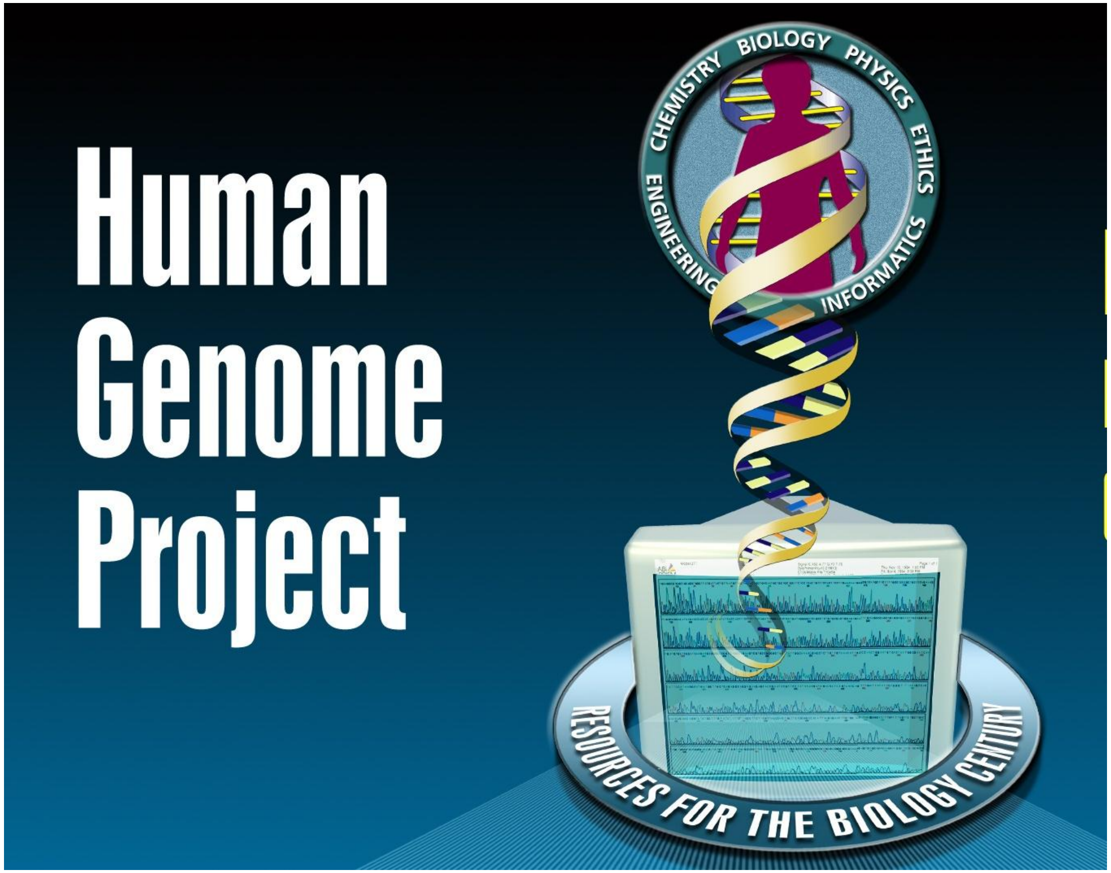
2001: HGP/Celera draft assembly published in Nature/Science

2003: HGP “complete” genome released

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

http://bit.ly/pNKUDJ

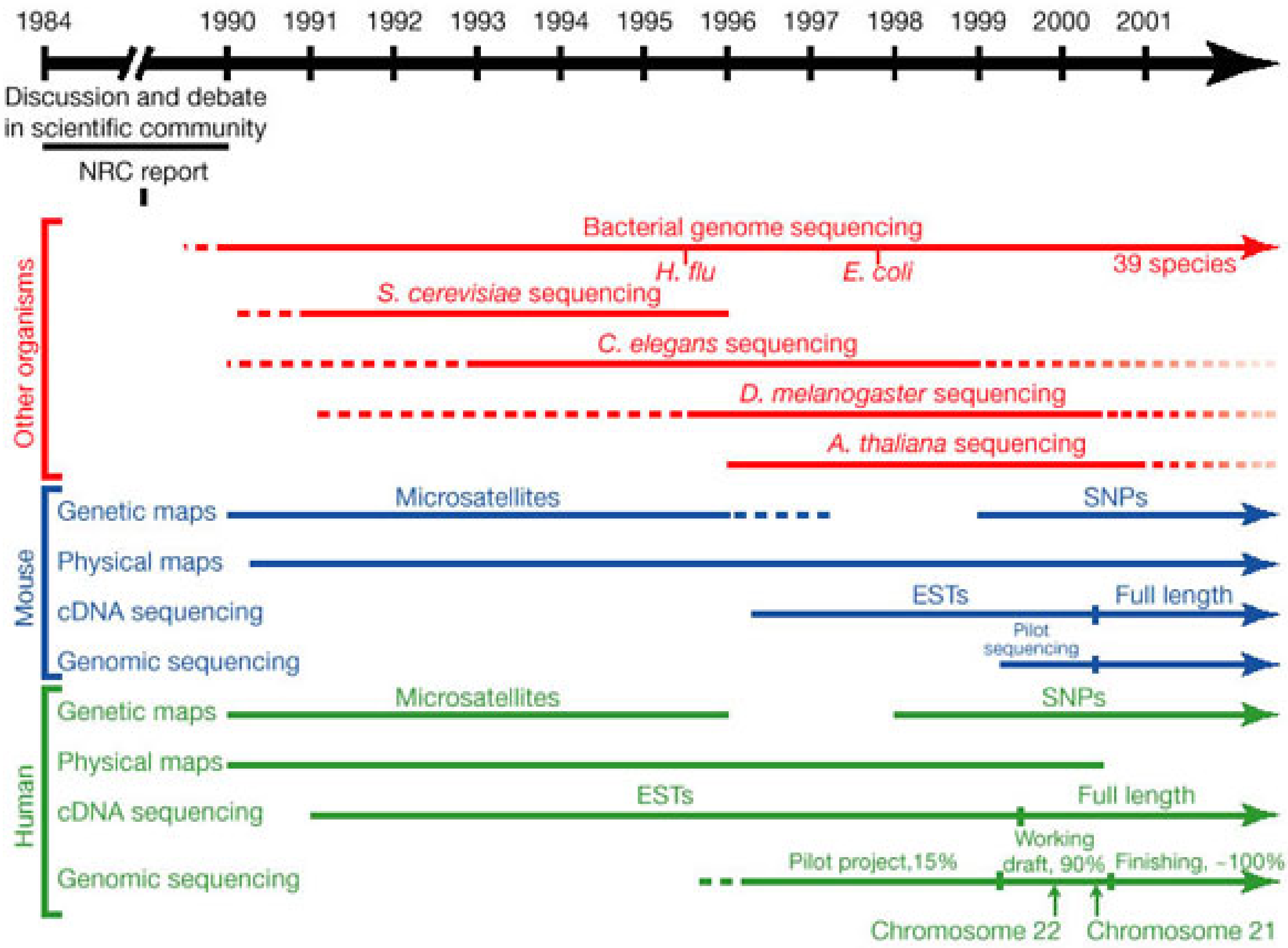
Human genome project



Human Genome Project

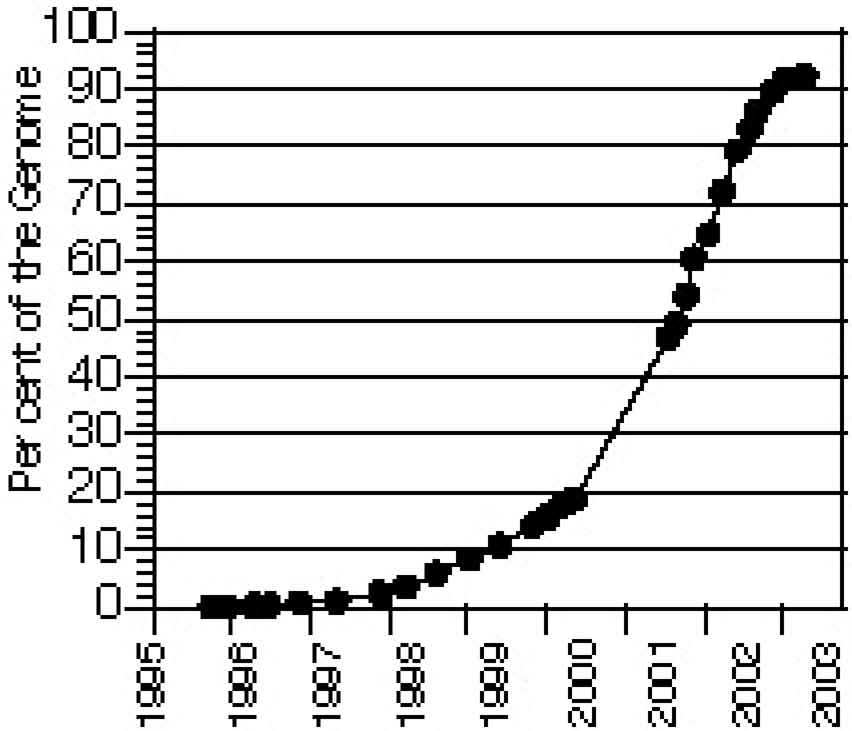
- One of the largest scientific endeavors
  - Target accuracy 1:10,000 bases
  - Started in 1990 by DoE and NIH
  - $3Billion and 15 years
  - Goal was to identify 25K genes and 3 billion bases
- Used the Sanger sequencing method
- Draft assembly done in 2000, complete genome by 2003, last chromosome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

Human Genome Project



http://bit.ly/q3Qsd5

Human Genome Project



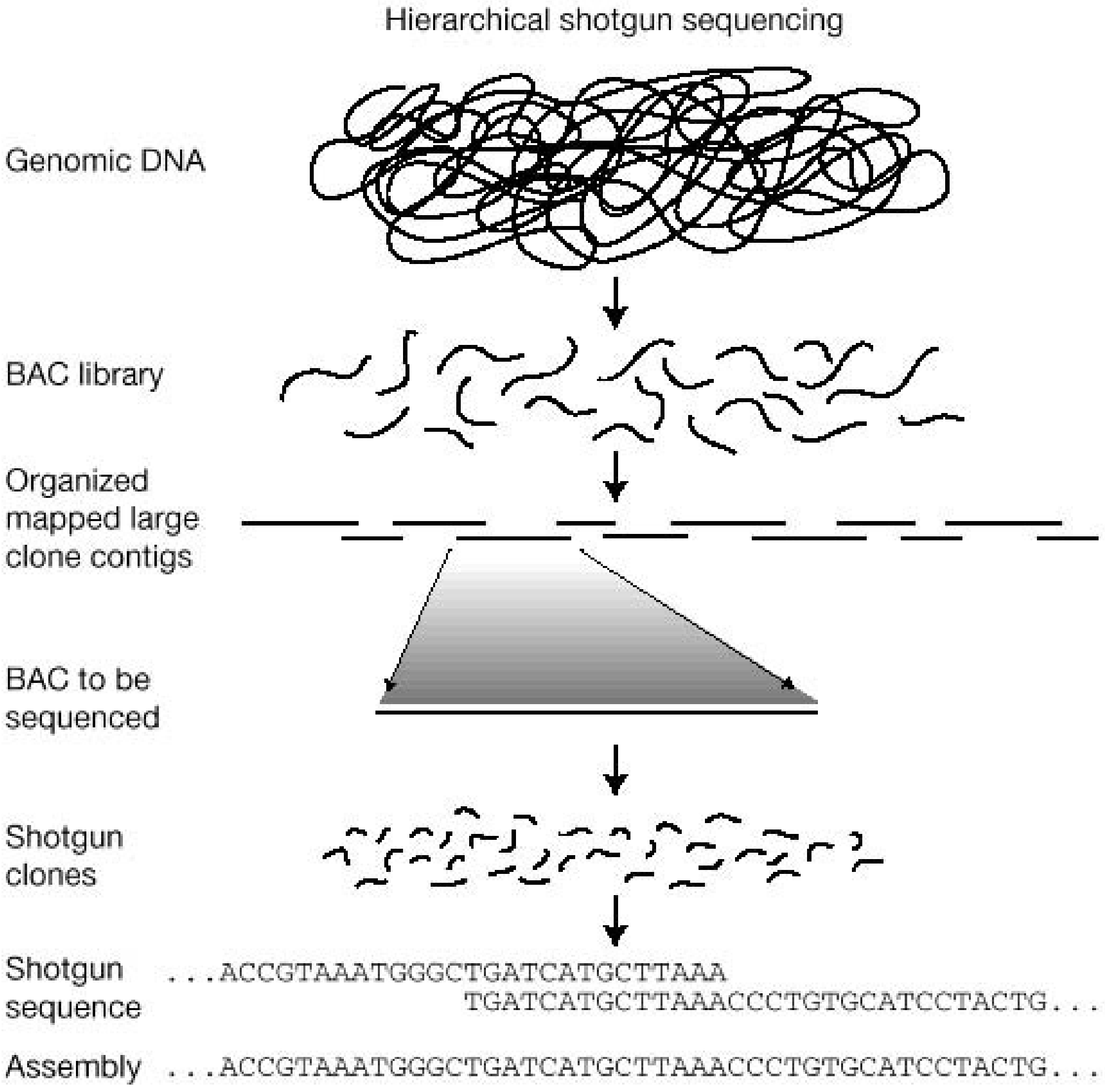
This blog post indicates ~2.86Gbase/3.1Gbase of the non-redundant genome has been sequenced in hg18 or ~**92%** centromeres, telomeres, and highly repetitive regions left

http://bit.ly/qML5Uq

How it was Accomplished

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

Hierarchical Shotgun Sequencing

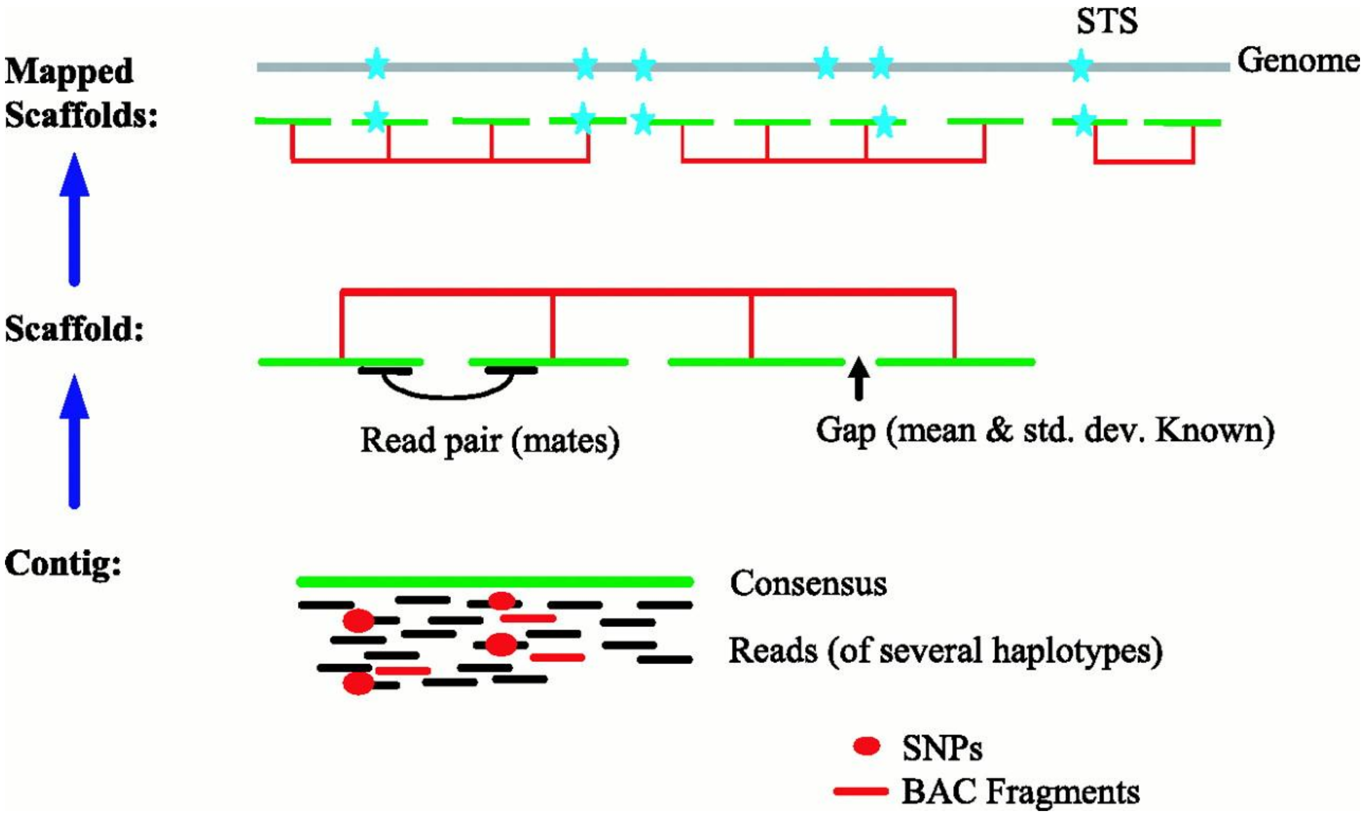

Hierarchical shotgun sequencing
Genomic DNA
BAC library
Organized mapped large clone contigs
BAC to be sequenced
Shotgun clones
Shotgun sequence
...ACCGTAAATGGGCTGATCATGCTTAAA
TGATCATGCTTAAACCCTGTGCATCCTACTG...
Assembly
...ACCGTAAATGGGCTGATCATGCTTAAACCCTGTGCATCCTACTG...

http://bit.ly/qM3Qbk

Shotgun Sequencing

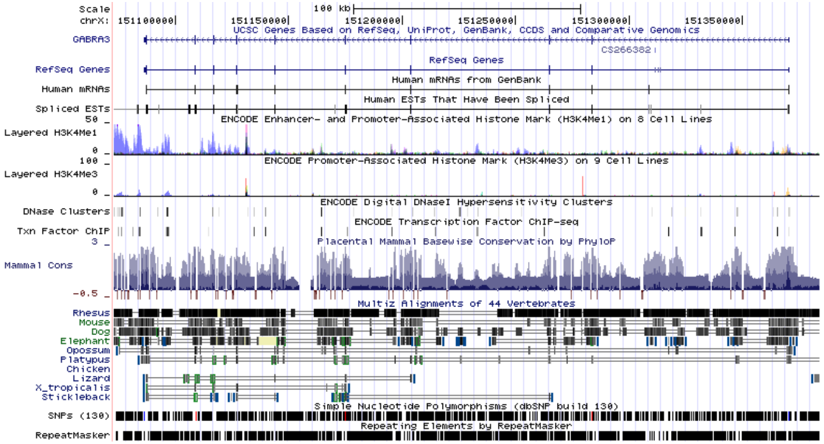
- Celera
  - Started in Sept 1999, goal was to do in $300M and 3 years what the public project was doing for $3B and 15 years!
  - Whole-genome shotgun sequencing
  - Used both whole-genome assembly and regional chromosome assembly
  - Incorporated data from the public project
  - Raised ethical concerns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

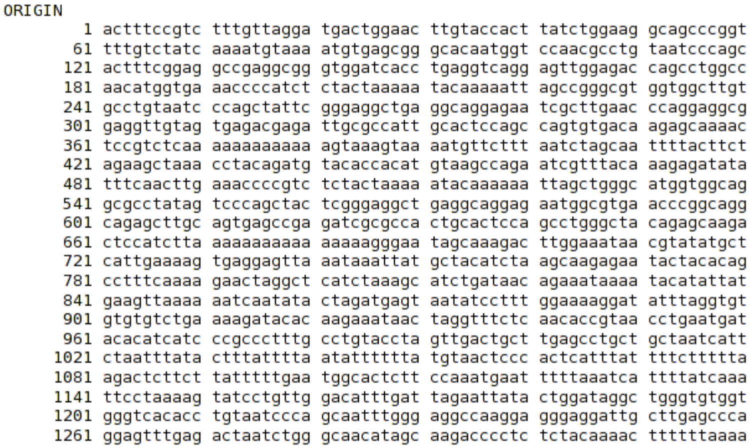
Celera Shotgun Sequencing



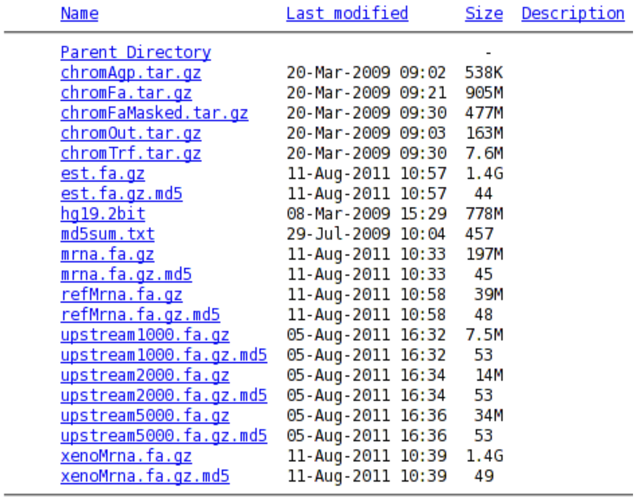
- Used paired-end strategy with variable insert size: 2, 10, and 50kbp

HGP Data Access

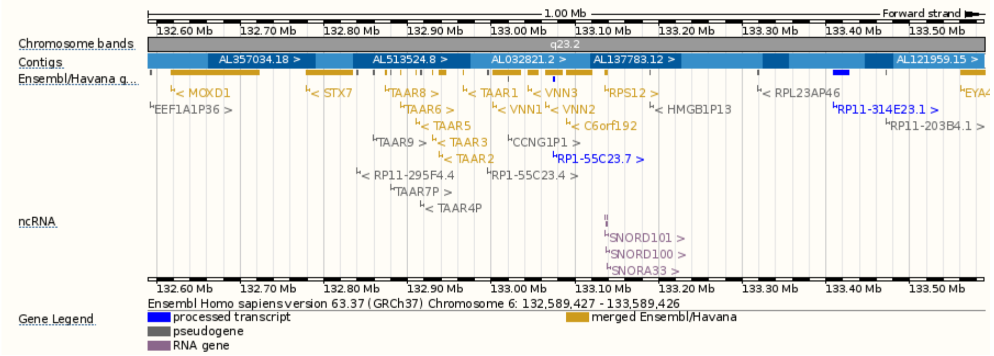

Scale
100 kb
chrX:
151100000
151150000
151200000
151250000
151300000
151350000
UCSC Genes Based on RefSeq, UniProt, GenBank, CCDS and Comparative Genomics
GABRA3
CS266382
RefSeq Genes
RefSeq Genes
Human mRNAs from GenBank
Human mRNAs
Human ESTs That Have Been Spliced
Spliced ESTs
50 _
ENCODE Enhancer- and Promoter-Associated Histone Mark (H3K4Me1) on 8 Cell Lines
Layered H3K4Me1
0 _
100 _
ENCODE Promoter-Associated Histone Mark (H3K4Me3) on 9 Cell Lines
Layered H3K4Me3
0 _
ENCODE Digital DNaseI Hypersensitivity Clusters
DNase Clusters
ENCODE Transcription Factor ChIP-seq
Txn Factor ChIP
3 _
Placental Mammal Basewise Conservation by PhyloP
Mammal Cons
-0.5 _
Multiz Alignments of 44 Vertebrates
Rhesus
Mouse
Dog
Elephant
Opossum
Platypus
Chicken
Lizard
X_tropicalis
Stickleback
Simple Nucleotide Polymorphisms (dbSNP build 130)
SNPs (130)
Repeating Elements by RepeatMasker
RepeatMasker



```
ORIGIN
        1 actttccgtc tttgttagga tgactggaac ttgtaccact tatctggaag gcagcccggt
       61 tttgtctatc aaaatgtaaa atgtgagcgg gcacaatggt ccaacgcctg taatcccagc
      121 actttcggag gccgaggcgg gtggatcacc tgaggtcagg agttggagac cagcctggcc
      181 aacatggtga aaccccatct ctactaaaaa tacaaaaatt agccgggcgt ggtggcttgt
      241 gcctgtaatc ccagctattc gggaggctga ggcaggagaa tcgcttgaac ccaggaggcg
      301 gaggttgtag tgagacgaga ttgcgccatt gcactccagc cagtgtgaca agagcaaaac
      361 tccgtctcaa aaaaaaaaaa agtaaagtaa aatgttcttt aatctagcaa ttttacttct
      421 agaagctaaa cctacagatg tacaccacat gtaagccaga atcgtttaca aagagatata
      481 tttcaacttg aaaccccgtc tctactaaaa atacaaaaaa ttagctgggc atggtggcag
      541 gcgcctatag tcccagctac tcgggaggct gaggcaggag aatggcgtga acccggcagg
      601 cagagcttgc agtgagccga gatcgcgcca ctgcactcca gcctgggcta cagagcaaga
      661 ctccatctta aaaaaaaaaa aaaaagggaa tagcaaagac ttggaaataa cgtatatgct
      721 cattgaaaag tgaggagtta aataaattat gctacatcta agcaagagaa tactacacag
      781 cctttcaaaa gaactaggct catctaaagc atctgataac agaaataaaa tacatattat
      841 gaagttaaaa aatcaatata ctagatgagt aatatccttt ggaaaaggat atttaggtgt
      901 gtgtgtctga aaagatacac aagaaataac taggtttctc aacaccgtaa cctgaatgat
      961 acacatcatc ccgccctttg cctgtaccta gttgactgct tgagcctgct gctaatcatt
     1021 ctaatttata ctttatttta atatttttta tgtaactccc actcatttat tttcttttta
     1081 agactcttct tatttttgaa tggcactctt ccaaatgaat ttttaaatca ttttatcaaa
     1141 ttcctaaaag tatcctgttg gacatttgat tagaattata ctggataggc tgggtgtggt
     1201 gggtcacacc tgtaatccca gcaatttggg aggccaagga gggaggattg cttgagccca
     1261 ggagtttgag actaatctgg gcaacatagc aagacccctc tctacaaaac tttttaaaa
```

