UNIVERSITY OF EXETER

Modern approaches to sequencing

(and some history)

Konrad Paszkiewicz

January 2015

Contents

- History of DNA
- Review of first generation sequencing techniques
- Short read second generation sequencing technology
- Third generation single molecule sequencing
- Nanopore sequencing

A history of DNA

“The double helix is indeed a remarkable molecule. Modern man is perhaps 50,000 years old, civilization has existed for scarcely 10,000 years and the United States for only just over 200 years; but DNA and RNA have been around for at least several billion years.

All that time the double helix has been there, and active, and yet we are the first creatures on Earth to become aware of its existence.”

Francis Crick (1916–2004)

The first person to isolate DNA

- Friedrich Miescher
  - Born with poor hearing
  - Father was a doctor and refused to allow Freidrich to become a priest
- Graduated as a doctor in 1868
  - Persuaded by his uncle not to become a practising doctor and instead pursue natural science
  - But he was reluctant…

Friedrich Miescher

Biology PhD angst in the 1800s

"I already had cause to regret that I had so little experience with mathematics and physics… For this reason many facts still remained obscure to me."

His uncle counselled:

"I believe you overestimate the importance of special training…"

Friedrich Miescher

1869 - First isolation of DNA

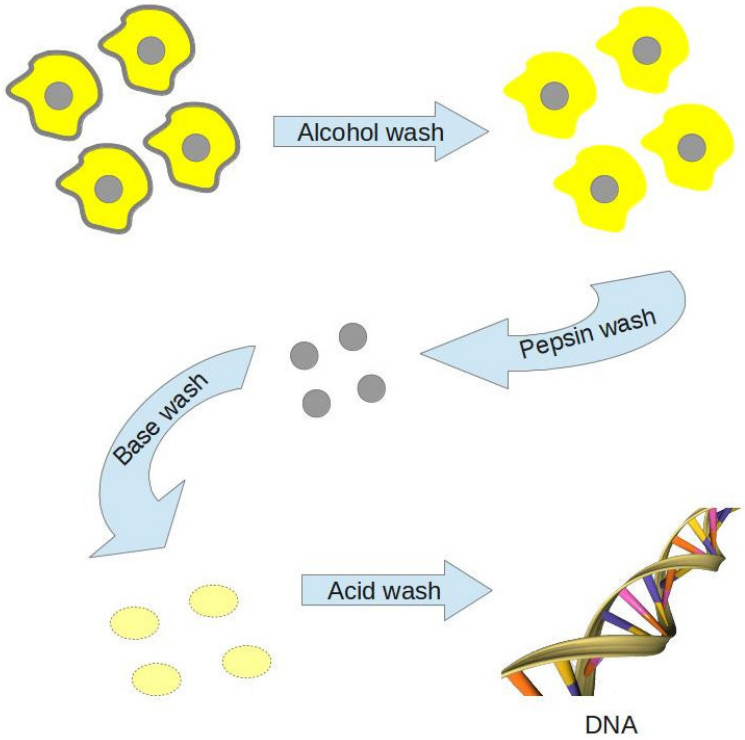
- Worked in Felix Hoppe-Seyler's laboratory in Tubingen, Germany
  - The founding father of biochemistry
- The lab was one of the first to crystallise haemoglobin and describe the interaction between haemoglobin and oxygen
- Freidrich extracted 'nuclein' on cold winter nights
  - Initially from human leukocytes extracted from bandage pus from the local hospital
  - Later from salmon sperm

Felix Hoppe-Seyler

Friedrich Miescher

Meischer's isolation technique

- Cells from surgical bandages or salmon sperm
- Alcohol to remove outer cell membrane
- Pepsin from pig stomachs
- Basic solution to dissolve nuclein in the nucleus
- Acid solution to precipitate the nuclein
- Difficult to do without also precipitating bound protein



http://www.howdoweknowit.com/2013/07/03/how-do-we-know-the-genetic-code-part-2/

Biology PhD angst in the 1800s

"I go at 5am to the laboratory and work in an unheated room. No solution can stand for more than 5 minutes… Often it goes on until late into the night."

His student remembered*:*

Friedrich failed to turn up for his own wedding. We went off to look for him. We found him quietly working in his laboratory.

Friedrich Miescher

1874 - First hints to composition

- By 1874 Meischer had determined that nuclein was
  - A four basic acid
  - High molecular weight
  - Nuclein was bound to ‘protamin’
- Came close to guessing its function
  - “If one wants to assume that a single substance is the specific cause of fertilisation, the one should undoubtedly first and foremost consider nuclein”
  - Discarded the idea because he thought it unlikely that nuclein could encode sufficient information

Friedrich Miescher

1881 - Discovering the composition of nuclein

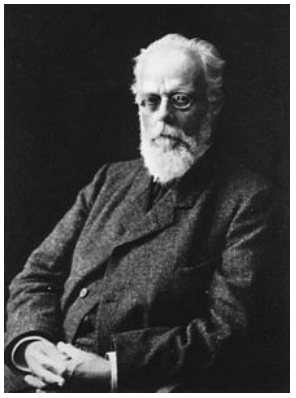
- Kossel worked in the same lab as Freidrich Miescher

- Discovered fundamental building blocks of nuclein
  - Purine and pyrimidine bases, sugar and phosphoric acid
  - Identified histone proteins and that nuclein was bound to histone in the nucleus
  - Inferred that nuclein was not used for energy storage but was linked to cell growth

Albrecht Kossel

1890s - Molecular basis of heredity

- Lots of theories
  - Stereo-isomers
  - Asymmetric atoms
  - Complex molecules
- Realisation that hereditary information is transmitted by one or more molecules
- 1893 August Weismann – germ plasm theory
- 1894 Eduard Strasburger- “nuclei from nuclei”



August Weismann

Eduard Strasburger

1900 - What we knew

Known

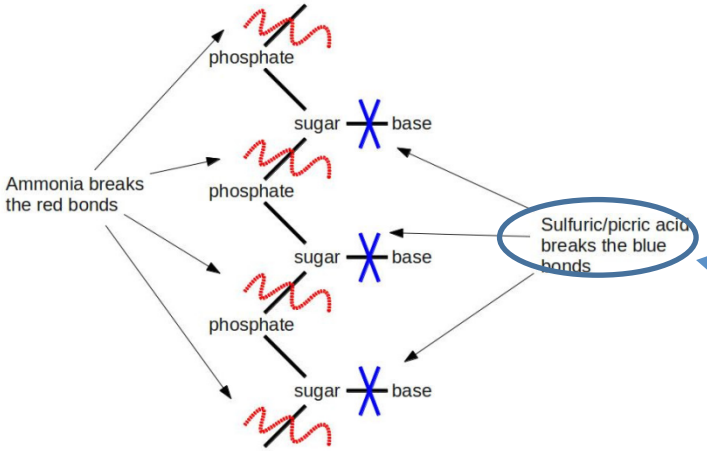
- Distinction between proteins and nucleic acids
- Somehow nuclein was involved in cell growth
- Somehow the nucleus was involved in cell division

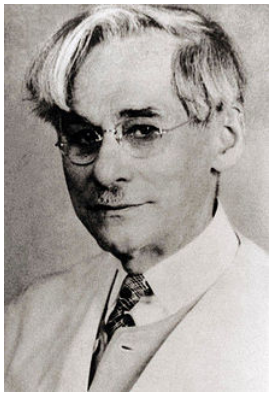
Unknown

- Mendel's lost laws
- Base composition of nucleic acids
- Role of the nucleus
- Distinction between RNA and DNA
- Significance of chromosomes
- That enzymes were proteins
- Most of biochemistry

1910s - More on the composition of DNA

- Determined relative composition of sugars, phosphate and sugars by hydrolysis of nucleic acid





Phoebus Levene

Close relative of TNT explosive

- Enabled the discovery of DNA and RNA bases
- Unfortunately, this method can destroy bases and bias results
- Made it impossible to compare composition between species
- Phoebus Levene proposed the tetranucleotide hypothesis
  - DNA consisted of repeating units of thymine, guanine adenine and cytosine
  - E.g. GACT GACT GACT
  - Convinced many that DNA could not be a carrier of hereditary information
  - Led to the assumption that DNA was just a structural component of cells

1910-30s - Chromosome theory of heredity

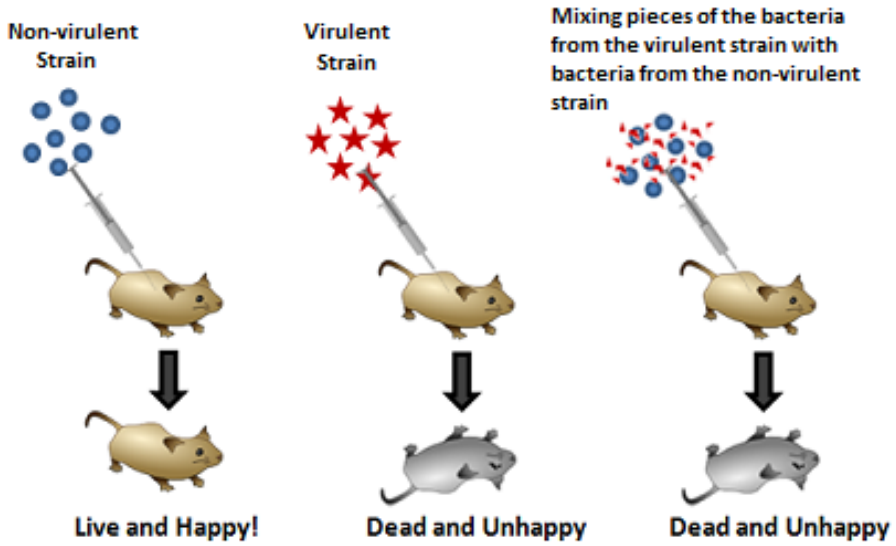
- Chromosome as a unit of heritability confirmed by Thomas Morgan
- Alfred Sturtevant creates the first genetic linkage map
- Genetic recombination shown to be caused by physical recombination of chromosomes by McClintock & Creighton

Thomas Morgan

Barbara McClintock

1928 - Inheritance of virulence

- Established that non-virulent pneumococci bacteria could be converted be made virulent by exposure to lysed virulent bacteria



- What was the 'transforming principle' which underlay this observation?

http://mic.sgmjournals.org/content/73/1/1.full.pdf

Frederick Griffiths

"Could do more with a kerosene tin and a primus stove than most men could do with a palace"

Hedley Wright

1944 – What is life?

- An ‘aperiodic solid crystal’ could code for an organism
- “A well-ordered association of atoms endowed with sufficient resistivity to keep its order permanently”
- Also placed living systems into a thermodynamic framework

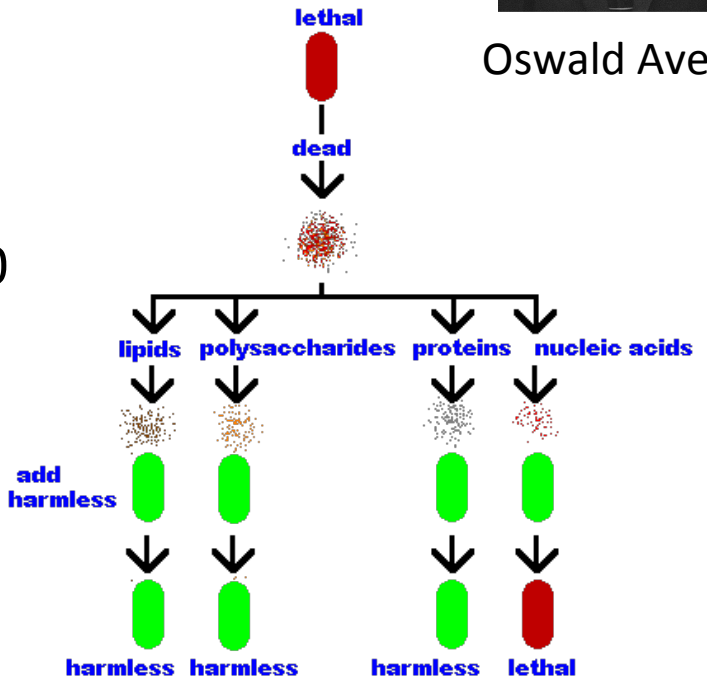
- Inspiration for Watson & Crick

Erwin Schrodinger

1944 – Establishing DNA as the transforming principle

- Separated cellular components and repeated Griffiths experiments
- Enabled by new 'ultra-centrifugation' technology
- Extended Griffiths work to prove that nucleic acids were the 'transforming principle'
- Also demonstrated that DNA, not RNA was the genetic material
- Incredibly small amounts – 1 in 600 million were sufficient to induce transformation

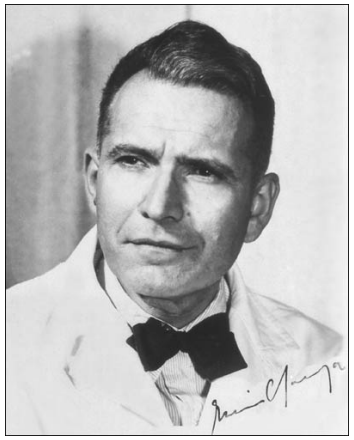
Oswald Avery



http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2135445/

1950 – Base composition between organisms

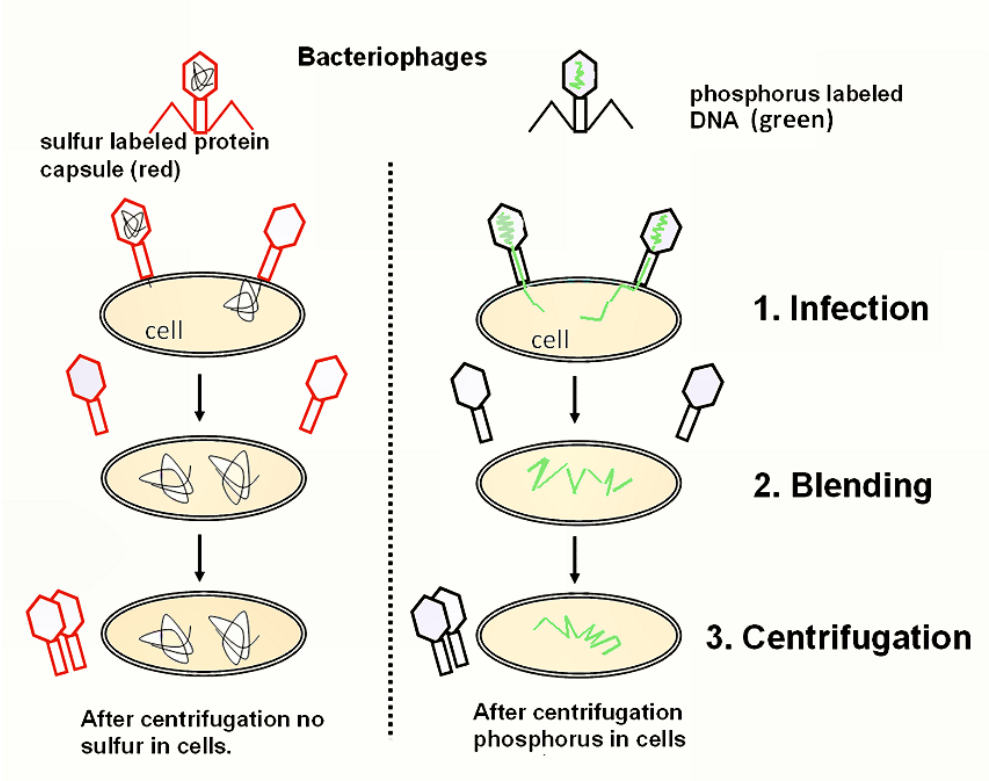
- Developed the base complementarity hypothesis
- Determined that the molar ratio of A/T and G/C were always very close to 1
- Relative proportions of bases varied between species but was the same within species
- Refuted Levene's 30 year-old tetranucleotide hypothesis



Erwin Chargaff

1952- Confirmation of Avery's experiment

- Bacteriophages infected bacteria by injecting DNA, not protein
- Finally confirmed the role of DNA as genetic material



Hershey Chase experiment

http://en.wikipedia.org/wiki/Hershey%E2%80%93Chase_experiment

1952 – X-ray diffraction patterns of DNA

- Wilkins, Franklin and Gosling
- Much improved X-ray diffraction patterns of the B-form of DNA
- Wilkins developed a method to obtain improved diffraction patterns using sodium thymonucleate to draw out long thin strands of DNA

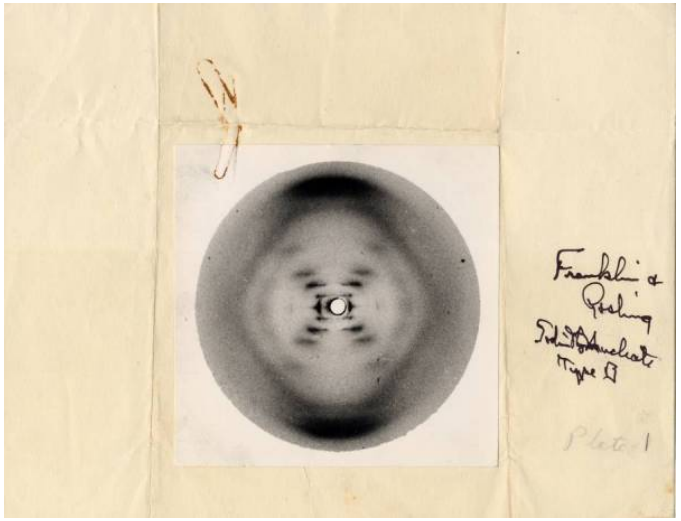
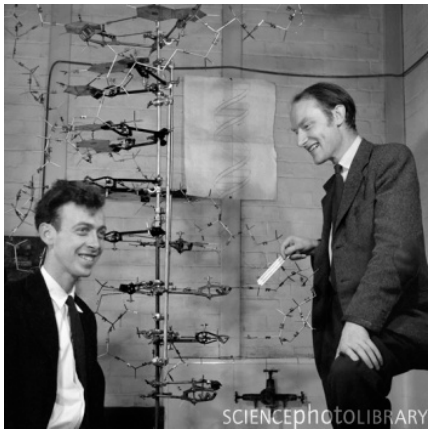


Photo Number 51

http://paulingblog.wordpress.com/2009/07/09/the-x-ray-crystallography-that-propelled-the-race-for-dna-astburys-pictures-vs-franklins-photo-51/

1953 – Watson & Crick obtain a structure for DNA

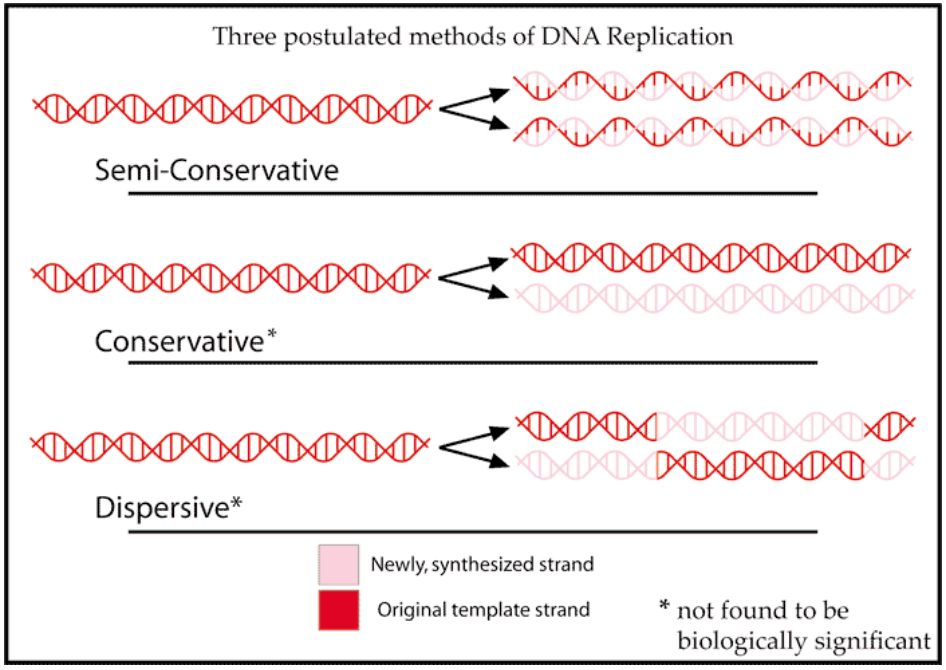
- B-model of DNA
- Relied upon data from Maurice Wilkins and Rosalind Franklin via Maz Perutz
- *"It has not escaped our notice that the specific pairing we have postulated immediately suggests a possible copying mechanism for the genetic material."*
- Broad acceptance of the structure did not occur until around 1960



Francis Crick & James Watson

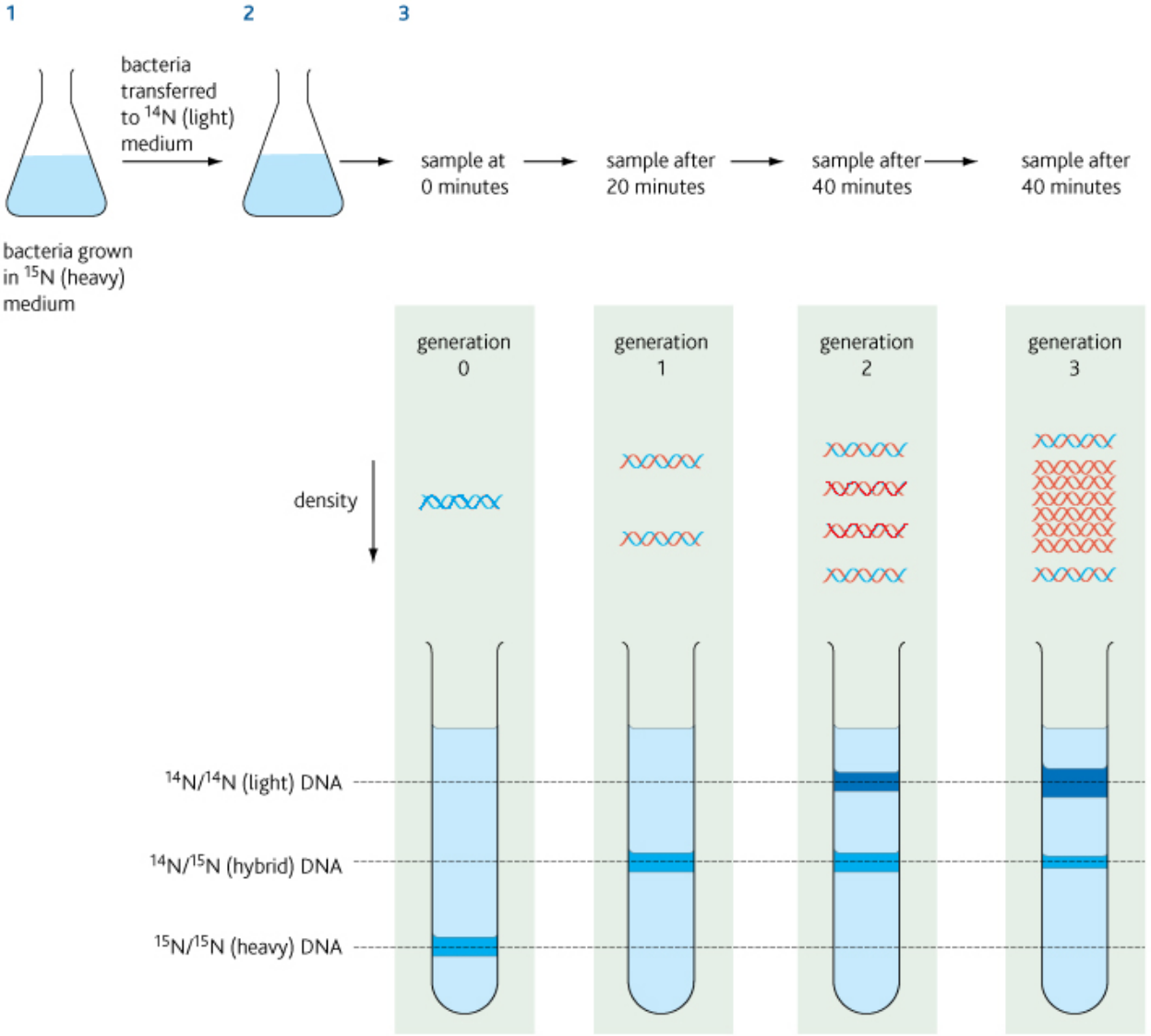
1958 – Evidence for the mechanism of DNA replication

- Meselson & Stahl
- Supported Watson & Crick's hypothesis of semi-conservative DNA replication



http://en.wikipedia.org/wiki/Meselson%E2%80%93Stahl_experiment

1958 – Evidence for the mechanism of DNA replication

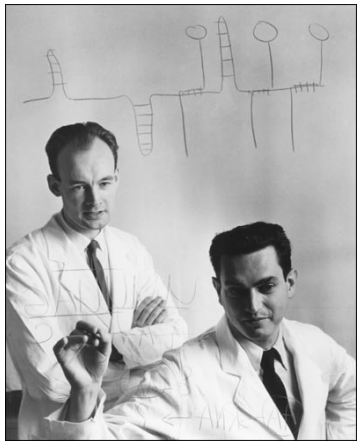


Other developments in molecular biology

- 1954 - George Gamow proposed a 3-letter code
- 1955 – Polynucleotide phosphorylase discovered
  - Enabled synthesis of homogeneous nucleotide polymers
- 1957 – Crick lays out 'central dogma'
- 1957-1963
  - RNA structure
  - Work on DNA-RNA hybridization
- 1960s
  - Crystal structures of tRNAs
  - Role in protein synthesis
  - Role of ribosomes

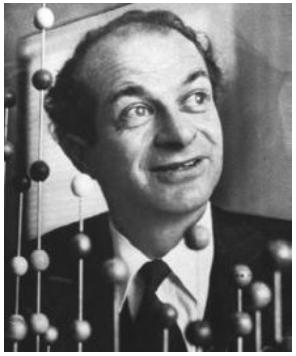
1961 - Deciphering the genetic code

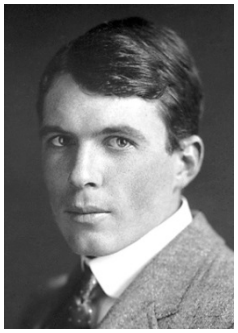
- How did DNA code for proteins?
- Nirenberg and Matthaei
- Used polynucleotide phophorylase to construct a poly-uracil polymer
- Added to a cell-free system containing ribosomes, nucleotides, amino acids, energy
- This produced an amino acid chain of phenylalanine
- Completed in mid 1960s by Har Gobind Khohrana



Other key figures

- Max Delbruck
  - Physicist who helped found molecular biology
- Salvador Luria
  - James Watson's PhD supervisor
  - Demonstrated with Delbruck that inheritance in bacteria was Darwinian and not Lamarkian
- Linus Pauling
  - Proposed triple helix model for DNA
- Lawrence Bragg
  - Hosted Watson & Crick
  - Rival of Pauling's
- Jerry Donohue, William Astbury, Raymond Gosling, John Randall, Fred Neufield, Herbert Wilson…





1962

- Nobel Prize awarded for Physiology or Medicine to Watson, Crick and Wilkins
- Rosalind Franklin died in 1958 of suspected radiation induced cancer

Russia's Usmanov to give back Watson's auctioned Nobel medal

The medal sold at auction for £3m

Russia's richest man has revealed that he bought US scientist James Watson's Nobel Prize gold medal, and intends to return it to him.

Related Stories

First generation sequencing

The development of sequencing methodologies

- What do we mean by ‘sequencing’?

- Determining the order and identity of chemical units in a polymer chain
  - Amino acids in the case of proteins
  - Nucleotides in the case of RNA and DNA

- Why do we do it?
  - 3D structure and function is dependent on sequence

1949 – Amino acids

- Sequenced bovine insulin
- Developed a method to label N-terminal amino acids
  - Enabled him to count four polypeptide chains
- Used hydrolysis and chromatography to identify fragments

Fred Sanger

1965 - RNA sequencing and structure

- Sequenced transfer RNA of alanine
- Used 2 ribonuclease enzymes to cleave the enzyme at specific motifs
- Chromatography
- 1968 Nobel prize

Robert Holley

Pancreatic RNase fragments: pG–G–G–C | G–U | G–U | G–G–C | G–U | A–G–U | A–G–C |

G–G–U | C | G–C | G–C | C | C | U | C | U | U | I–G–C |

Taka-Diastase T1 fragments: pG | G | G | C–G | U–G | U–G | G | C–G | U–A–G | U–A–G | C–G | G | U–C–G | C–G | C(C,C,U)C–U–U–I | G | C–

RNase (*continued*) G–G–T | ψ | C | G–C | I–ψ | G–G–G–A–G–A–G–U* | C | U | C | C |

G–A–U | U | C | C | G–G–A–C | U | C | G–U | C | C | A–C | C | AOH

T1 (*continued*) G | G | T–ψ–C–G | C–I–ψ–G | G | G | A–G | A–G | U*(C,U)C–C–G |

A–U–U–C–C–G | G | A–C–U–C–G | U–C–C–A–C–C–AOH

FIG. 7. One of many possible arrangements of the pancreatic RNase and RNase T1 digest fragments that shows the overlaps between the two digests. The RNA molecule is accounted for by the 16 oligonucleotide sequences indicated by the *solid lines*. Only the positions of the two terminal sequences are known. *Vertical lines* indicate the position of enzymatic attack. The *asterisk* indicates that the uridine may be partially substituted by DiHU.