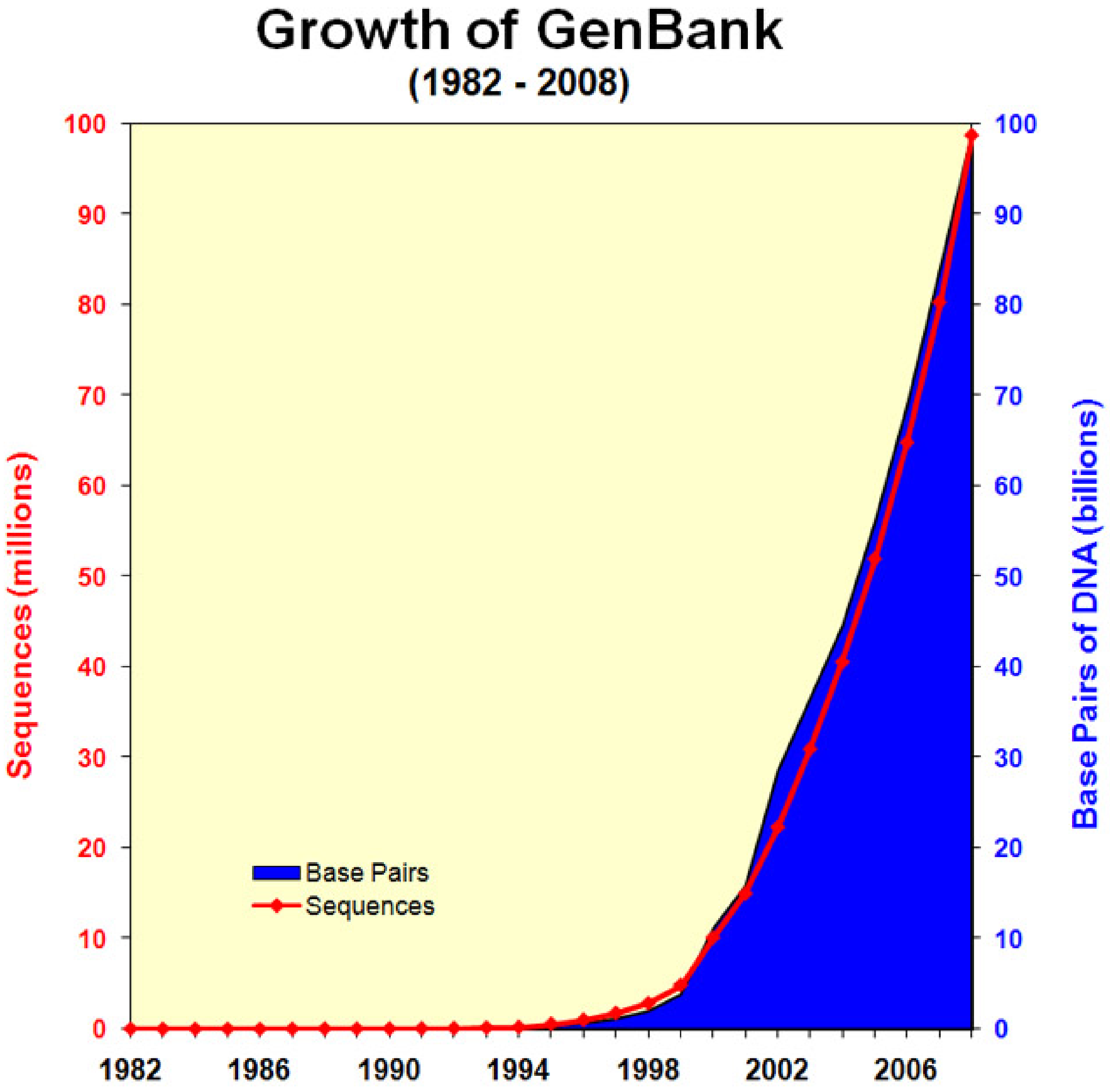


Name	Last modified	Size	Description
Parent Directory		-	
chromAgp.tar.gz	20-Mar-2009 09:02	538K	
chromFa.tar.gz	20-Mar-2009 09:21	905M	
chromFaMasked.tar.gz	20-Mar-2009 09:30	477M	
chromOut.tar.gz	20-Mar-2009 09:03	163M	
chromTrf.tar.gz	20-Mar-2009 09:30	7.6M	
est.fa.gz	11-Aug-2011 10:57	1.4G	
est.fa.gz.md5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



Ensembl Homo sapiens version 63.37 (GRCh37) Chromosome 6: 132,589,427 - 133,589,426

Results in GenBank, UCSC, Ensembl & others



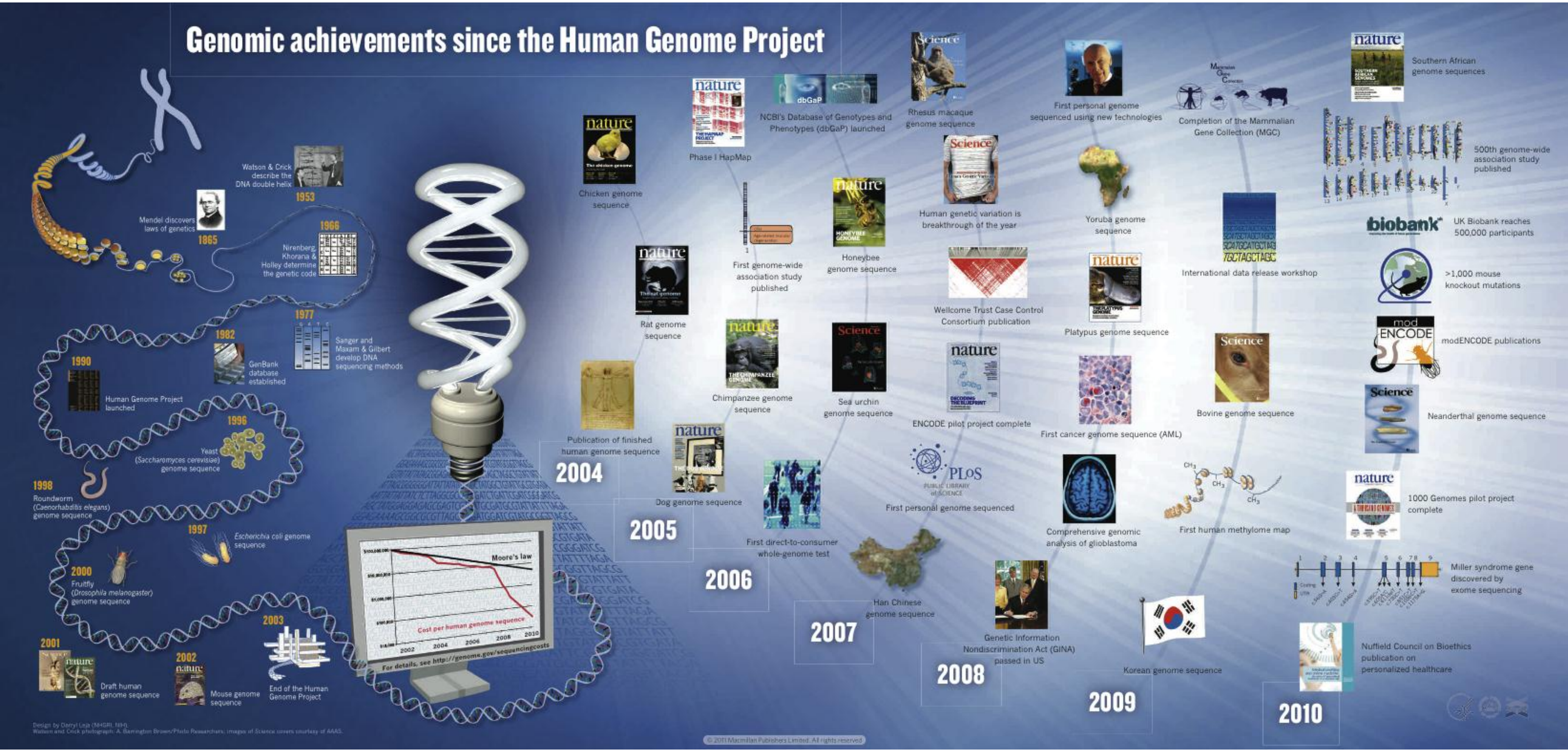
June 2011 Release 129,178,292,958 bases

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

20110813 http://www.ncbi.nlm.nih.gov/sites/genome

Achievements Since the HGP


Genomic achievements since the Human Genome Project
1865
Mendel discovers laws of genetics
1953
Watson & Crick describe the DNA double helix
1966
Nirenberg, Khorana & Holley determine the genetic code
1977
Sanger and Maxam & Gilbert develop DNA sequencing methods
1982
GenBank database established
1990
Human Genome Project launched
1996
Yeast (Saccharomyces cerevisiae) genome sequence
1998
Roundworm (Caenorhabditis elegans) genome sequence
1997
Escherichia coli genome sequence
2000
Fruitfly (Drosophila melanogaster) genome sequence
2001
Draft human genome sequence
2002
Mouse genome sequence
2003
End of the Human Genome Project
$100,000,000
$10,000,000
$1,000,000
$100,000
$10,000
2002
2004
2006
2008
2010
Moore's law
Cost per human genome sequence
For details, see http://genome.gov/sequencingcosts
2004
Chicken genome sequence
Rat genome sequence
Publication of finished human genome sequence
2005
Phase I HapMap
Chimpanzee genome sequence
Dog genome sequence
2006
NCBI's Database of Genotypes and Phenotypes (dbGaP) launched
First genome-wide association study published
Honeybee genome sequence
Sea urchin genome sequence
First direct-to-consumer whole-genome test
2007
Rhesus macaque genome sequence
Human genetic variation is breakthrough of the year
Wellcome Trust Case Control Consortium publication
ENCODE pilot project complete
First personal genome sequenced
Han Chinese genome sequence
2008
First personal genome sequenced using new technologies
Yoruba genome sequence
Platypus genome sequence
First cancer genome sequence (AML)
Comprehensive genomic analysis of glioblastoma
Genetic Information Nondiscrimination Act (GINA) passed in US
2009
Completion of the Mammalian Gene Collection (MGC)
International data release workshop
Bovine genome sequence
First human methylome map
Korean genome sequence
2010
Southern African genome sequences
500th genome-wide association study published
UK Biobank reaches 500,000 participants
>1,000 mouse knockout mutations
modENCODE publications
Neanderthal genome sequence
1000 Genomes pilot project complete
Miller syndrome gene discovered by exome sequencing
Nuffield Council on Bioethics publication on personalized healthcare

Design by Darryl Leja (NHGRI, NIH).
Watson and Crick photograph: A. Barrington Brown/Photo Researchers; images of Science covers courtesy of AAAS.

ED Green *et al. Nature* 470, 204-213 (2011) doi:10.1038/nature09764

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 $67 billion to the US economy

http://www.genome.gov/27544383

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

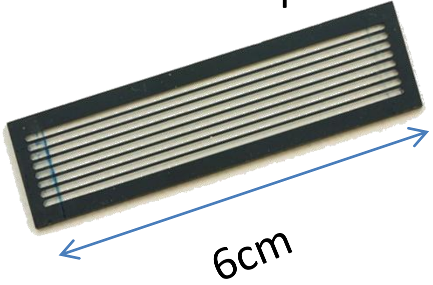
Illumina HiSeq Key Features

- 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)

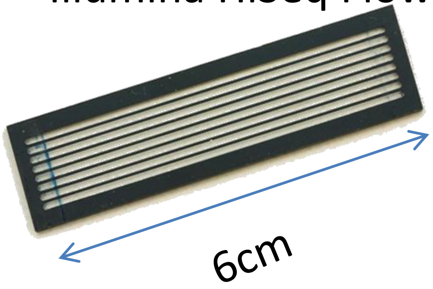
Illumina HiSeq Flowcell



Illumina HiSeq setup

Automated sample preparation

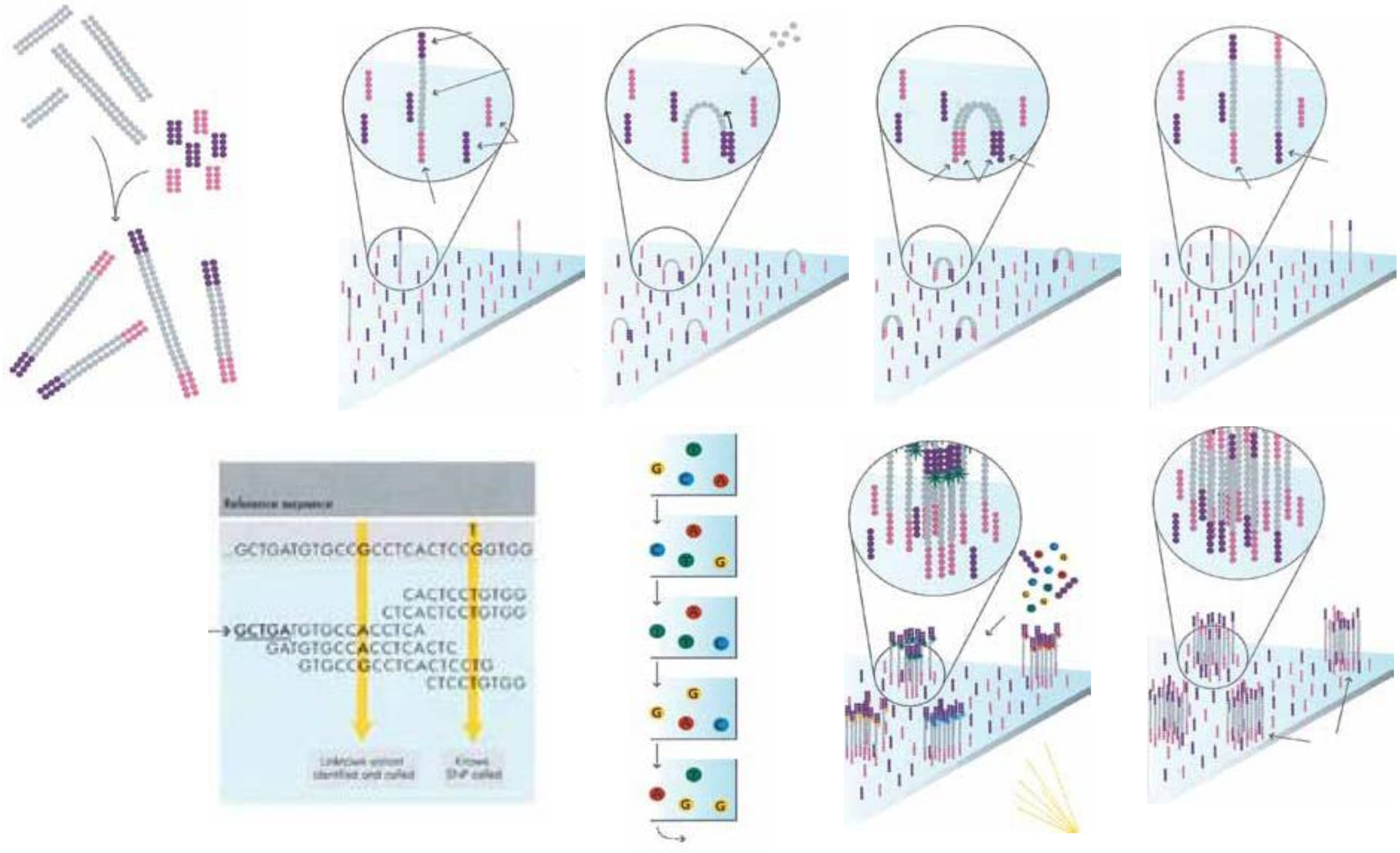
Illumina HiSeq Flowcell



cBot Cluster generation

HiSeq 2500

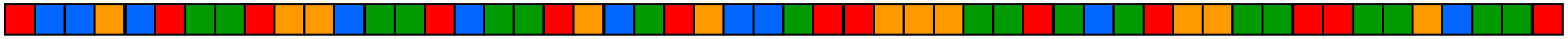
Illumina Sequencing



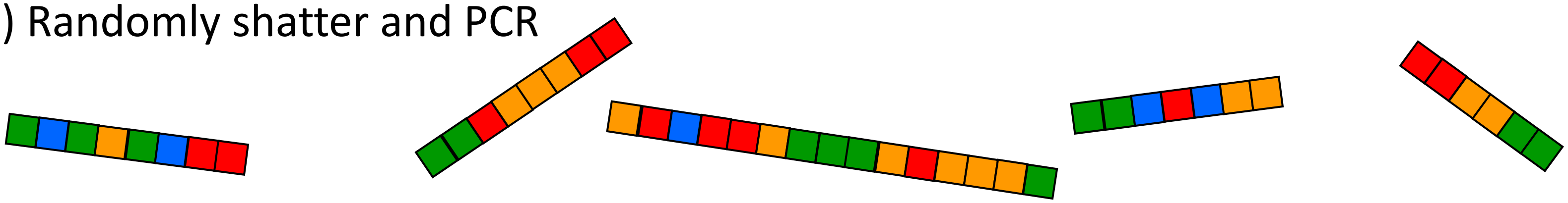
http://www.illumina.com/Documents/products/techspotlights/techspotlight_sequencing.pdf

DNA sample preparation (over-simplified)

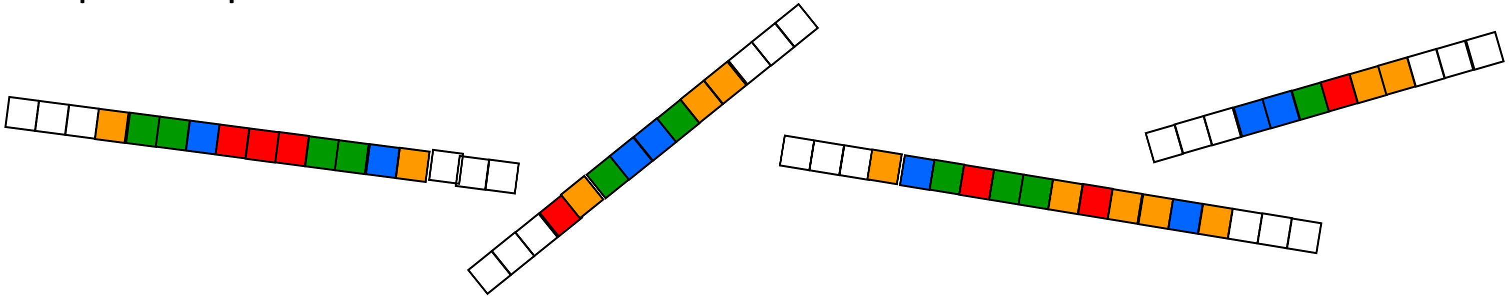
1) Extract DNA



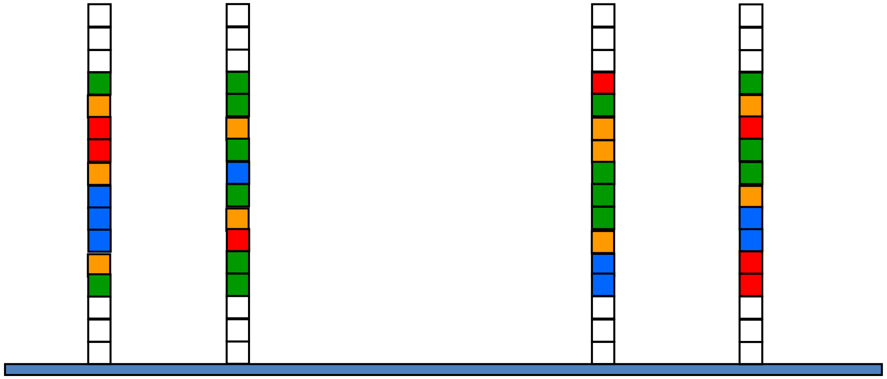
2) Randomly shatter and PCR



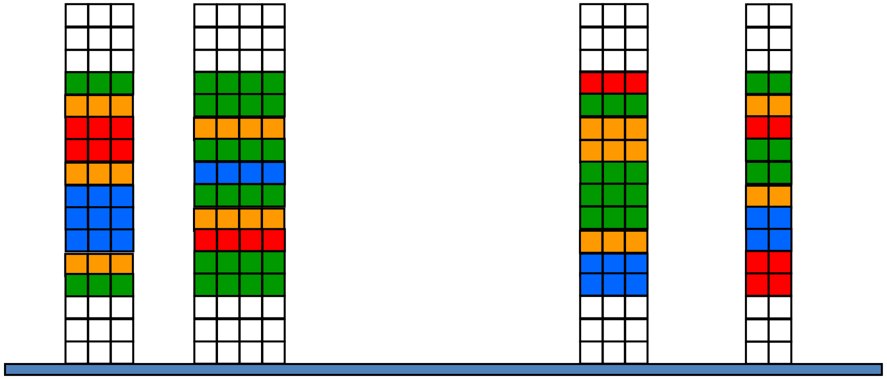
3) Attach adapter sequence □□□



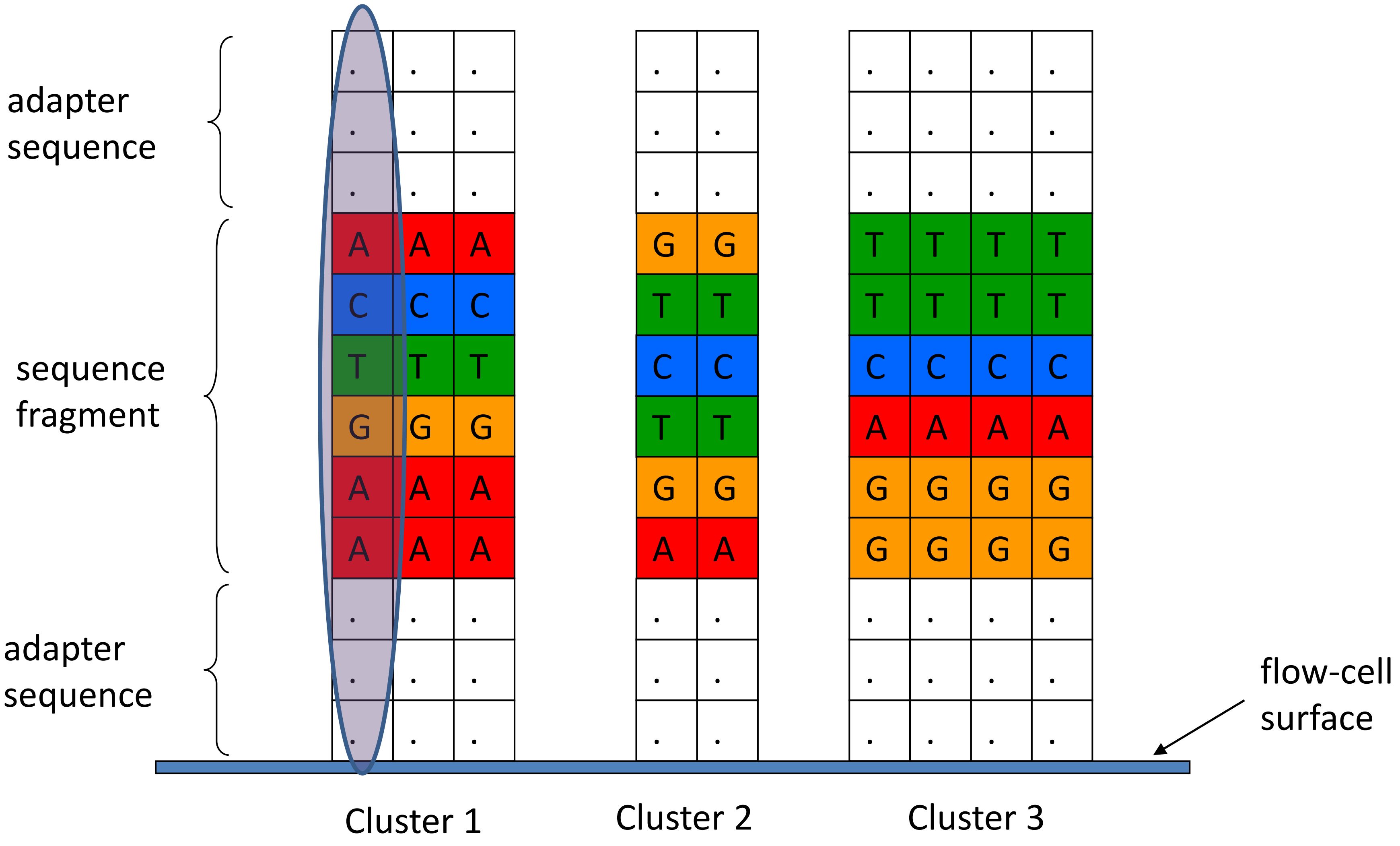
4) Attach to flow-cell surface



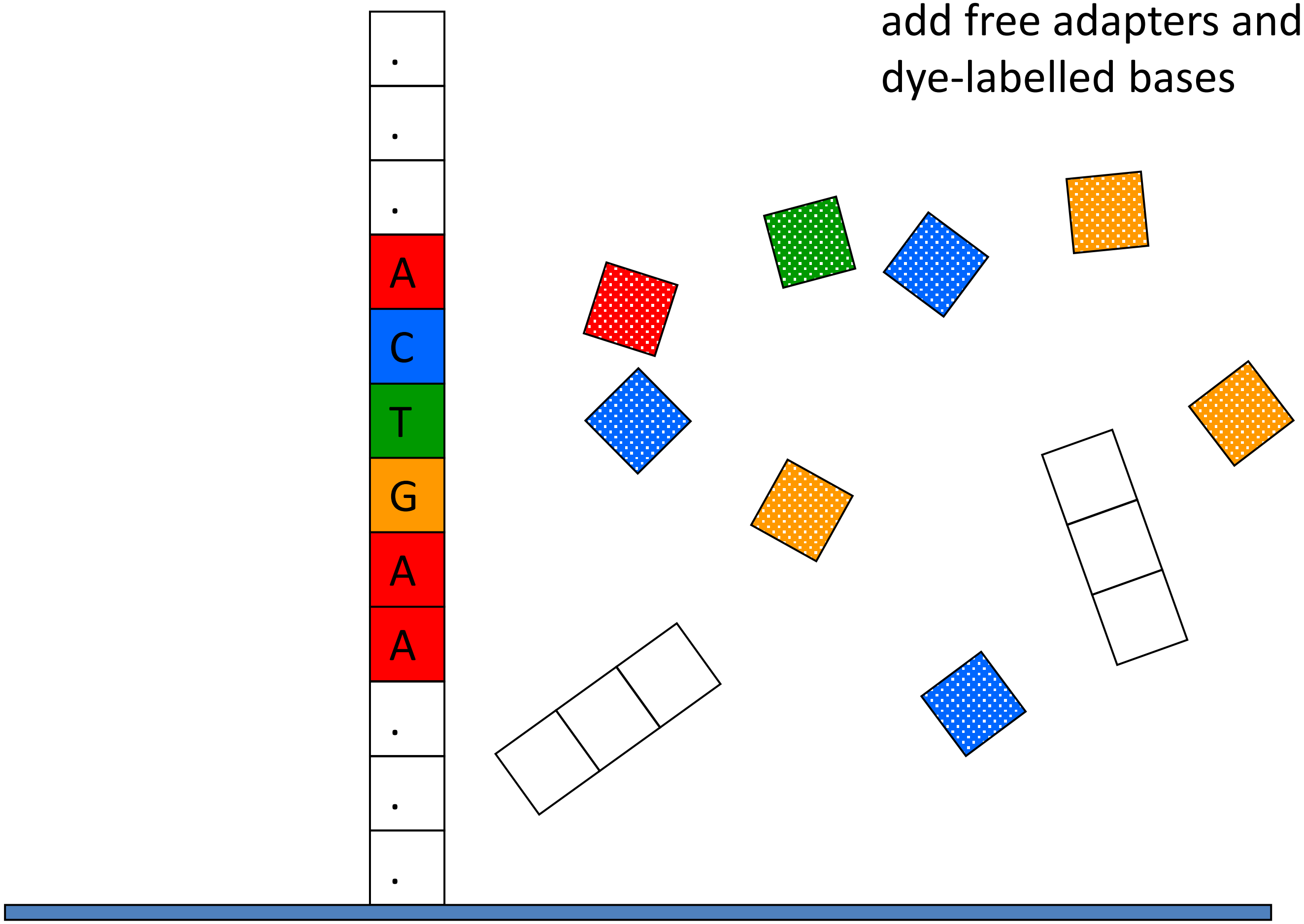
5) PCR-amplify into clusters



Sequence clusters on the flow cell

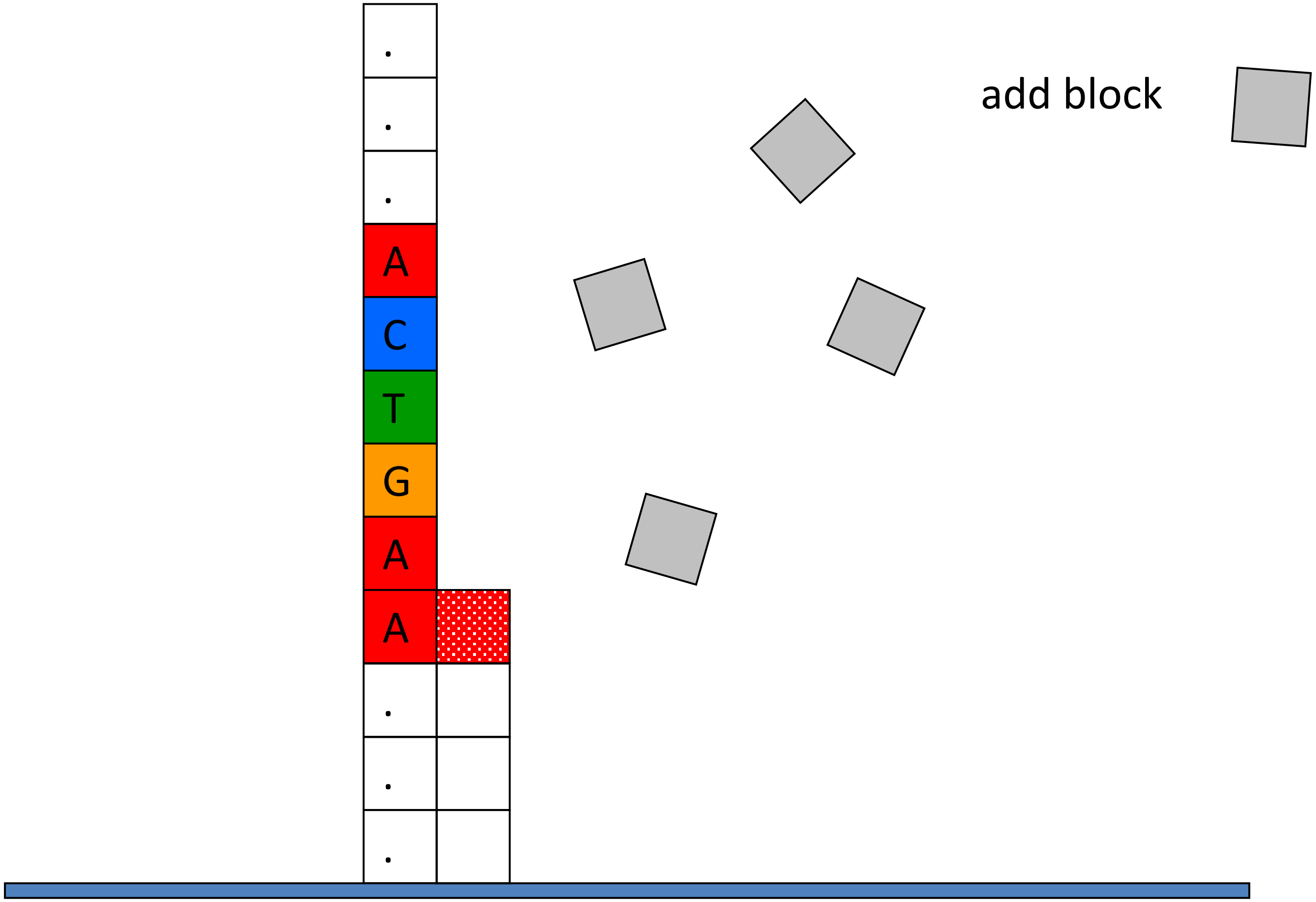


Sequencing cycle 1

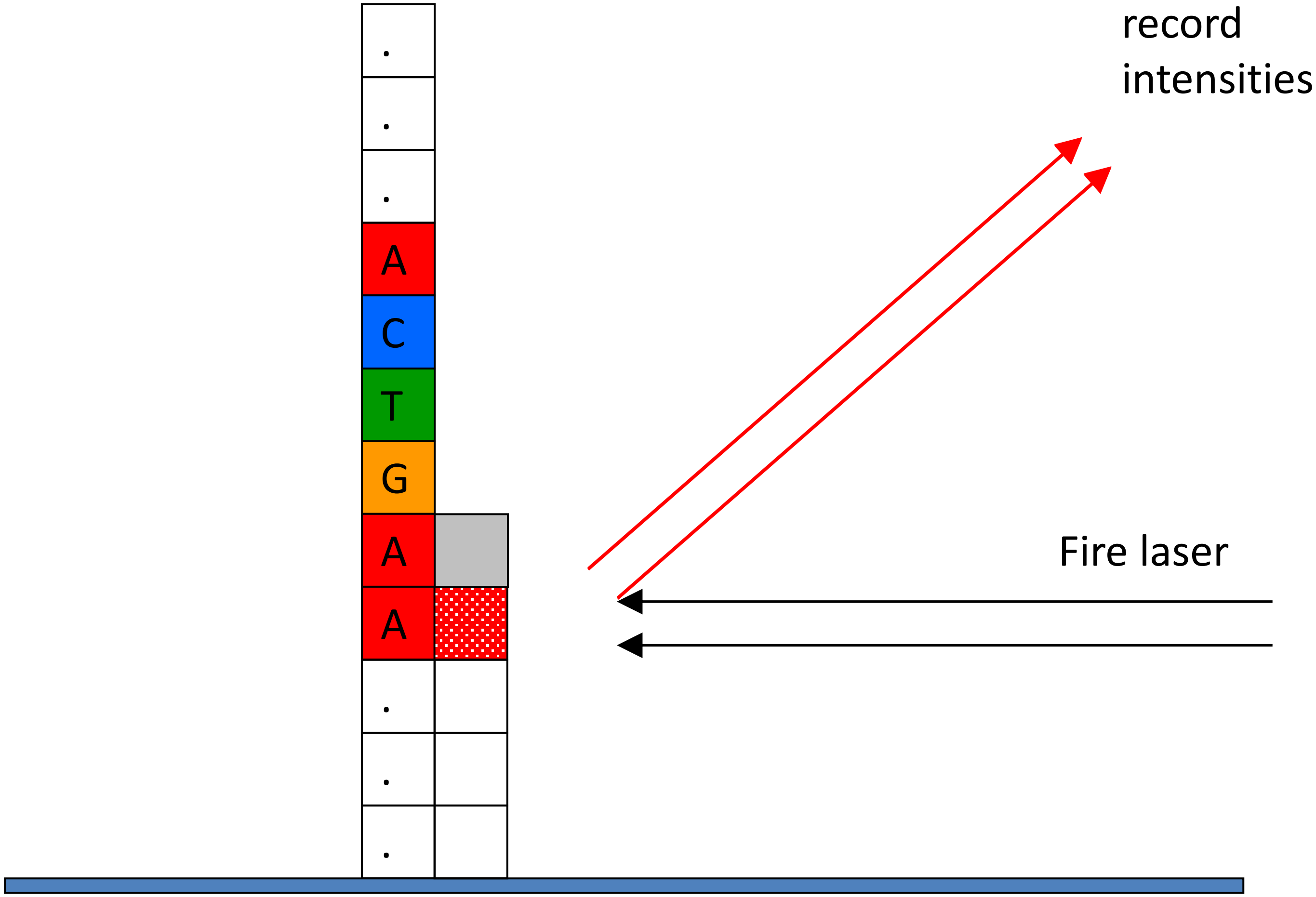


add free adapters and
dye-labelled bases

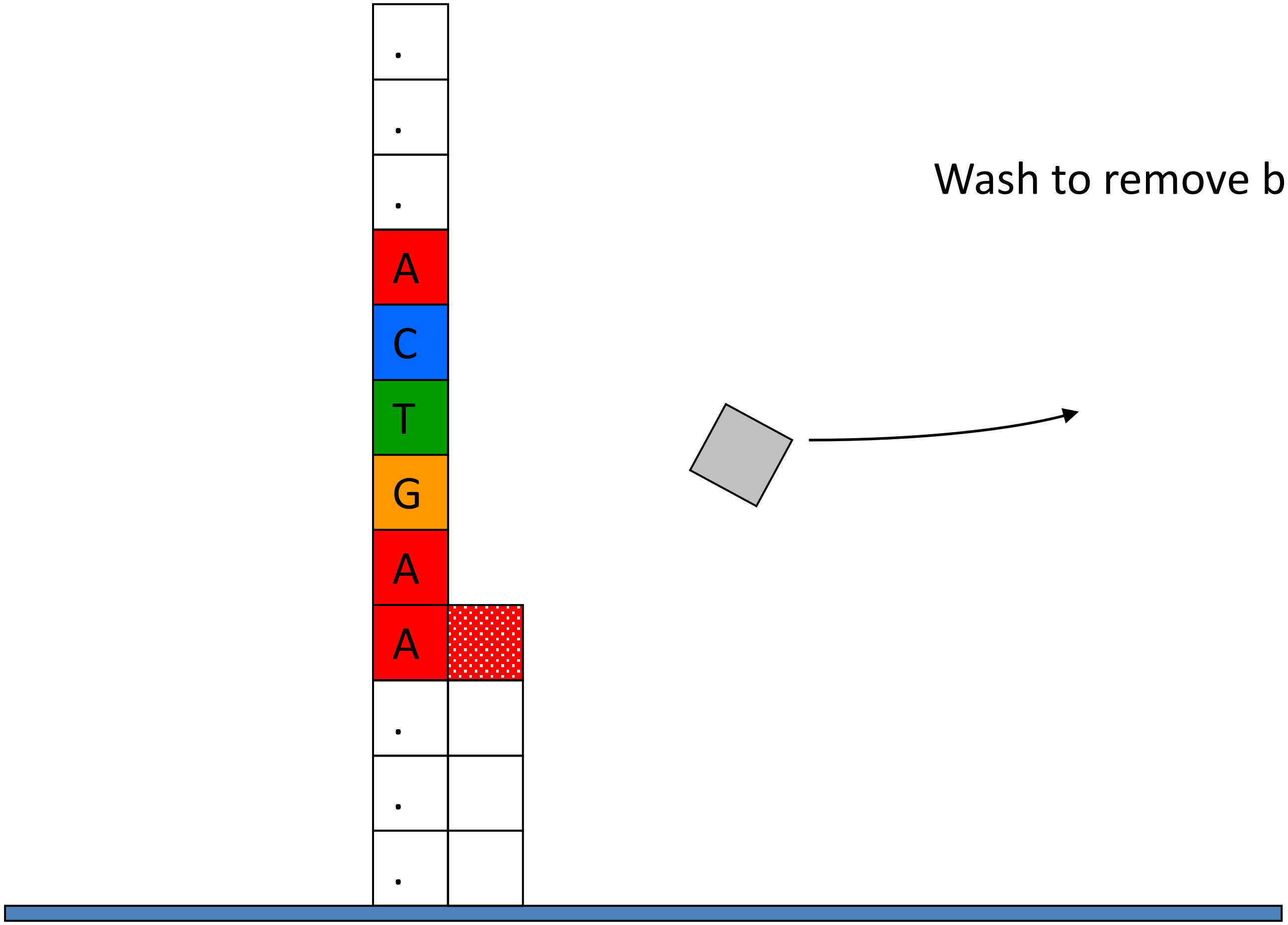
Sequencing cycle 1



Sequencing cycle 1

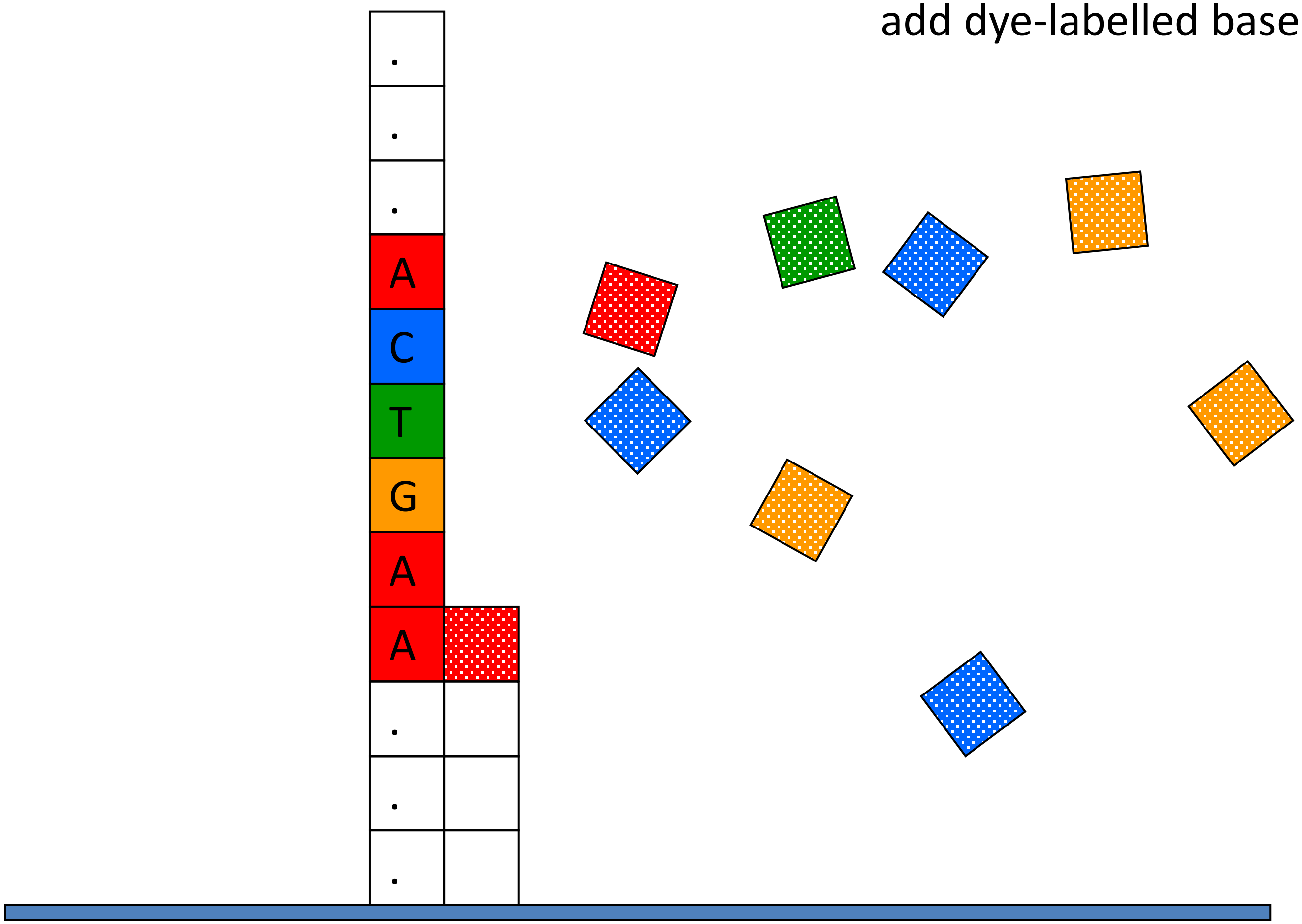


Sequencing cycle 1



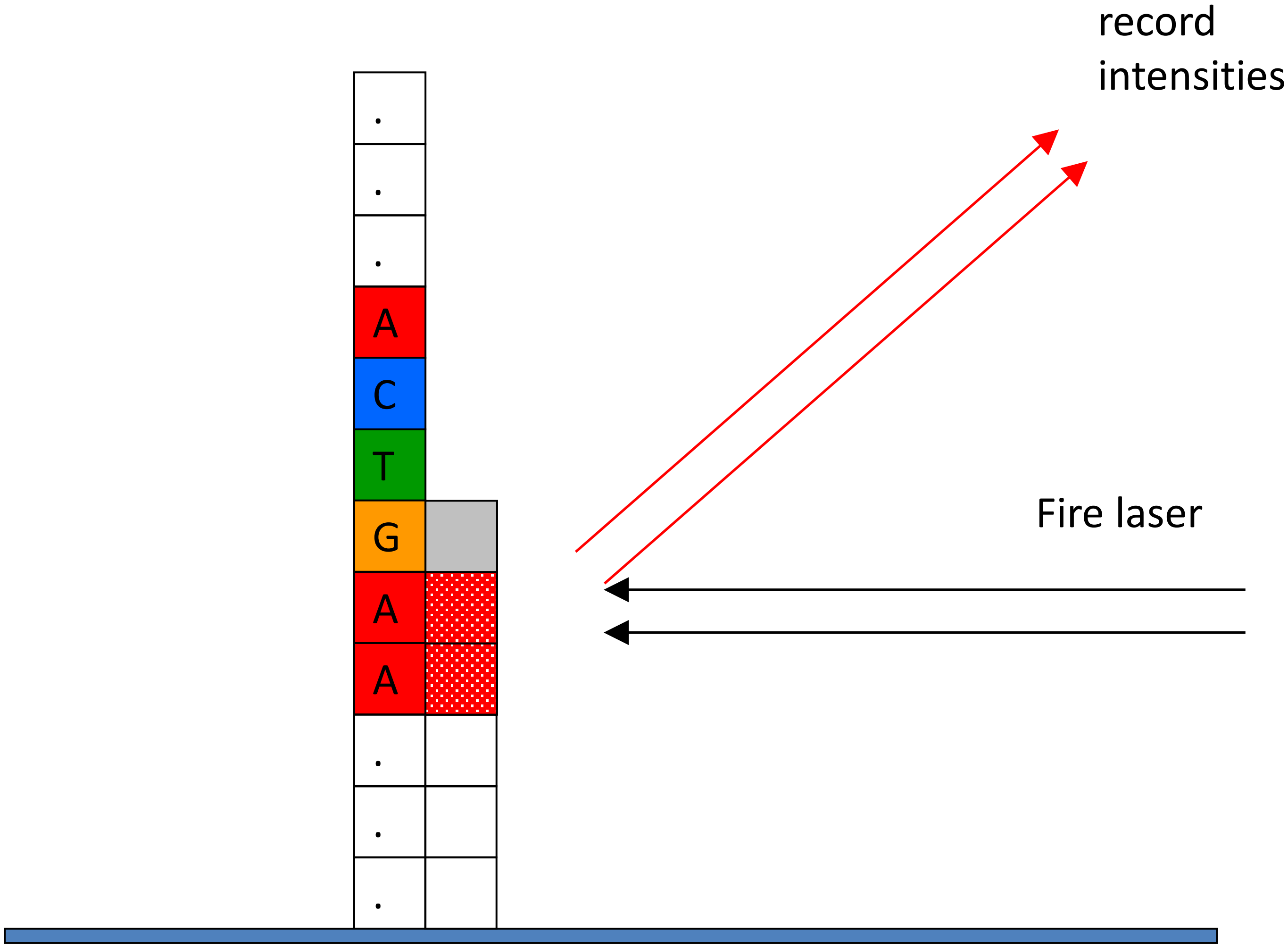
Wash to remove block

Sequencing cycle 2

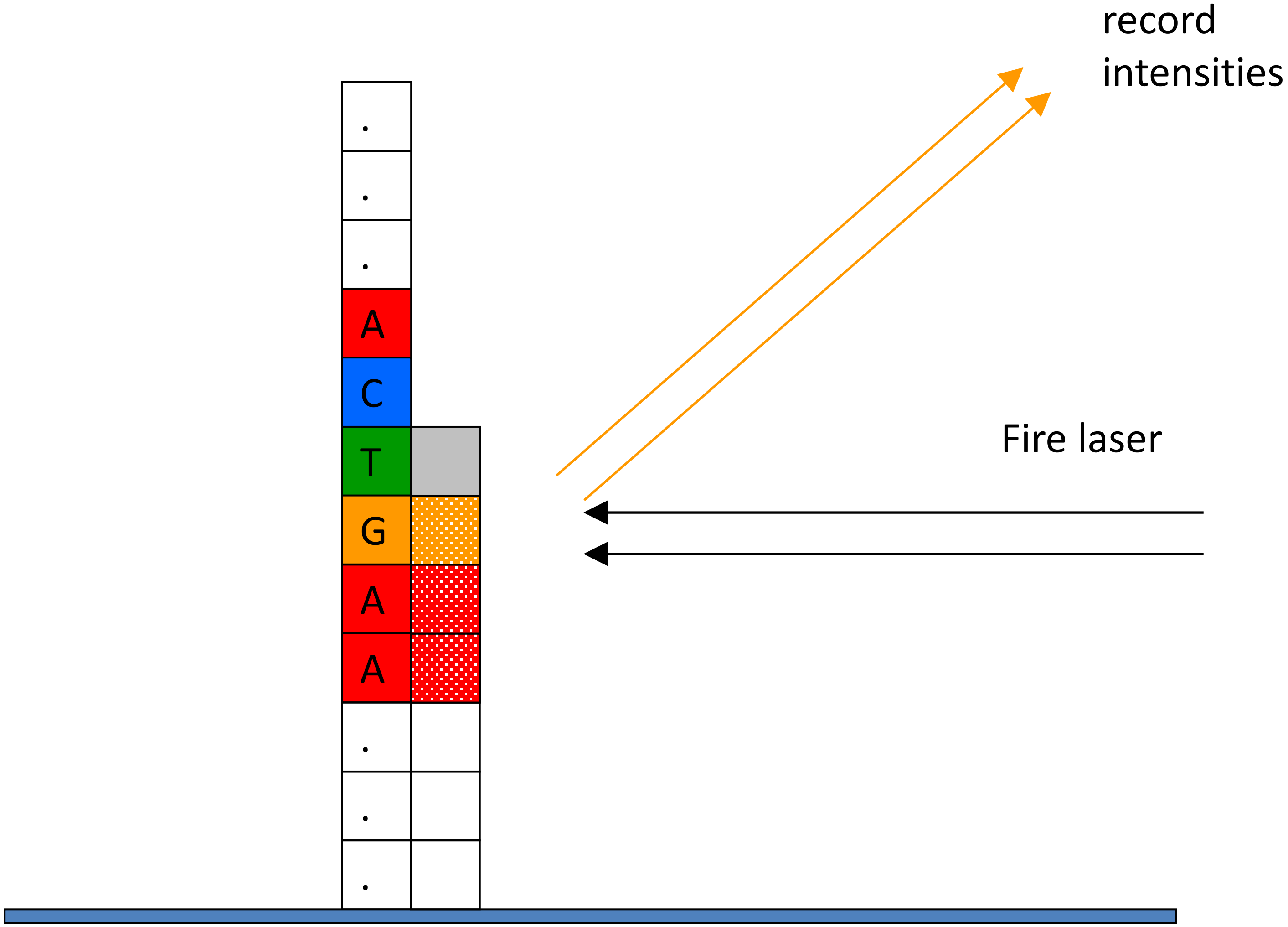


add dye-labelled bases

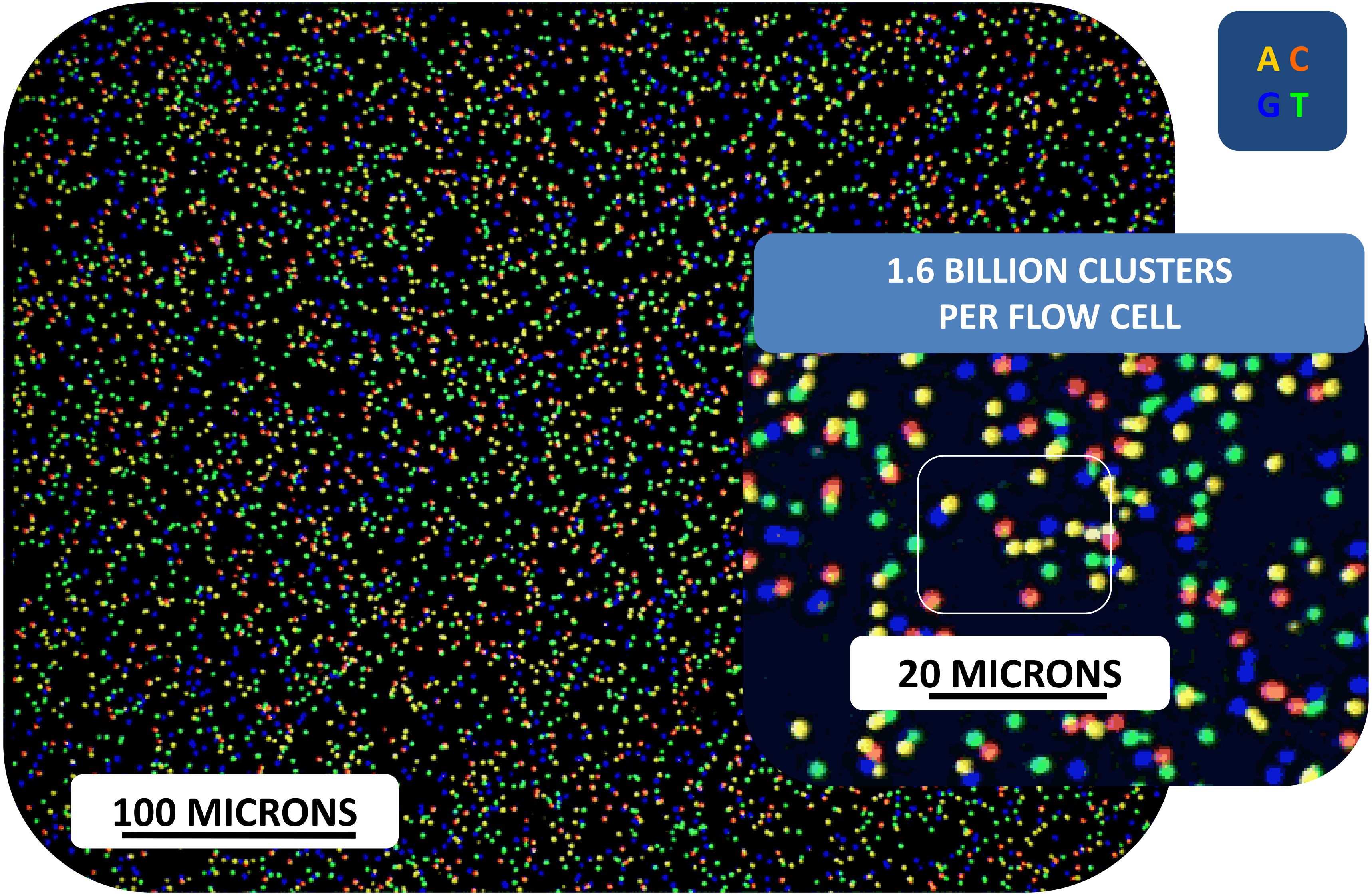
Sequencing cycle 2



Sequencing cycle 3

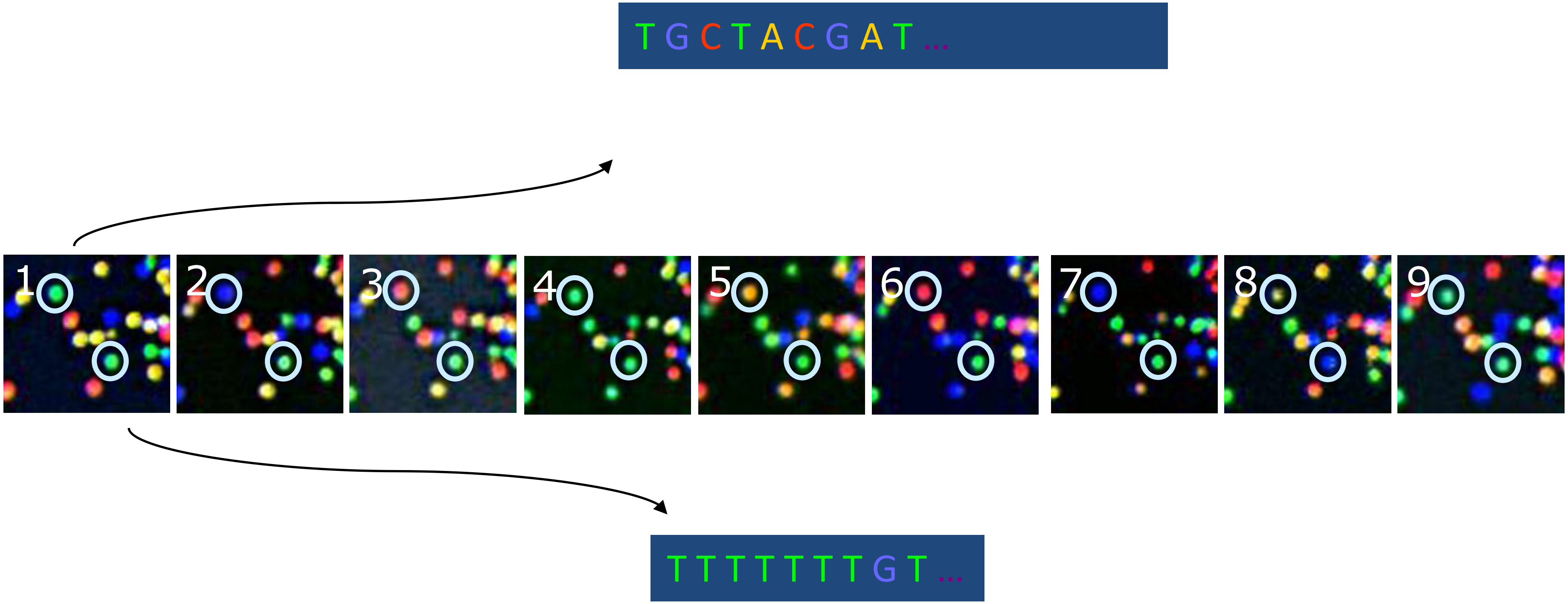


Illumina Sequencing : How it looks



1.6 BILLION CLUSTERS
PER FLOW CELL

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

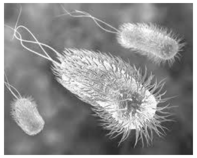
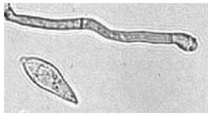
HiSeq 2500
2 lanes

2 day run time

Comparison

APPLICATION	RAPID RUN MODE	HIGH OUTPUT MODE
ChIP-Seq **Transcription Factor** **1 x 36 bp**	40 Samples 7 Hours	200 Samples 2 Days