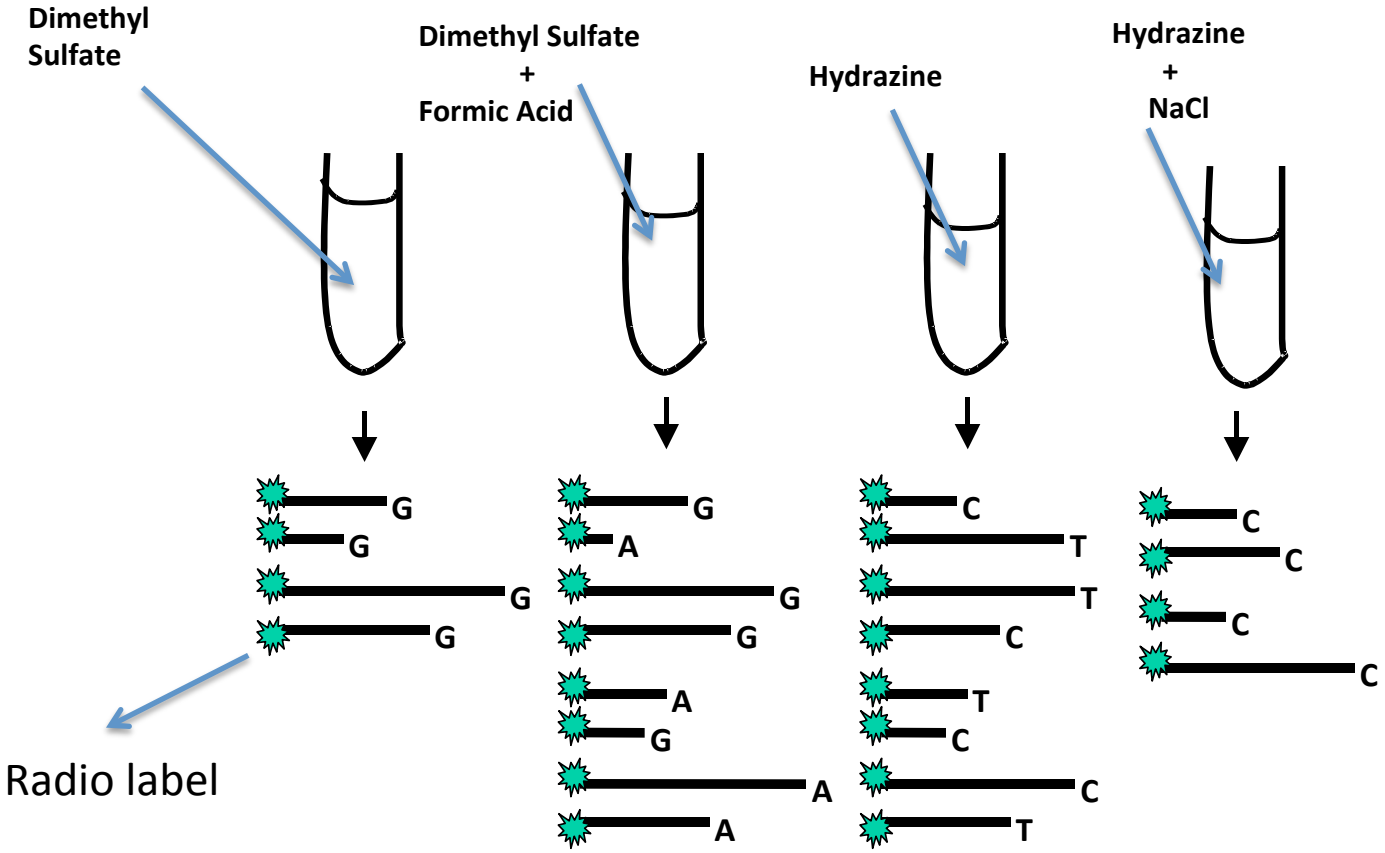
http://www.sciencemag.org/content/147/3664/1462

1975 - The dawn DNA sequencing

- Between 1975-1977 three methods of DNA sequencing were published

- Fred Sanger's Plus/Minus method
- Maxam-Gilbert
- Fred Sanger's chain termination method

Maxam-Gilbert Sequencing



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

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

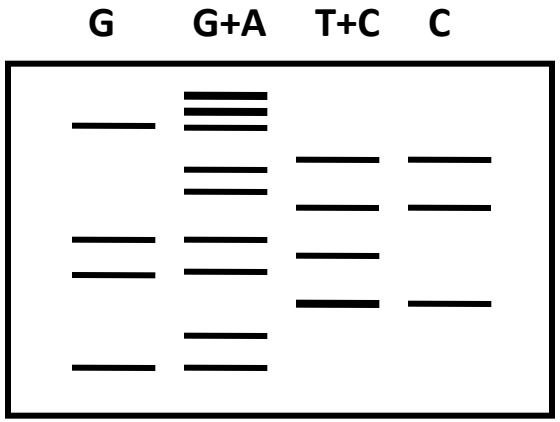
A Guanine/Adenine Cleavage (2). Dimethyl sulfate methylates the guanines in DNA at the N7 position and the adenines at the N3 (3). The glycosidic bond of a methylated purine is unstable (3, 4) and breaks easily on heating at neutral pH, leaving the sugar free. Treatment with 0.1 M alkali at 90° then 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

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′).

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 φX174)

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 φX174 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 φX174 (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.

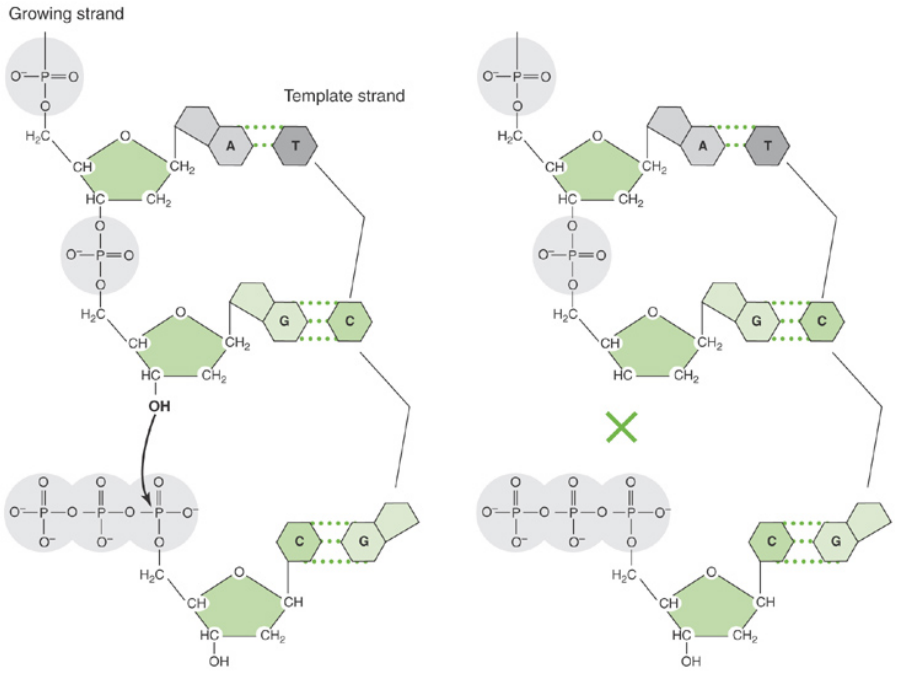
METHODS

Sanger Sequencing

- Uses a mixture of radio-labelled di-deoxy (ddNTP) and dexoy (dNTP) nucleotides to terminate base incorporation as soon as a ddNTP is encountered
- 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 (about 100:1), 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

Sanger sequencing

AGCTGCCCG

ddATP +
four dNTPs

ddA
dAdGdCdTdGdCdCdCdG

ddCTP +
four dNTPs

dAdGddC
dAdGdCdTdGddC
dAdGdCdTdGdCddC
dAdGdCdTdGdCdCddC

ddGTP +
four dNTPs

dAddG
dAdGdCdTddG
dAdGdCdTdGdCdCdCddG

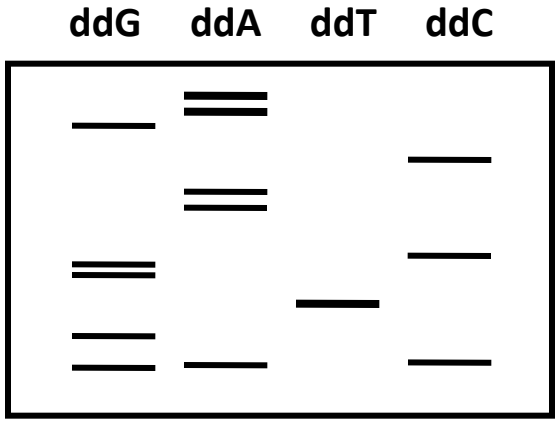
ddTTP +
four dNTPs

dAdGdCddT
dAdGdCdTdGdCdCdCdG

Sanger di-deoxy method

Longer fragments

Shortest fragments



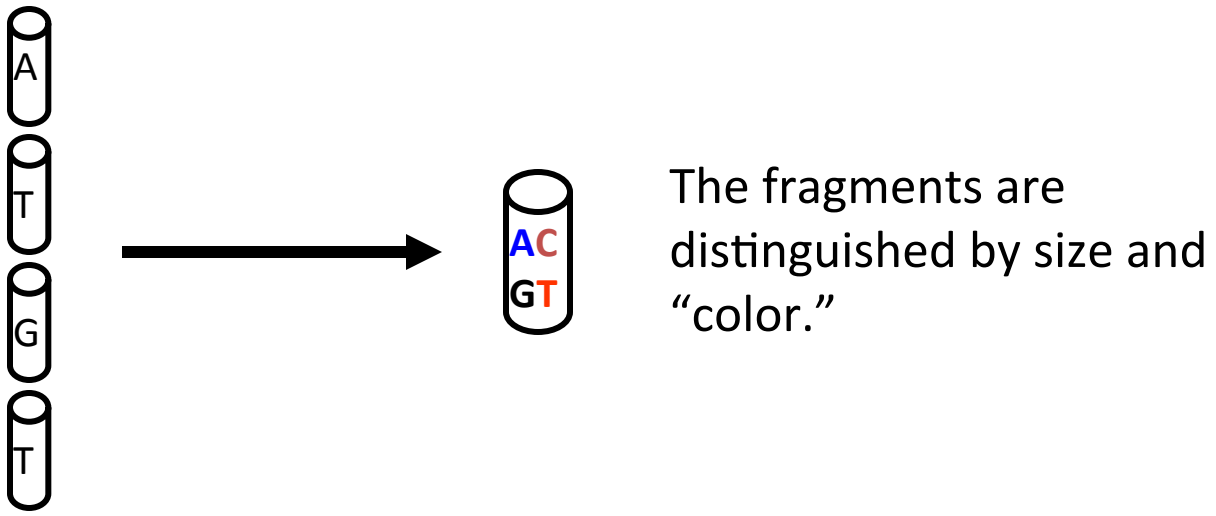
3′
A
A
G
C
A
A
C
G
T
G
C
A
G
5′

1985: Automating Sanger Sequencing

- Disadvantages of Sanger sequencing
  - Labour intensive
  - Used radioactive labels
  - Interpretation/analysis was subjective
- Difficult to scale up
- Leroy Hood, Michael Hunkpiller developed an automated method utilising:
  - Fluorescent labels instead of radioactivity
  - Utilise computerised algorithms to analyse data
  - Robotics

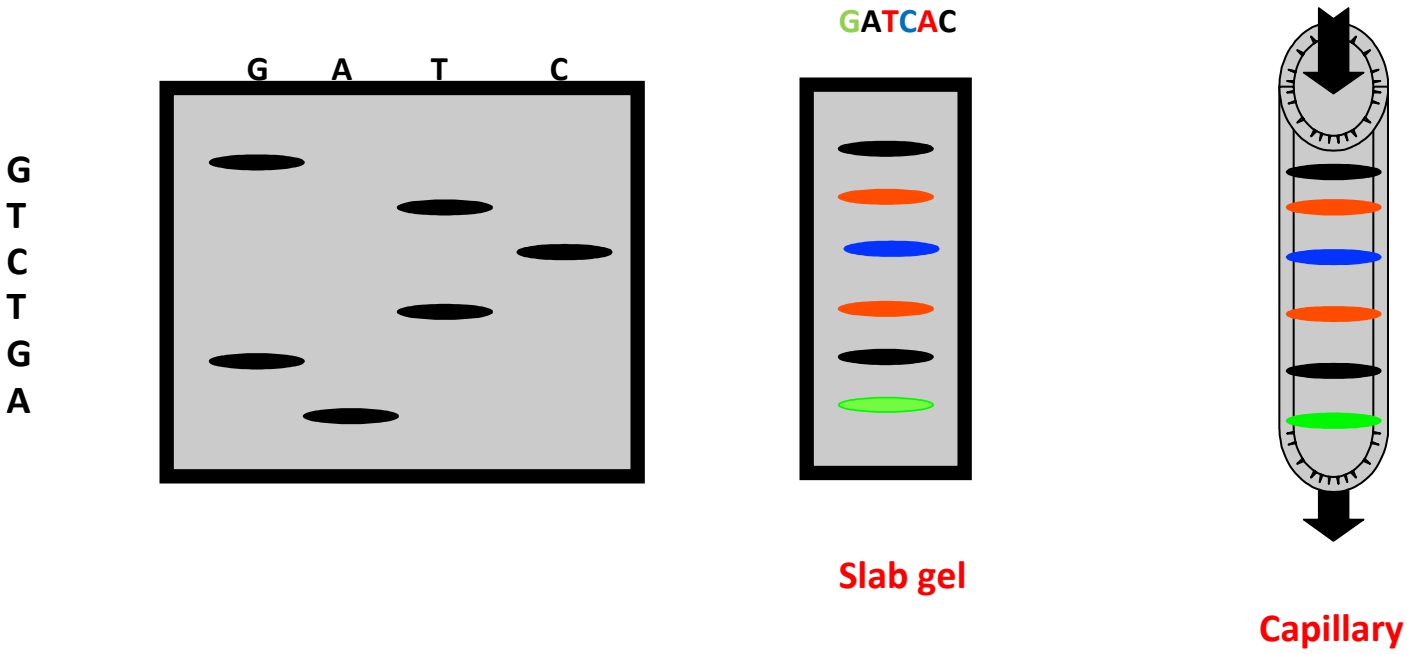
Dye Terminator Sequencing

- 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.



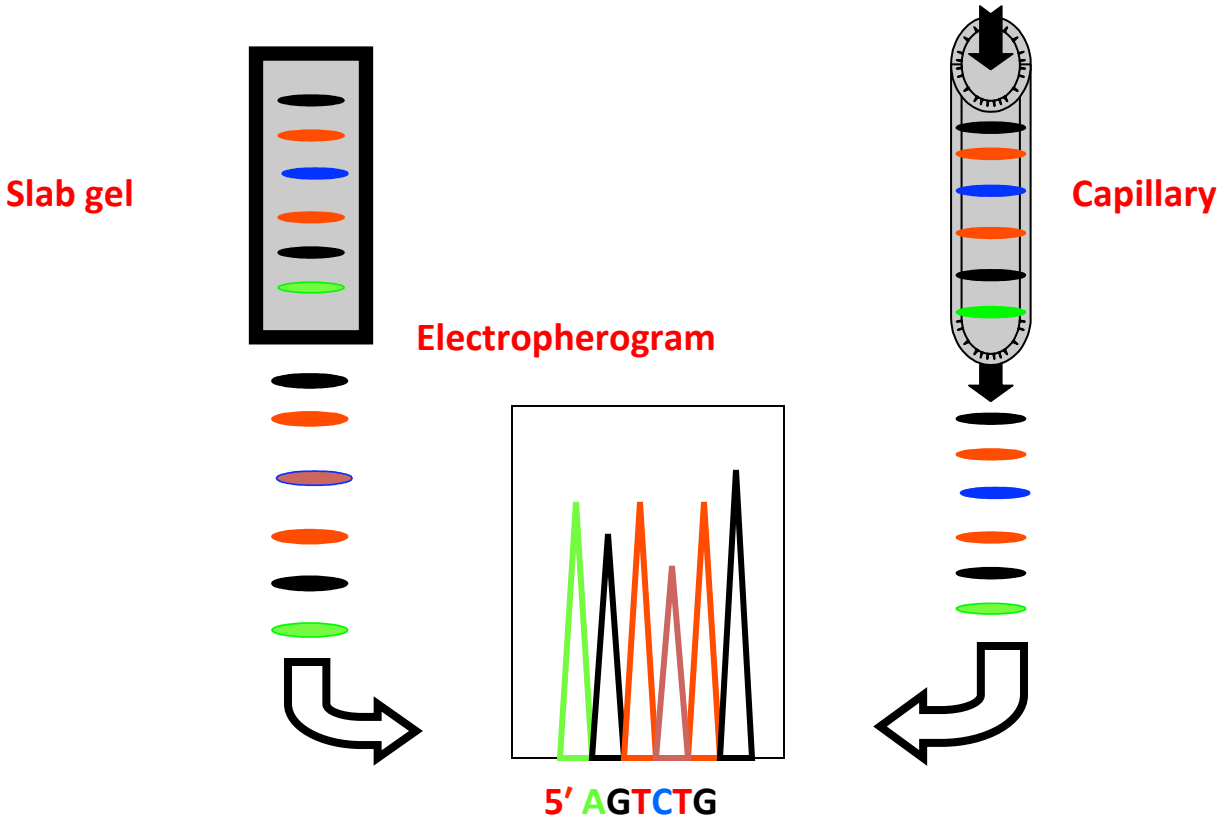
Dye Terminator Sequencing

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

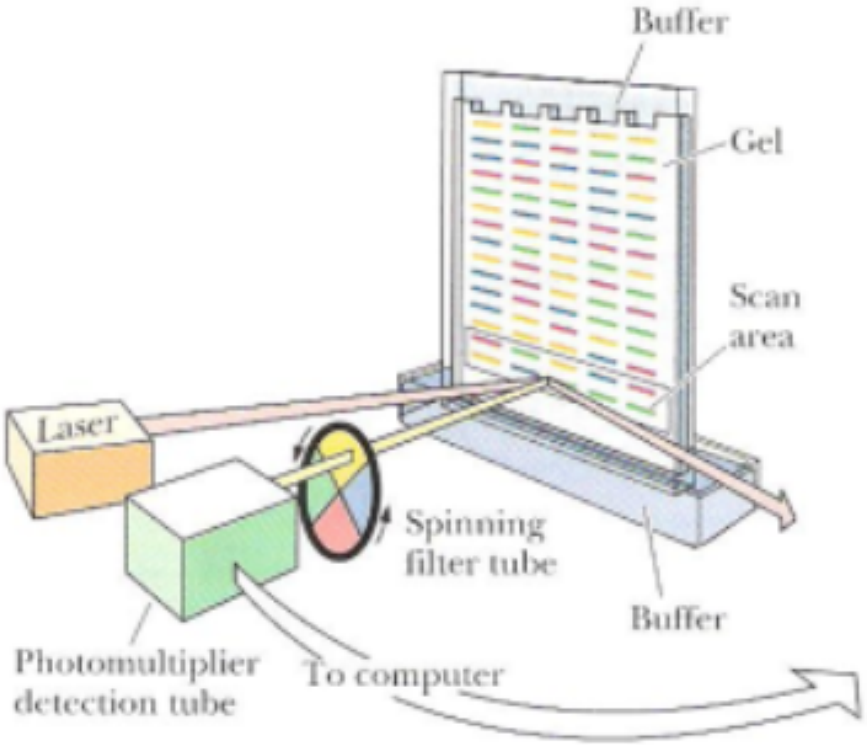


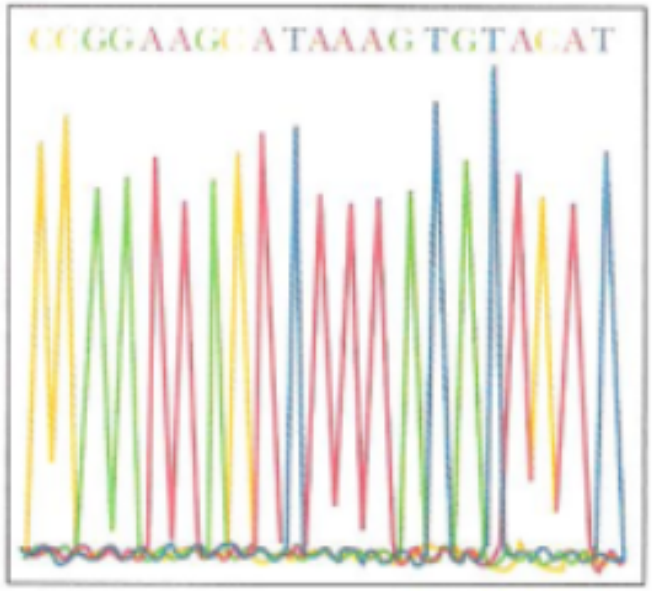
Dye Terminator Sequencing

- The DNA ladder is read on an electropherogram.



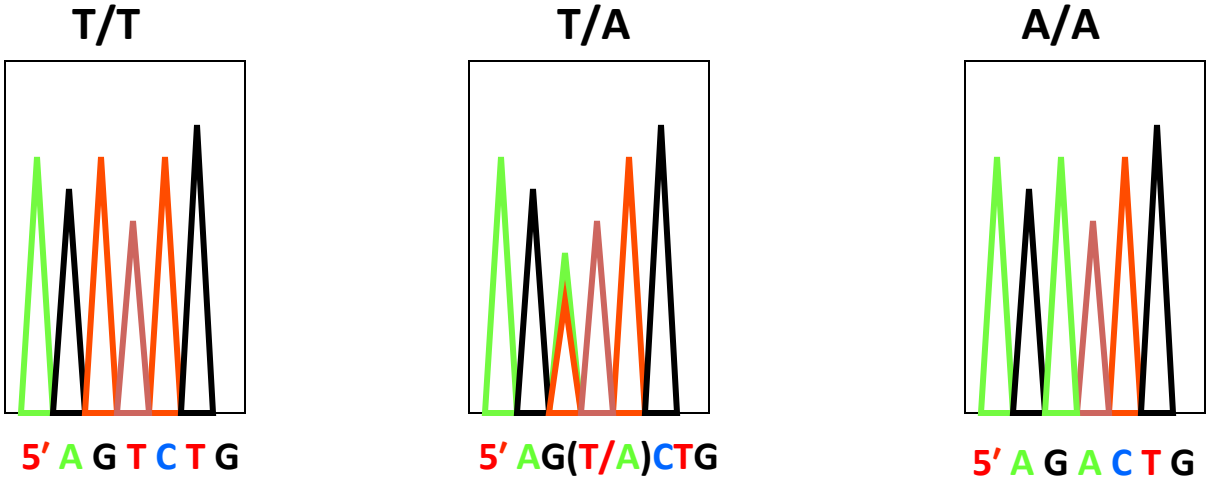
Automated Version of the Dideoxy Method





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 very high quality bases - 99.9%: ~600bp 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

The challenge of DNA sequencing

- 1869 – First DNA isolated
- 1944 – Establishing DNA as the genetic material
- 1953 – Double helix discovered
- 1957 – Central dogma proposed
- 1961 – Genetic code deciphered
- 1965 - First RNA sequence determined
- 1977 – Sanger DNA sequencing method published

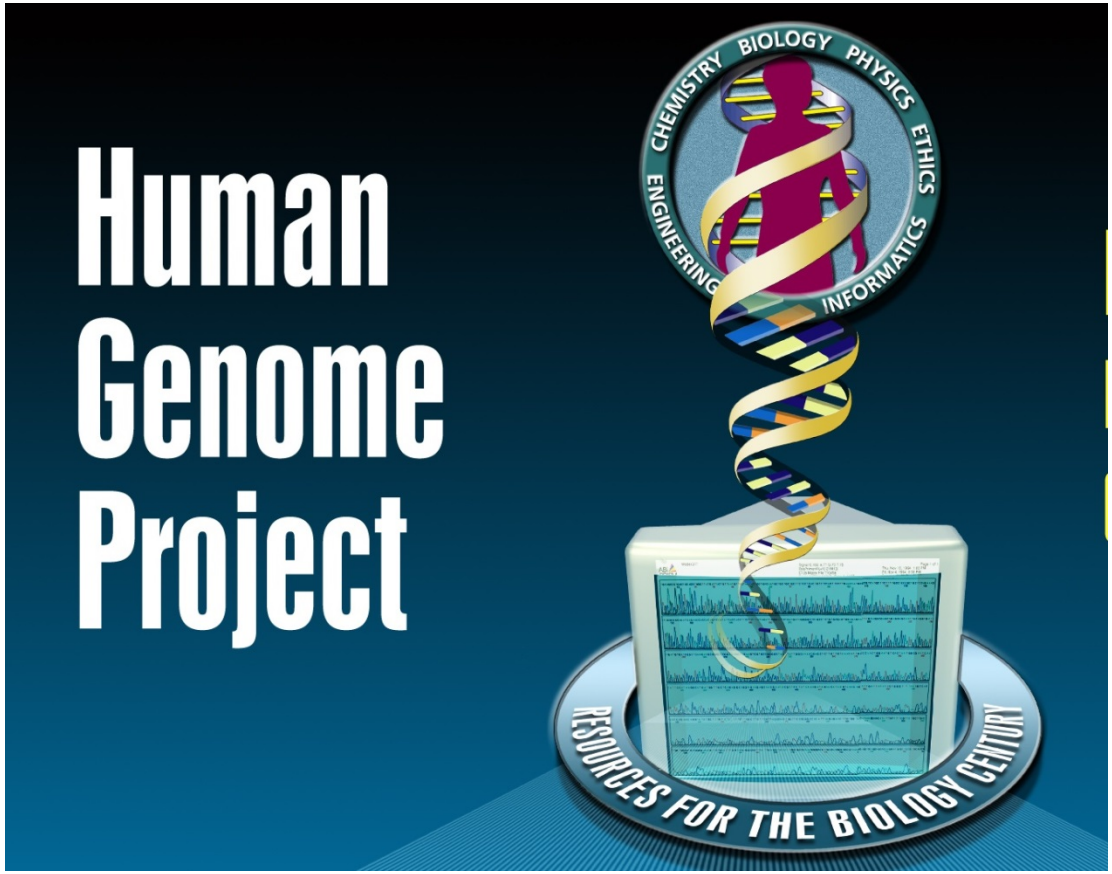
- How did we come to know so much about a molecule without being able to read it?
- Why did it take so long?