mRNA-Seq **2 x 50 bp**	24 Samples 16 Hours	120 Samples 5 Days
TruSeq Exome Seq **62 MB Region** **100x Coverage** **2 x 100 bp**	15 Samples 27 Hours	85 Samples 12 Days
Human Whole Genome **>30x Coverage** **2 x 100 bp**	1 Sample 27 Hours	5 Samples 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)

Other equipment required (optional)

Agilent Bravo liquid handling robot
£85k

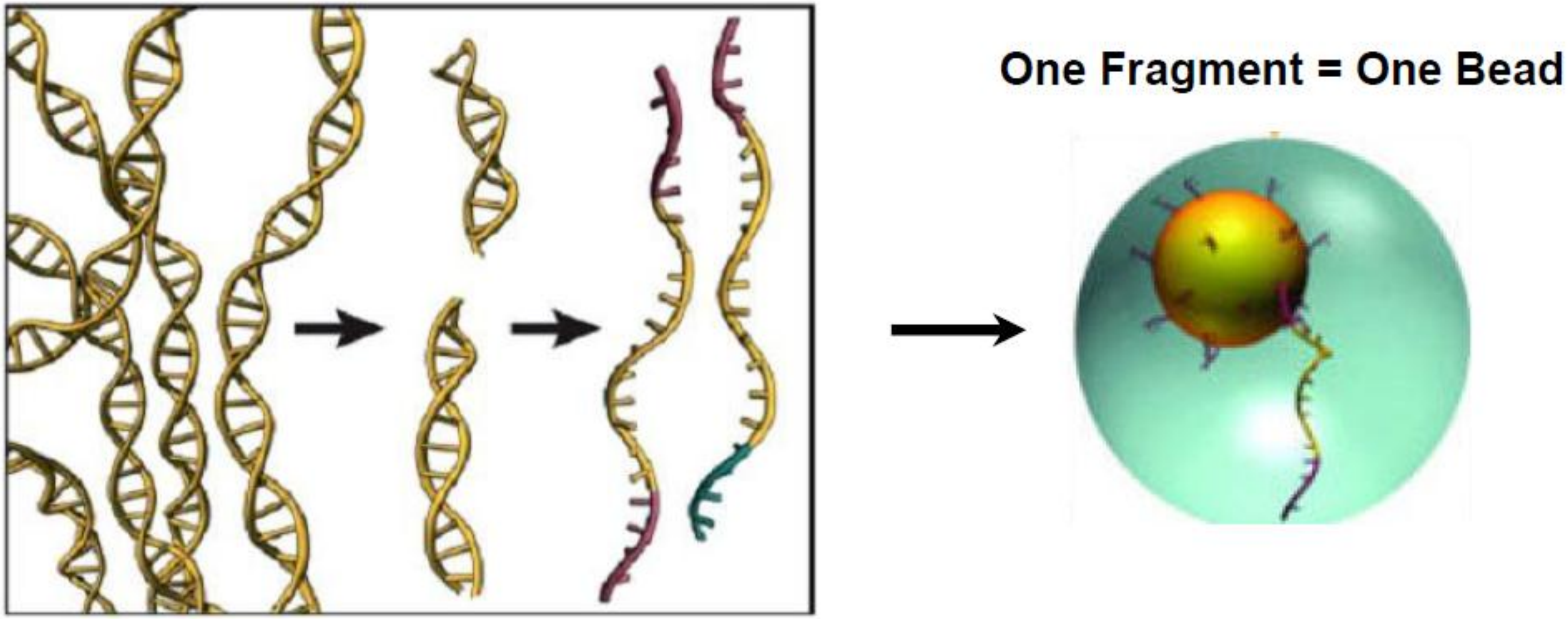
Agilent Tapestation
£30k

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

454 Step 1: Sample preparation

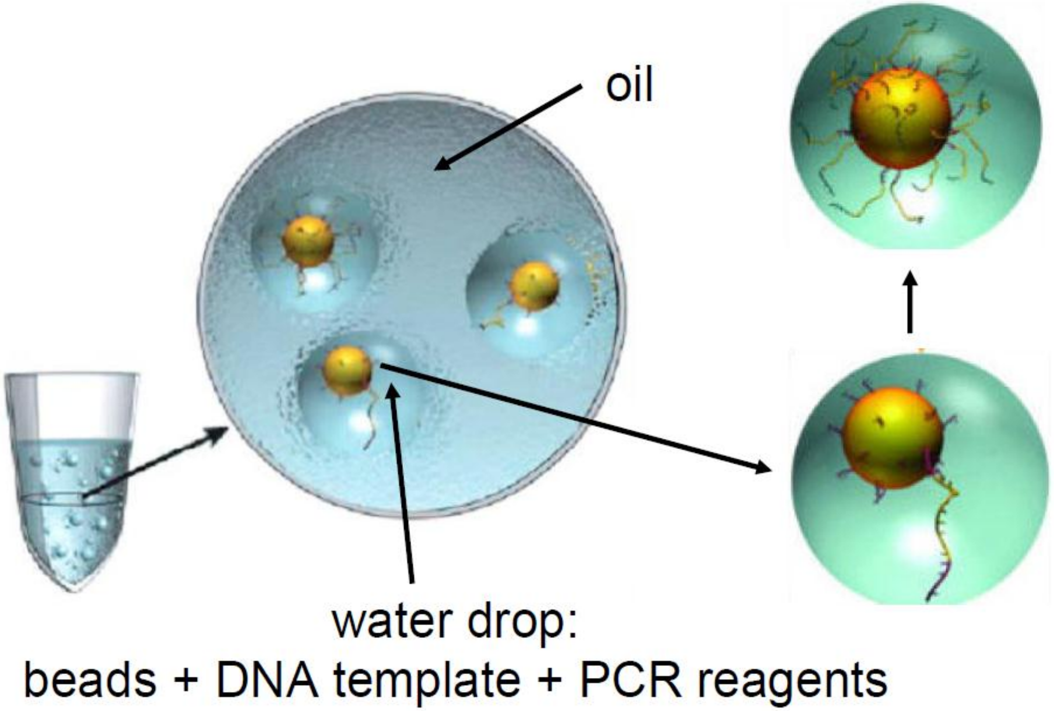


1. Genomic DNA is isolated and fragmented.
2. Adaptors are ligated to single stranded DNA
3. This forms a library

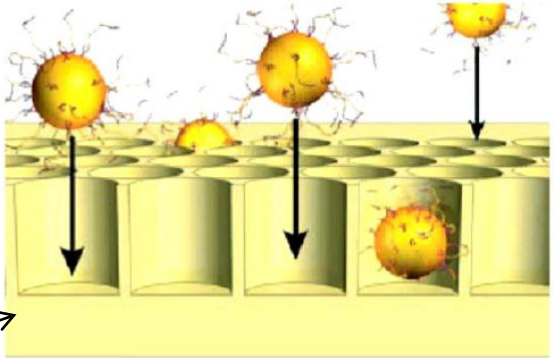
4. The single stranded DNA library is immobilised onto proprietary DNA capture beads

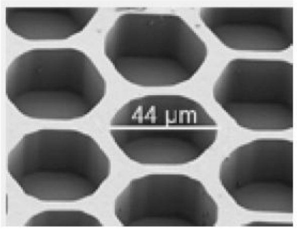
454 Step 2: Amplification

Water-based emulsion PCR



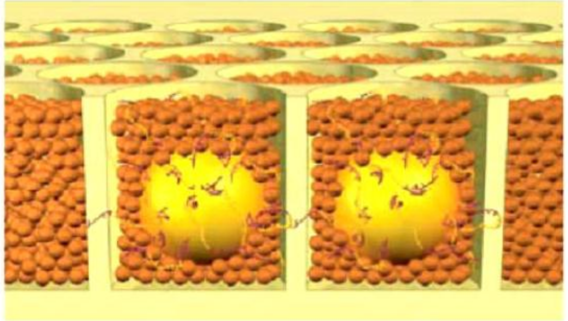
454 Step 3: Load emPCR products

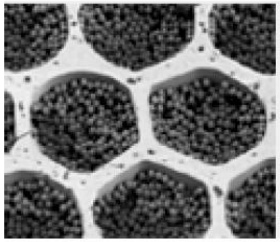




Picotitre plate

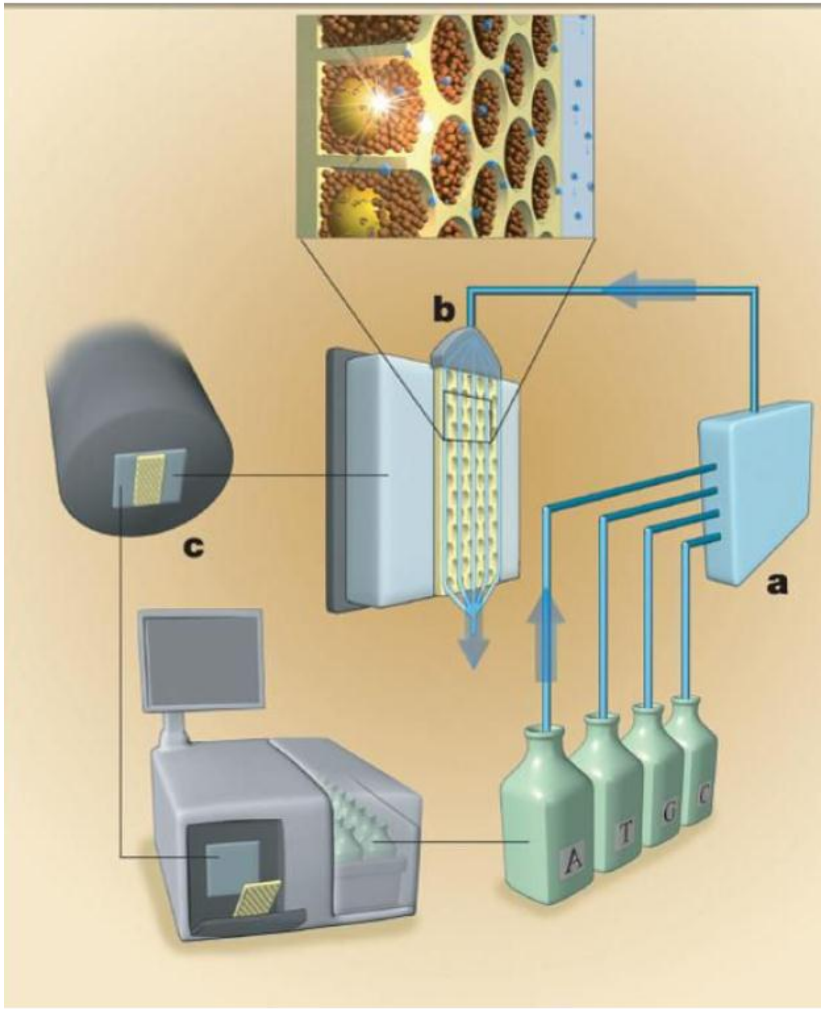
- enrich for DNA + beads
- diameter of the wells allows for only 1 bead/well





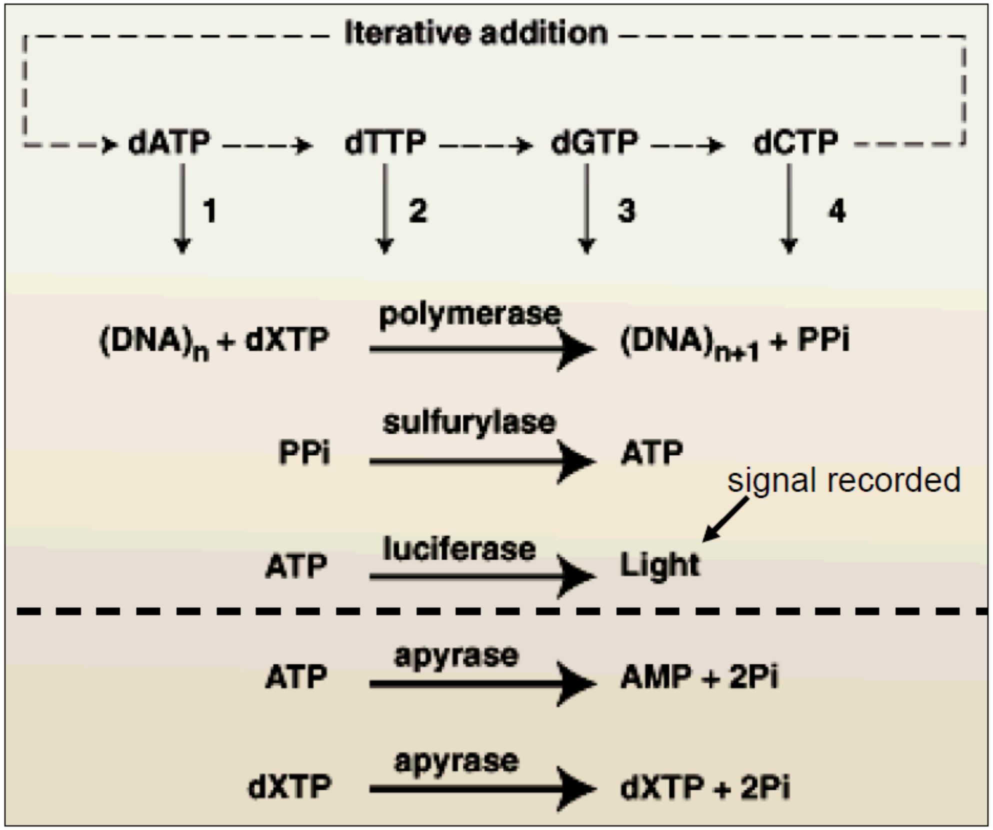
Smaller beads (red) carrying immobilized enzymes required for pyrophosphate sequencing are deposited into each well.

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

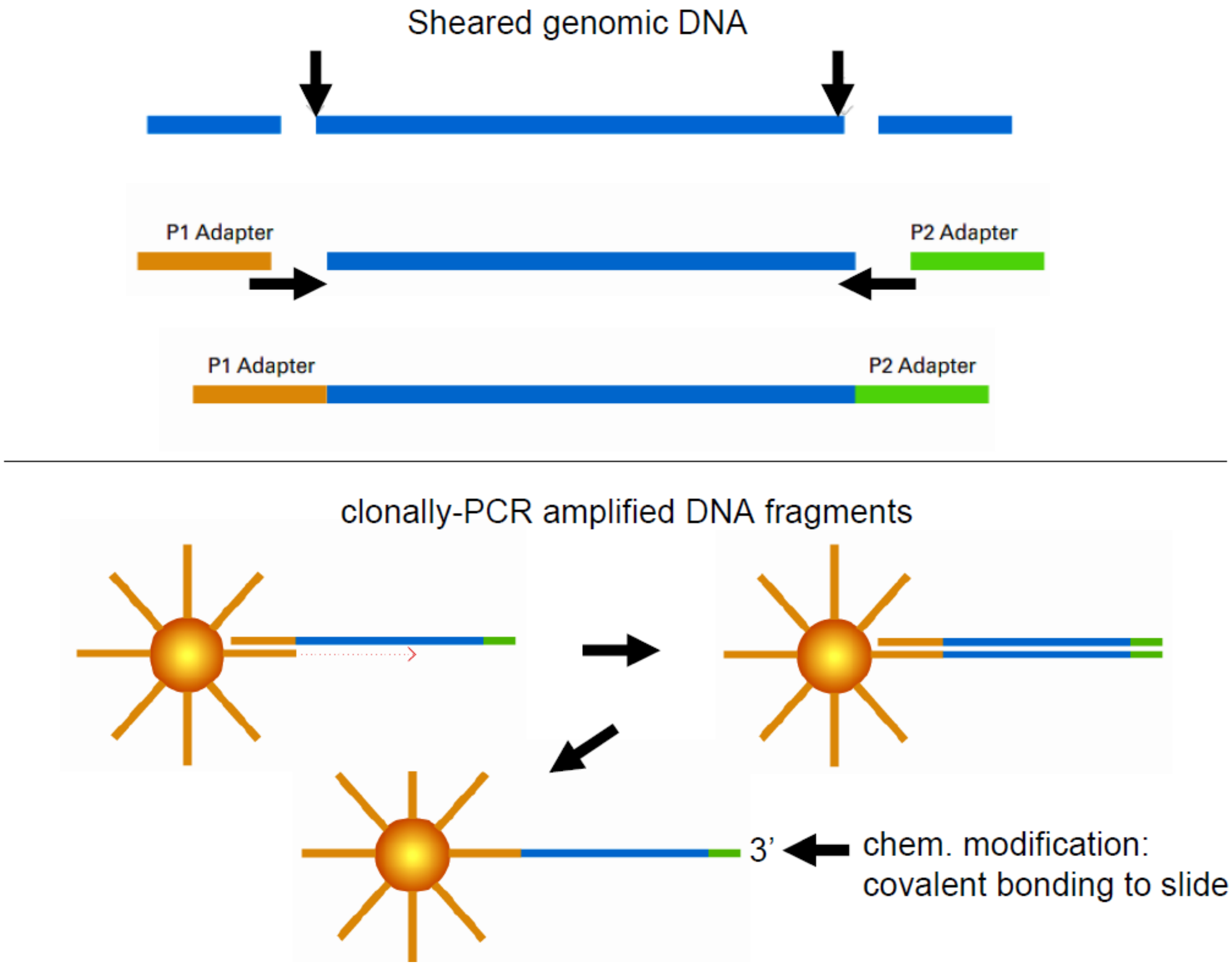
BUT – Only one colour is emitted

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



Sheared genomic DNA

P1 Adapter

P2 Adapter

P1 Adapter

P2 Adapter

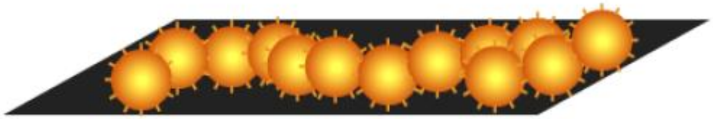
clonally-PCR amplified DNA fragments

3'

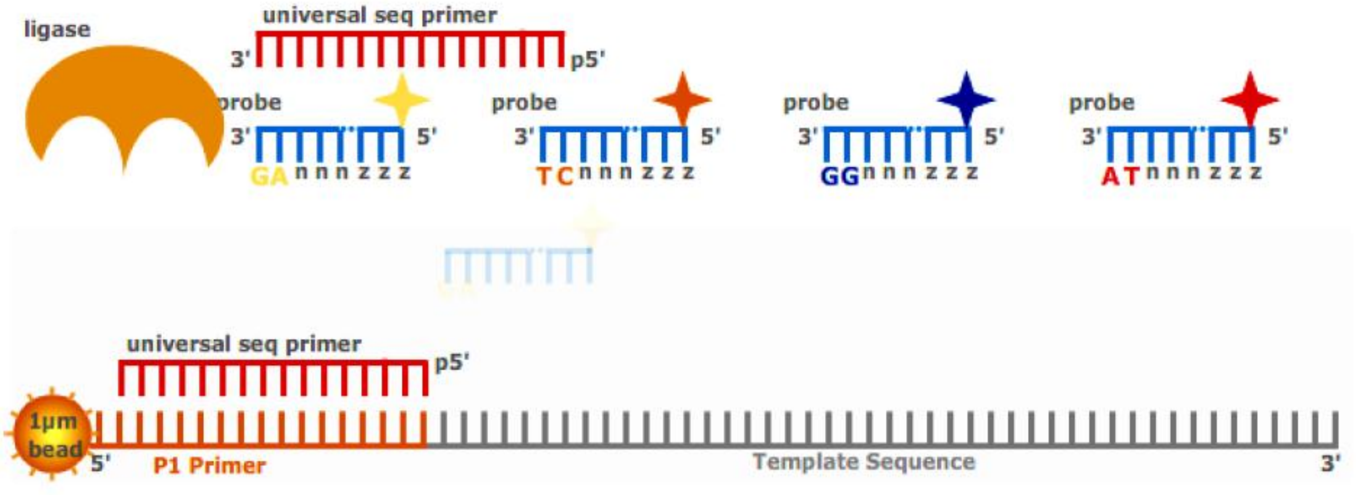
chem. modification:
covalent bonding to slide

SOLiD: Step 2 Attach beads

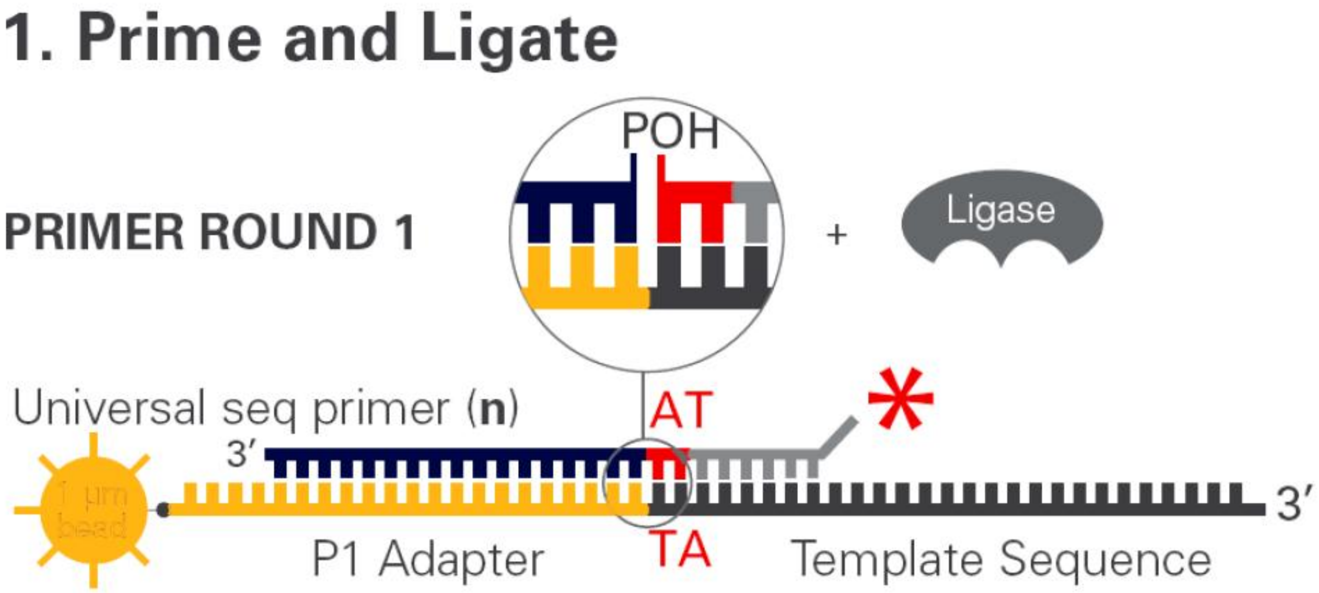
3’-modified beads
deposited onto glass slide



Sequential ligation with dye-labeled **oligonucleotides**

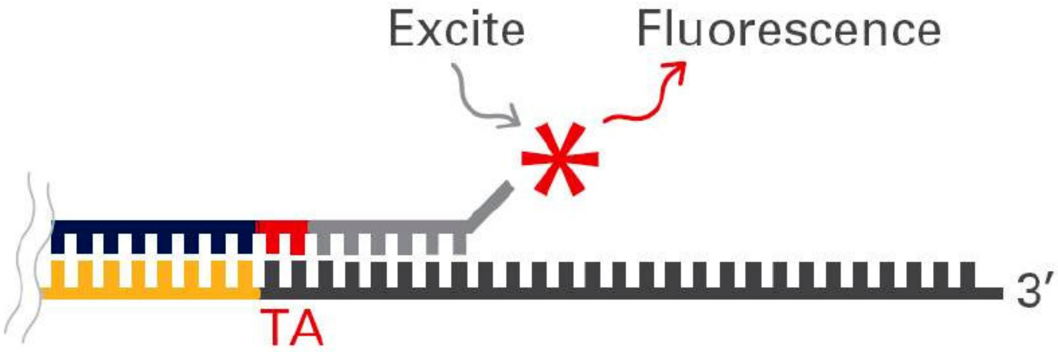


SOLiD: Step 3 Sequencing 1

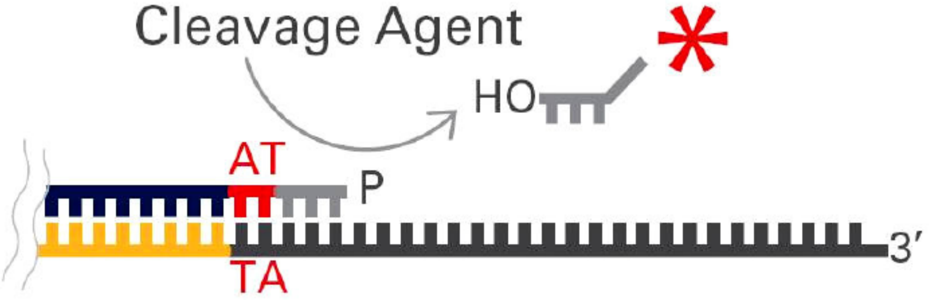


SOLiD: Step 3 Sequencing 2

2. Image

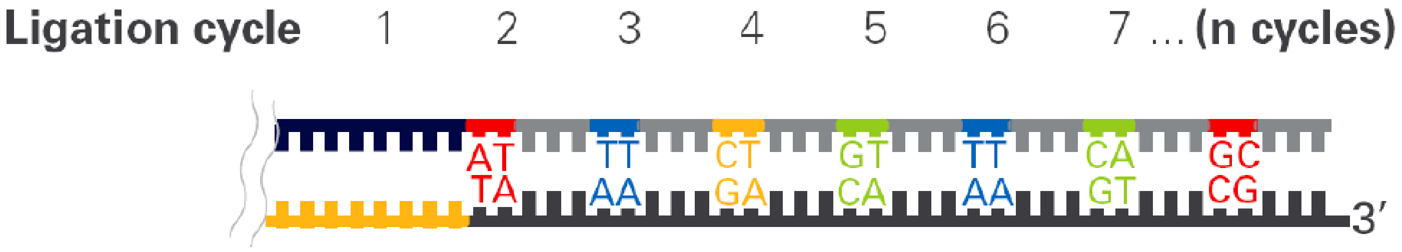


3. Cleave off Fluor



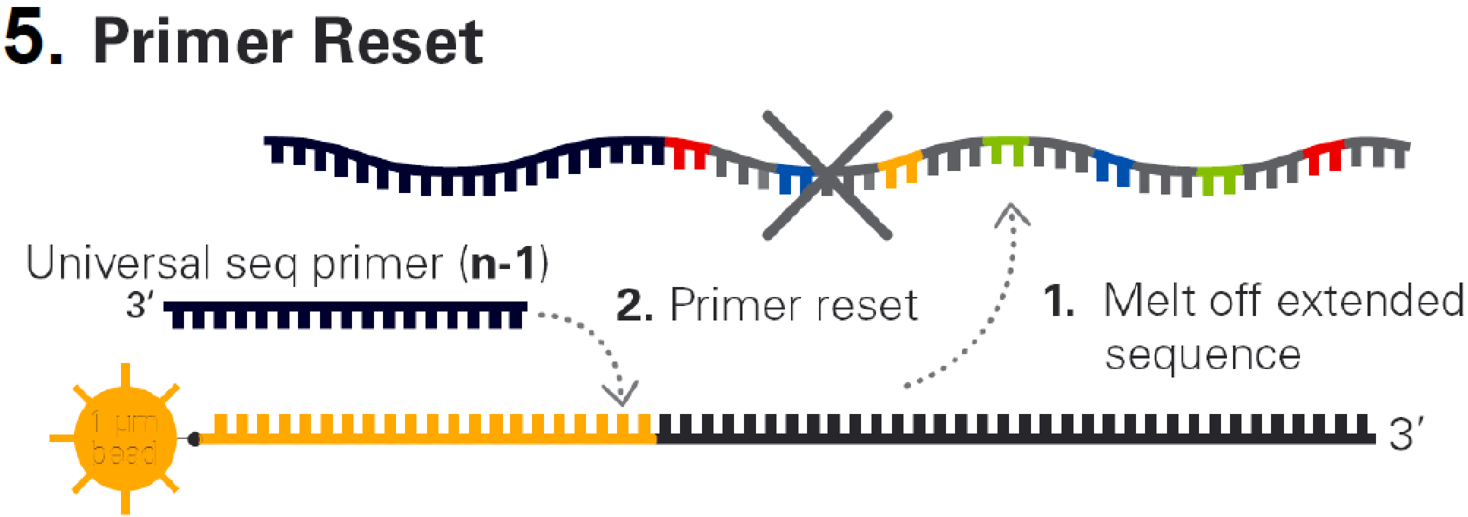
SOLiD Step 3 Sequencing 3

4. Repeat steps 1-4 to Extend Sequence



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

Universal seq primer (**n-1**)

3′

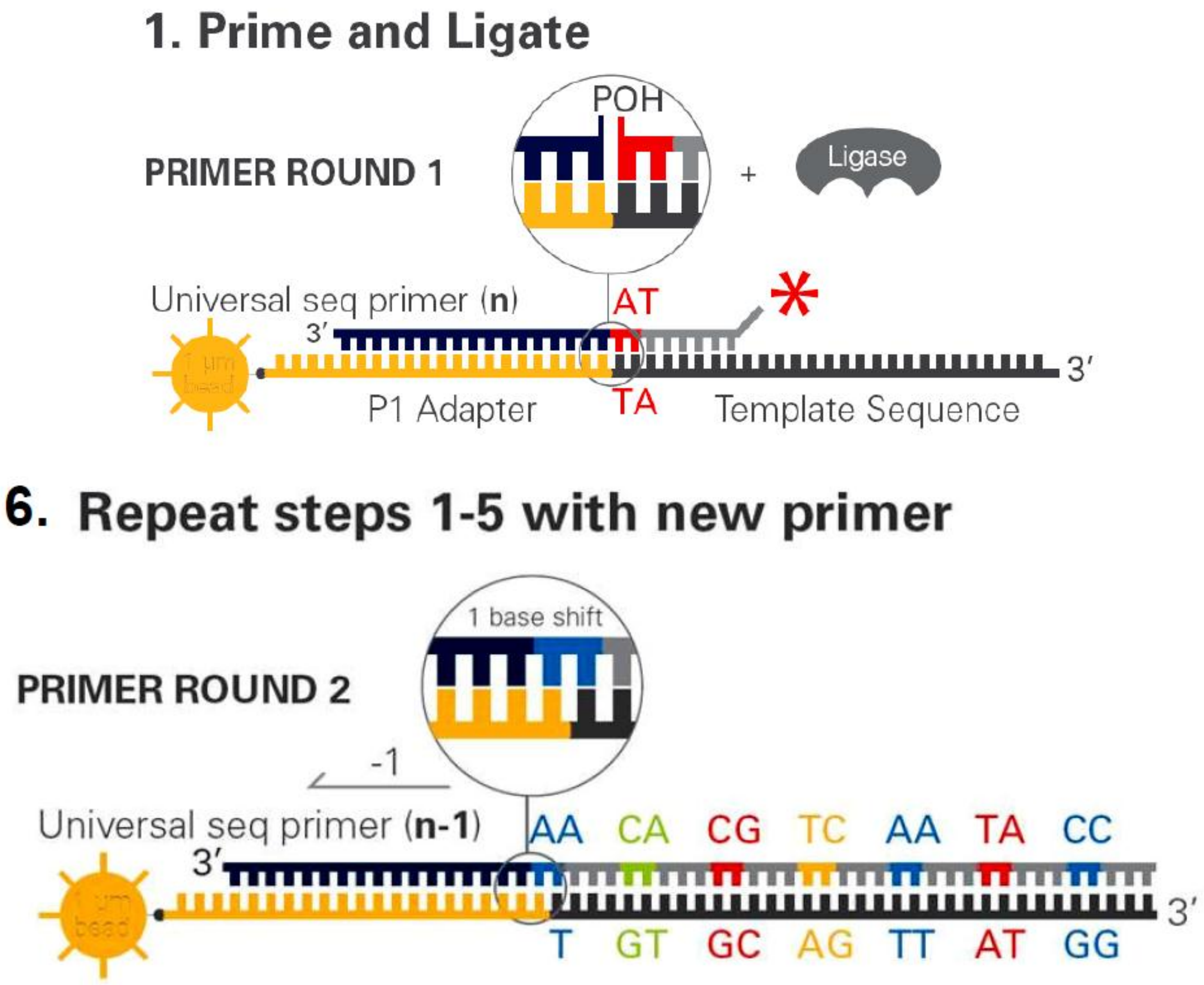
2. Primer reset

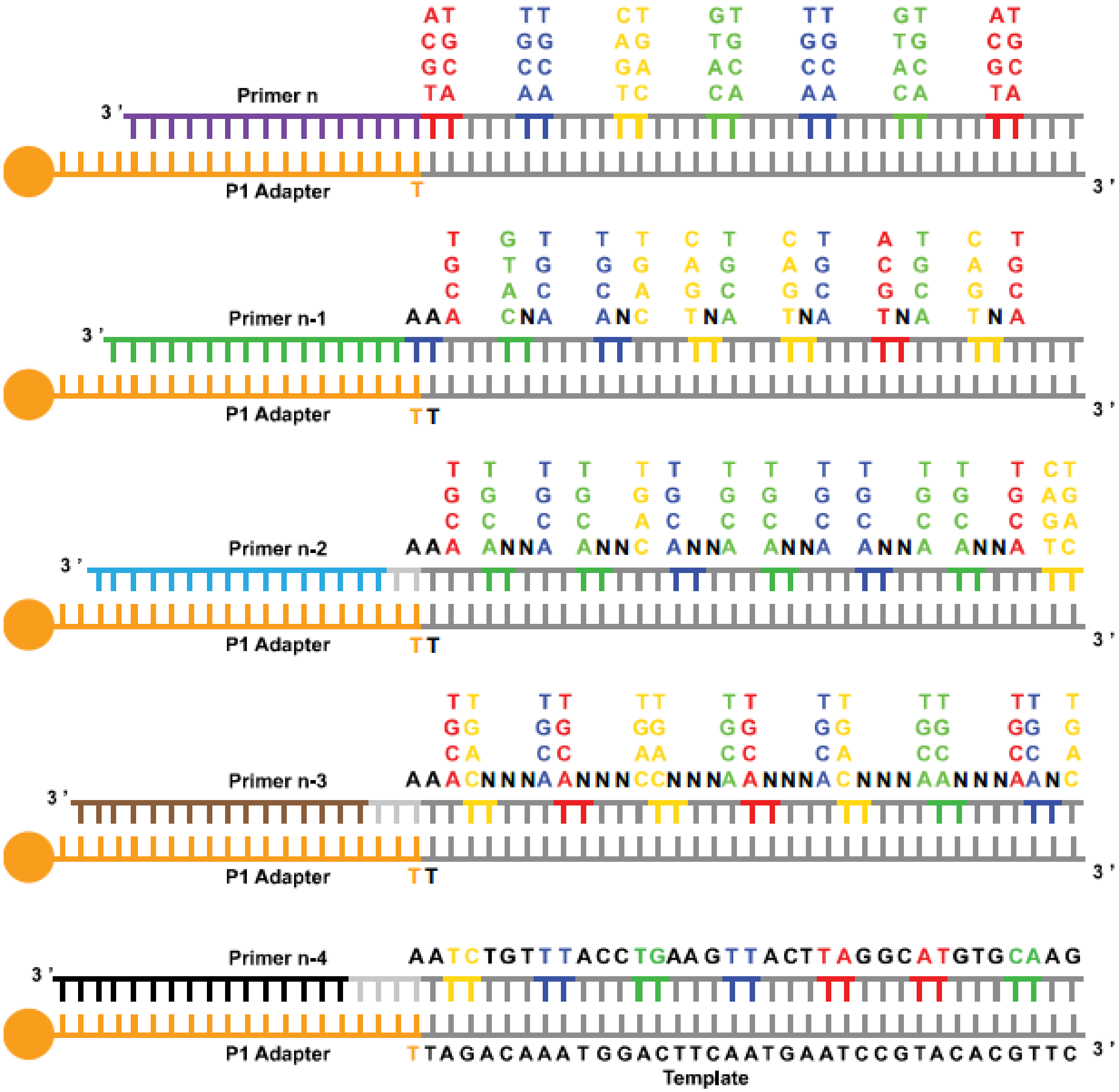
1. Melt off extended sequence

1 μm bead

3′

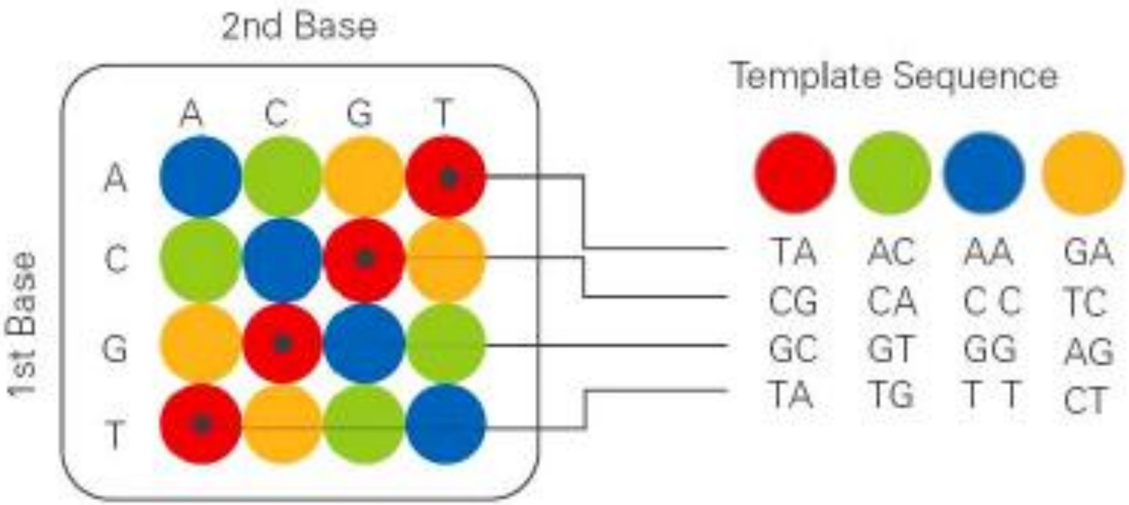
SOLiD Step 3 Sequencing 5




Primer n
3 '
P1 Adapter
T
3 '
Primer n-1
3 '
P1 Adapter
TT
3 '
Primer n-2
3 '
P1 Adapter
TT
3 '
Primer n-3
3 '
P1 Adapter
TT
3 '
Primer n-4
AATCTGTTTACCTGAAGTTACTTAGGCATGTGCAAG
3 '
P1 Adapter
TTAGACAAATGGACTTCAATGAATCCGTACACGTTC
3 '
Template