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

Clyde A. Hutchison III
DNA sequencing: bench to bedside and beyond Nucl. Acids Res. (2007) 35 (18): 6227-6237 a

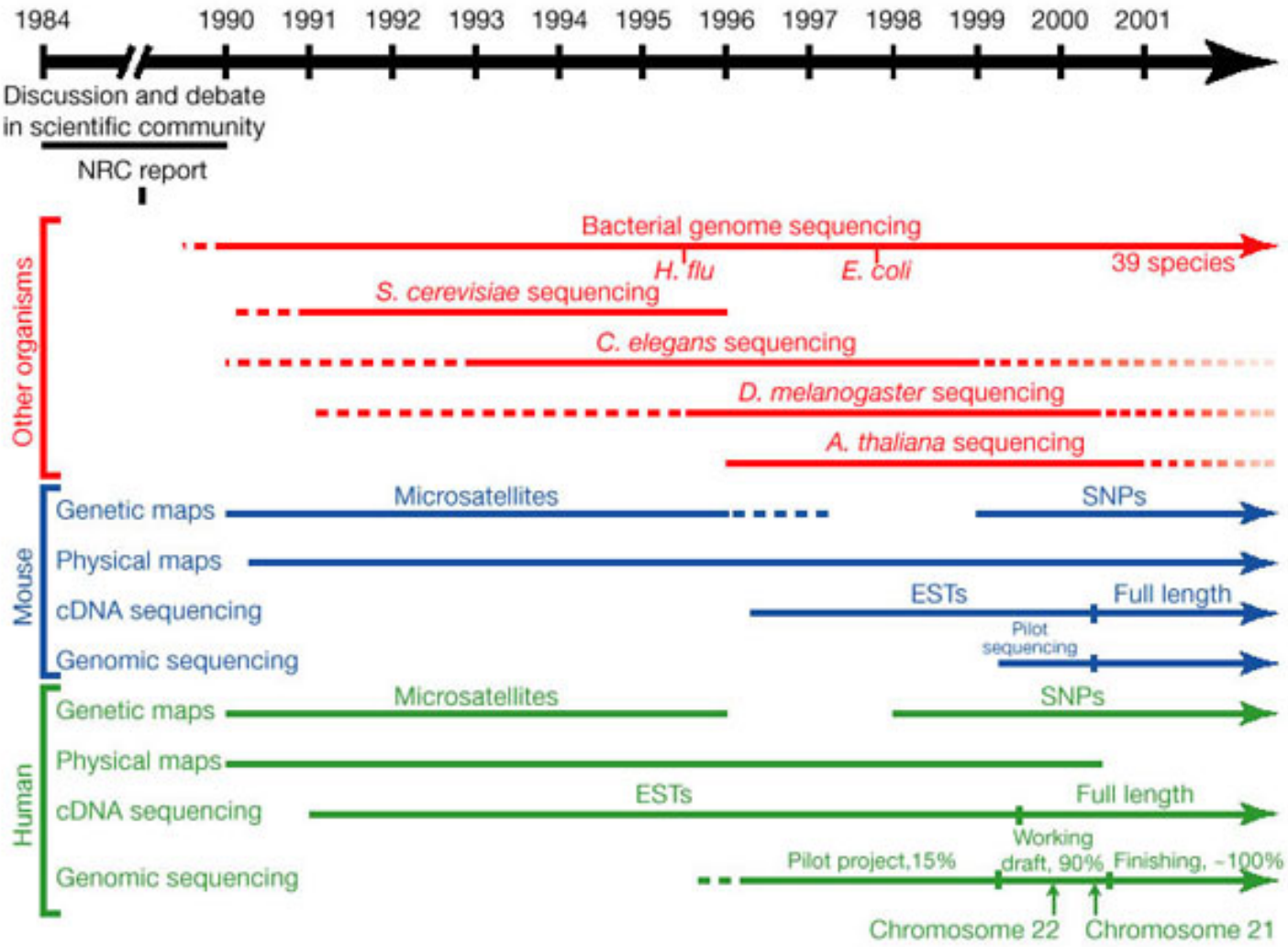
Human genome project



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
- Still being improved

Human Genome Project

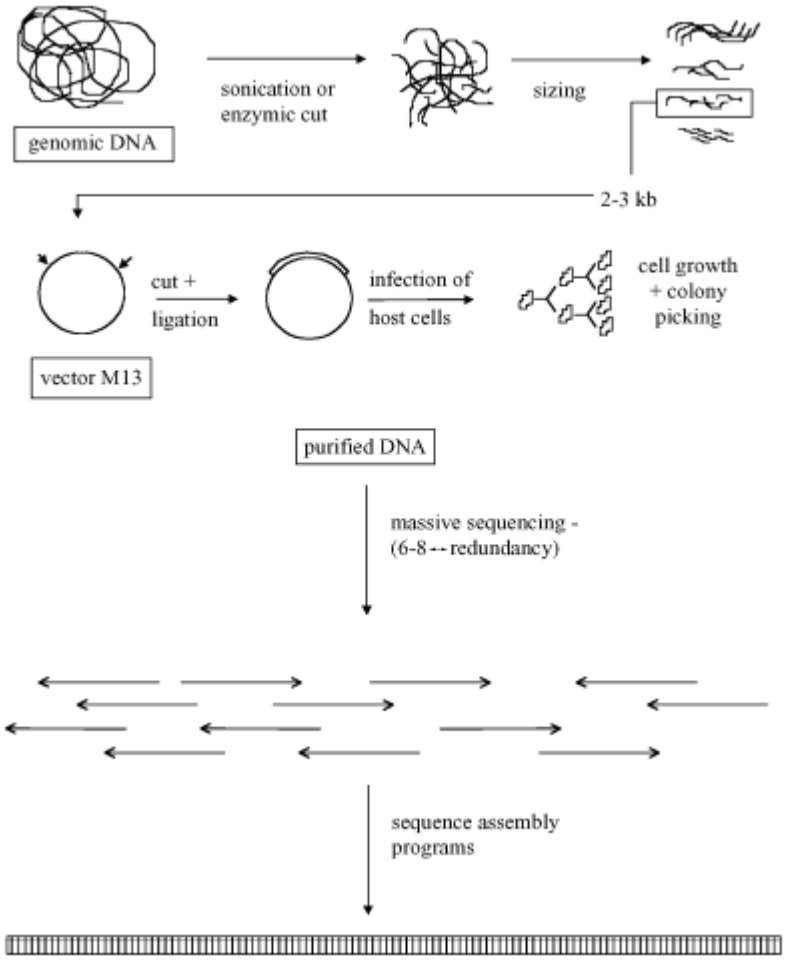


http://bit.ly/q3Qsd5

How it was Accomplished

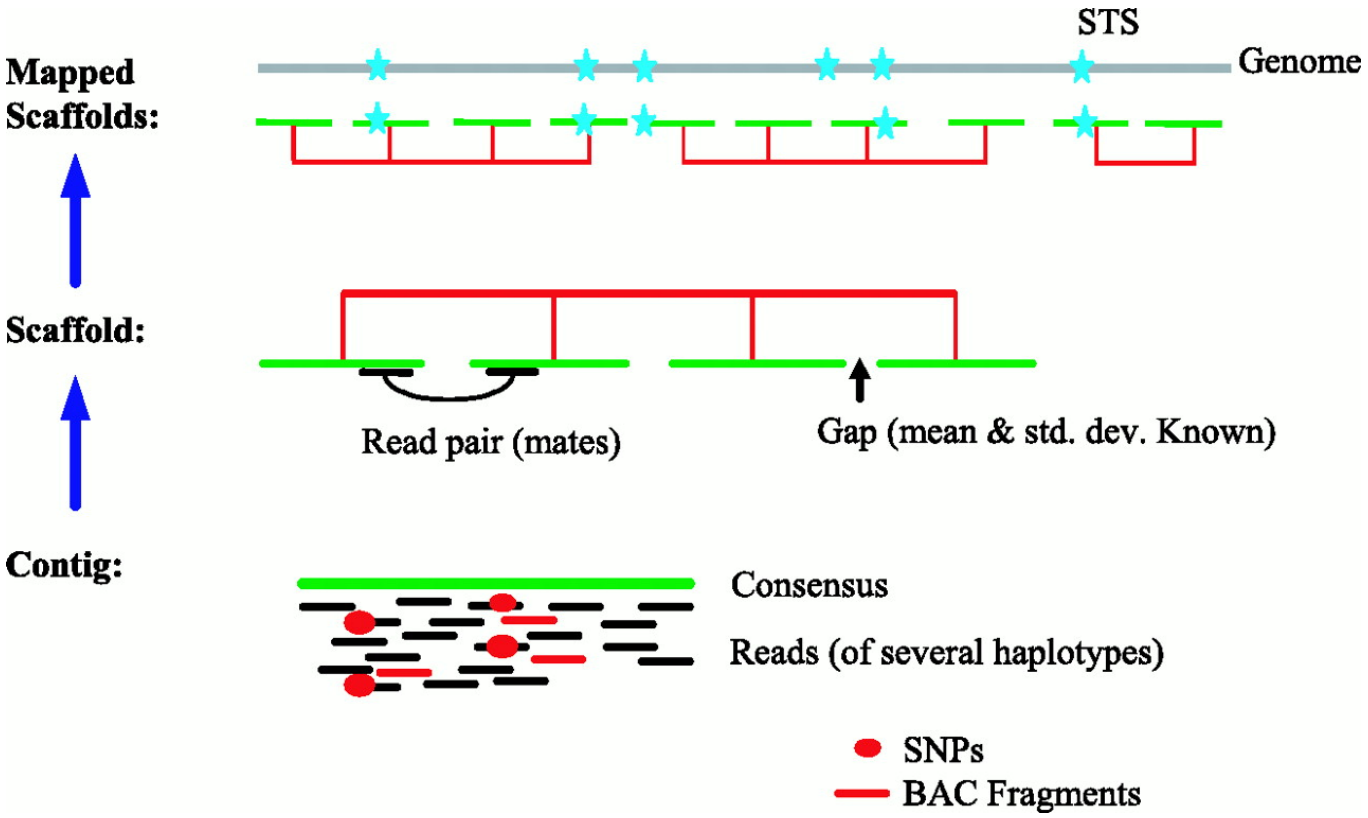
- Public Project
  - Hierarchical shotgun approach
  - Large segments of DNA were cloned via BACs and located along the chromosome
  - These BACs where shotgun sequenced
- Celera
  - Pure shotgun sequencing
  - Used public data (released daily) to help with assembly

Hierarchical Sequencing



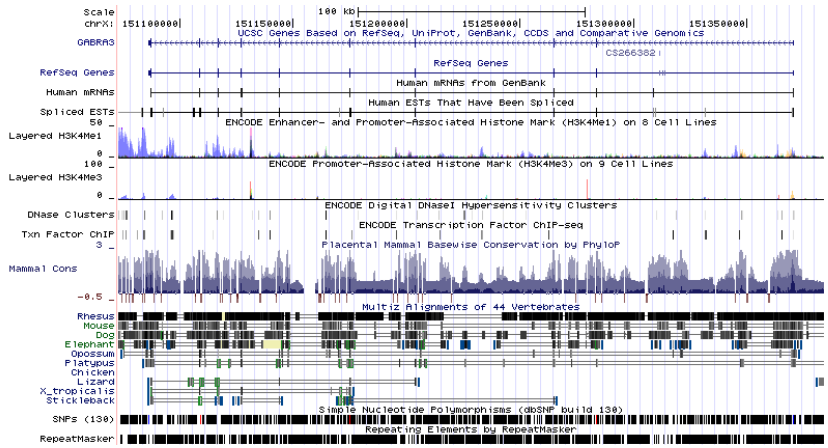
http://www.icb.uncu.edu.ar/upload/dnasequencing.pdf

Celera Shotgun Sequencing



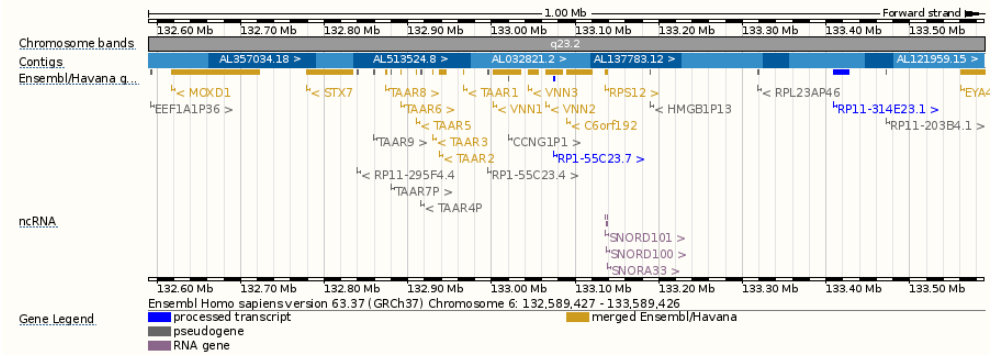
- Used paired-end strategy with variable insert size: 2, 10, and 50kbp

HGP Data Access




```
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 aaacccctgtc tctactaaaa atacaaaaaa ttagctgggc atggtggcag
      541 gcgcctatag tcccagctac tcgggaggct gaggcaggag aatggcgtga acccggcagg
      601 cagagcttgc agtgagccga gatcgcgcca ctgcactcca gcctgggcta cagagcaaga
      661 ctccatctta aaaaaaaaaa aaaagggaa 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 ttttttaaaa
```

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	

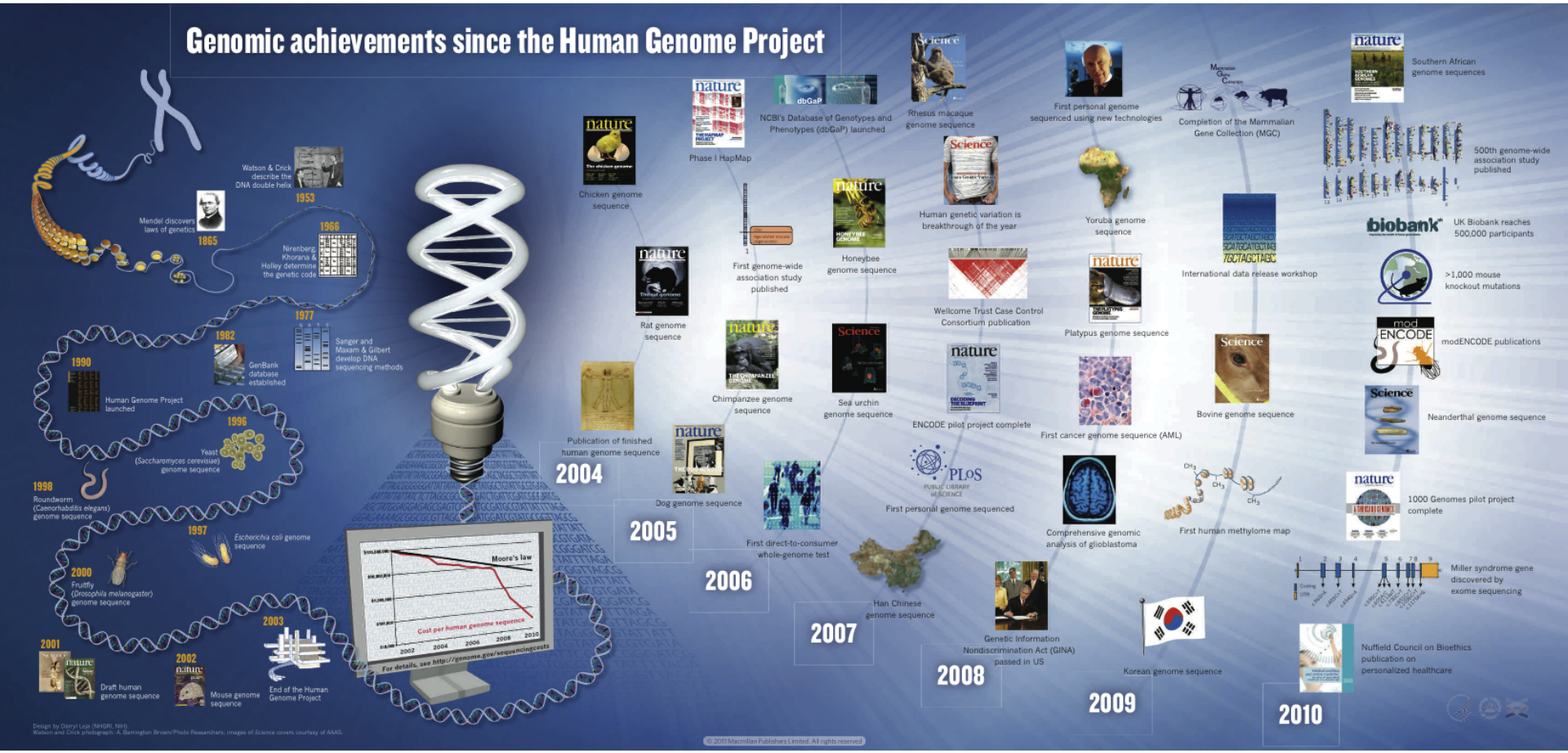


Results in GenBank, UCSC, Ensembl & others

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

Achievements Since the HGP



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 at least $67 billion to the US economy

http://www.genome.gov/27544383

2004 onwards:
Beyond 1 species, 1 genome

- Cost of producing a single genome could vary from $10,000s to $100,000s using capillary sequencers
- Labour intensive methodology

- New methods were required to lower the overall cost per genome

Second generation short read technologies

Large scale sequencing 2006

PRODUCTION

Rooms of equipment
Subcloning > picking > prepping
35 FTEs
3-4 weeks

SEQUENCING

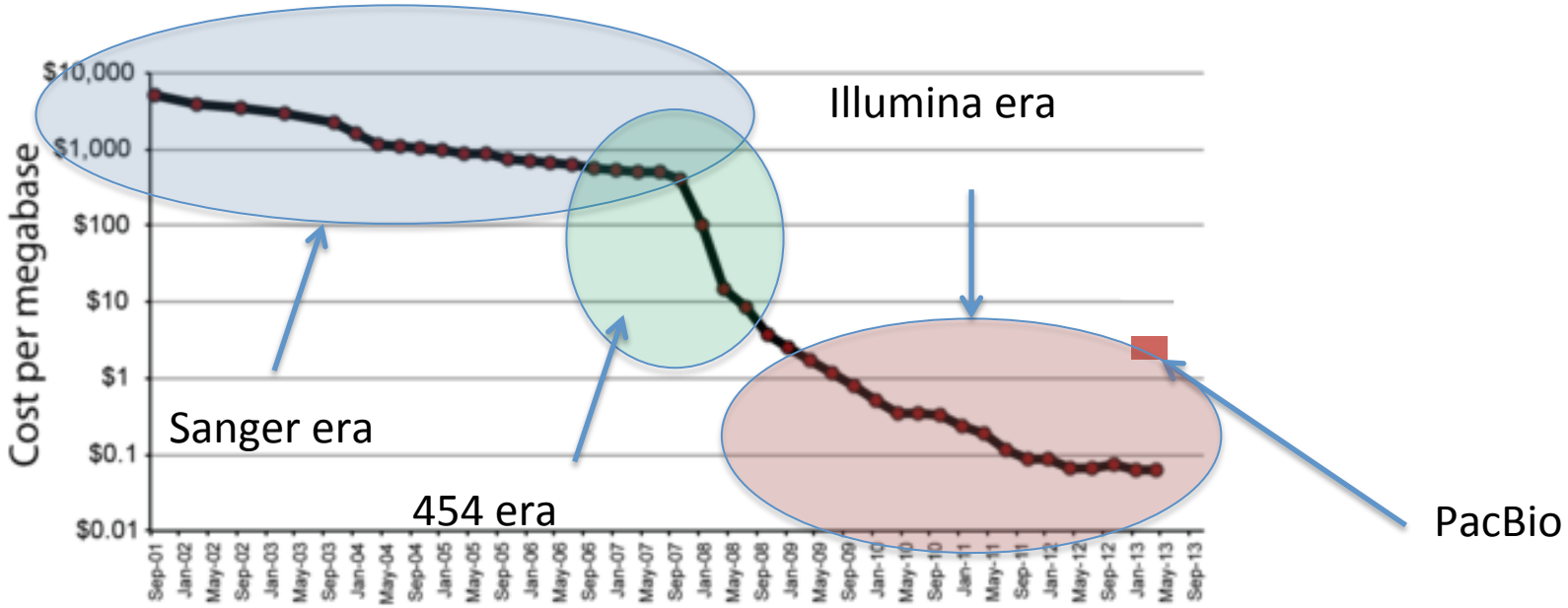
74x Capillary Sequencers
10 FTEs
15-40 runs per day
1-2Mb per instrument per day
120Mb total capacity per day

Large scale sequencing 2008

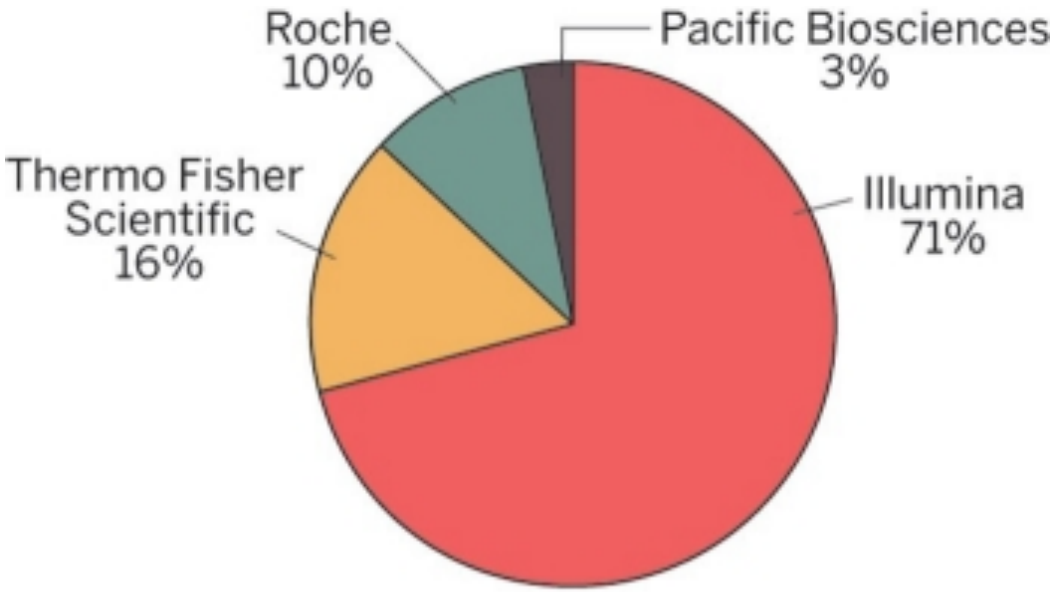
Illumina GAII sequencer

Key advantages over Sanger Sequencing

- Hugely reduced labour requirements
- Cost per sequence
- Reduced time to result
- Decentralisation



Market share



World market in 2013 = $1.3 billion

Fun fact

- Clive Brown
- Formerly director of Computational Biology at Solexa (Illumina)
- Chief Technology Officer at Oxford Nanopore

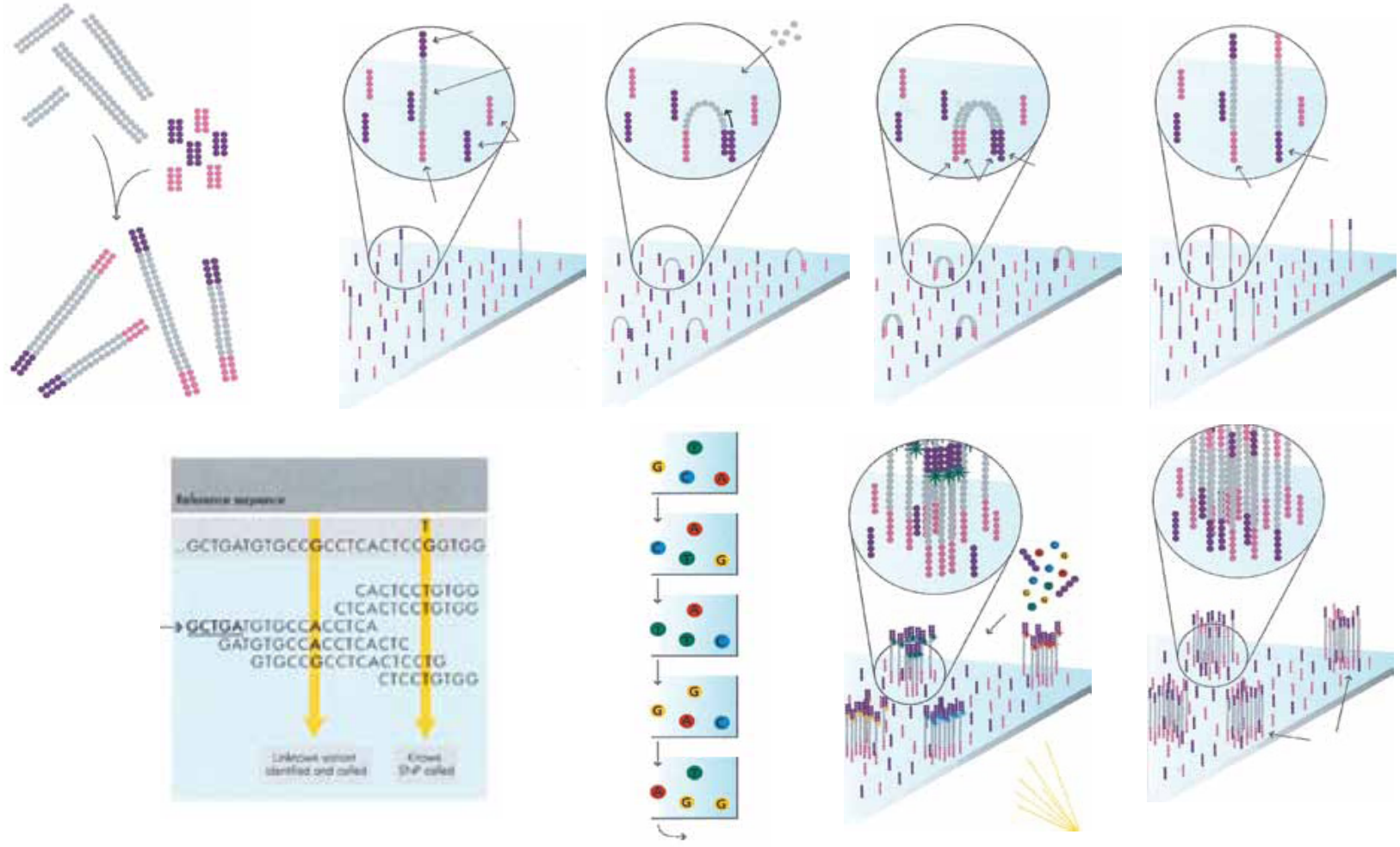
Illumina Sequencing By Synthesis

Illumina HiSeq

Illumina NextSeq

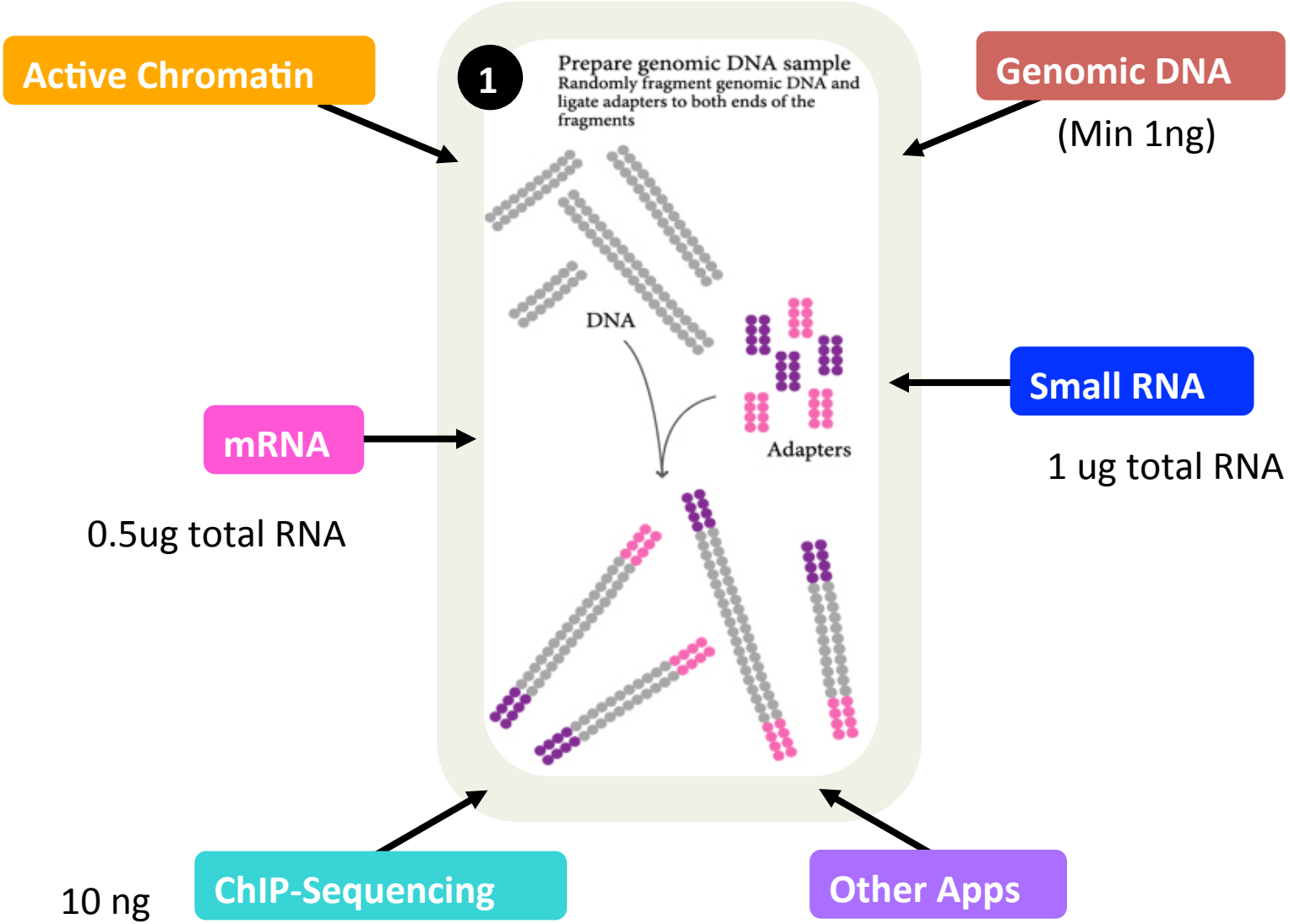
Illumina MiSeq

Illumina Sequencing

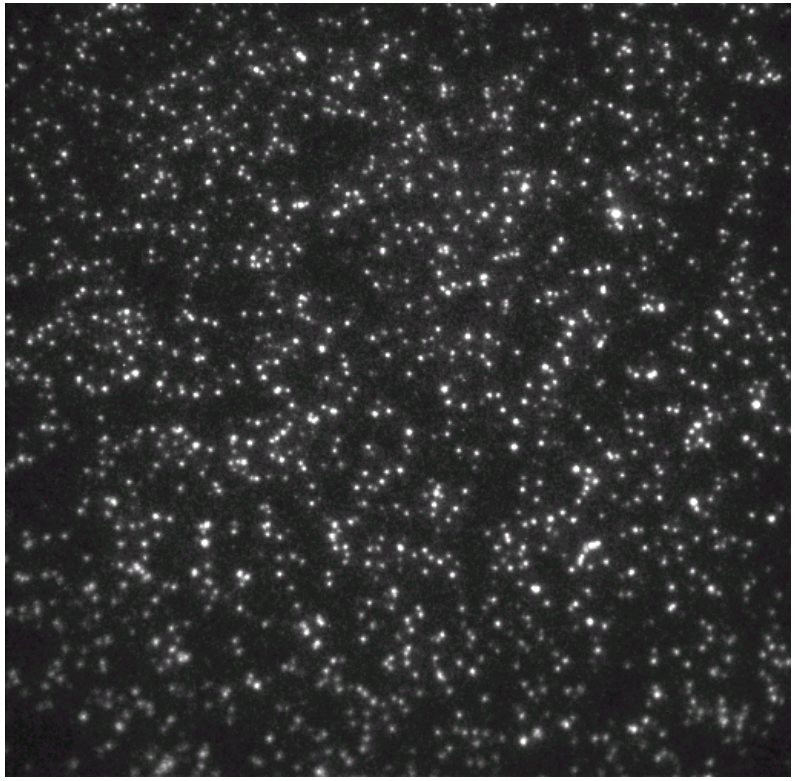


http://www.illumina.com/Documents/products/techspotlights/techspotlight_sequencing.pdf