SoLID Colour space

Possible dinucleotides encoded by each color



2nd Base

1st Base

	A	C	G	T
A	🔵	🟢	🟠	🔴
C	🟢	🔵	🔴	🟠
G	🟠	🔴	🔵	🟢
T	🔴	🟠	🟢	🔵

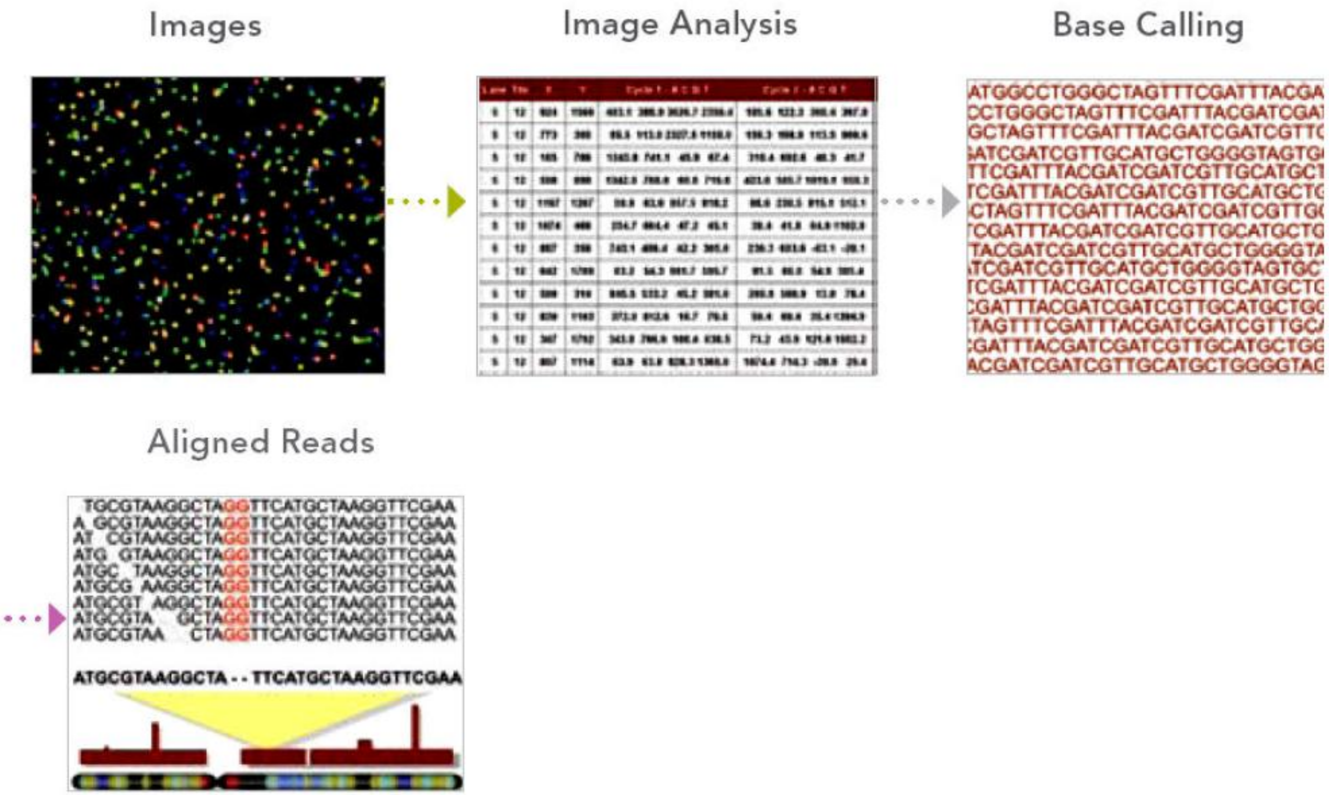
Template Sequence

🔴	🟢	🔵	🟠
TA	AC	AA	GA
CG	CA	C C	TC
GC	GT	GG	AG
TA	TG	T T	CT

Common features

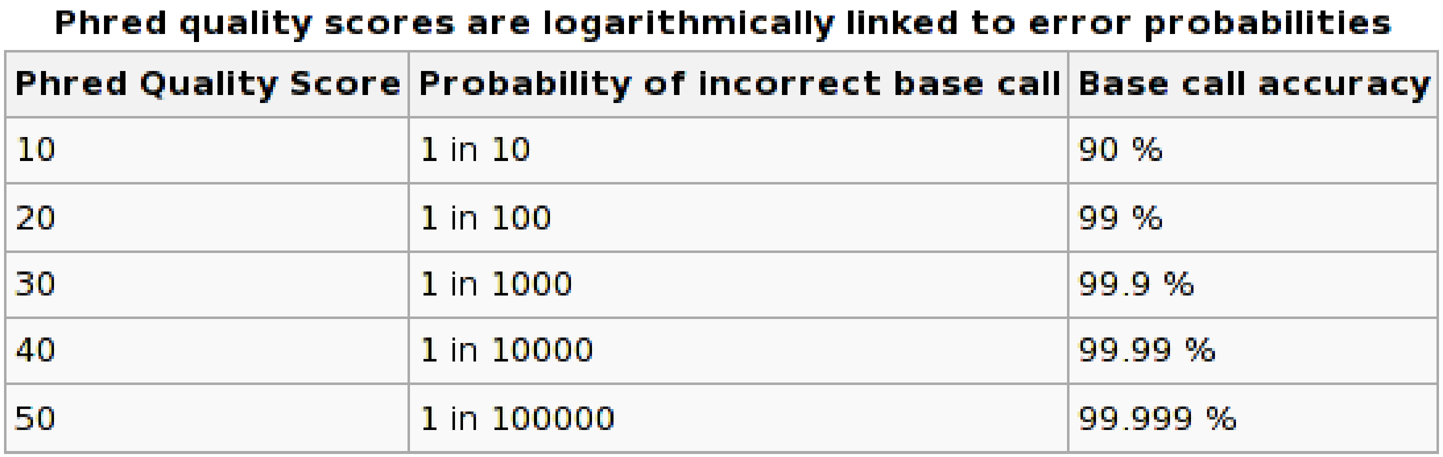
- All 3 platforms share the following:
  - Adaptor sequences to fix probes to a surface/bead
  - Amplification
  - Use of fluorescent probes and CCD devices
  - Capable of paired-end reads
  - Post-processing software to determine image quality
  - Shorter read lengths compared to traditional capillary based sequencers
  - Much higher data volumes (~Gb)
  - Sequence a human genome in a matter of days

Common features



Phred Score

- Phred program: http://en.wikipedia.org/wiki/Phred_base_calling
- $Q = -10 \log10(P)$
- $P = 10^{\wedge}(-Q/10)$



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 (60% instruments)
- Market splitting 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 prices

Life Technologies Ion Torrent

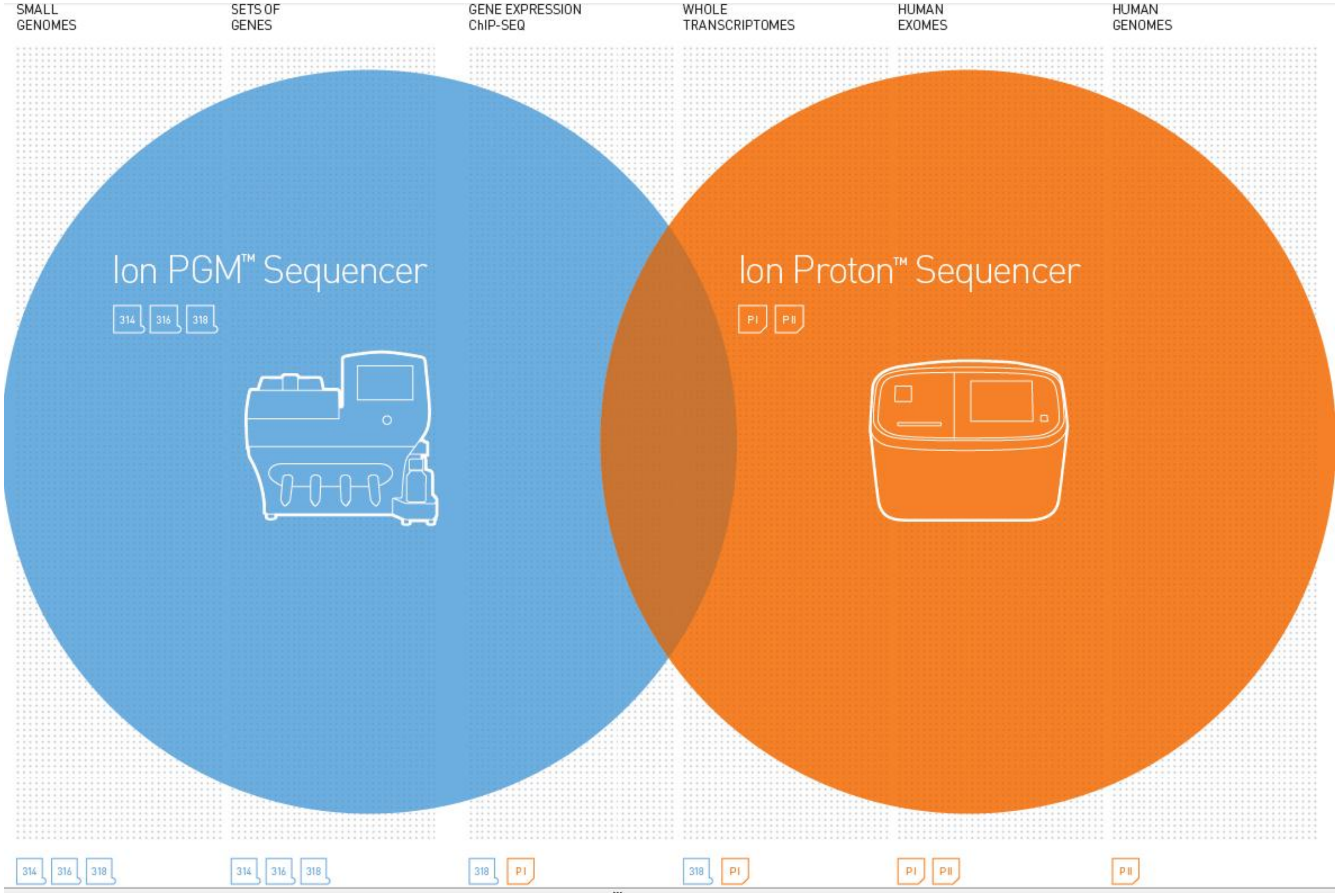
454-like chemistry without dye-labelled nucleotides

- 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**

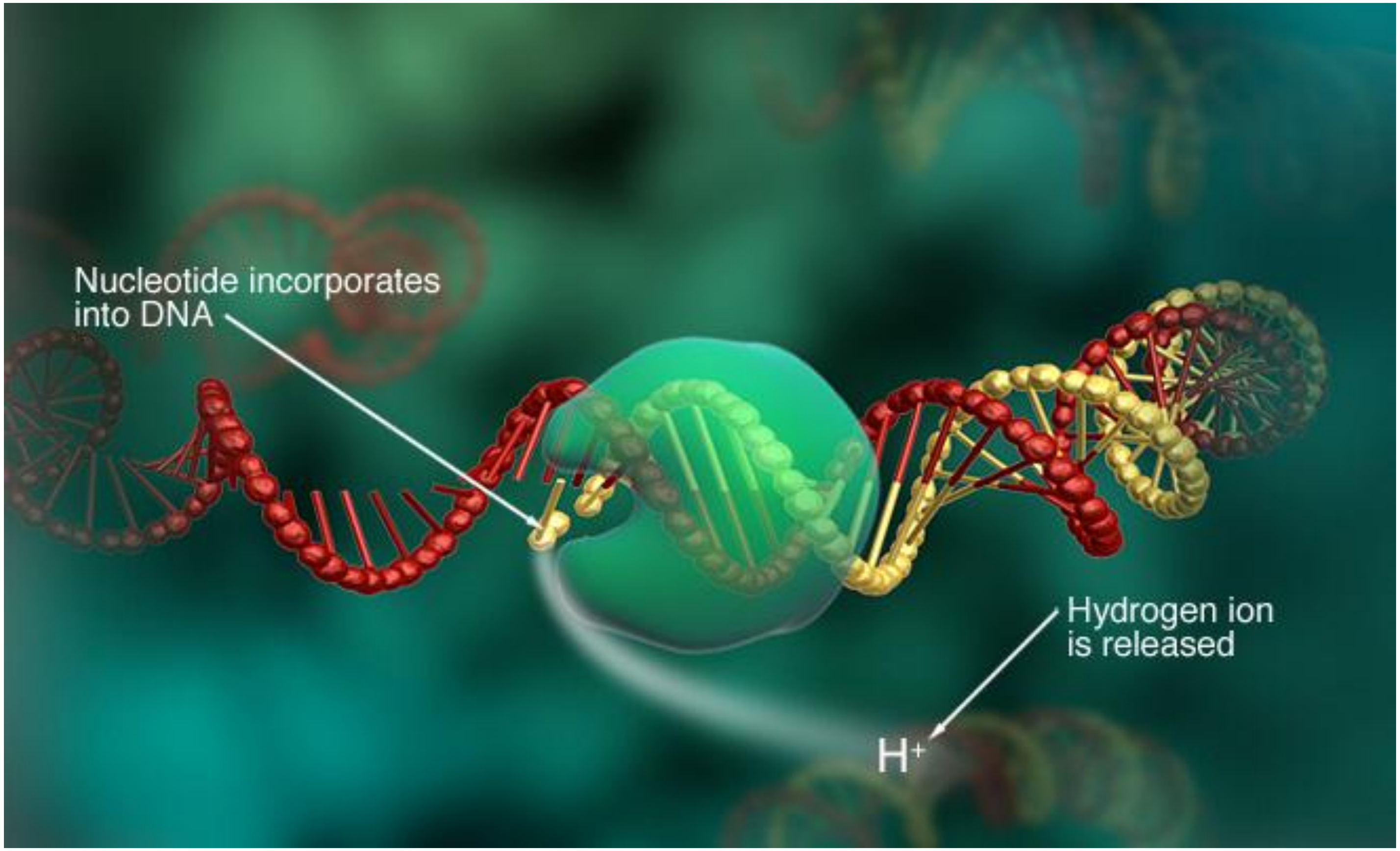
Life Technologies Ion Proton

- 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

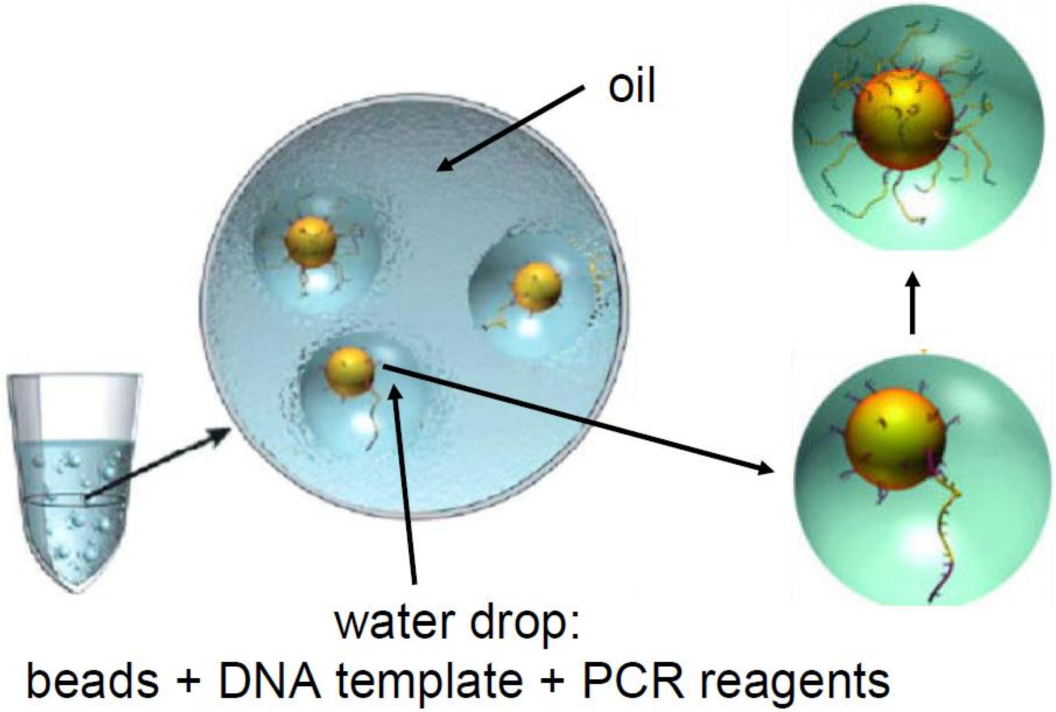


Ion Torrent

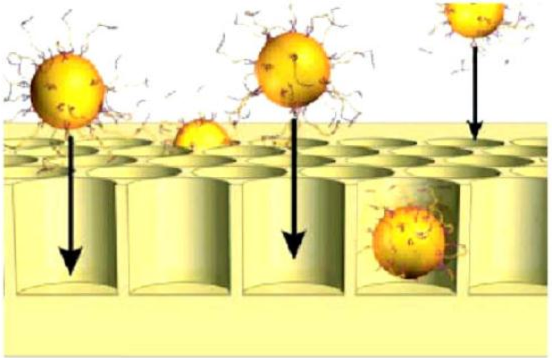

Nucleotide incorporates into DNA
Hydrogen ion is released
H+

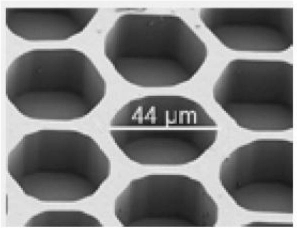
Library prep

- 454 style library using emulsion PCR

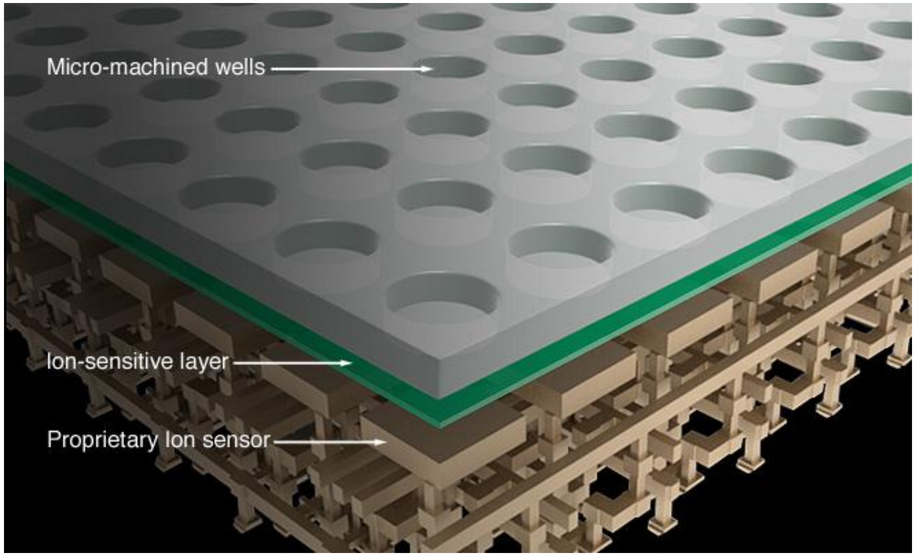


Ion Torrent

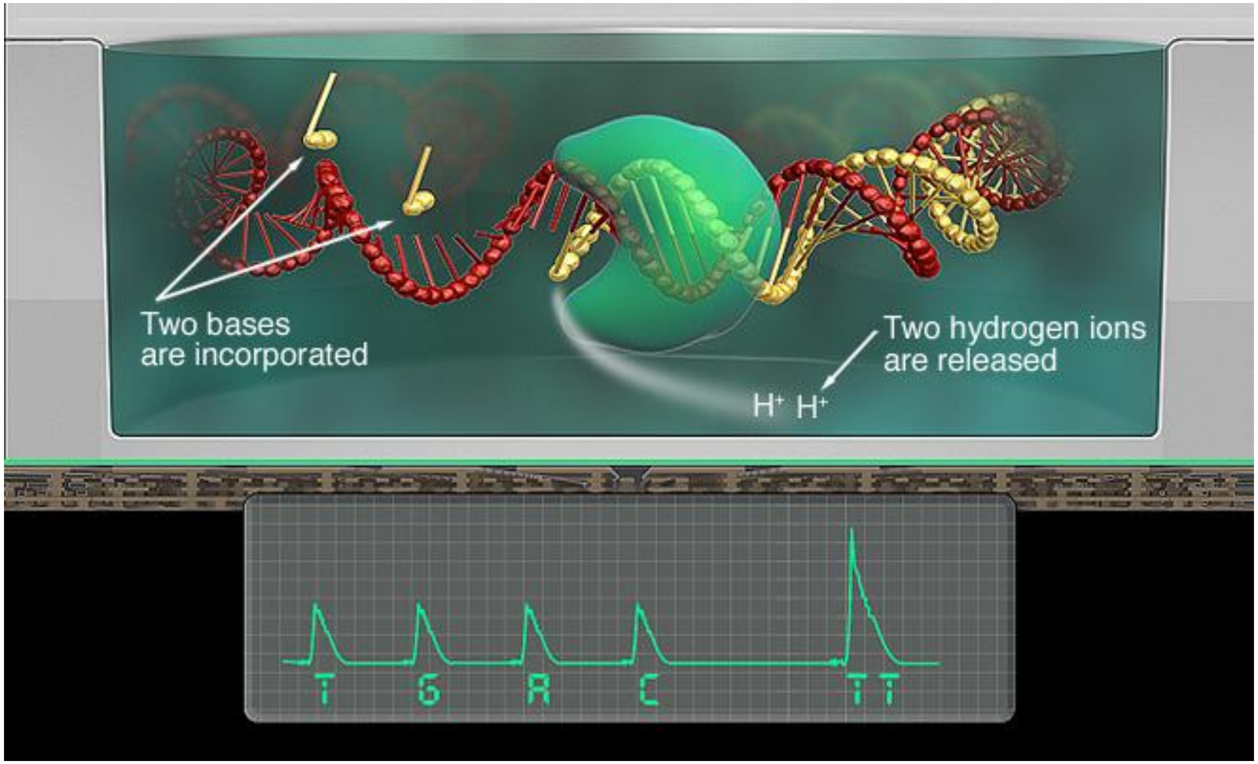




- enrich for DNA + beads
- diameter of the wells allows for only 1 bead/well



Ion System


Two bases
are incorporated
Two hydrogen ions
are released
H⁺ H⁺
T
G
A
C
TT

Benchtop sequencers

Ion Proton (P1 chip)