Step 1: Sample Preparation



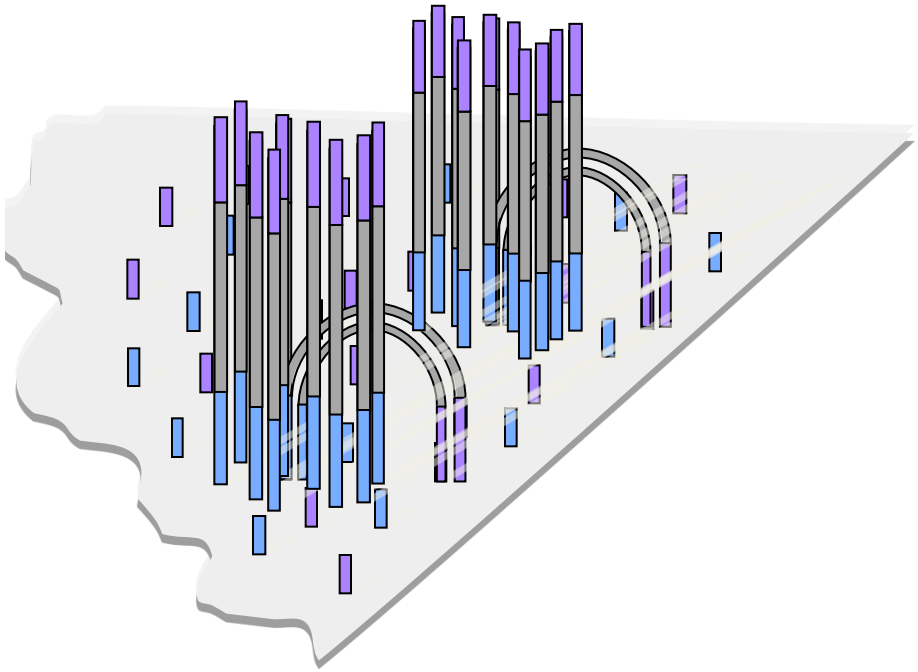
Step 2: Clonal Single Molecule Arrays



100um

Random array of clusters

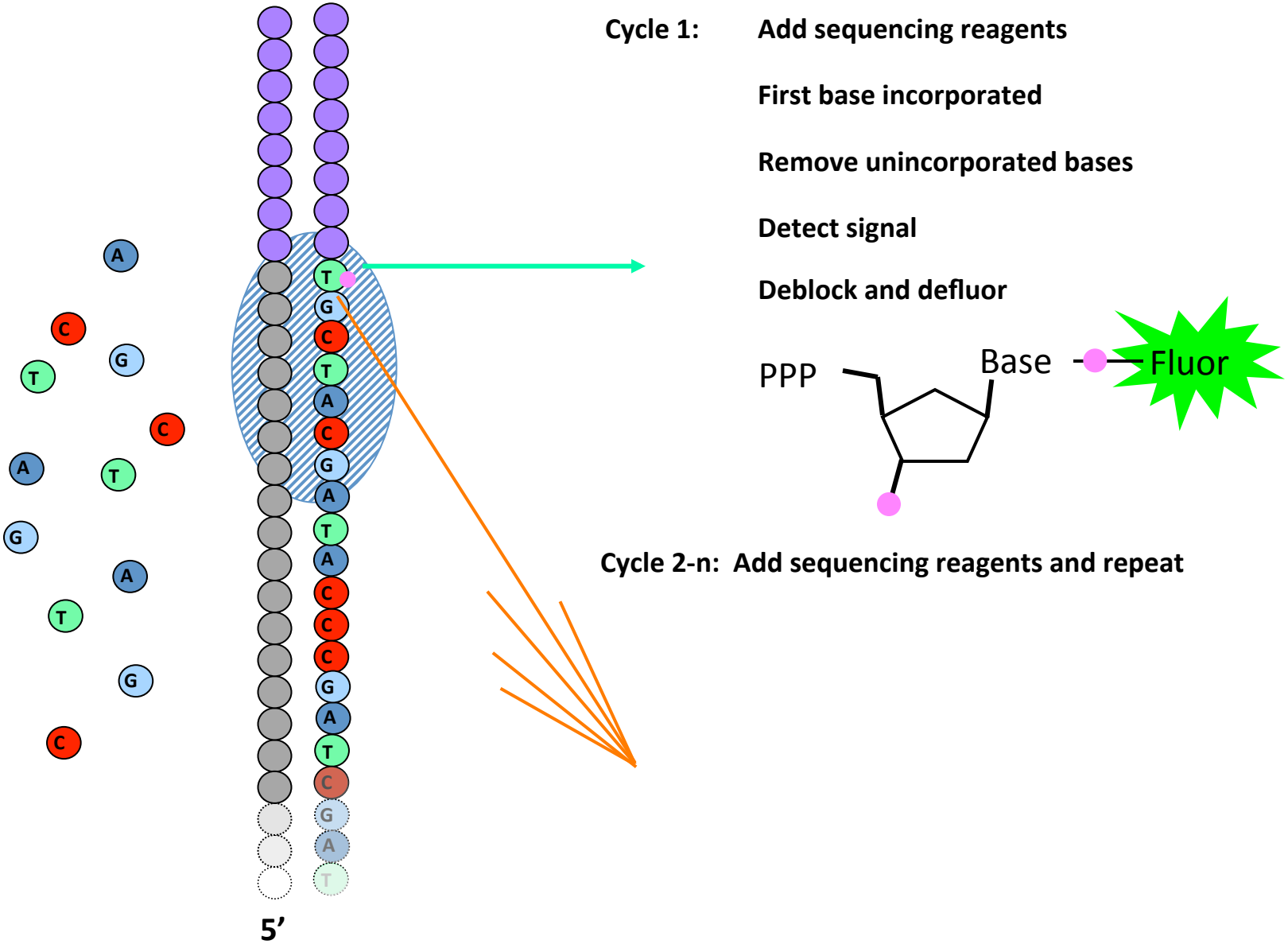
Attach single molecules to surface
Amplify to form clusters



~1000 molecules per ~ 1 um cluster
~2 billion clusters per flowcell

1 cluster = 1 sequence

Step 3: Sequencing By Synthesis (SBS)



Cycle 1: **Add sequencing reagents**

First base incorporated

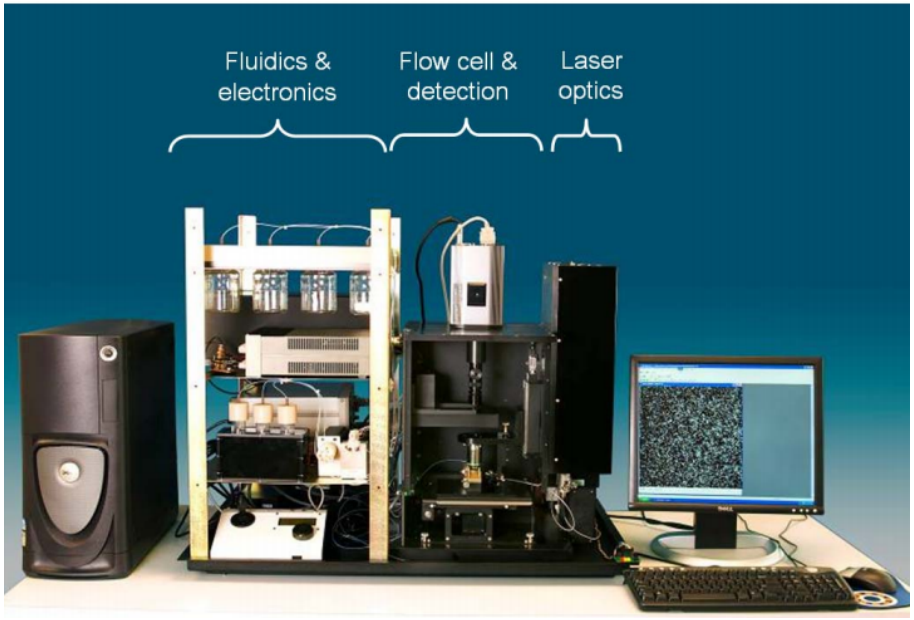
Remove unincorporated bases

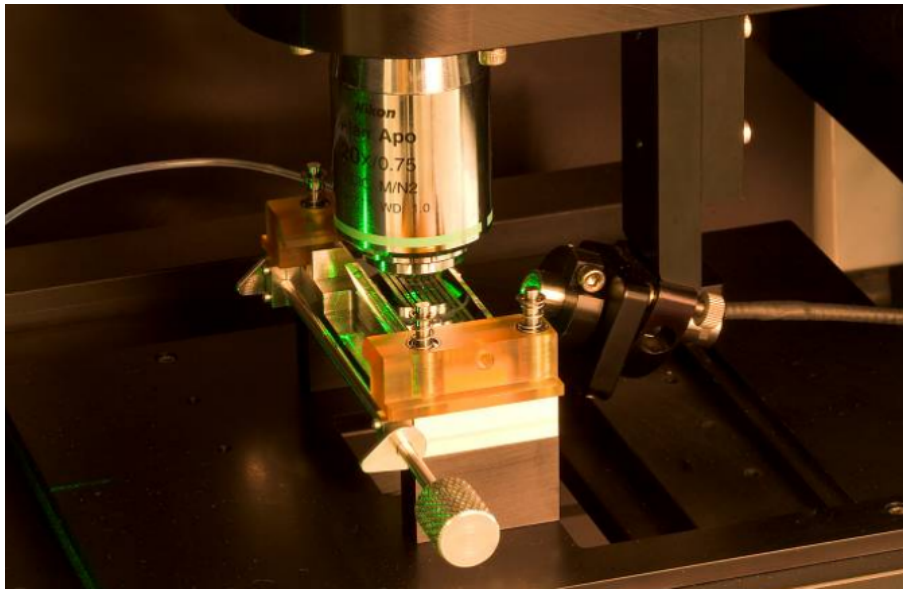
Detect signal

Deblock and defluor

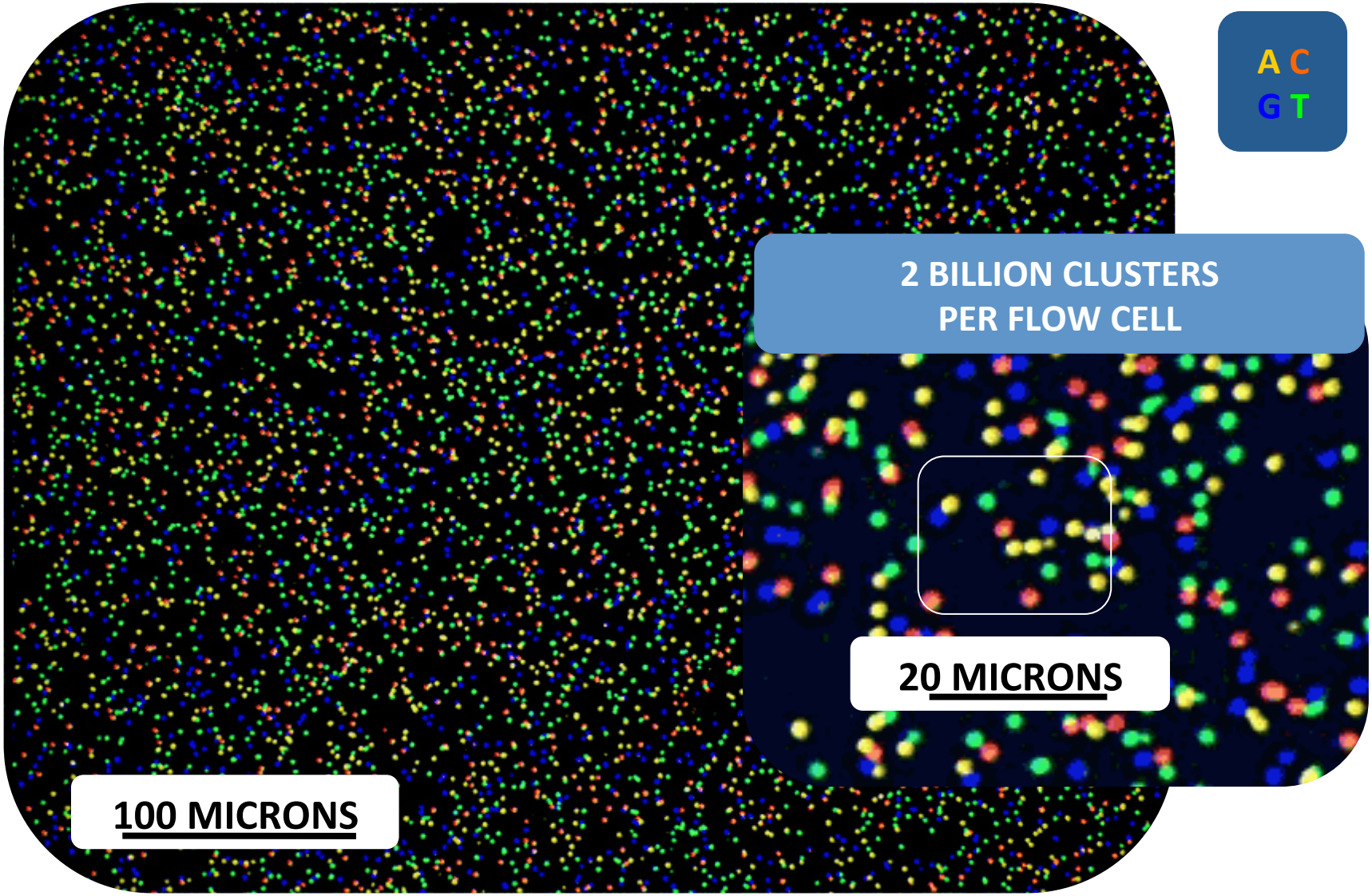
Cycle 2-n: Add sequencing reagents and repeat

Under the hood:

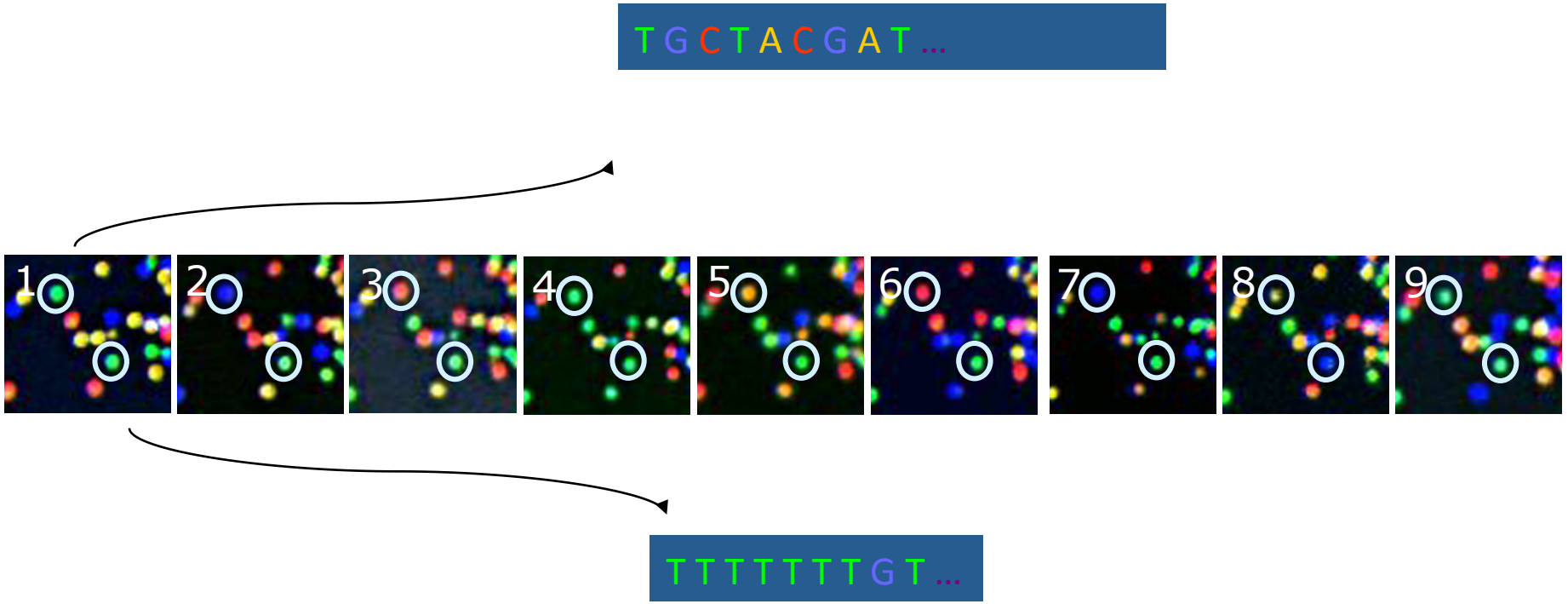




Illumina Sequencing : How it looks



Base calling from raw data

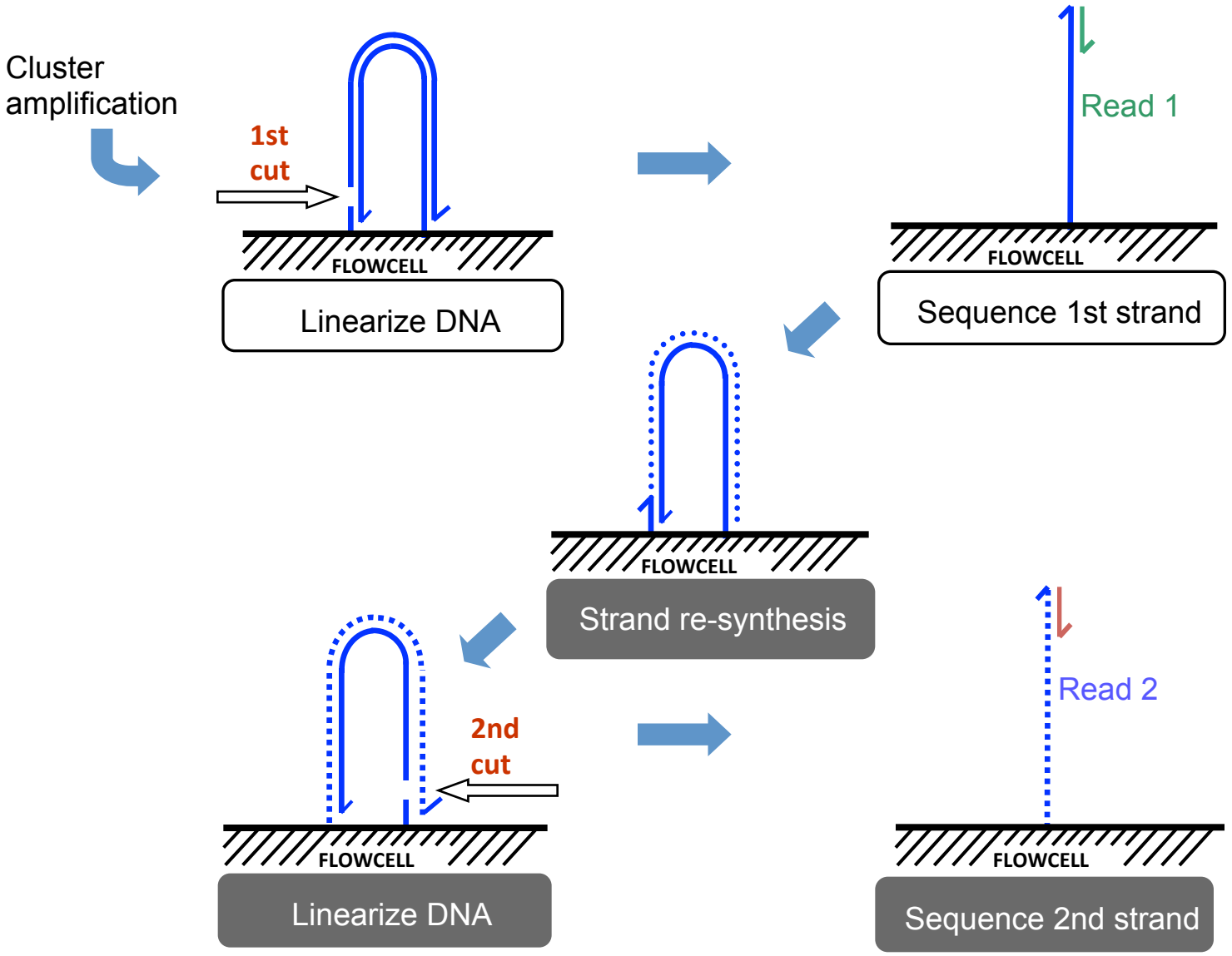


The identity of each base of a cluster is read off from sequential images.

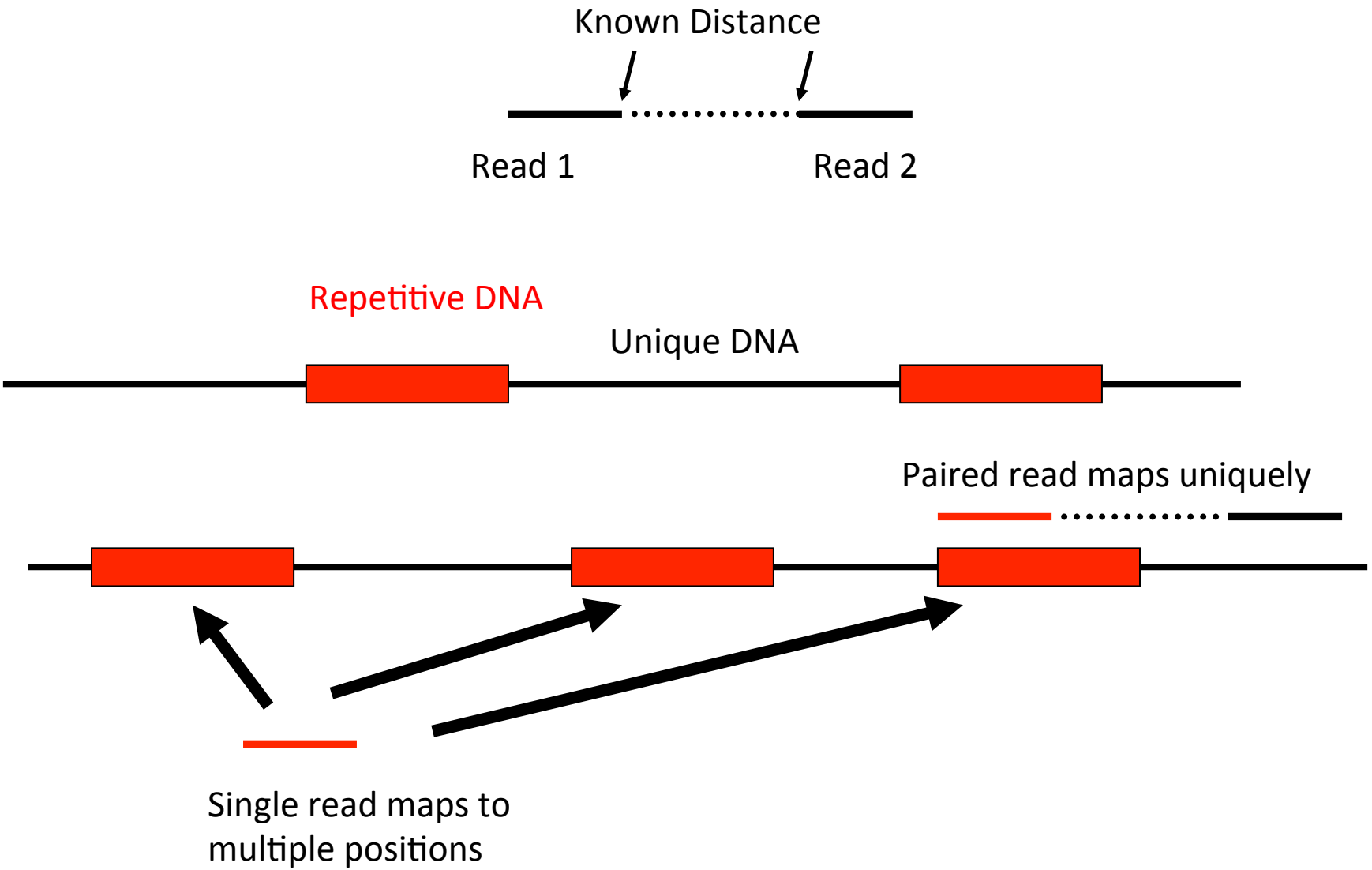
Paired-End Sequencing

- Provides distance relationship between two reads
- Important for many applications
  - Characterise insertions, deletions, copy number variants, rearrangements
  - Required for *de novo* genome assembly
  - Enables sequencing across repeats
  - Useful for isoform inference, transcriptome assembly etc.

Illumina Paired-End Sequencing

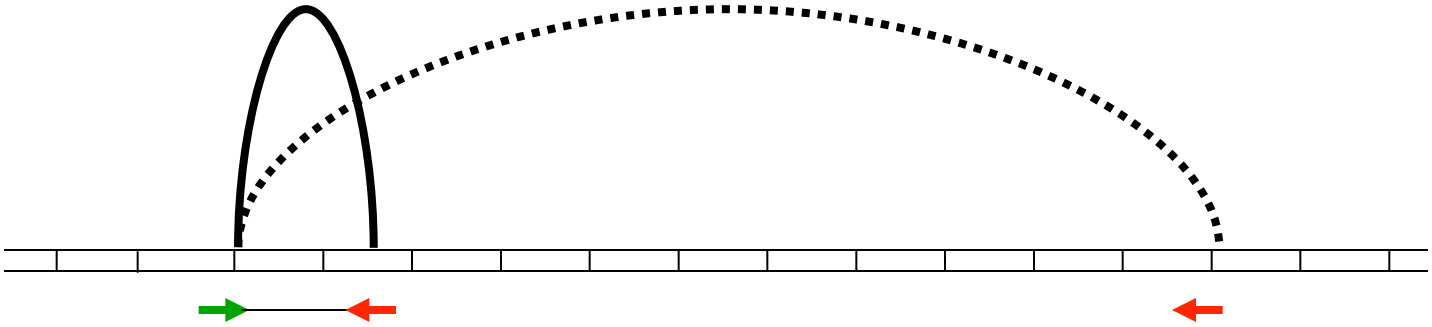


Paired-end reads are important



Working with Paired Reads

- Applicable to different fragment size ranges
  - up to ~800 bp for standard paired-libraries
  - 2 - 20kb mate-pair libraries



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

Illumina platforms

Illumina HiSeq

- 500Gbase/flowcell
- 8 human genomes
- 7 day run time
- High output or rapid run mode
- Read lengths up to 250bp
- Requires large numbers of samples (or large genomes) to obtain lowest cost

- 4 colour chemsitry
- £750,000 incl 3 year servicing

Illumina NextSeq 500

- 90Gbase/flowcell
- 1 human genome
- 2 day run time
- High output or rapid run mode
- Read lengths up to 150bp

- 2-colour chemistry

- £250,000 incl 3 year servicing

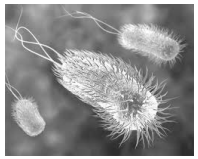
Illumina MiSeq

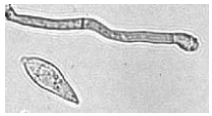
- 15Gbase/flowcell
- 2 day run time
- Read lengths up to 300bp

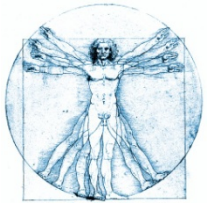
- 4 colour chemsitry

- £120,000 incl 3 year servicing

What does this mean?







Rapid run	High output run
48 genomes (£200 per sample)	96 genomes/lane (<£100 per sample)
10 genomes (£400 per sample)	10 genomes/lane (£250 per sample)
8 genomes (£500 per sample)	8 genomes/lane (£300 per sample)
	1 genome (£3000)

UNIVERSITY OF EXETER

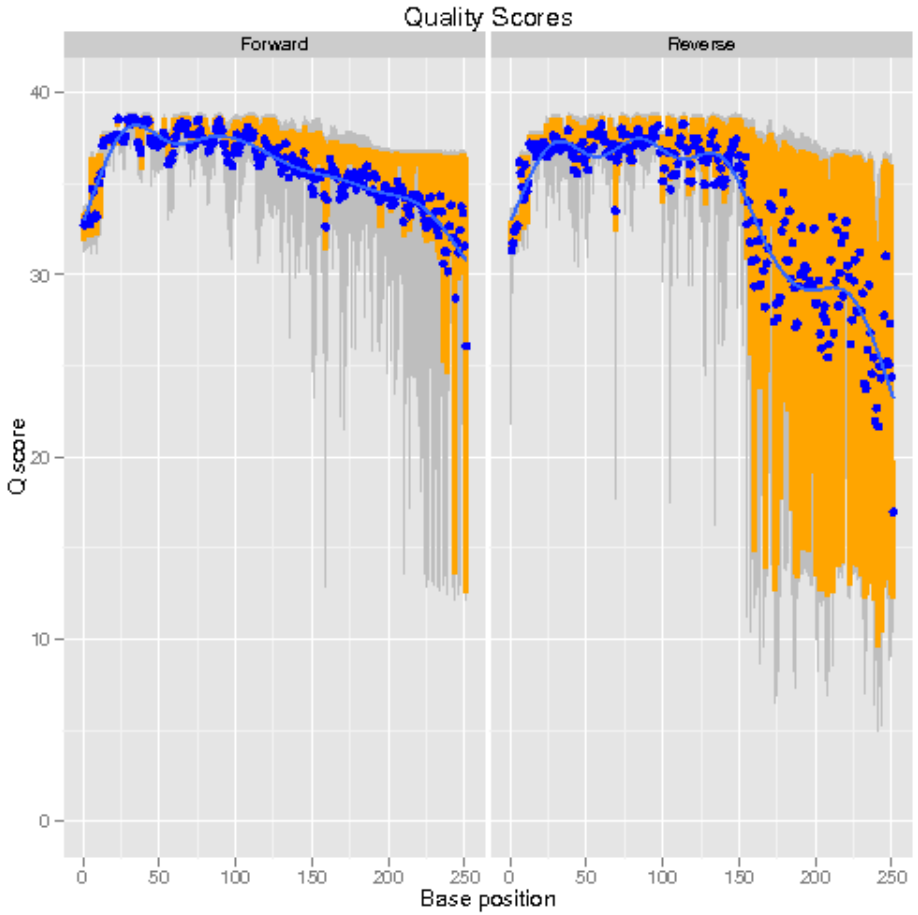
Potential issues with Illumina sequencing

- Specific motifs which are difficult to sequence
  - GGC motif
  - Inverted repeats

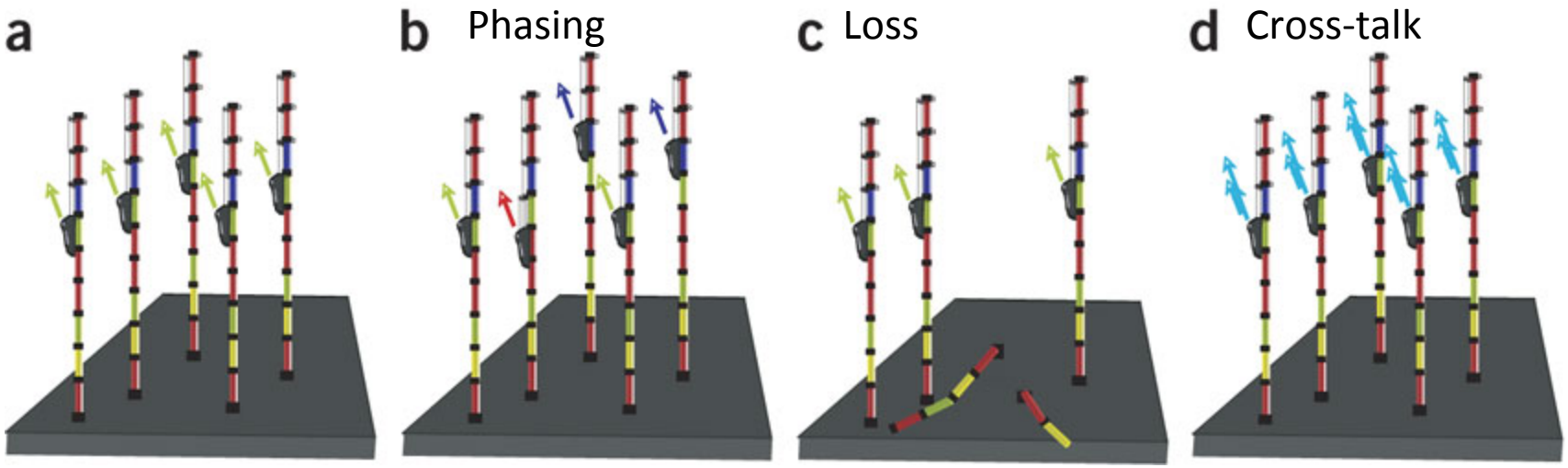
- Now mostly resolved
  - Low diversity sequences
    - 16S/amplicon sequences
    - Custom adaptors with barcodes at 5' end
    - Now a much reduced problem thanks to software updates
  - GC/AT bias
    - GC clusters are smaller than AT
    - (less of a problem post June 2011)

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

Why do quality scores drop towards the end of a read?



3 main factors



Schematic representation of main Illumina noise factors.
(a–d) A DNA cluster comprises identical DNA templates (colored boxes) that are attached to the flow cell. Nascent strands (black boxes) and DNA polymerase (black ovals) are depicted.
(a) In the ideal situation, after several cycles the signal (green arrows) is strong, coherent and corresponds to the interrogated position.
(b) Phasing noise introduces lagging (blue arrows) and leading (red arrow) nascent strands, which transmit a mixture of signals.
(c) Fading is attributed to loss of material that reduces the signal intensity (c).
(d) Changes in the fluorophore cross-talk cause misinterpretation of the received signal (blue arrows; d). For simplicity, the noise factors are presented separately from each other.

http://arep.med.harvard.edu/pdf/Fuller_09.pdf

Erlich et al. Nature Methods 5: 679-682 (2008)

New Illumina developments

- 2x250bp reads (HiSeq rapid run mode)
- Ordered flowcells (HiSeq X-Ten)
- 10kb synthetic reads (approx. 5-6 million per lane)
  - Useful for phasing of haplotypes and denovo assembly

New Illumina Developments

	 HiSeq 2500		 HiSeq 3000	 HiSeq 4000
Run Mode	Rapid Run	High-Output	N/A	N/A
Flow Cells per Run	1 or 2	1 or 2	1	1 or 2
Output Range	10-300 Gb	50-1000 Gb	125-750 Gb	125-1500 Gb
Run Time	7-60 hours	<1-6 days	<1-3.5 days	<1-3.5 days
Reads per Flow Cell†	300 million	2 billion	2.5 billion	2.5 billion
Maximum Read Length	2 x 250 bp	2 x 125 bp	2 x 150 bp	2 x 150 bp

Limits to Illumina technology

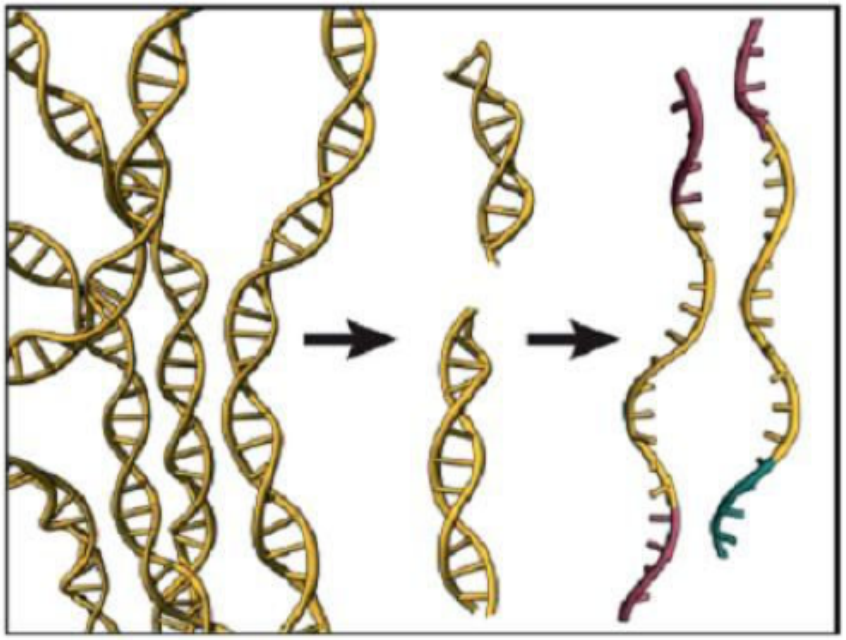
- Limitations:
  - Reagent degradation
  - Dephasing
    - Leads to higher error rates
    - A 1% loss of signal or polymerase error every cycle leads to only 35% correct signal after 100 cycles
  - Sequencing time is always governed by the cyclic nature of the instrument (one base at a time)
    - Ideally dispense with incorporate, image, wash cycles
  - Size of fragments which can be clustered on the flowcell
    - Read lengths beyond the size of the DNA fragment are useless
    - Inefficient clustering >800bp

454 and Ion Torrent

Fun fact

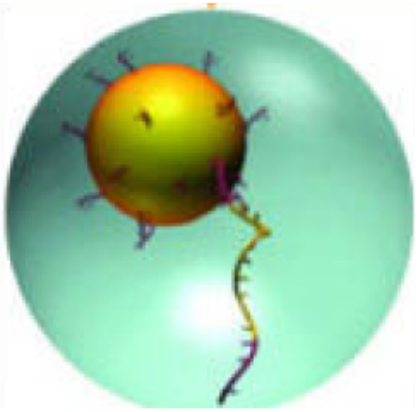
- Jonathan Rothberg
- Set up 454 in late 90s
- Sold to Roche in 2007
- Founded Ion Torrent in 2007
- Superseded 454
- Sold to Life Tech in 2010

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

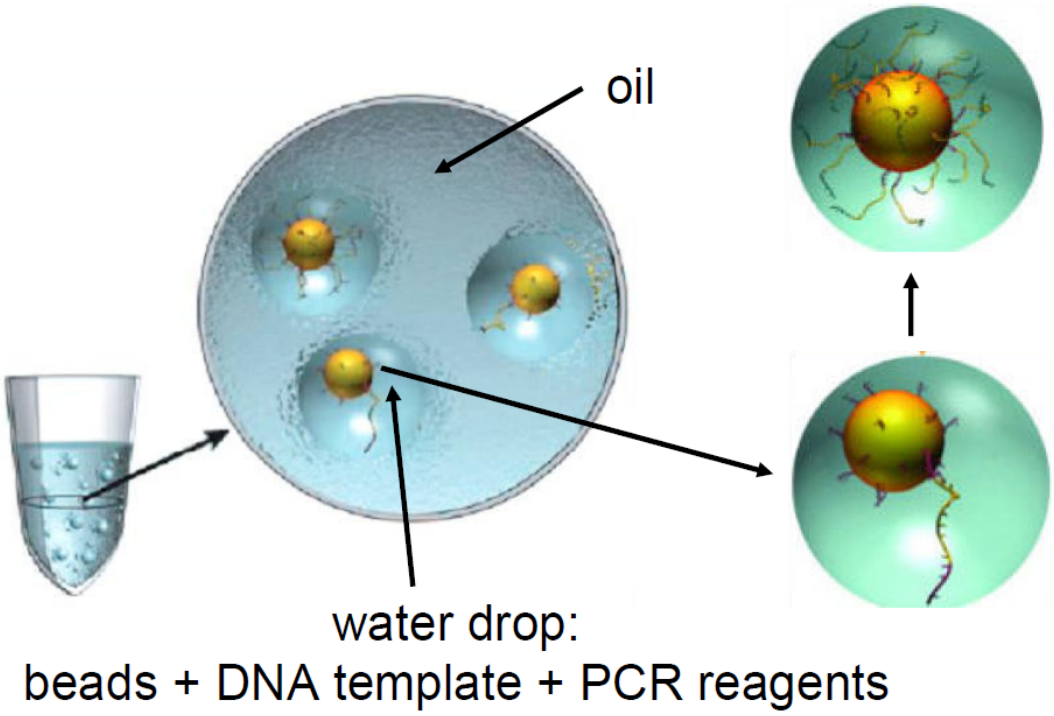
One Fragment = One Bead



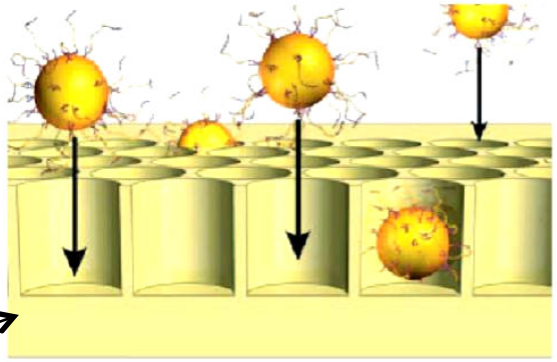
4. The single stranded DNA library is immobilised onto proprietary DNA capture beads

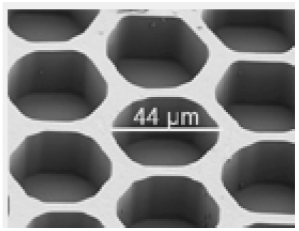
454 Step 2: Amplification

Water-based emulsion PCR



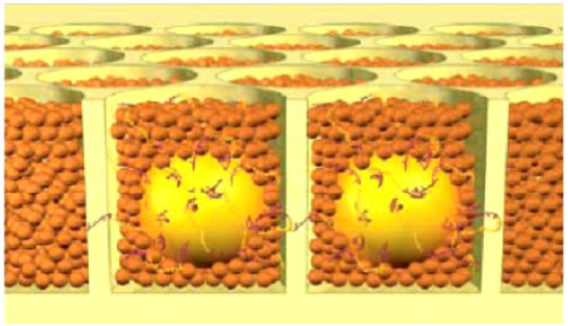
454 Step 3: Load emPCR products

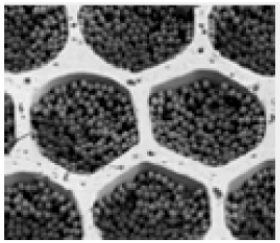




Picotitre plate

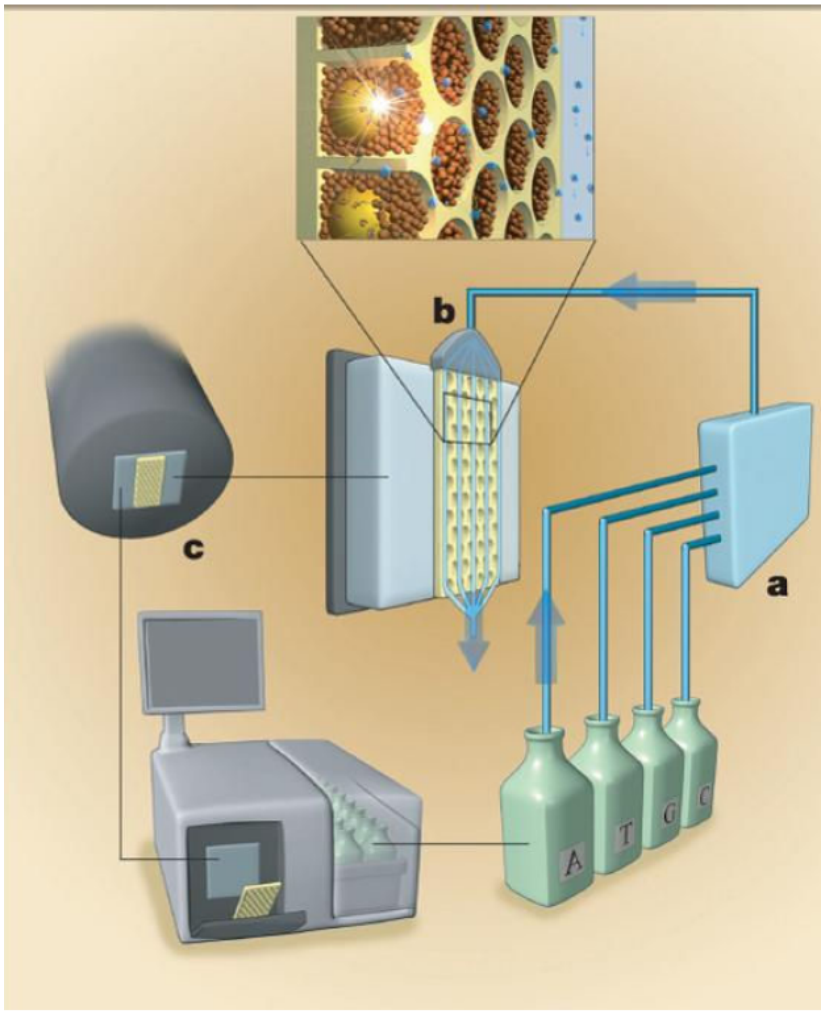
- enrich for DNA + beads
- diameter of the wells allows for only 1 bead/well





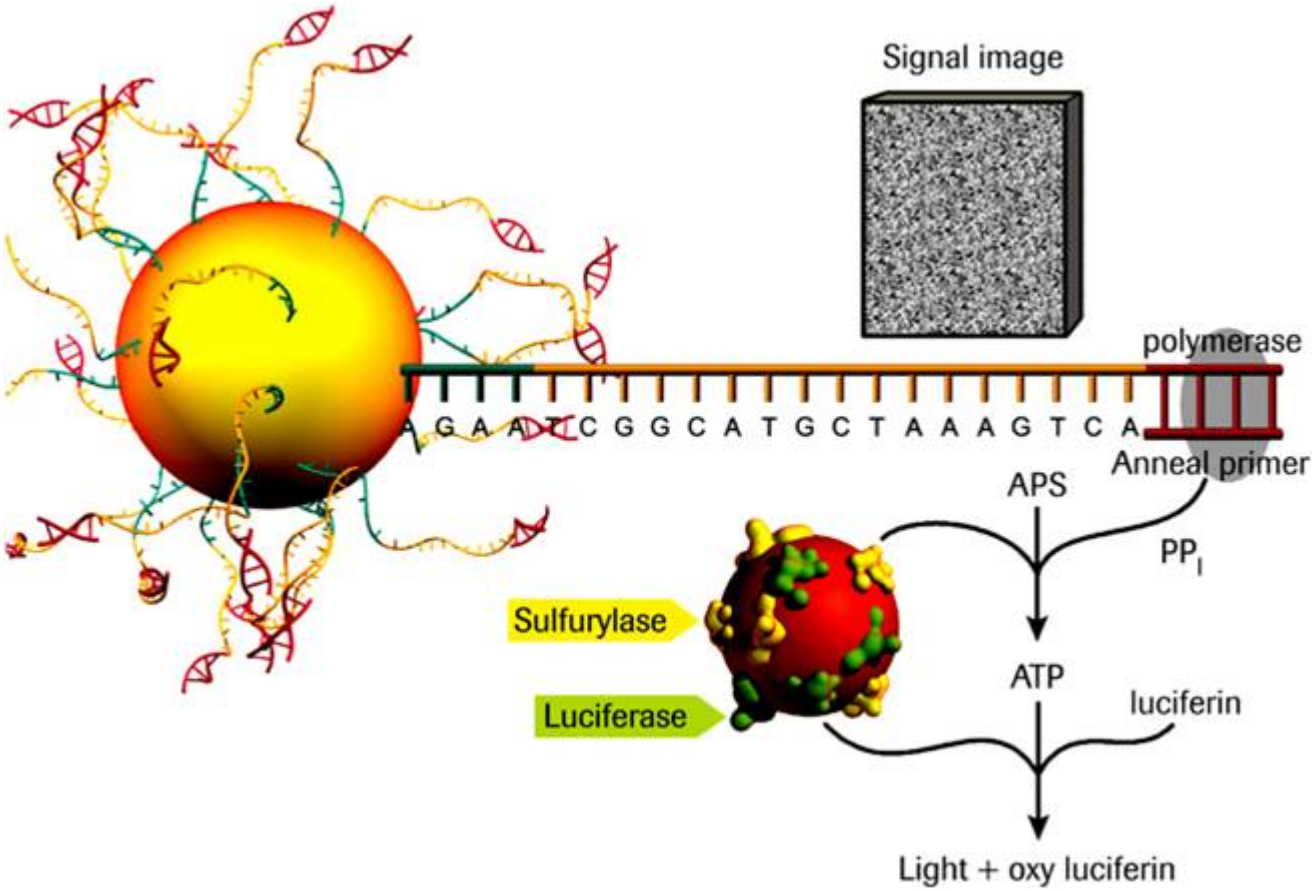
Smaller beads (red) carrying immobilized enzymes required for pyrophosphate sequencing are deposited into each well.

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



Life Technology Ion Torrent

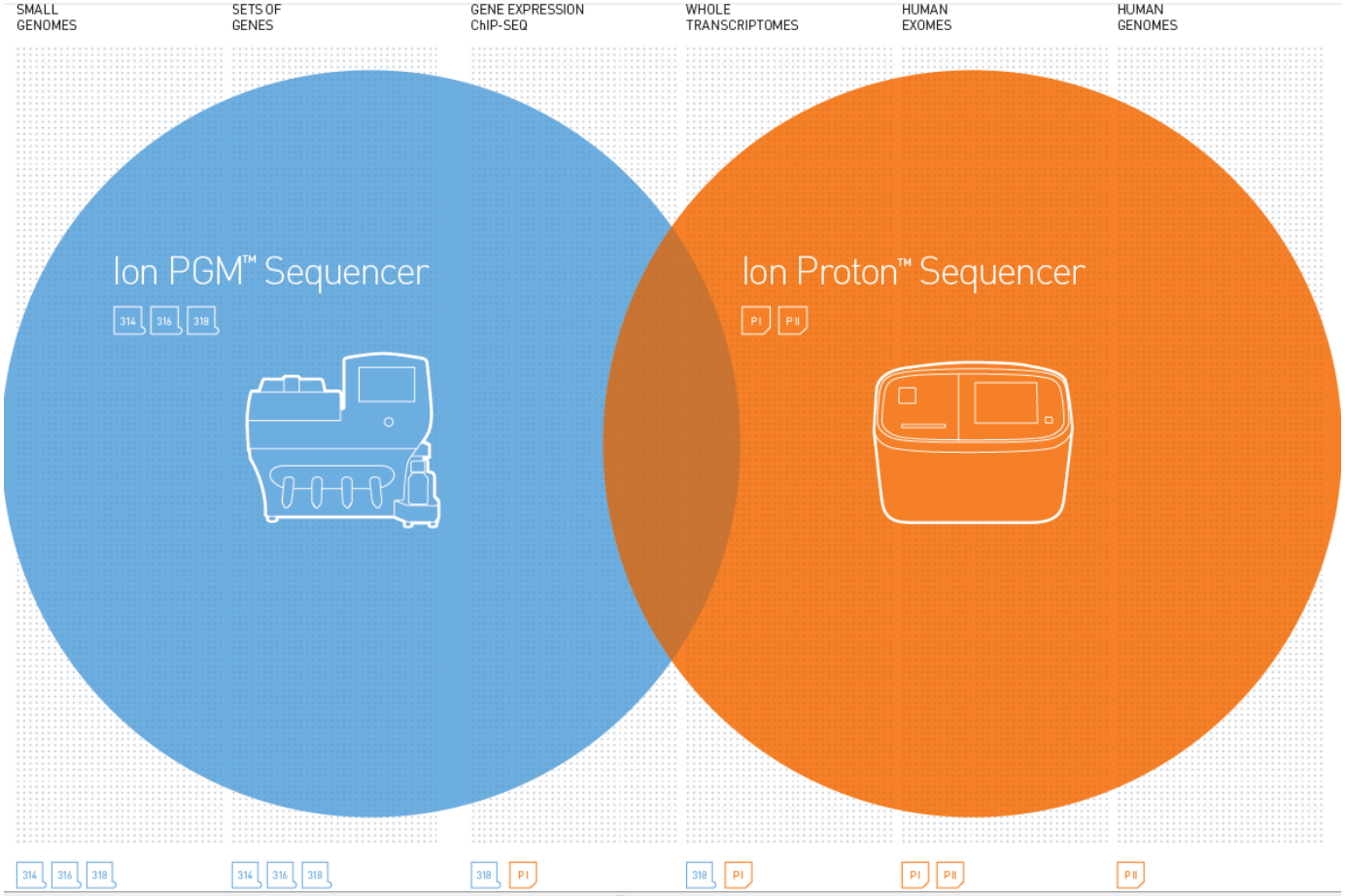
454-like chemistry without dye-labelled nucleotides

- 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

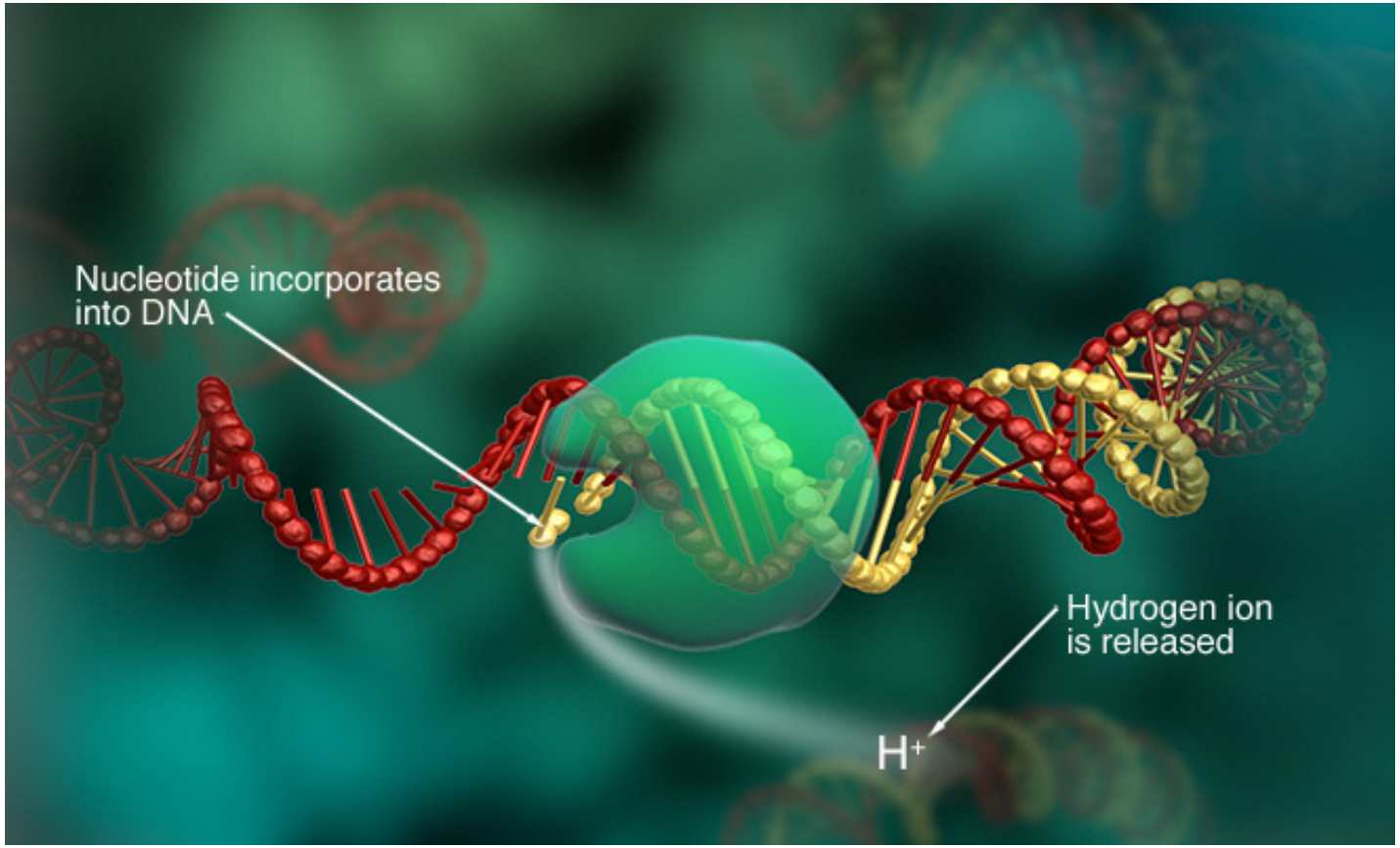
- 454-like chemistry without dye-labelled nucleotides
- No optics, CMOS chip sensor
- Up to 200bp reads (single-end)
- 2 hour run-time (+8 hours on One Touch)
- Output is dependent on chip type (P1 or P2 coming soon)
- 60-80 million reads (P1)
- **$1500 per run**
- **$150K for the instrument, plus $75k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