- 60-80M reads
- Up to single-end 200 base pair runs
- 16Gb/run
- 4 hour run time
- **$???/run**
- **$???K instrument**

- **One touch system required**

Illumina MiSeq

- 30M reads
- 2X250bp
- 7.5 Gbase/run
- Run 36 hours
- **$800 / 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 papers

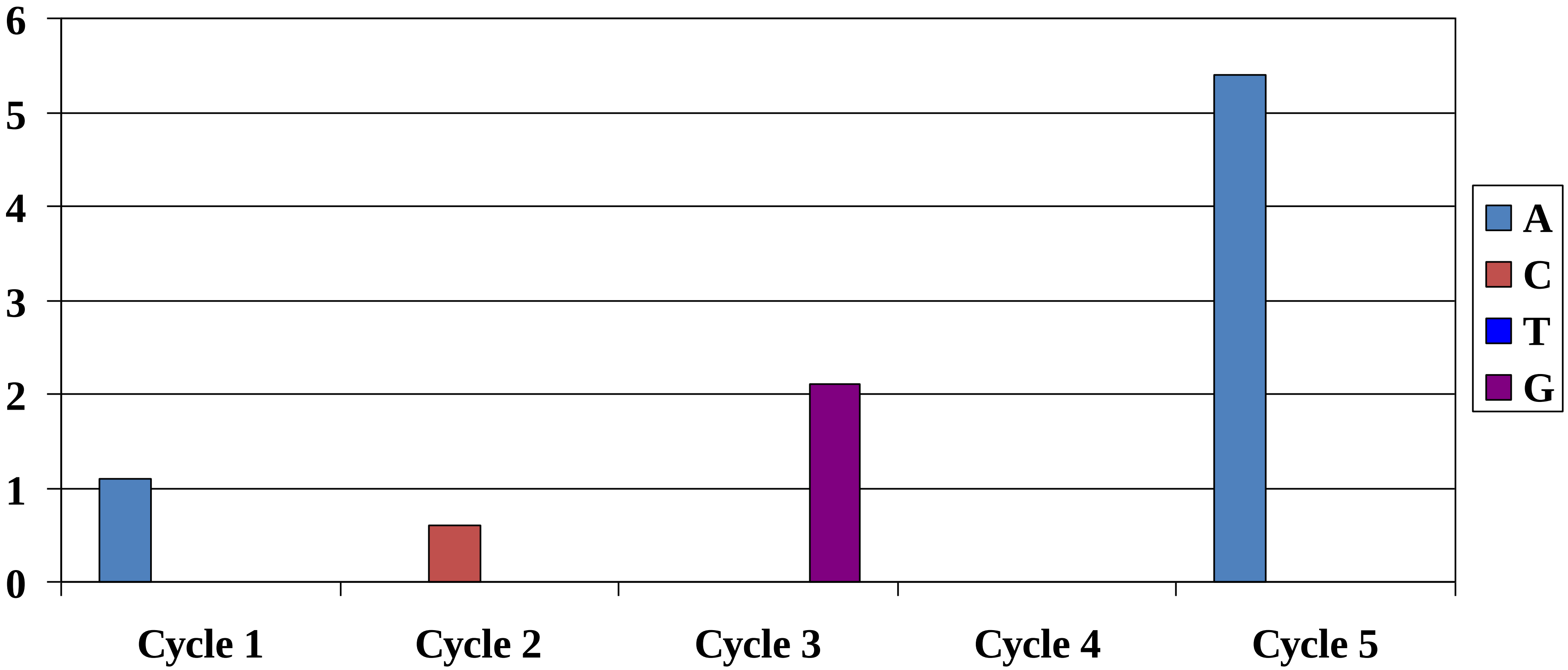
- Loman, N. J., Misra, R. V, Dallman, T. J., Constantinidou, C., Gharbia, S. E., Wain, J., & Pallen, M. J. (2012). Performance comparison of benchtop high-throughput sequencing platforms. *Nature biotechnology*, *30*(5), 434–9. doi:10.1038/nbt.2198

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
- 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



A **?c** **TT** **-** **AAAAA?a**

- Different between signal of 1 and signal of 2 = **100%.**
- Different between signal of 5 and 6 is **20%** so errors more likely after eg. AAAAA.

Third generation sequencers

Third generation sequencers

- My definition: Single-molecule sequencing
- Currently only PacBio RS is commercially available
- Others include Oxford Nanopore, GnuBio, Raindance

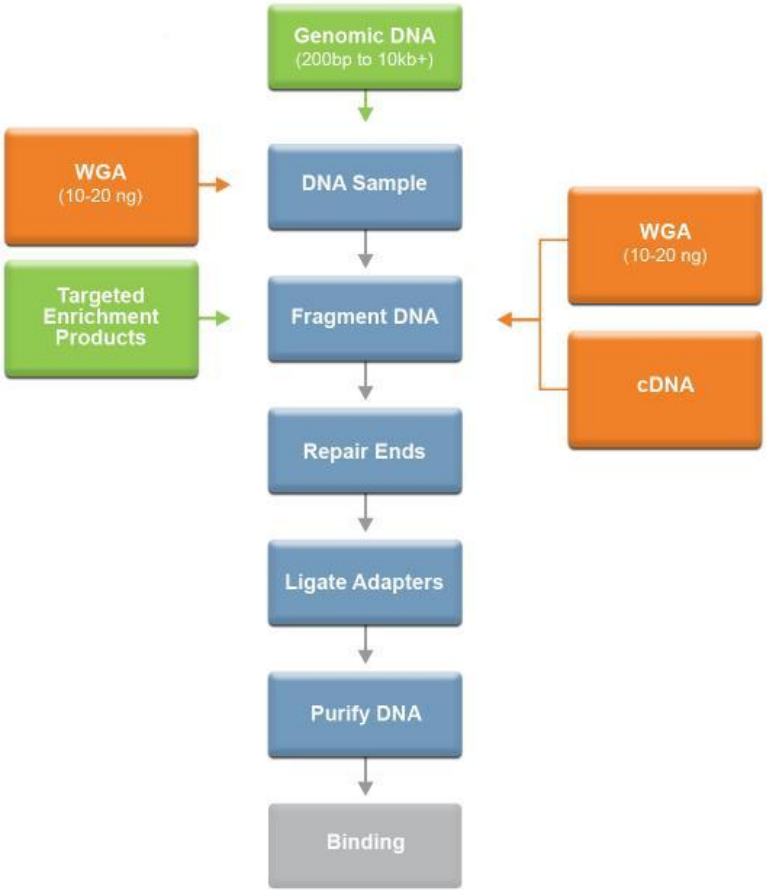
Pacific Biosciences RS

Introduction

- Based on monitoring a single molecule of DNA polymerase within a zero mode waveguide (ZMW)
- 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 flurorescent label for milliseconds -> nucleotide recorded

Library prep

Sample Preparation



Genomic DNA (200bp to 10kb+)

WGA (10-20 ng)

DNA Sample

Targeted Enrichment Products

Fragment DNA

WGA (10-20 ng)

cDNA

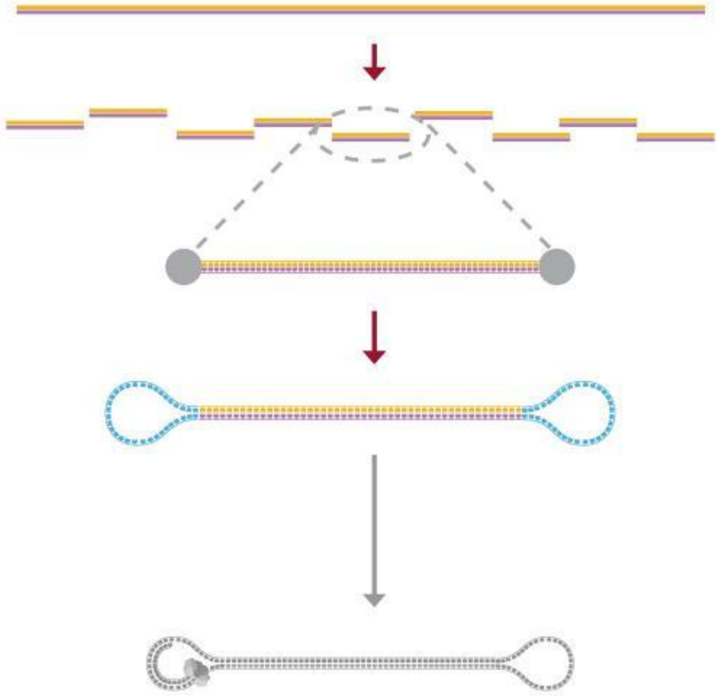
Repair Ends

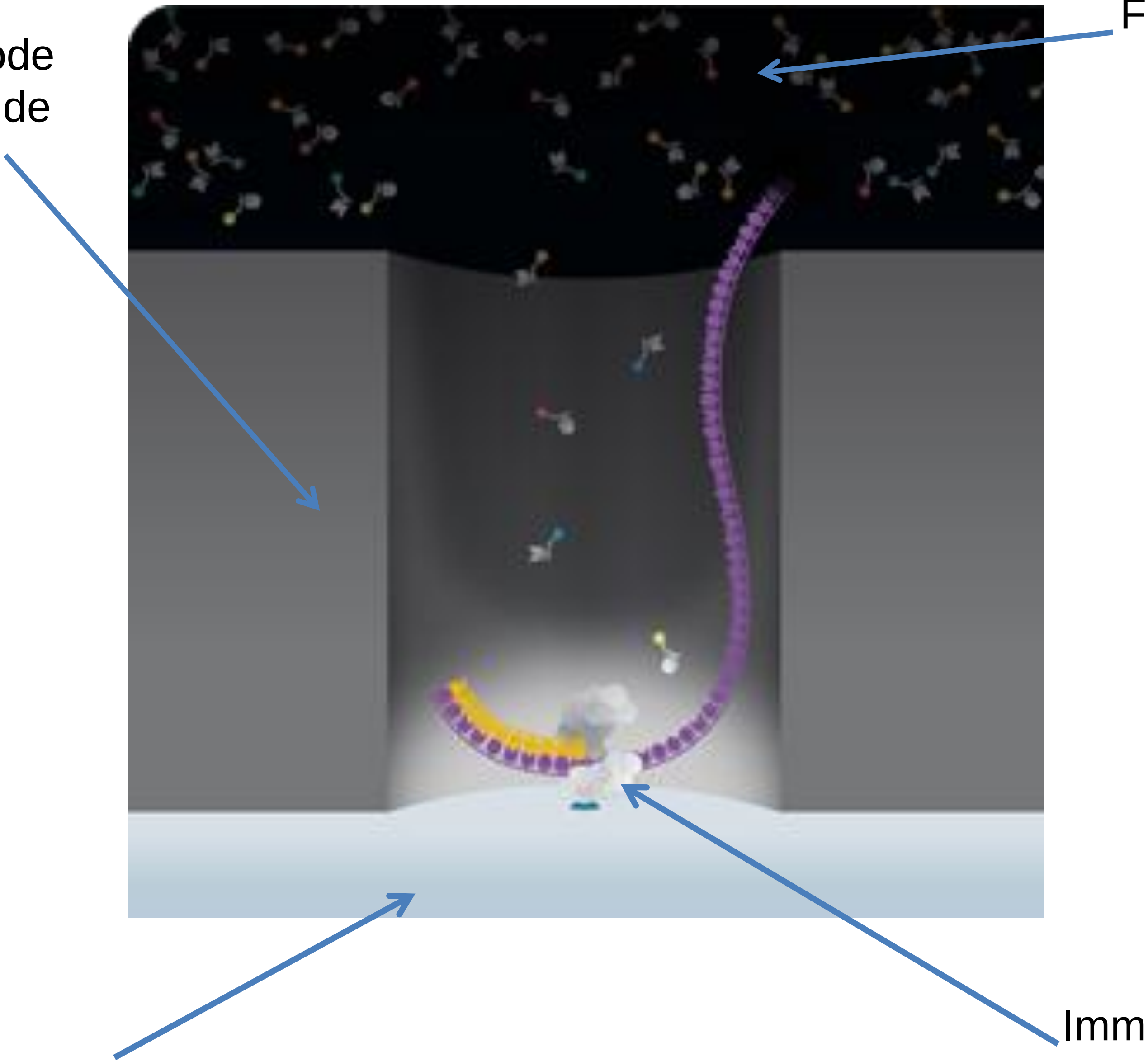
Ligate Adapters

Purify DNA

Binding

Building of SMRTbell





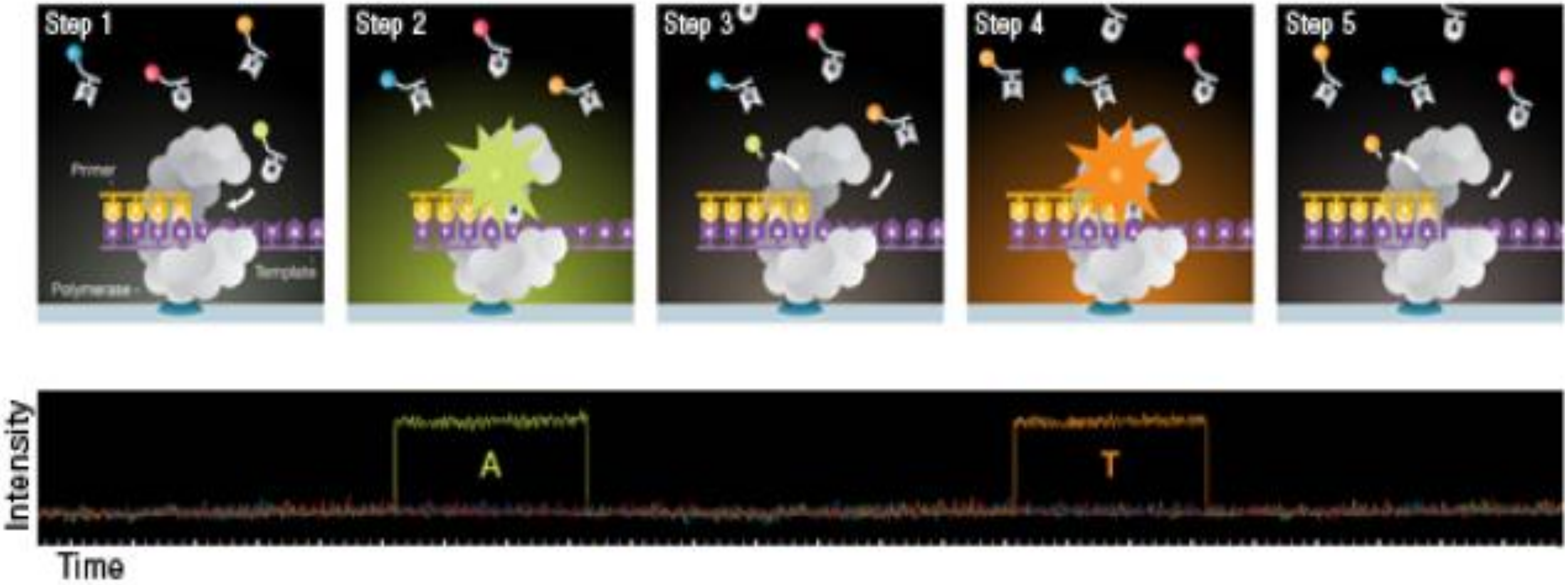
Zero mode waveguide

Free nucleotides

Immobilised DNA polymerase

Laser and detector

Observing a single polymerase

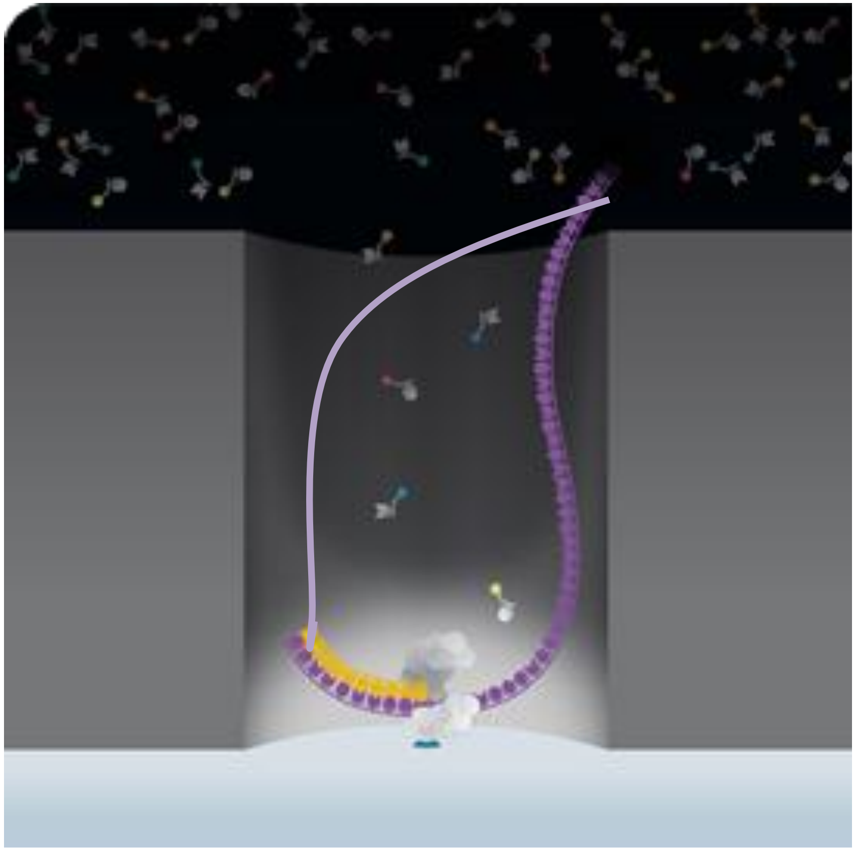


Novel applications

- Epigenetic changes (e.g. Methylation) affect the amount of time a fluorophore is held by the polymerase

- Circularise each DNA fragment and sequence continuously

Circular sequencing



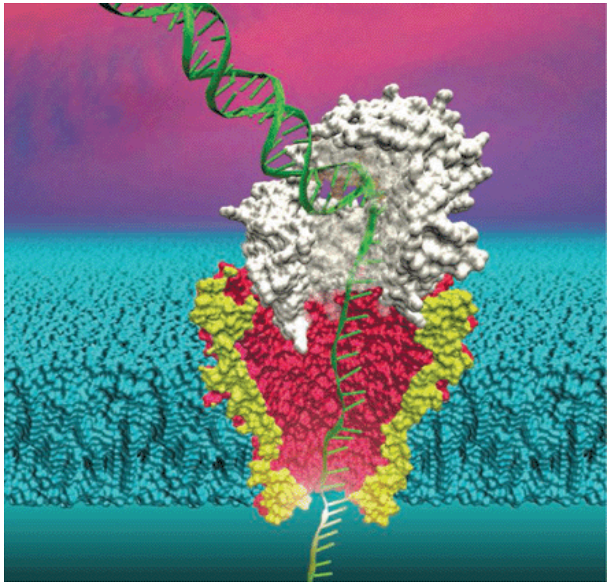
Pacific Biosciences

- Advantages
  - Longer reads lengths (200bp-10kb) (but only 200-500bp initially)
  - 40 minute run time
  - Same molecule can be sequenced repeatedly
  - Epigenetic modifications can be detected
- Disadvantages
  - Library prep required (but only 10-20ng needed)
  - Enzyme based
  - Only 20k-75k reads per run initially (~10-100Mb yield)
  - High (15%) error rate per run (but multiple runs reduce this)
  - $750k machine plus expensive reagents

Bioinformatics Implications

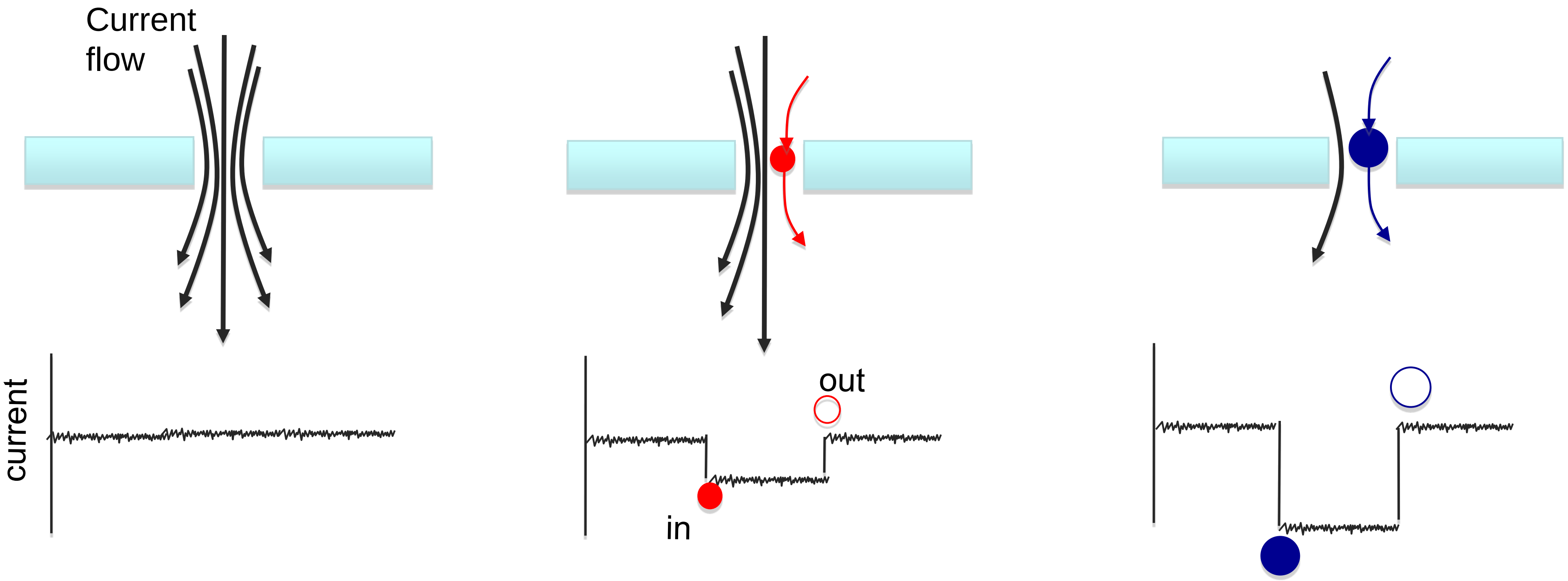
- Relatively low data output limits practical widespread use
- Can obtain some 10kb fragments
- Best used in conjunction with Illumina reads to correct high error rate

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



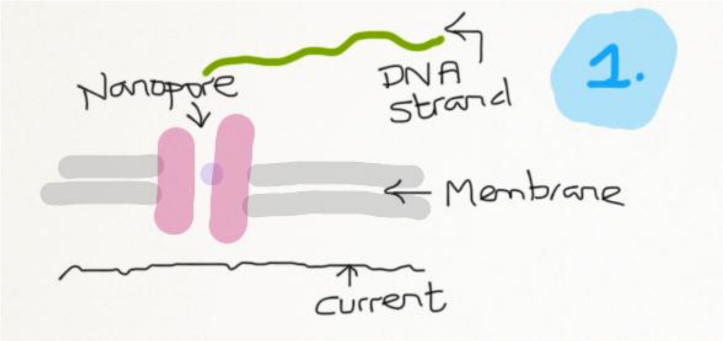
What is a nanopore?