Ion Torrent vs Ion Proton



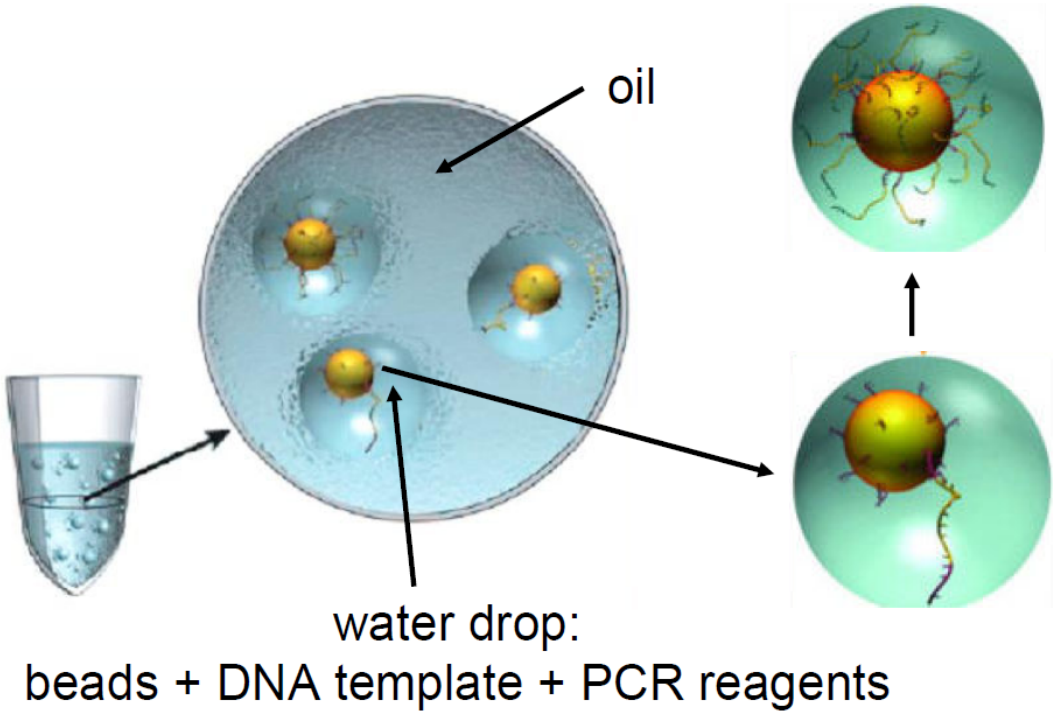
UNIVERSITY OF EXETER

Ion Torrent

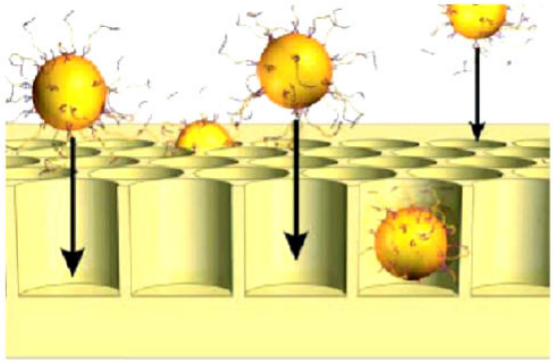

Nucleotide incorporates into DNA
Hydrogen ion is released
H+

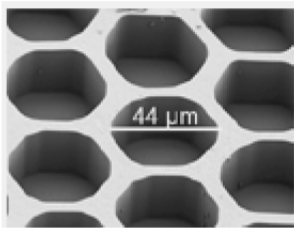
Library prep

- 454 style library using emulsion PCR

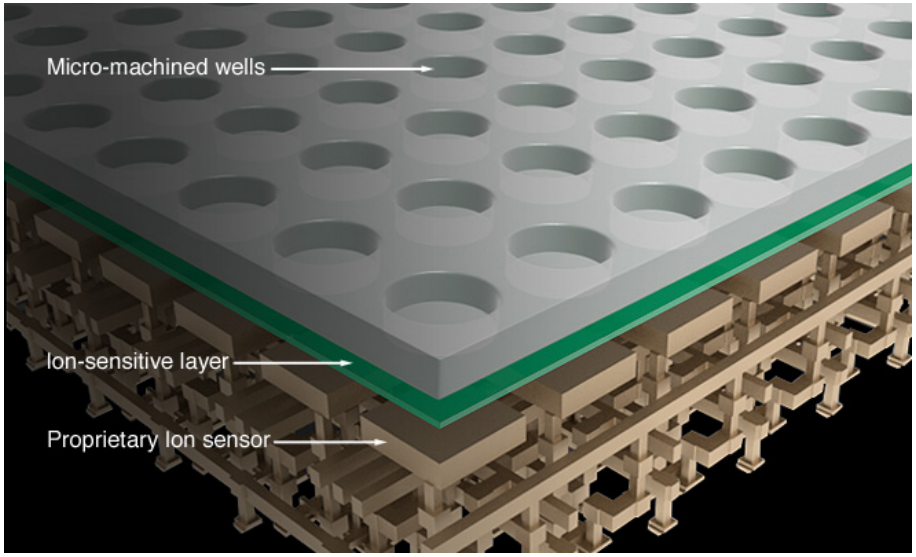


Ion Torrent

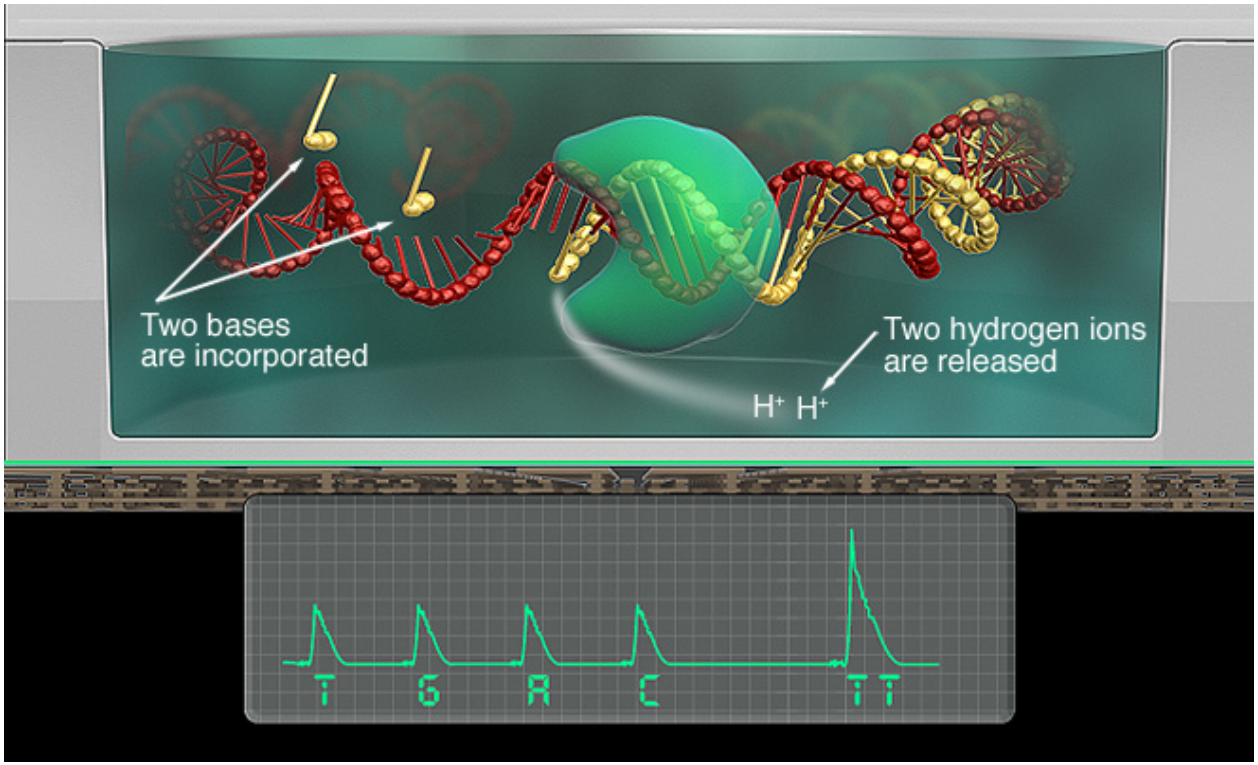




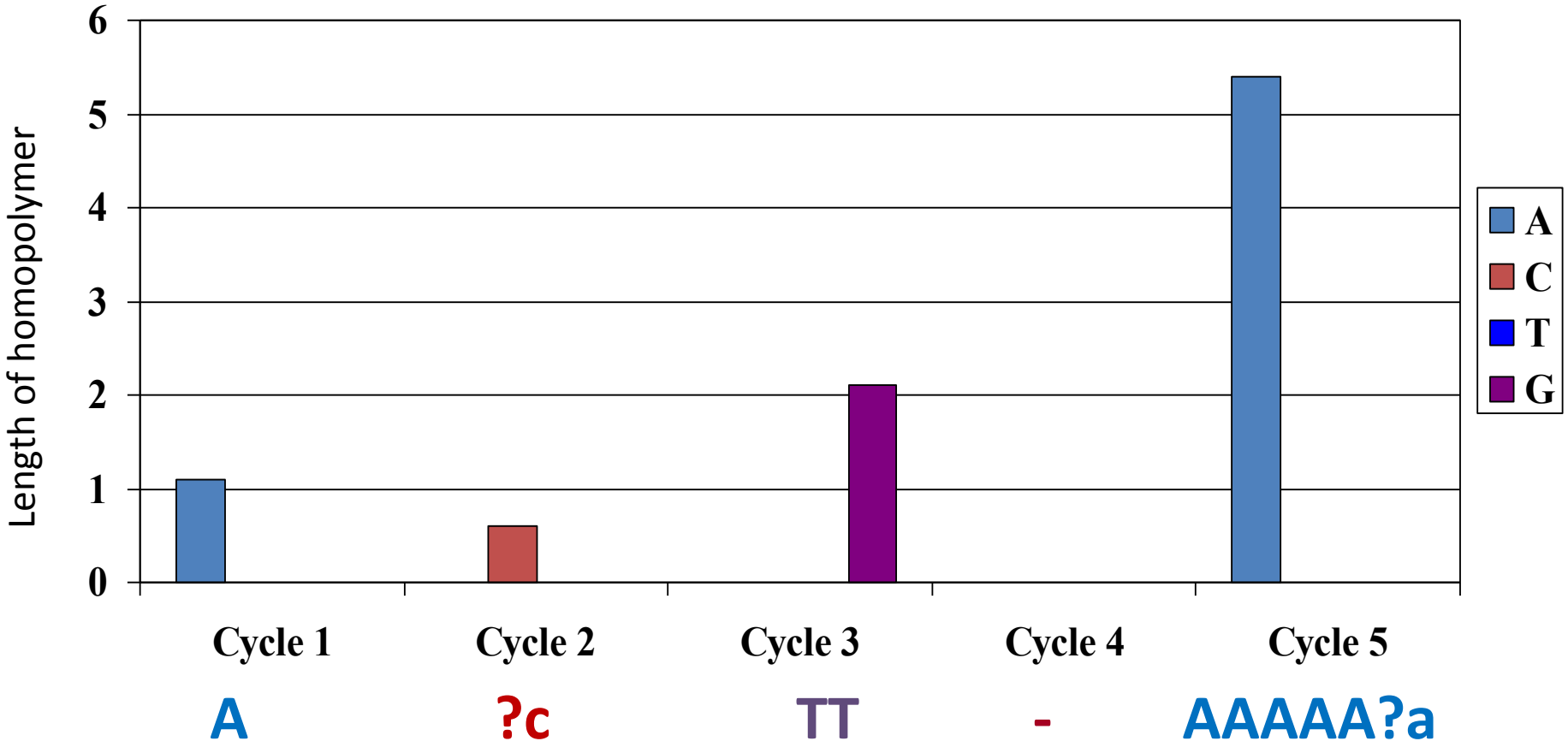
- enrich for DNA + beads
- diameter of the wells allows for only 1 bead/well



Ion System


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

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 ANTTAAAAAAA or ATTAAAA…. Etc

Third generation sequencers

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

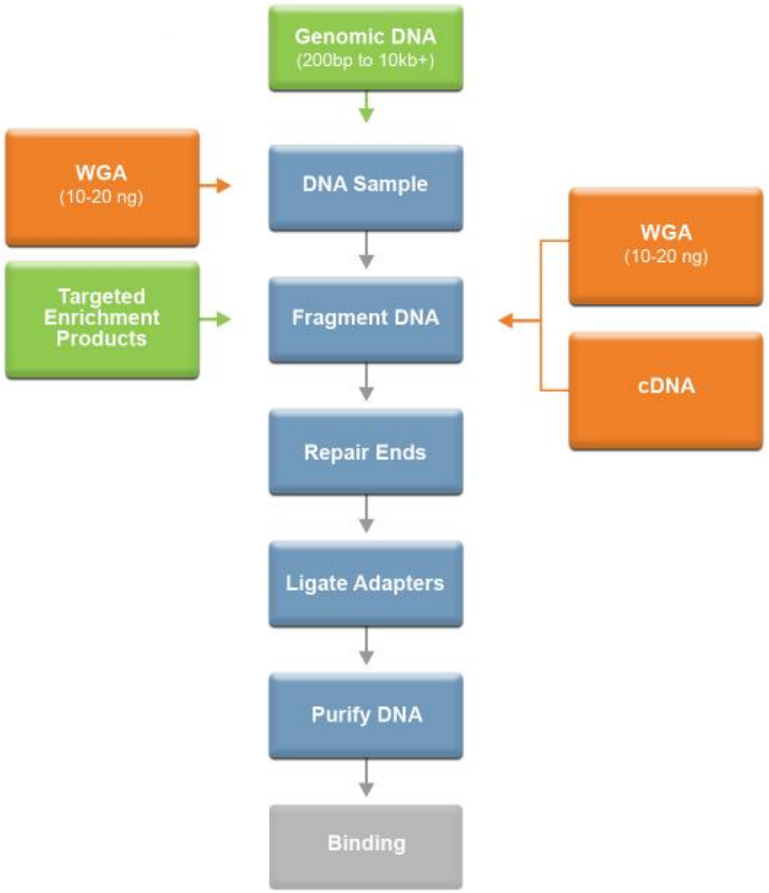
Pacific Biosciences RS II

Introduction

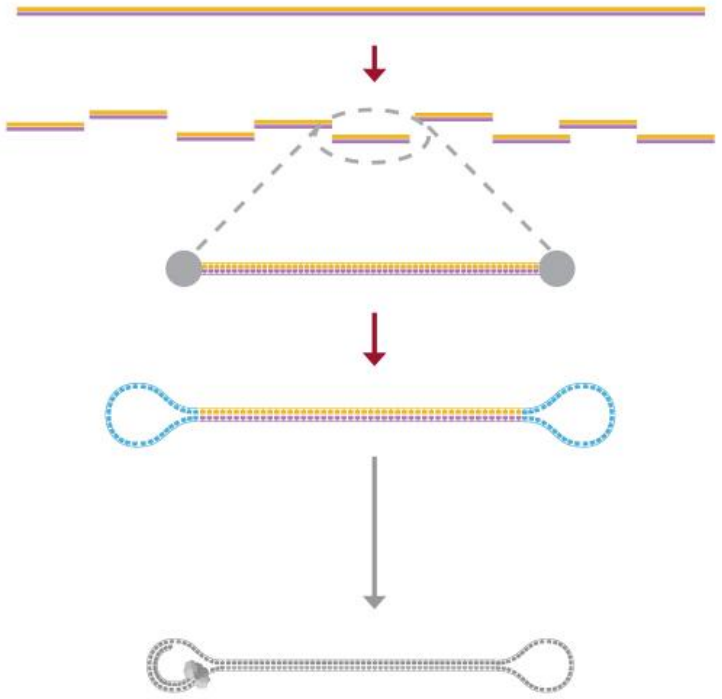
- Based on monitoring a single molecule of DNA polymerase within a zero mode waveguide (ZMW)
  - 150,000 ZMWs on a SMRT flowcell on PacBio RSII
- Nucleotides with fluorophore attached to phosphate (rather than base) diffuse in and out of ZMW (microseconds)
- As polymerase attaches complementary nucleotide,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

Library prep

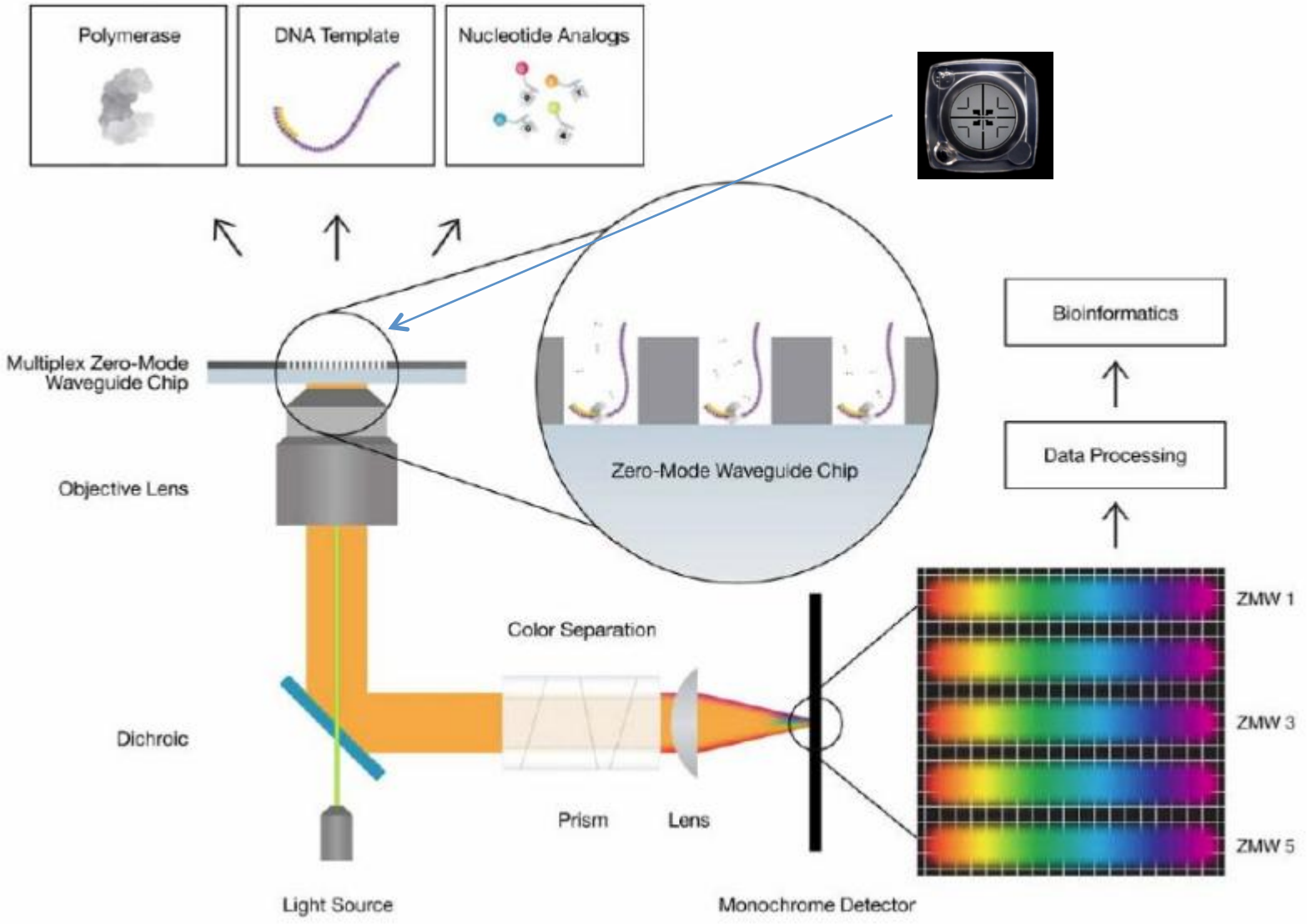
Sample Preparation

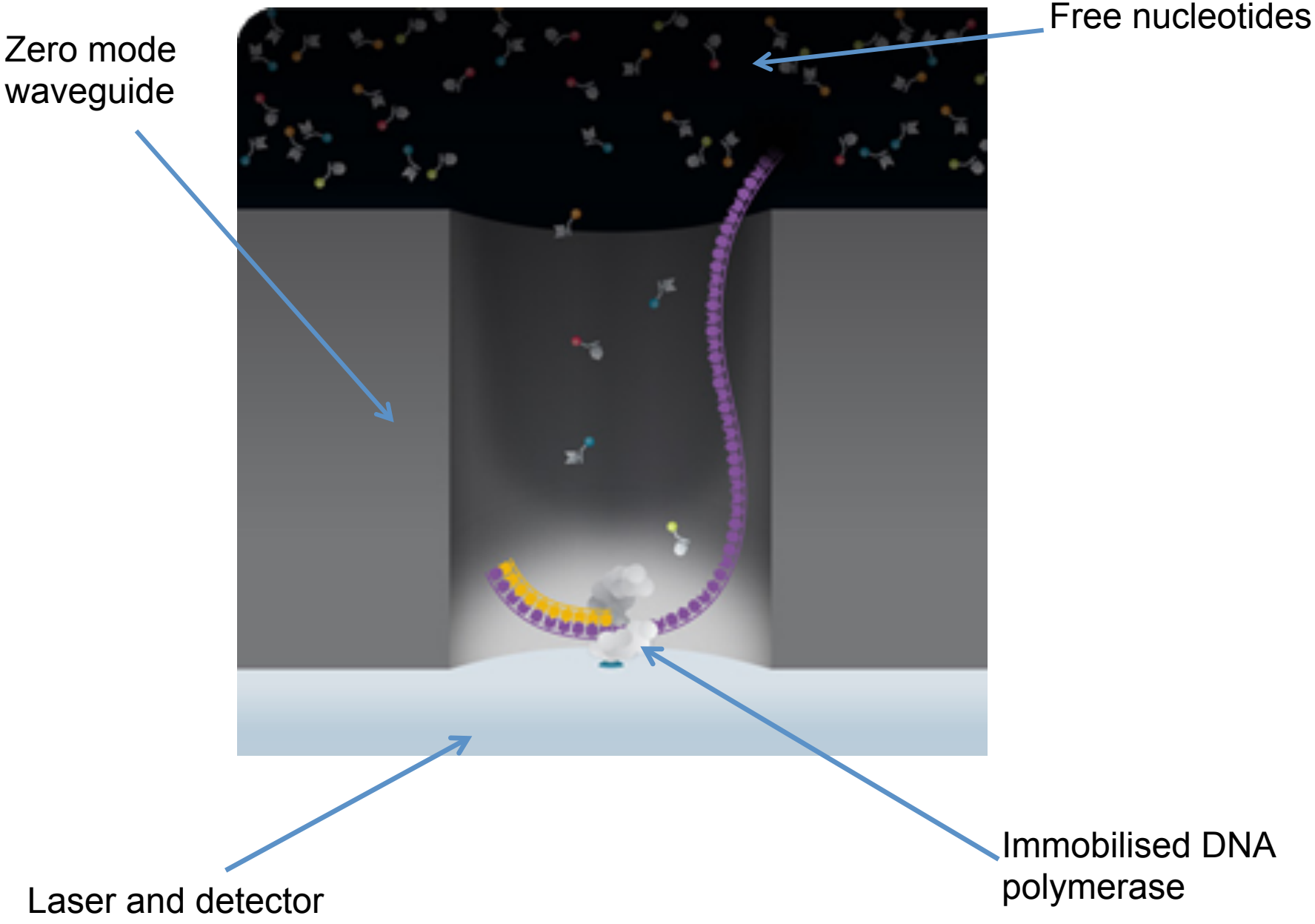


Building of SMRTbell

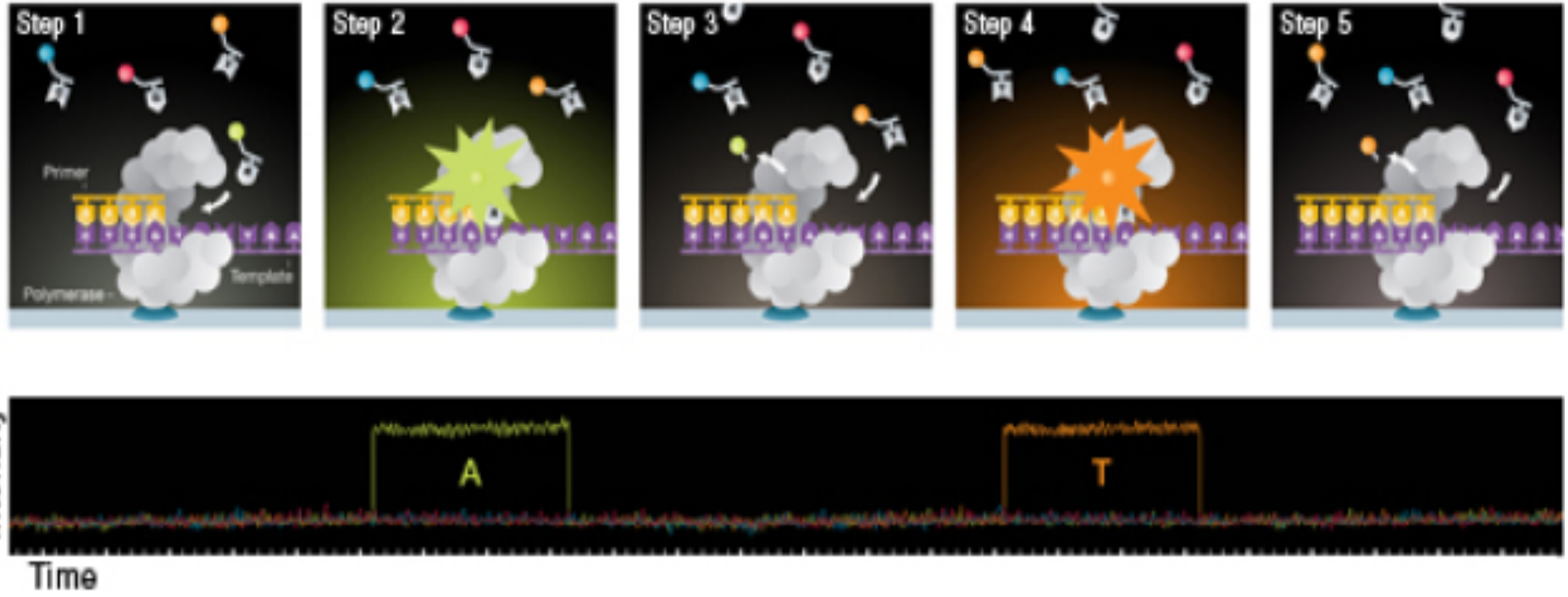


SMRT Cell

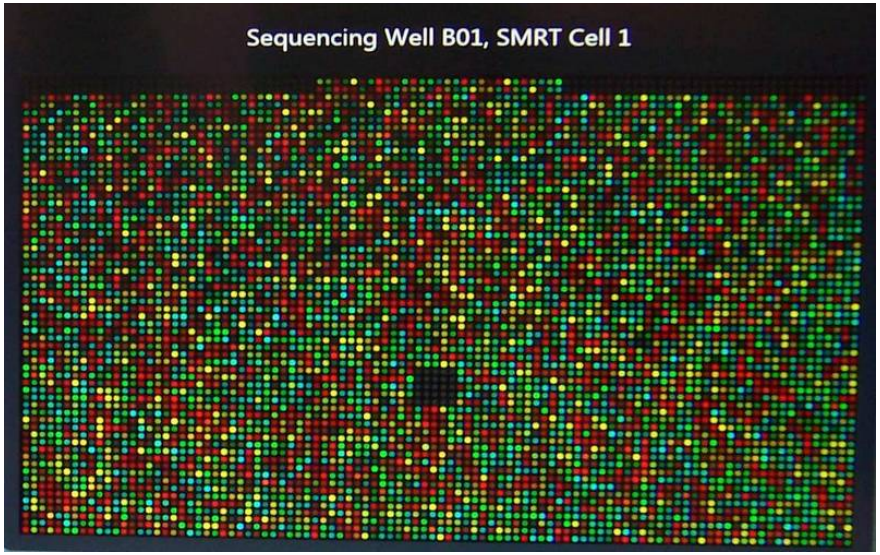

Polymerase
DNA Template
Nucleotide Analogs
Multiplex Zero-Mode Waveguide Chip
Objective Lens
Zero-Mode Waveguide Chip
Bioinformatics
Data Processing
Color Separation
Dichroic
Prism
Lens
ZMW 1
ZMW 3
ZMW 5
Light Source
Monochrome Detector


Zero mode waveguide
Free nucleotides
Laser and detector
Immobilised DNA polymerase

Observing a single polymerase

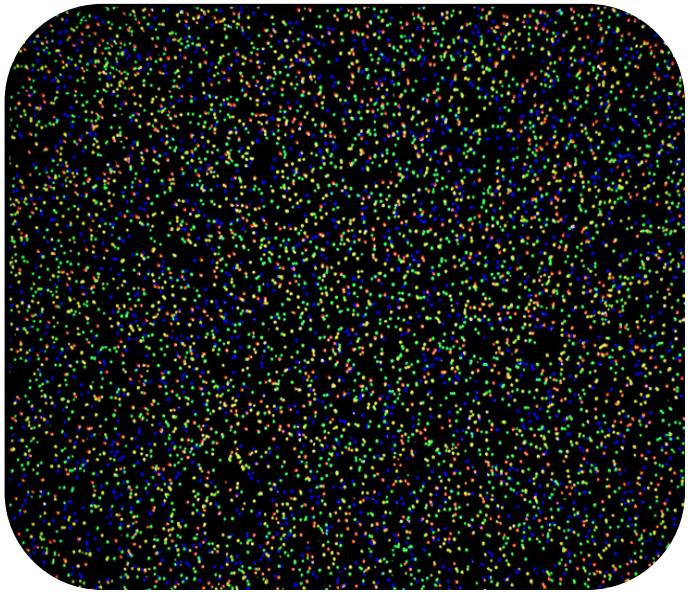


What it looks like



PacBio ZMWs with single DNA strand

Ordered



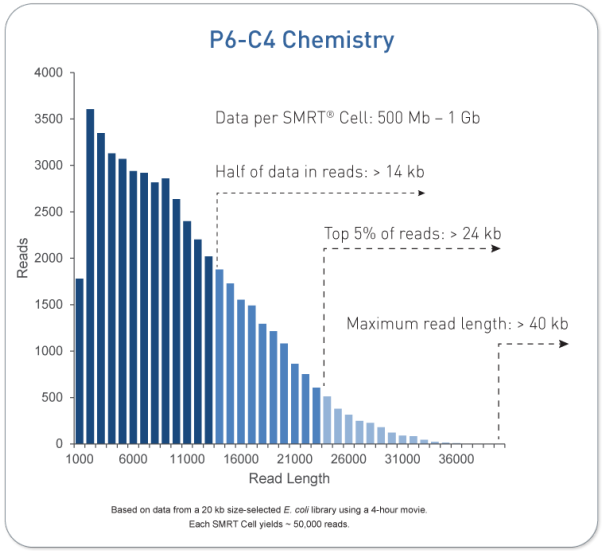
Illumina DNA mono-colonal clusters

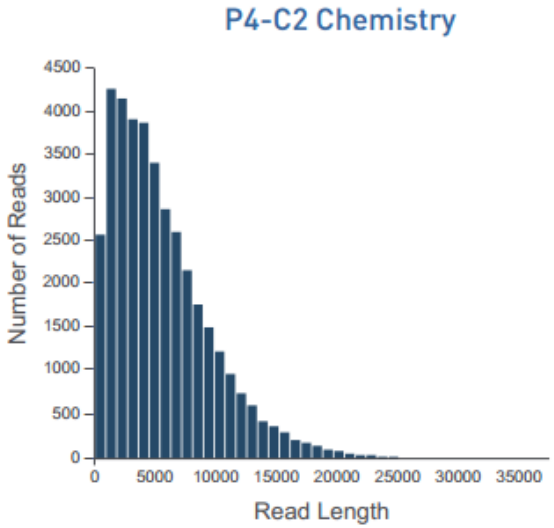
Unordered

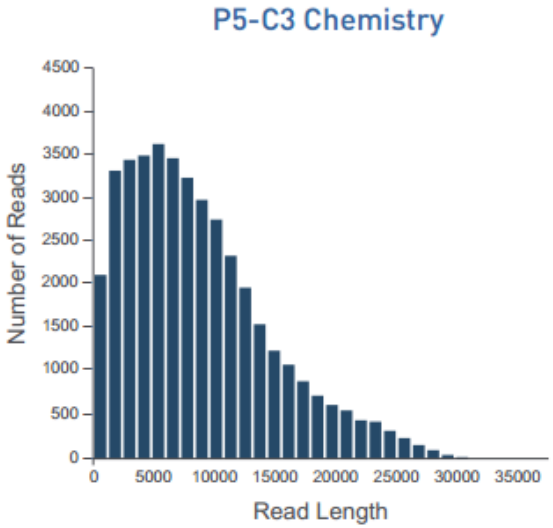
Output statistics

- Approximately 50,000 sequences per SMRT flowcell
- 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 12kb up to 20-40kb

Read lengths





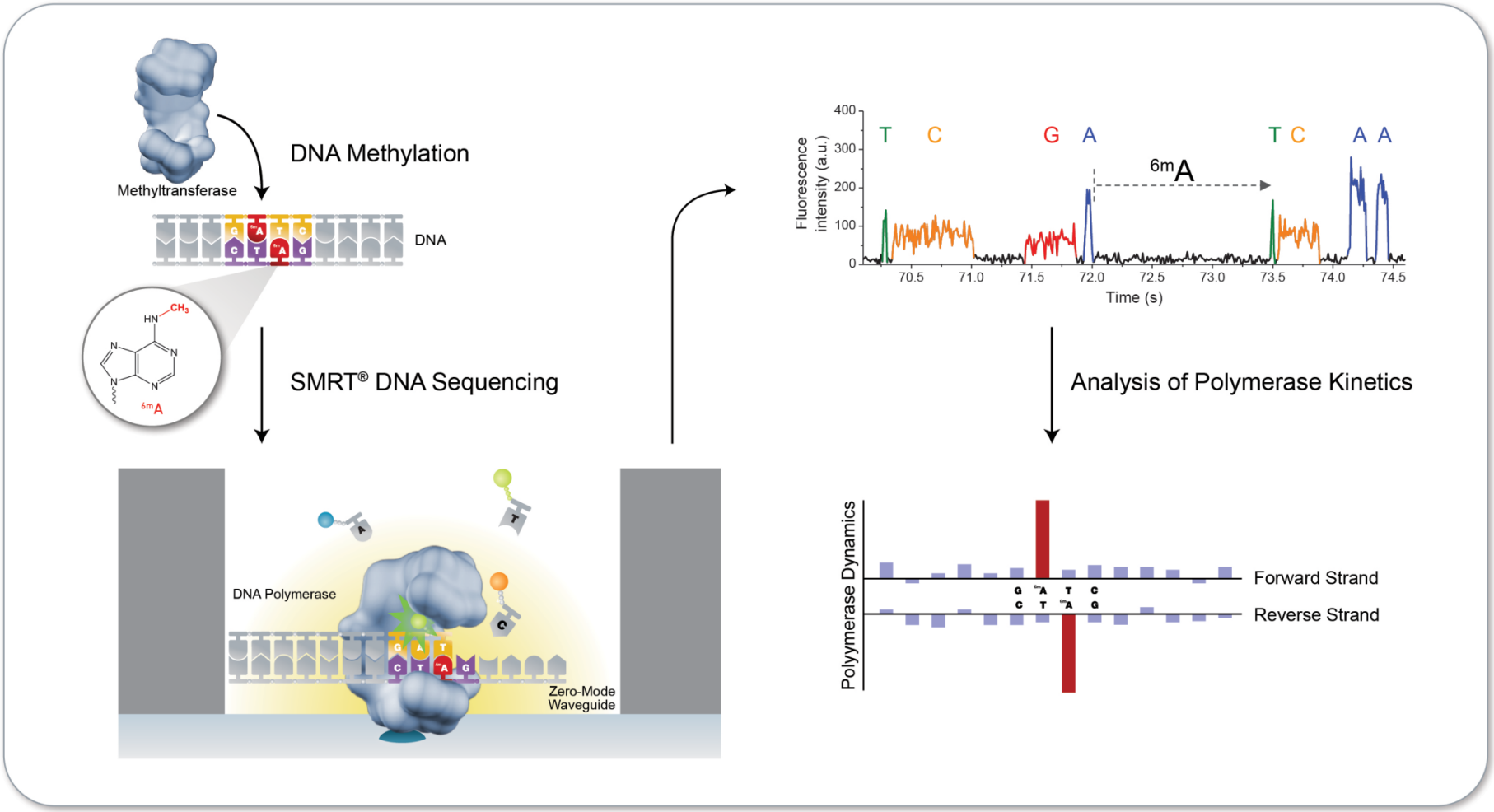


Novel applications

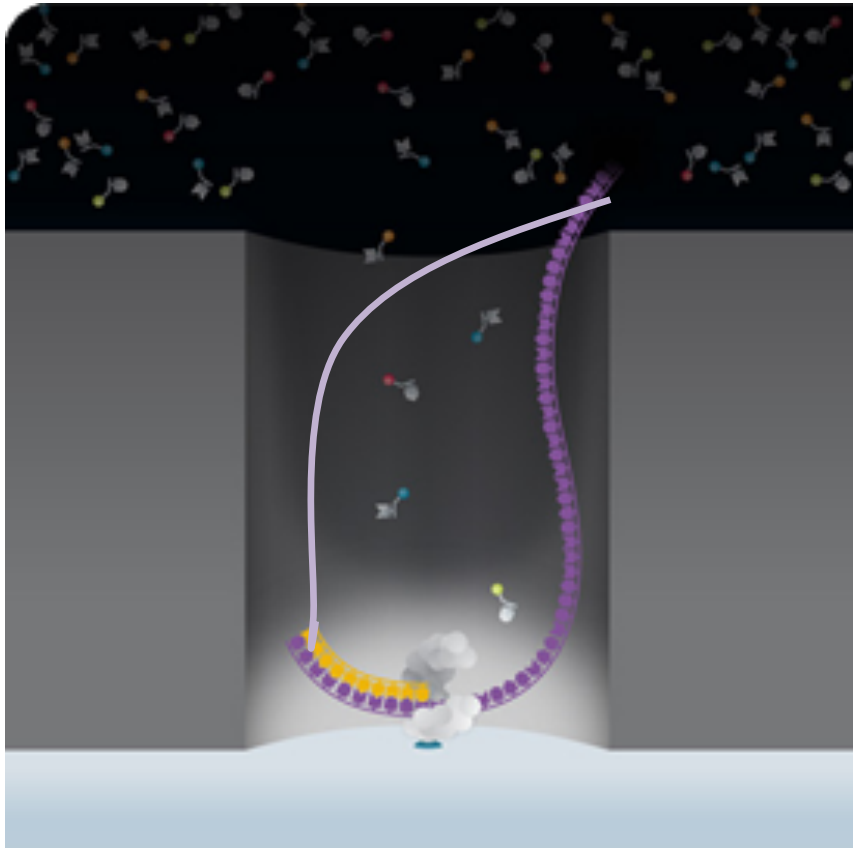
- Epigenetic changes (e.g. Methylation) affect the amount of time a fluorophore is held by the polymerase

- Circularise each DNA fragment and sequence continuously

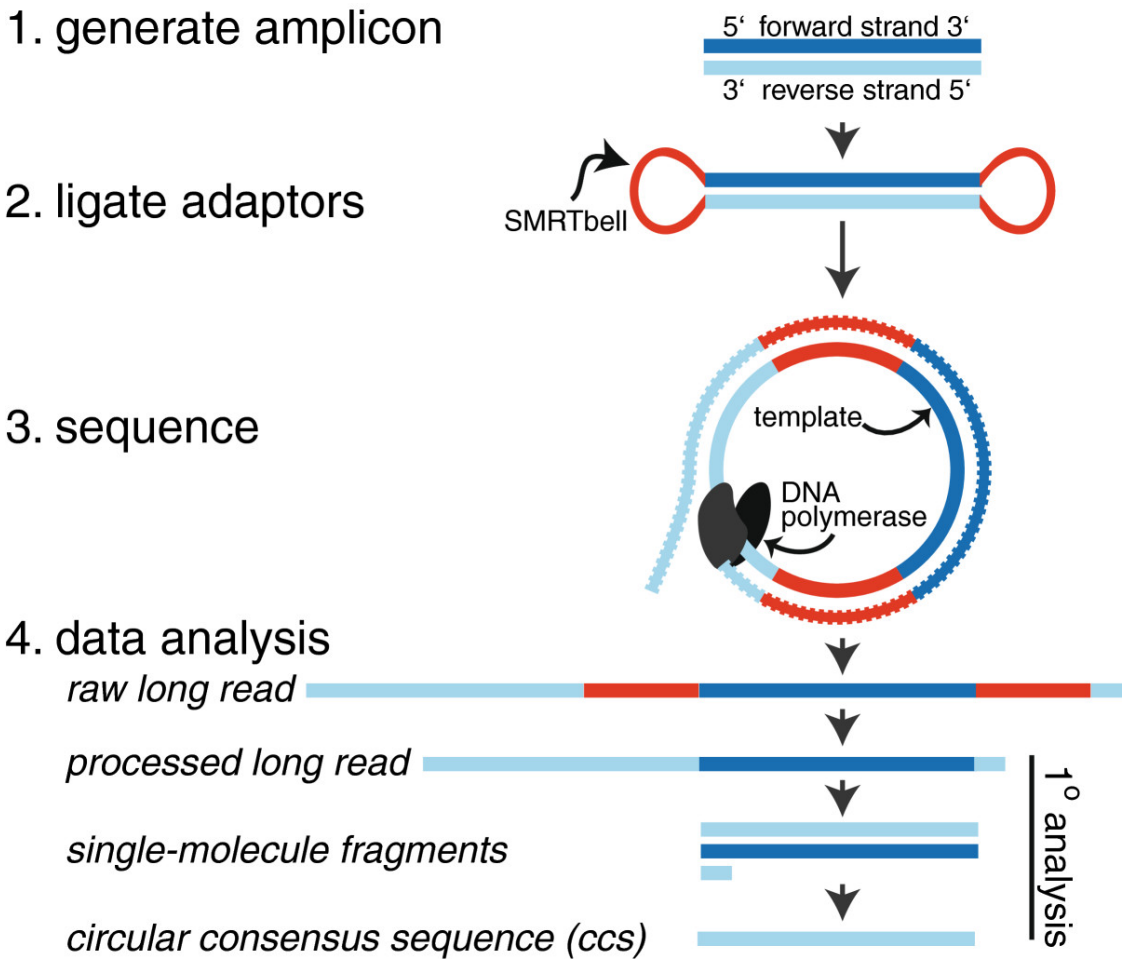
Epigenetic changes



Circular consensus sequencing

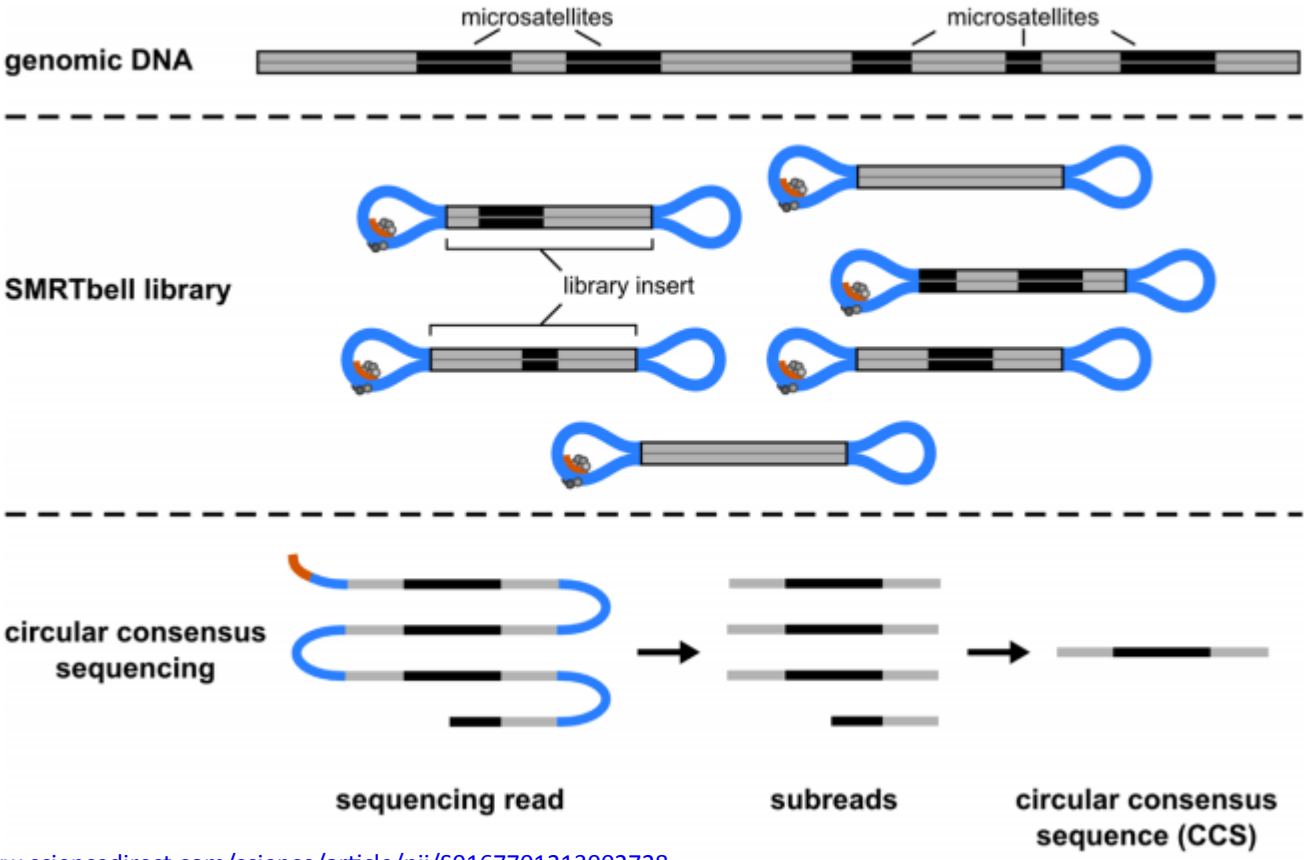


Circular consensus sequencing



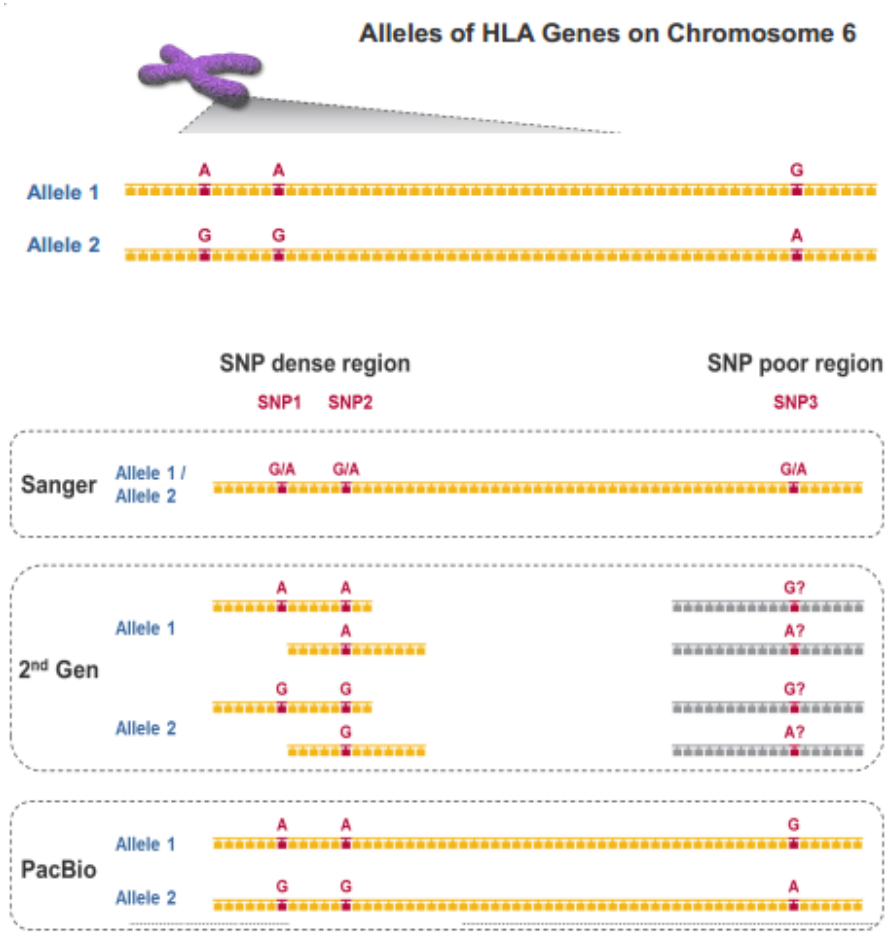
Fichot, E. B., & Norman, R. S. (2013). Microbial phylogenetic profiling with the Pacific Biosciences sequencing platform. *Microbiome*, *1*(1), 10. doi:10.1186/2049-2618-1-10

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.microbiomejournal.com/content/1/1/10

Long reads to sequence highly repetitive loci



http://www.pacificbiosciences.com/pdf/Poster_AlleleLevelSequencing_PhasingFulllengthHLAClassIandIIGenes.pdf

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 7-8 times to reduce error rates < 1%
- 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 14kb up to 40kb with P6-C4 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 still required (micrograms needed)
  - If you use PCR based methods – you are no longer seqeuncing true single molecules
  - Still enzyme based
  - Only 50,000 reads/run. 400-500Mb yield
  - High (10-15%) error rate per run (but consensus can reduce <1%)
  - $900k machine

Bioinformatics Implications

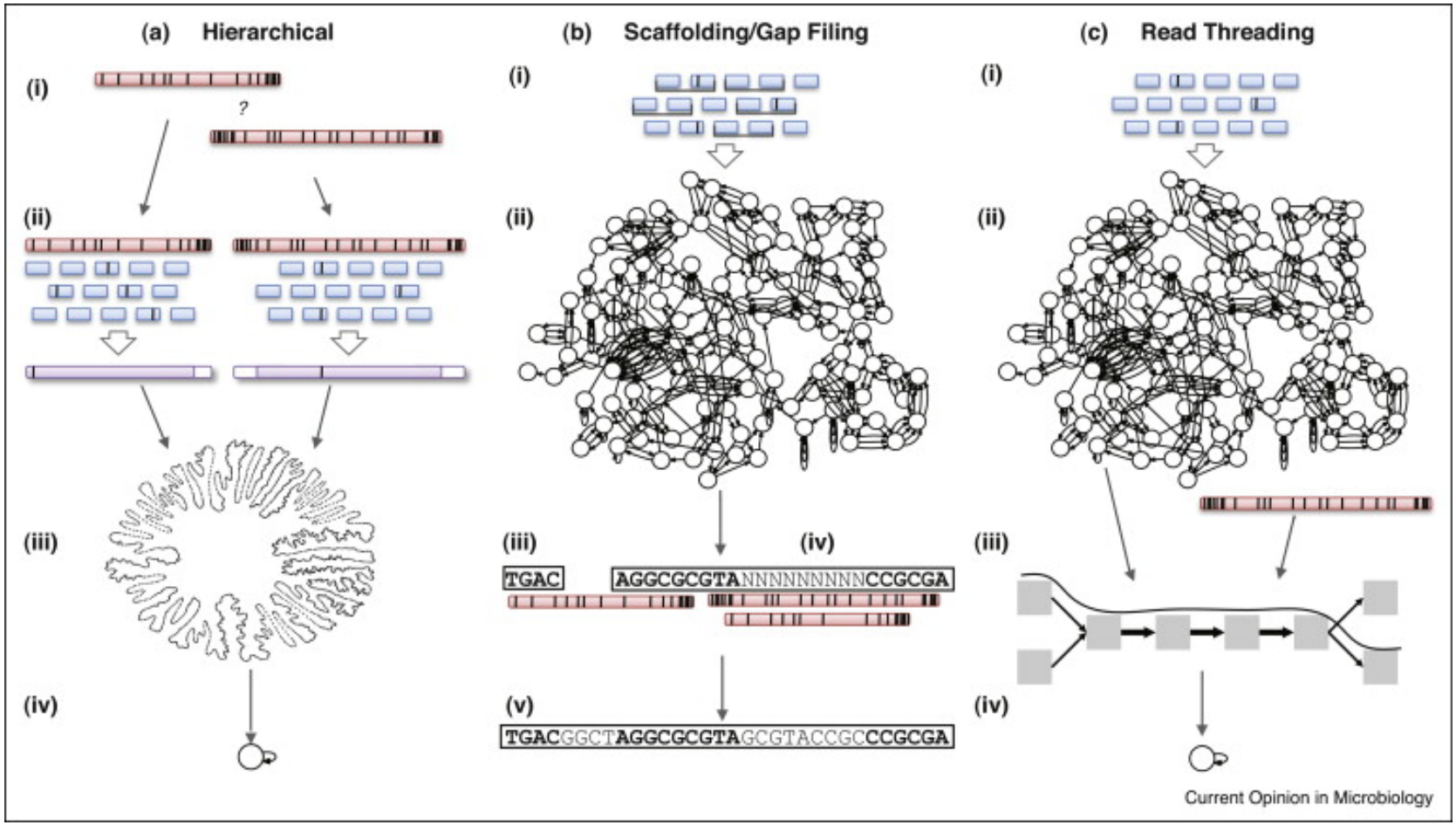
- Relatively low data and high per base cost limits practical widespread use
- Can obtain useful 20-40kb fragments (C6 chemistry)
- Best used in conjunction with error correction algorithms utilising shorter PacBio reads (or Illumina data)
- Excellent to assist scaffolding of genomes
- Able to generate complete bacterial genomes
- Has been used to generate higher eukaryote genomes (e.g. Drosophila) but cost is prohibitive

Sergey Koren, Adam M Phillippy, One chromosome, one contig: complete microbial genomes from long-read sequencing and assembly, Current Opinion in Microbiology, Volume 23, February 2015, Pages 110-120, ISSN 1369-5274, http://dx.doi.org/10.1016/j.mib.2014.11.014. (http://www.sciencedirect.com/science/article/pii/S1369527414001817)

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

Genome assembly methods



http://www.sciencedirect.com/science/article/pii/S1369527414001817

PacBio training resources

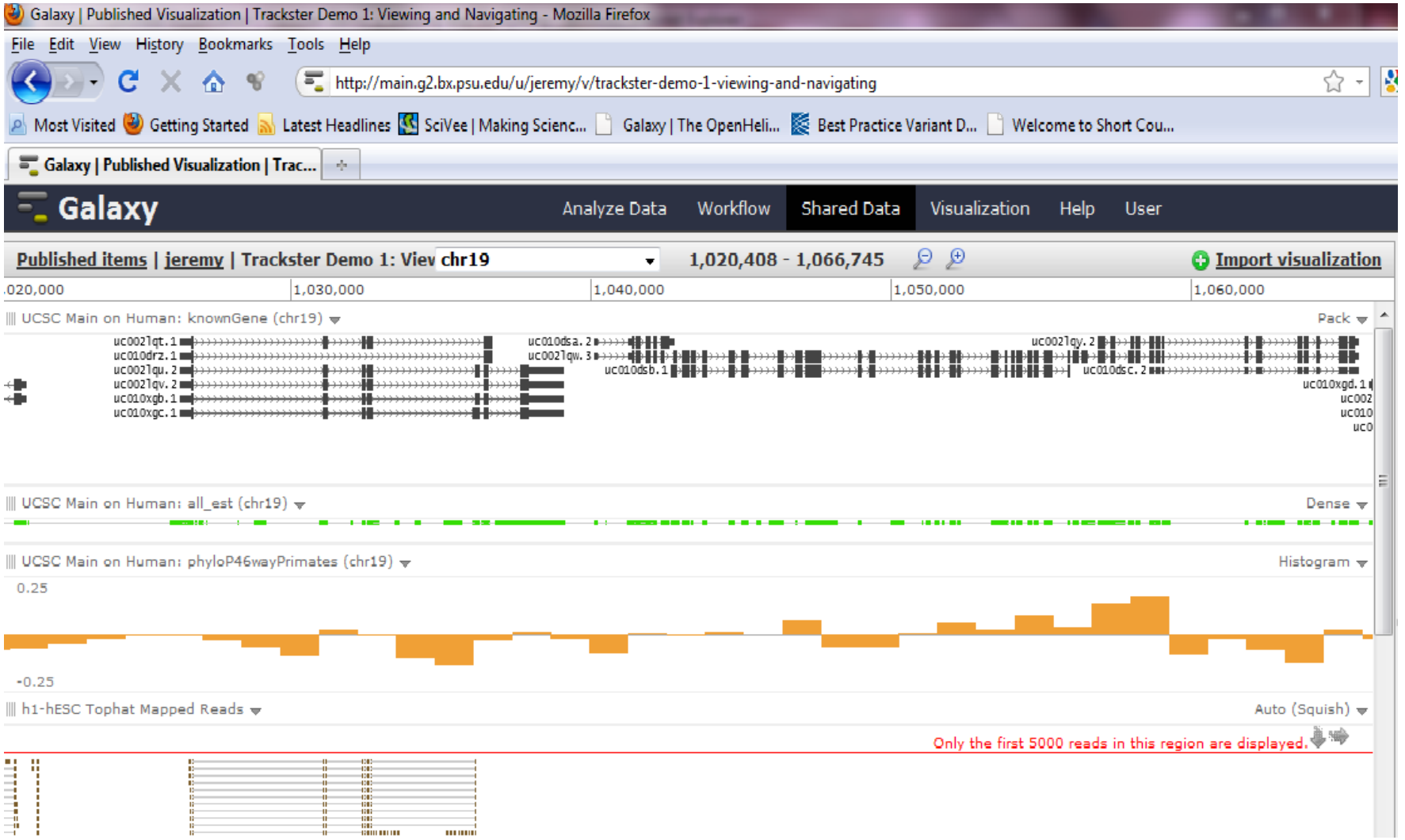
- https://github.com/PacificBiosciences/Bioinformatics-Training/wiki

Data processing

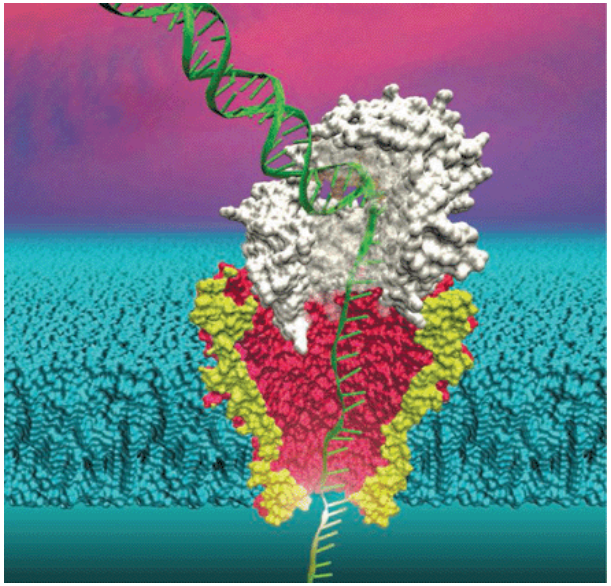
konrad@zeus:~/emmatest/raw_illumina_reads

```
#**+,**.*-/1473*,*(**********+-31*51/031,*3301/7*-3/0+**&**#%*.**.*.0.&*#*(%
@IL29_5550:3:1:2664:952#1/1
NATAATTCTTCACTCTTTCTCCAACAGCTTCTTTAAACAAACCTTCTATCTTAGCTATCNCTTCTTTGTTTTCTTC
+
#033313373>>>>>>>>>>>>>>>>>>>>>>>><><>>>>>>>>>>>:>>>>>>5577#17337>>>>>>>>>>>
@IL29_5550:3:1:3076:931#1/1
NATACATCAATAGTGTTATCTCTATCTGGACGACCTATATCCTTTTTAAATTTTAAGTTNAATACTTTAGGATGTT
+
#777737705>>>>>>>>>>>>>><><>>><>:>><>>>>>>>>>>>>>>>>>>>7153#5.3(4:9:96><<>>7
@IL29_5550:3:1:3195:946#1/1
NAGATAATGTATCAGTAACCATAACATCAGGTGCACCTACAAGTAAATCATTACCTCTCNTTACACATCCACCATC
+
#-+******.>>>7:67777><<7<40773(..300-.5/666/663674354447371#*0/34>>>>>99/98>
@IL29_5550:3:1:3412:940#1/1
NGATATTAACTTTATCAAACCAGGTGTAGGGGAAACGACAAGGGTTTTACTTAGAAGAGNTCCGTATAAGATTCTT
+
#/0-.03435>>>><>>:>>>>><>67677><<<><>><<9966:>>>>>>>>>>37/3#33.03:>>>:<>>><<
@IL29_5550:3:1:3748:930#1/1
NTTTATAAAATGAGATTGGTCGTAATAACCAATATCTTGTGCTAAAGTTTCAAATGGTCNTTTATTGTTTTGTACT
+
#11477/73,>>>>>>>><>73.0311411>>>>>><>>><<>>><>>><,33.15,73#7..7366666><>>>:
@IL29_5550:3:1:3843:938#1/1
NCAAAATCTTTTGCAGCTTTTTCAAAAGTCTCTACCATGTGTTTAGTTTTAACATCTAANTTATTTTCTTTTAGAC
+
#703377544>><>>>><>>>>>>>>>>>>>><>>><>>><>>>>6776(.77/4144.#37043>>><>>>><>3
@IL29_5550:3:1:3884:930#1/1
NCAACTTTCACAGTTAAGACTGACCAAGAAACAGGACAAACTATAATATCTGGTATGGGNGAATTACACTTAGAGA
+
#507075441>>:<<>><<>>>>>>>>>>>>>>>9>>>>>>>>:>>>>>><>>>>1754#371.1<>>>><>>>><
```