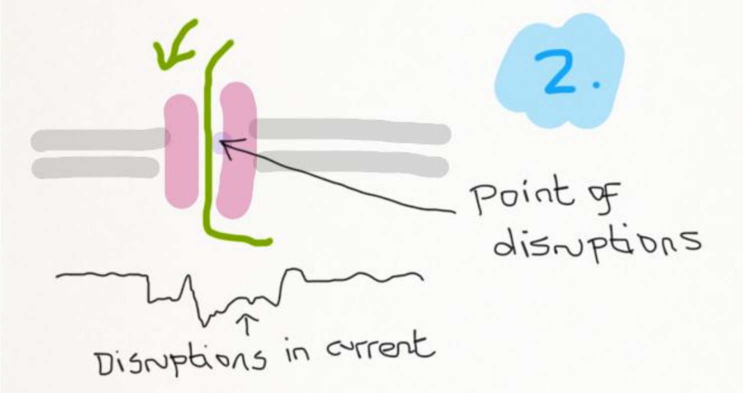
- Either biological or synthetic

- Biological
  - Lipid bilayers with alpha-haemlolysin 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

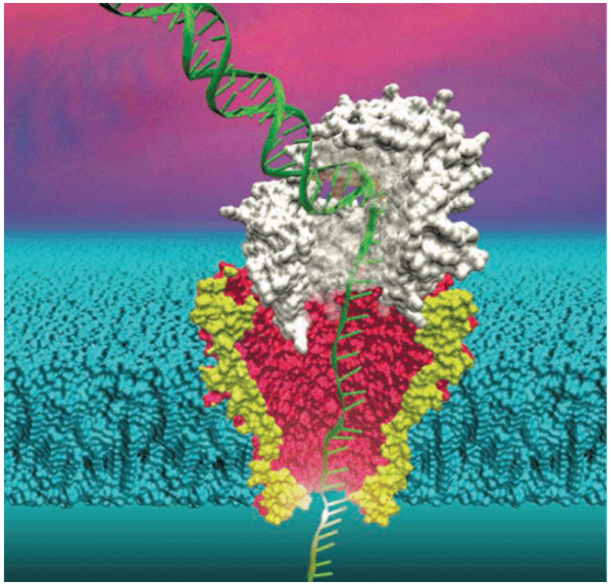




http://thenerdyvet.com/category/tech/

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



Approaches to simplify nanopore sequencing

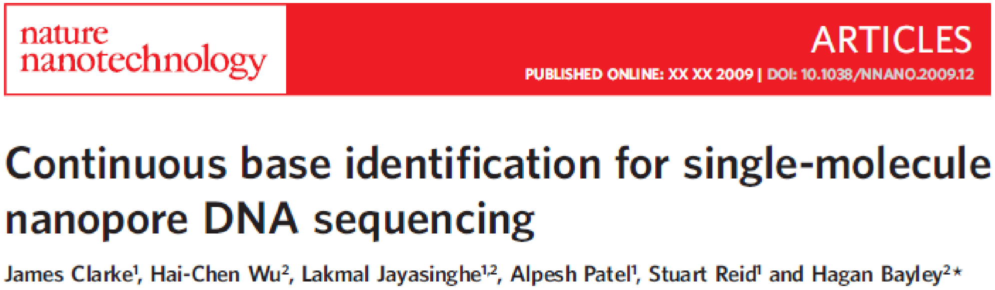
- Slow down movement of bases through nanopore
  - Use an enzyme to chop DNA up and sequence individual bases as they pass through a pore
  - And/or use an enzyme to slow the progress of DNA through a pore
  - Monitor capacitative changes in the bilayer
- Hybridize labels to single stranded DNA
  - Force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

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

Oxford nanopore

- Company which appears closest to commercialisation
- Two approaches to sequencing
  - Strand sequencing
  - Exo-nuclease sequencing
- Both use synthetic membranes compatible with alpha-haemolysin derived pores

Nucleotide Recognition

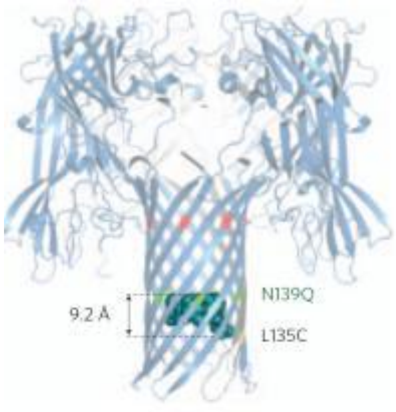


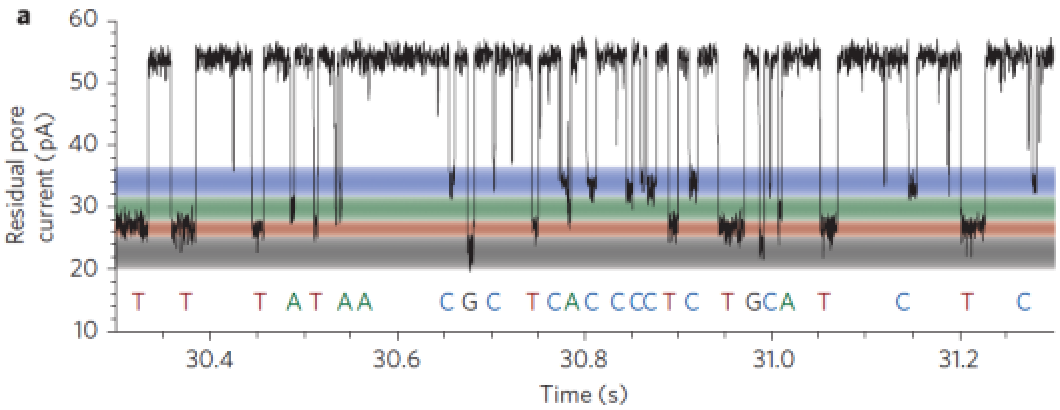
nature nanotechnology | ARTICLES

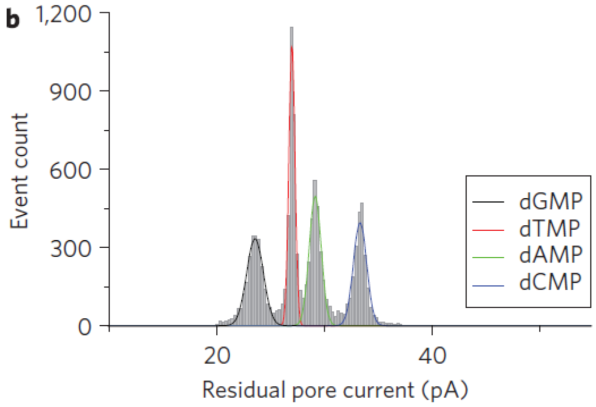
PUBLISHED ONLINE: XX XX 2009 | DOI: 10.1038/NNANO.2009.12

Continuous base identification for single-molecule nanopore DNA sequencing

James Clarke[1], Hai-Chen Wu[2], Lakmal Jayasinghe[1,2], Alpesh Patel[1], Stuart Reid[1] and Hagan Bayley[2]*

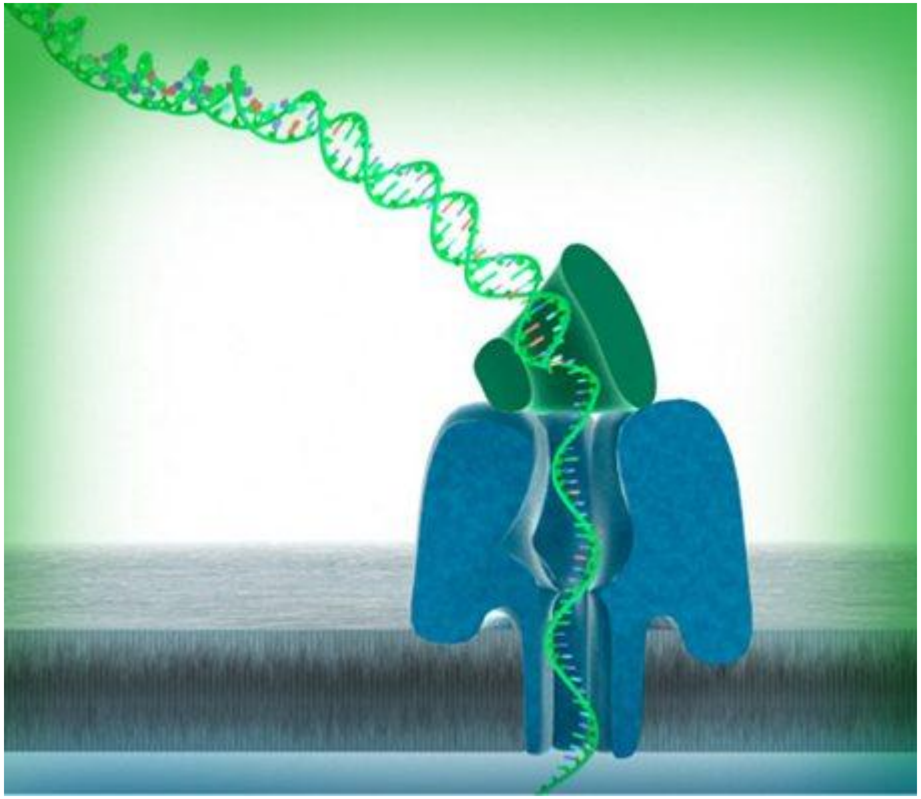




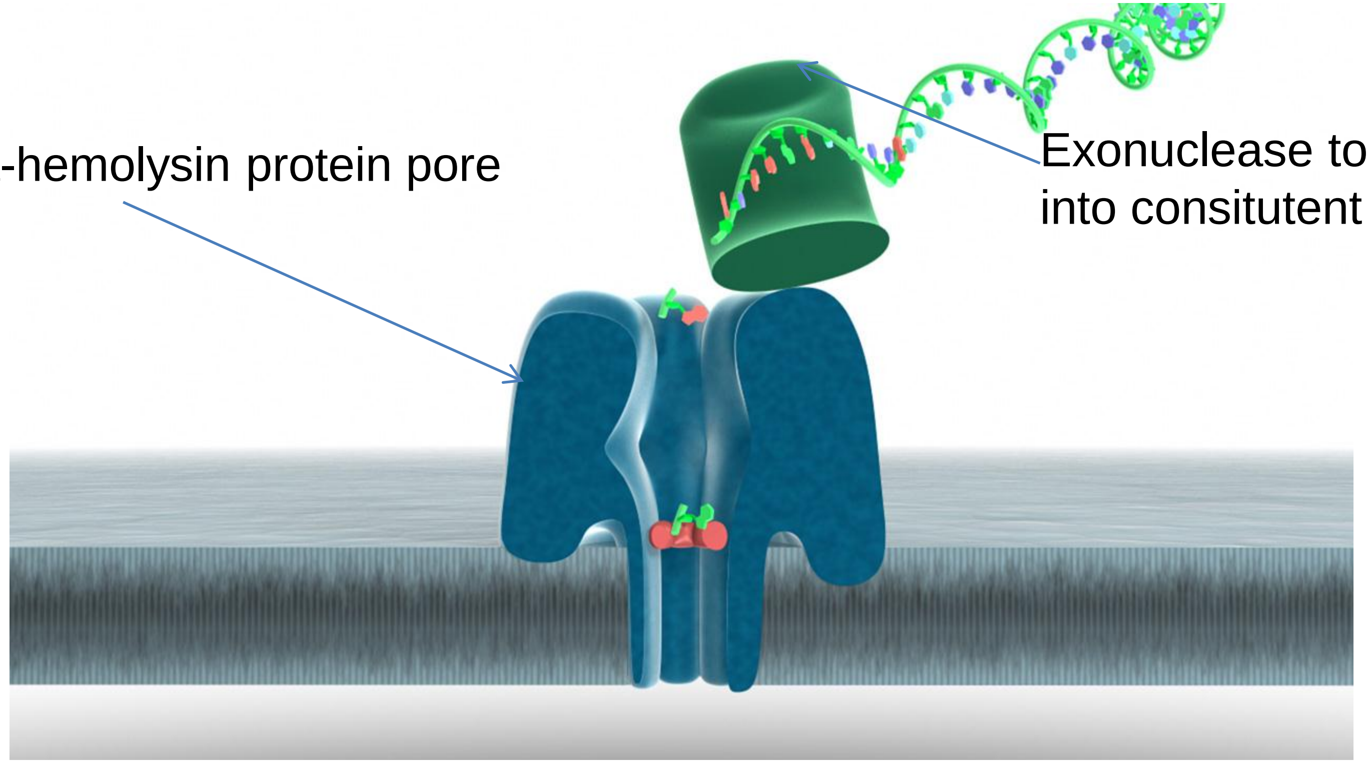


Strand-sequencing

- Used in the recently advertised GRIDIon and MinIon systems

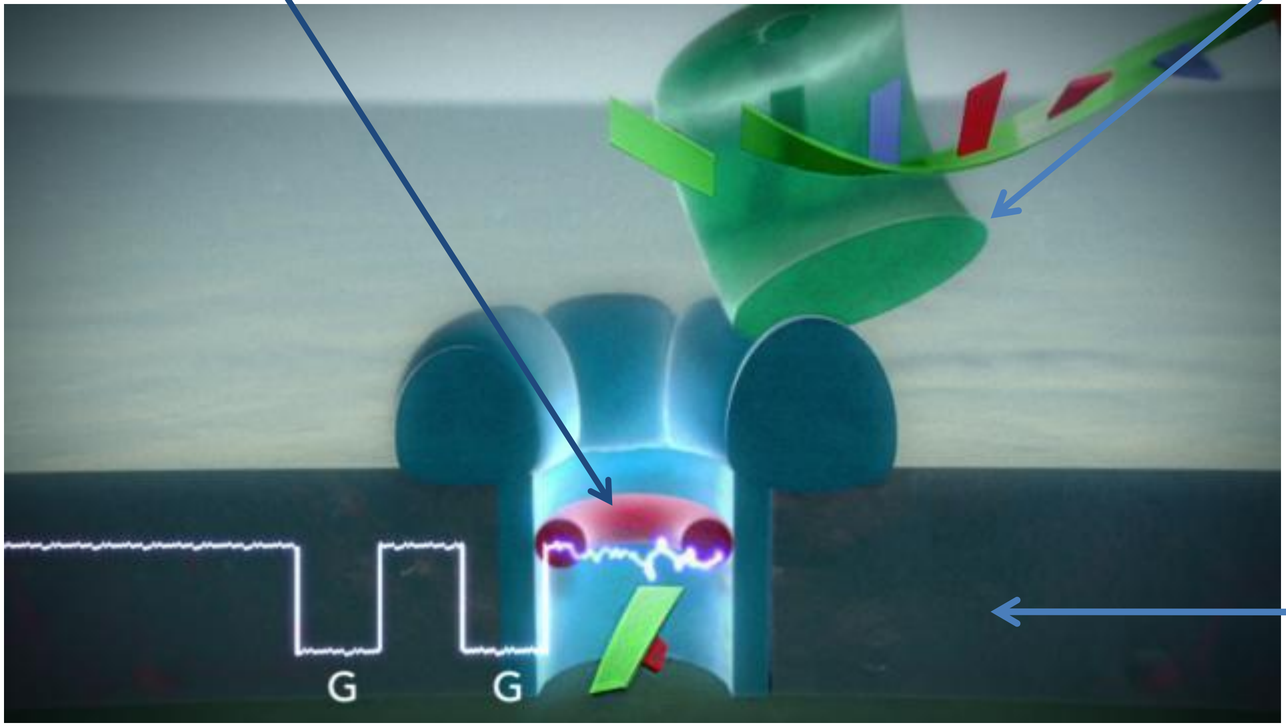


Exonuclease sequencing



Alpha-hemolysin protein pore

Exonuclease to chop DNA into consitutent nucleotides



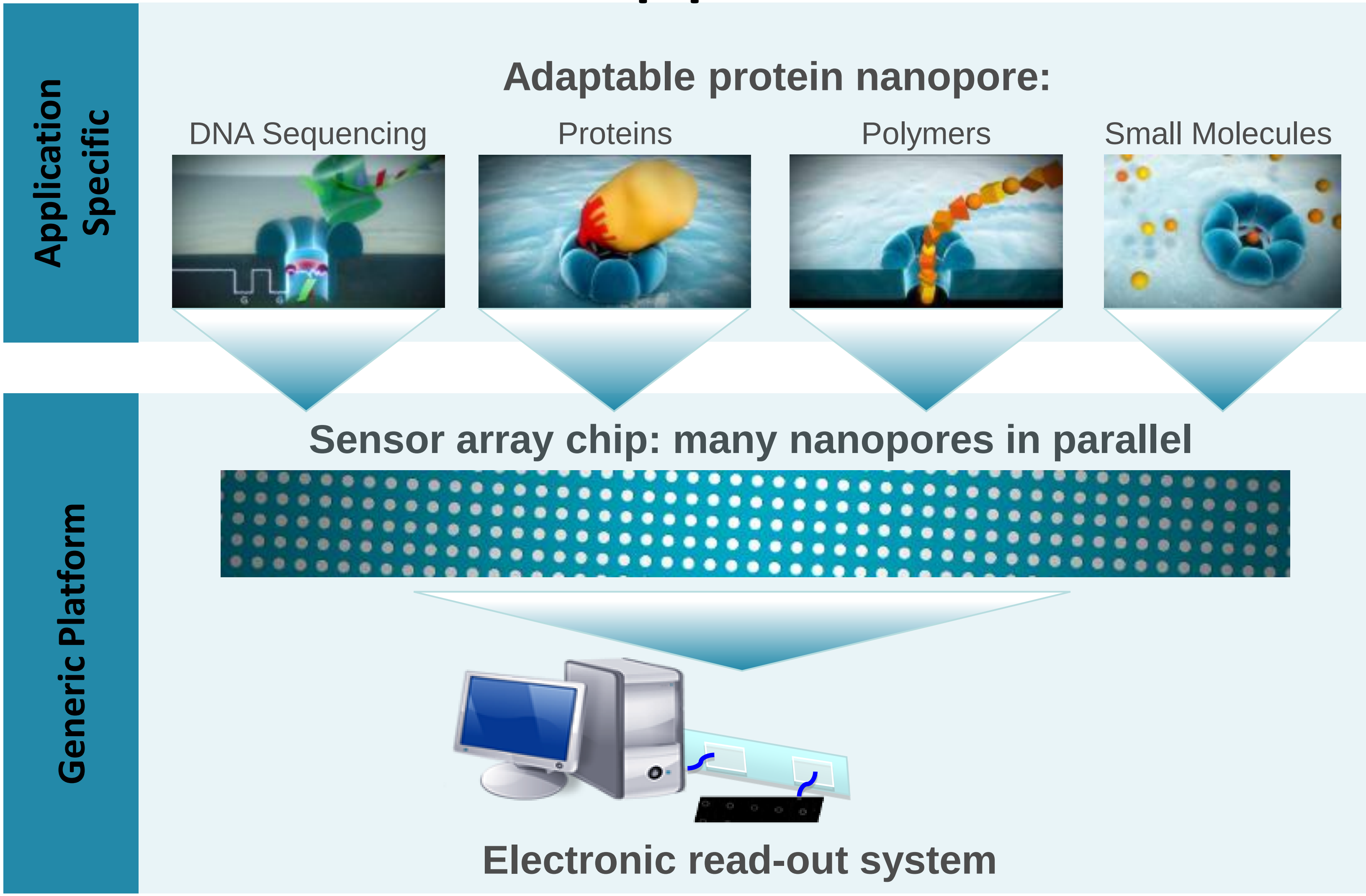
Cyclodextrin molecule inside alpha hemolysin

Exonuclease

Lipid bilayer

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

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

Platforms

- GridION for sequencing centres
  - Human genome in 2 hours for around $1000
  - No estimated pricing of instrument

- MinIon for individuals
  - $900 for 2000 pore chip
  - Assuming 10kb reads – 20Mb
  - 4% error rate in trials

Oxford nanopore

- Advantages
  - No library prep required
  - Long reads lengths (1kb-100kb)
  - Protein –> solid-state upgrades may eliminate reagent costs (3-5 years)
  - Fast turn around
  - Could measure epigenetic modifications and other molecules
- Disadvantages
  - Potentially non-stochastic errors (i.e. some sequences harder to sequence accurately)
  - Difficult to see how the same molecule could be sequenced repeatedly

Bioinformatics Implications

- Could prove to be yet another step change as with 2nd generation sequencing
- Can obtain >10kb fragments
- Error profiles will be crucial to determining success
- Longer read lengths may make alieviate some bionformatics headaches
- …it may lead to different bottlenecks

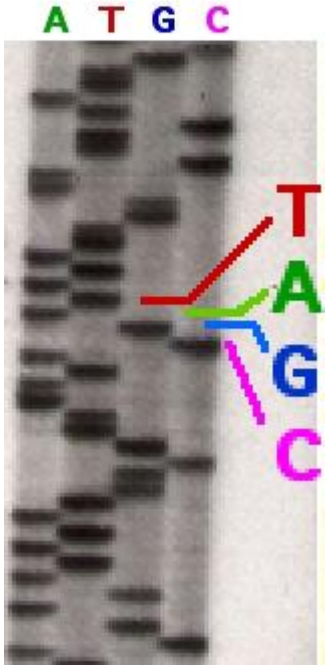
Useful papers/videos

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

Sequencing – back on the benchtop

Full circle

- 1980

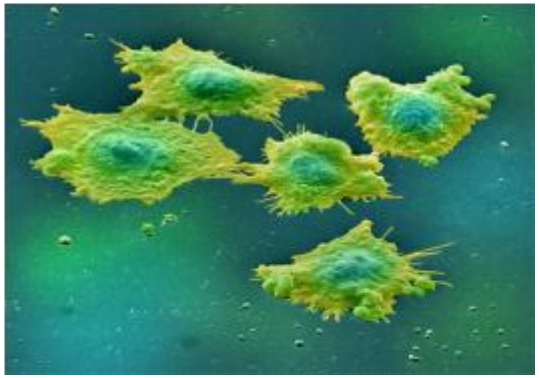


- 2000

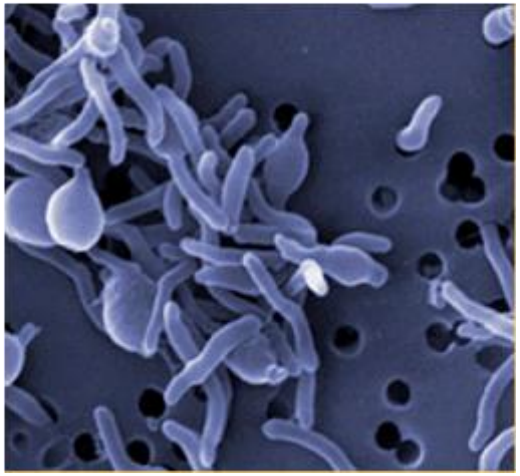
- 2020?

Ultimately: will we sequence every person?

Every cancer:
Accurate diagnosis and
targeted treatment?



Every baby:
Lifetime 'baseline' resource,
disease prevention?



Every infectious agent:
Control of disease
spread and resistance

Final note

- Sequencing means nothing without biological and environmental context
- The current revolution in sequencing may reveal that personalised medicine may not be the cure-all

Intratumor Heterogeneity and Branched Evolution Revealed by Multiregion Sequencing

Marco Gerlinger, M.D., Andrew J. Rowan, B.Sc., Stuart Horswell, M.Math., James Larkin, M.D., Ph.D., David Endesfelder, Dip.Math., Eva Gronroos, Ph.D., Pierre Martinez, Ph.D., Nicholas Matthews, B.Sc., Aengus Stewart, M.Sc., Patrick Tarpey, Ph.D., Ignacio Varela, Ph.D., Benjamin Phillimore, B.Sc., Sharmin Begum, M.Sc., Neil Q. McDonald, Ph.D., Adam Butler, B.Sc., David Jones, M.Sc., Keiran Raine, M.Sc., Calli Latimer, B.Sc., Claudio R. Santos, Ph.D., Mahrokh Nohadani, H.N.C., Aron C. Eklund, Ph.D., Bradley Spencer-Dene, Ph.D., Graham Clark, B.Sc., Lisa Pickering, M.D., Ph.D., Gordon Stamp, M.D., Martin Gore, M.D., Ph.D., Zoltan Szallasi, M.D., Julian Downward, Ph.D., P. Andrew Futreal, Ph.D., and Charles Swanton, M.D., Ph.D.

N Engl J Med 2012; 366:883-892March 8, 2012DOI: 10.1056/

Thanks to:

Audrey Farbos

Karen Moore

Christine Sambles

Wellcome Trust

Contact me:

k.h.paszkiewicz@exeter.ac.uk

Supported by
wellcome trust