Data processing

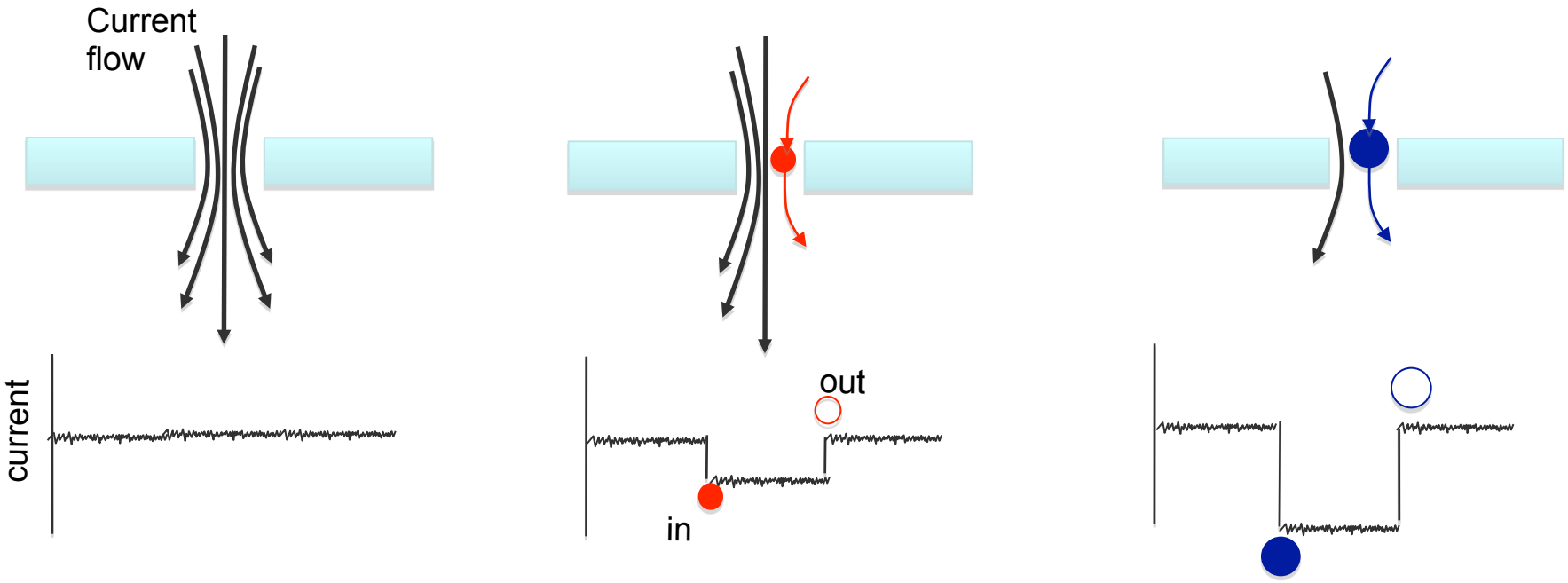


Nanopore sequencing

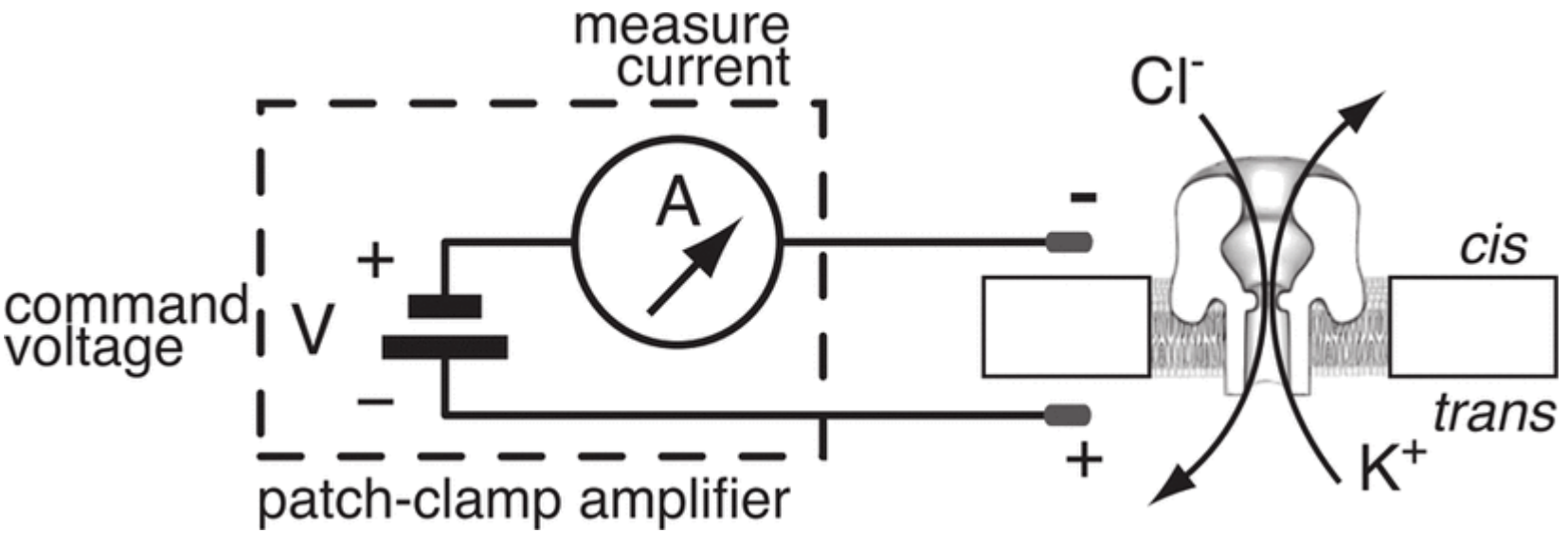


What is a nanopore?

- 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



Detection



http://ieeexplore.ieee.org/xpls/icp.jsp?arnumber=5626570

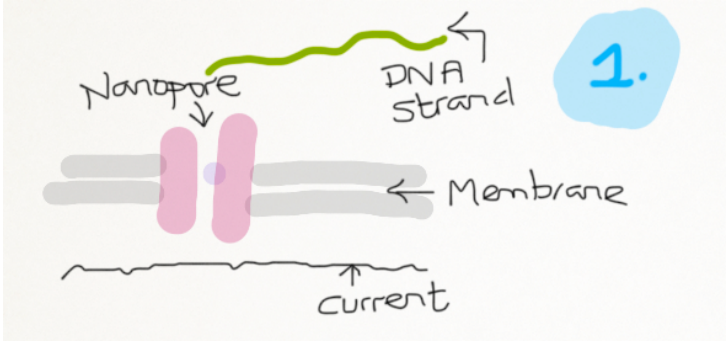
Types of pore

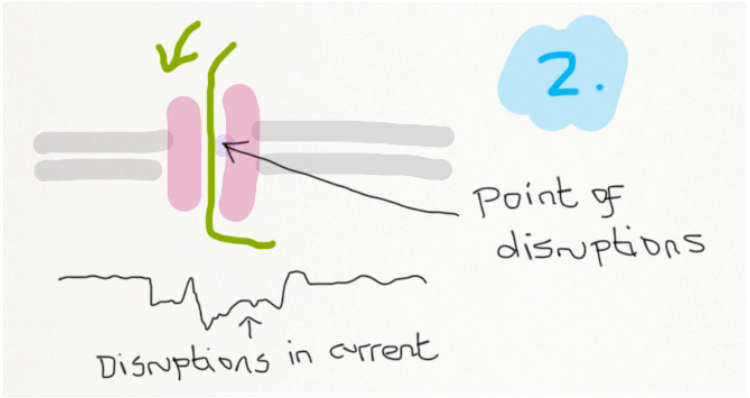
- Either biological or synthetic

- 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 DNA 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

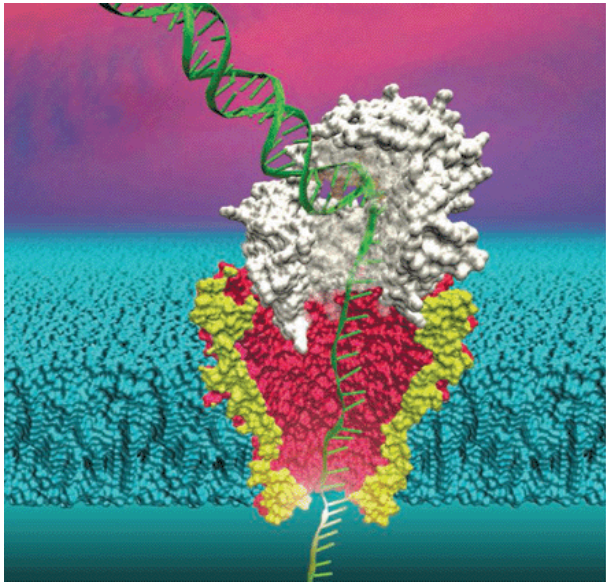




http://thenerdyvet.com/category/tech/

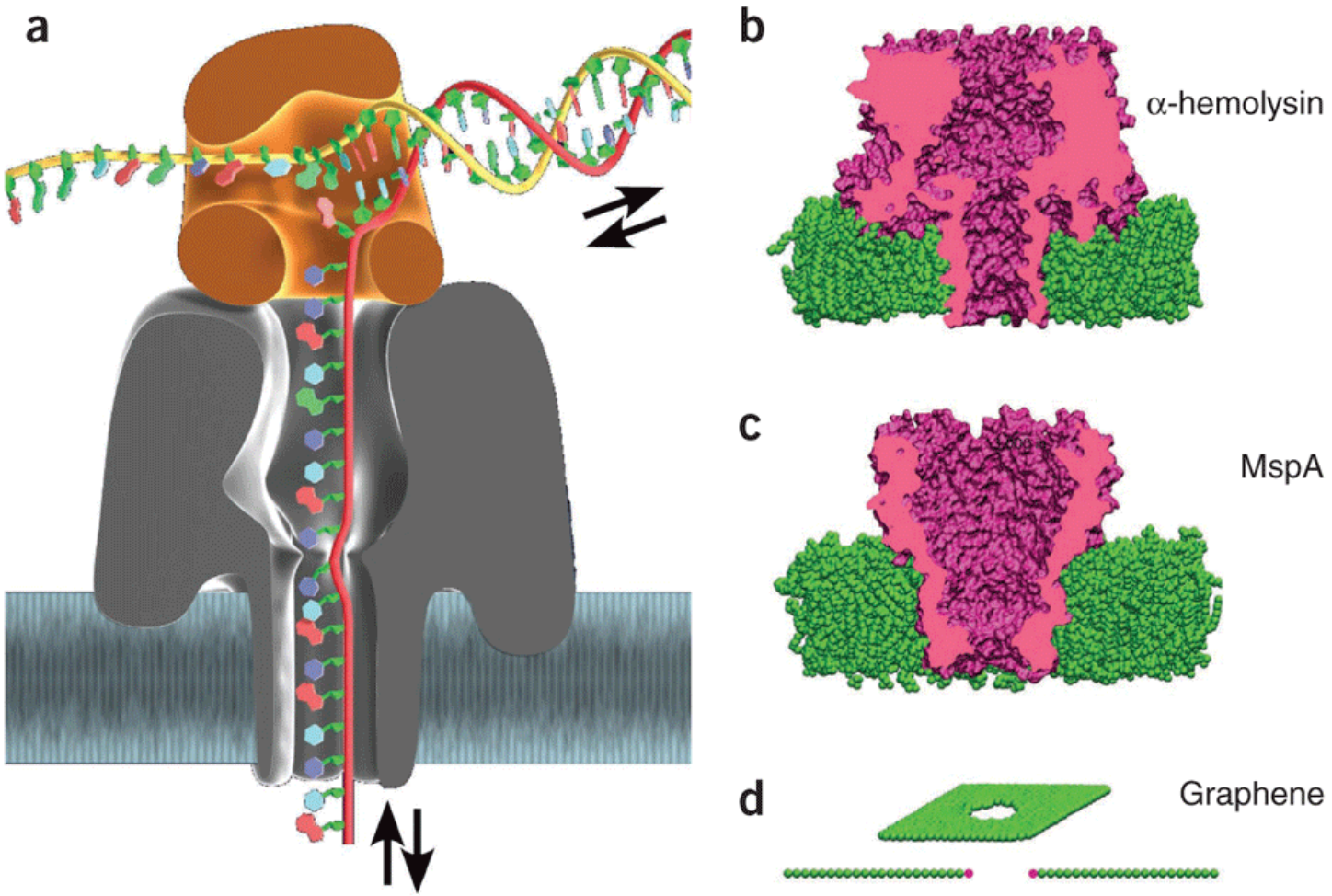
Nanopore sequencing

- 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
  - Search space for DNA to find a pore is large



http://omicfrontiers.com/2014/04/10/nanopore

Nanopore sequencing



http://www.nature.com/nbt/journal/v30/n4/full/nbt.2181.html

Nucleotide Recognition

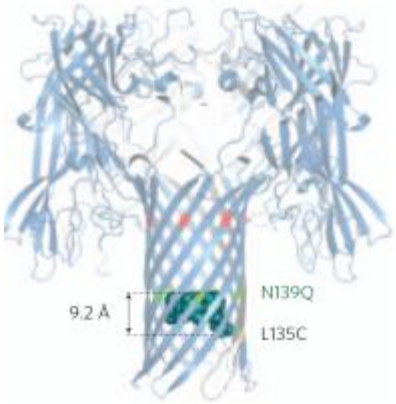
nature nanotechnology

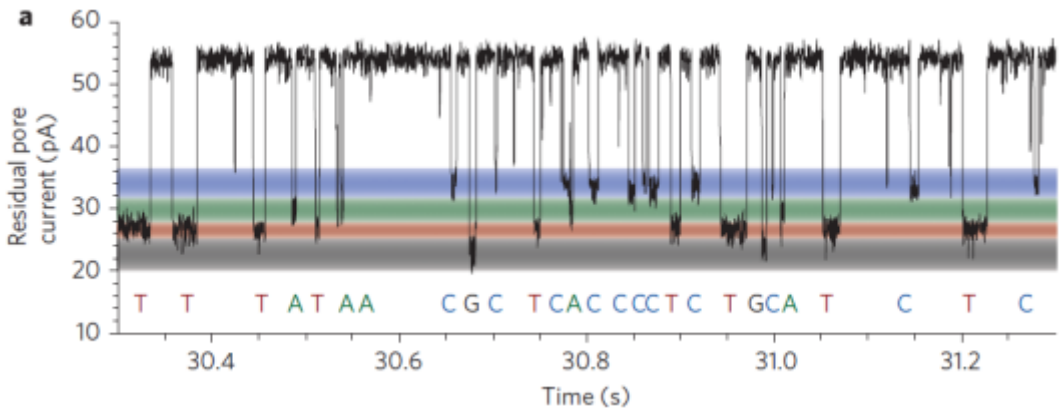
ARTICLES

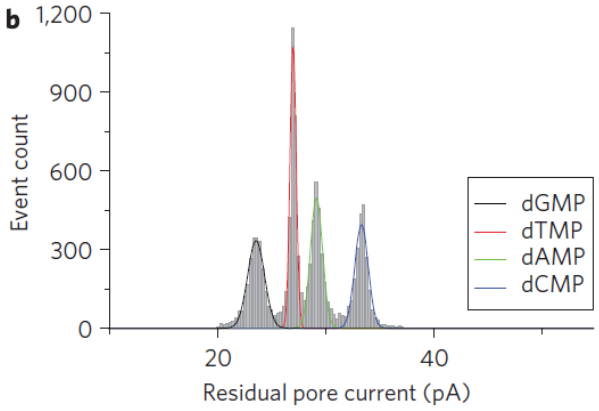
PUBLISHED ONLINE: XX XX 2009 | DOI: 10.1038/NNANO.2009.12

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]*







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 capacitive 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

Companies involved

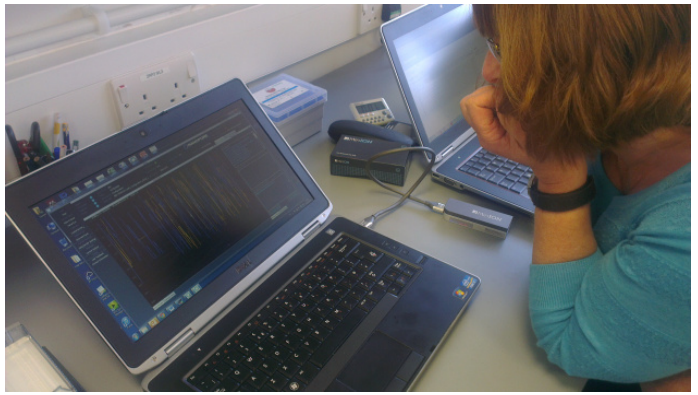
- ONT is closest to commercialisation
- Two approaches to sequencing
  - Exo-nuclease sequencing (originally part of a co-marketing agreement with Illumina)
  - Strand sequencing (now in commercialisation)

- Strand sequencing method is being commercialised by Oxford Nanopore

Oxford Nanopore MAP programme

- MinIon Access Programme
- Provides access to several flowcells and reagents for library preparation
- Beta testing programme – round 1
- Round 1 closed to new applicants
- Round 2 may re-open applications

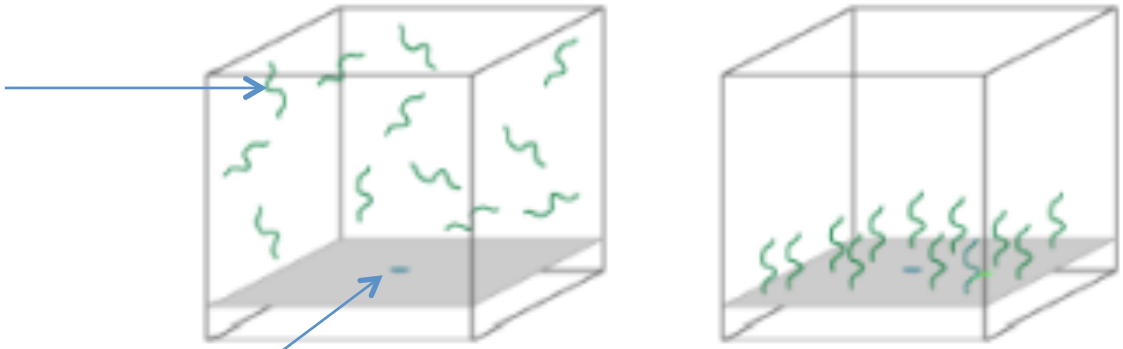
Oxford Nanopore



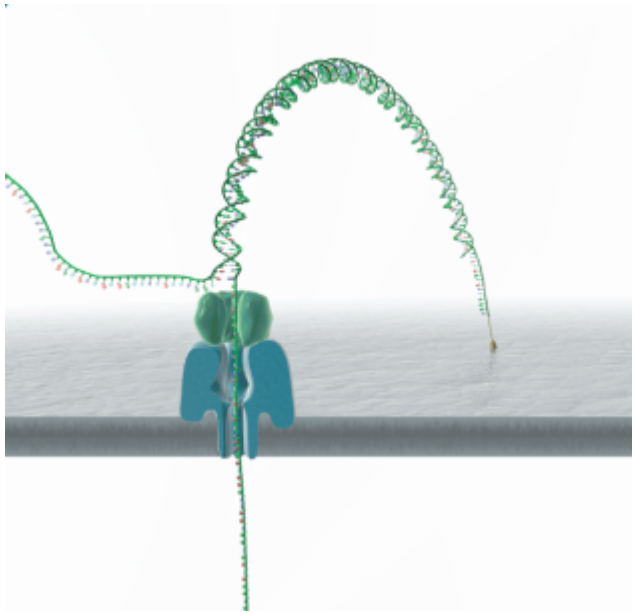
Oxford Nanopore platforms

DNA binding to membrane

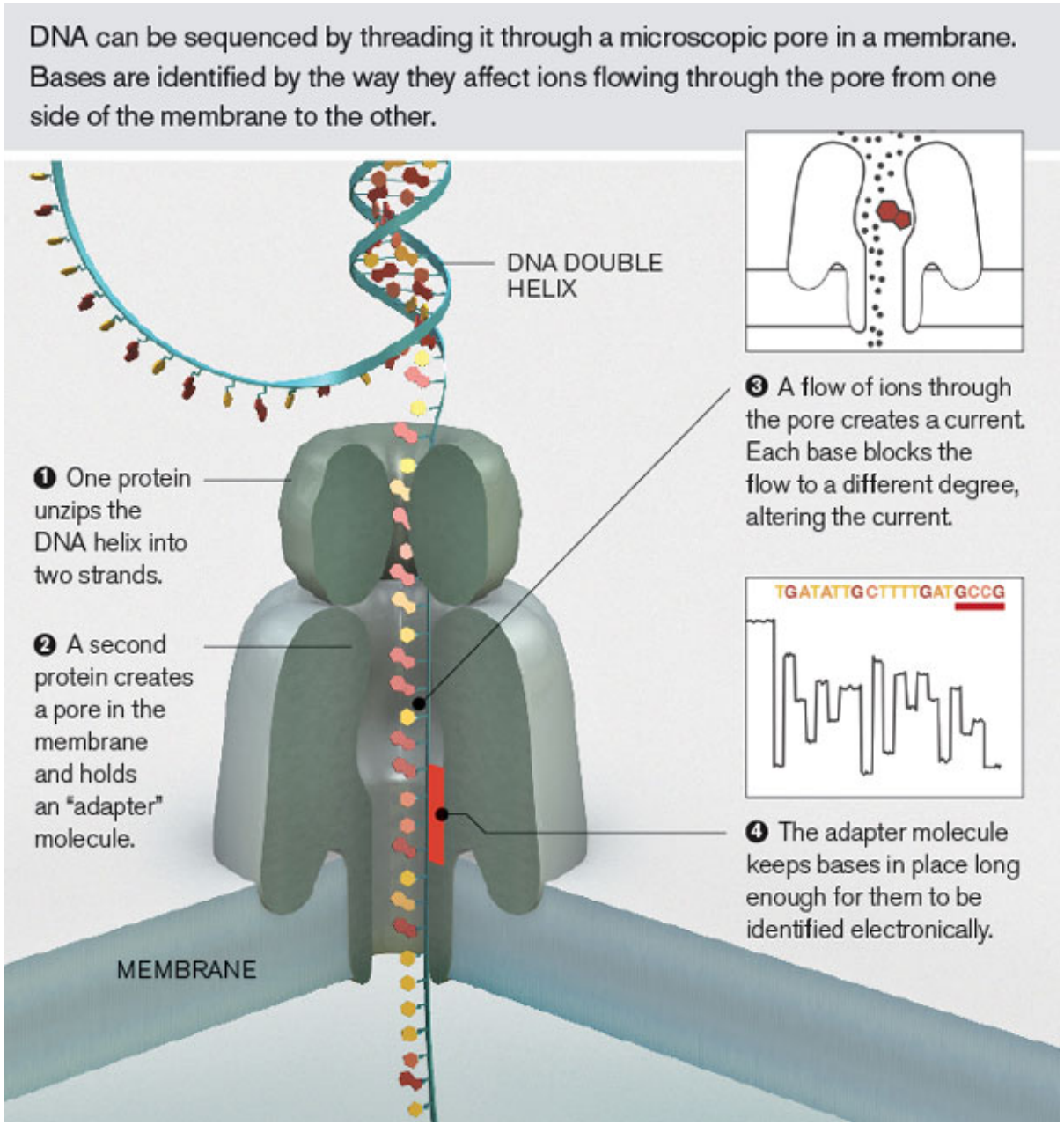
Double stranded DNA fragment



Pore



Oxford Nanopore principle



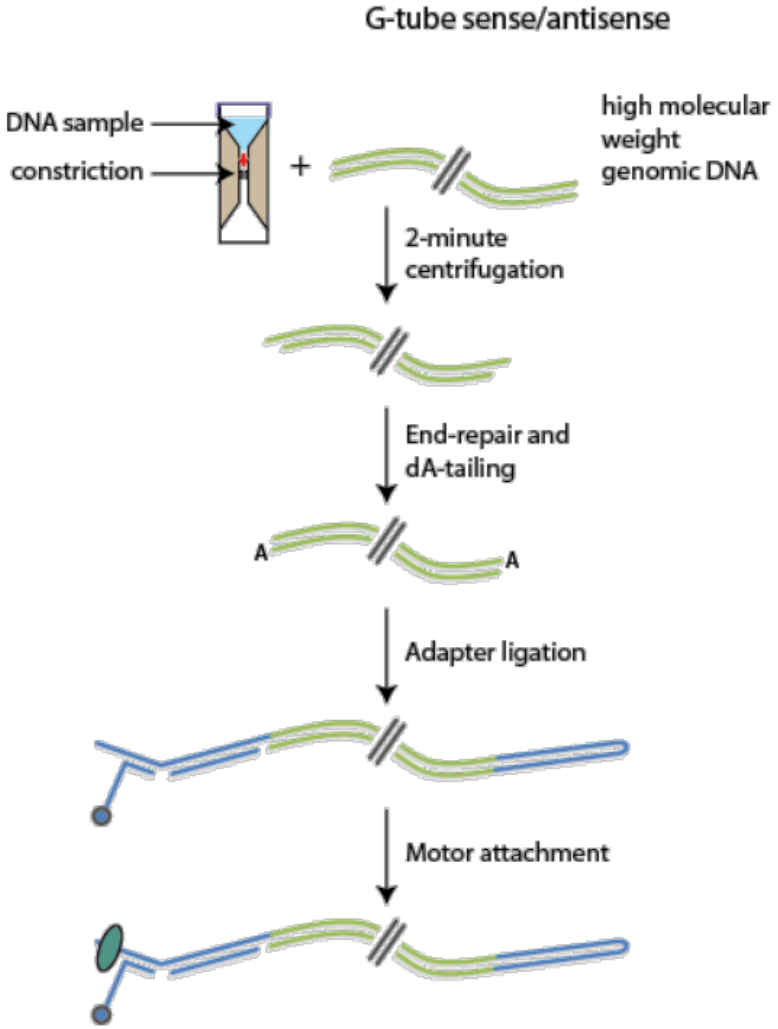
http://www2.technologyreview.com/article/427677/nanopore-sequencing/

MinIon features

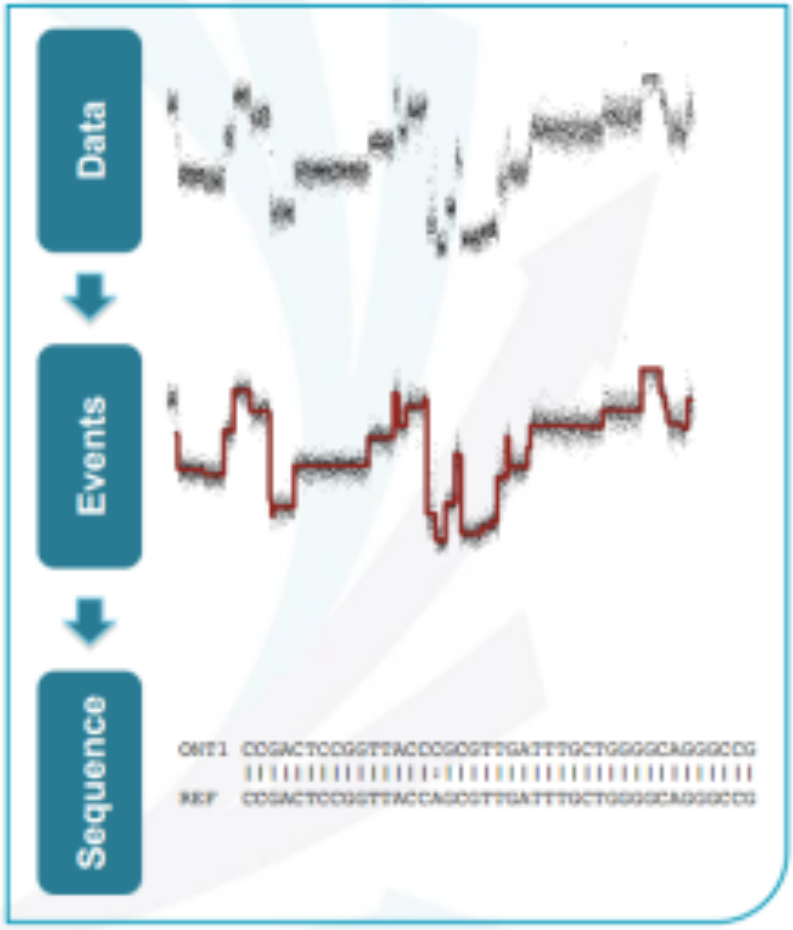
- 512 pores
- Library preparation is required
- Read lengths up to 40-50kb
  - Limited by input DNA
- 30Mb-500Mb output
  - Very variable

Library preparation

- Input requirements
- Upwards of 5ug of DNA
- Issues – keeping long DNA fragments



Challenge of basecalling



- Hidden Markov model
- Only four options per transition
- Pore type = distinct kmer length

- Form probabilistic path through measured states currents and transitions
  - e.g. Viterbi algorithm

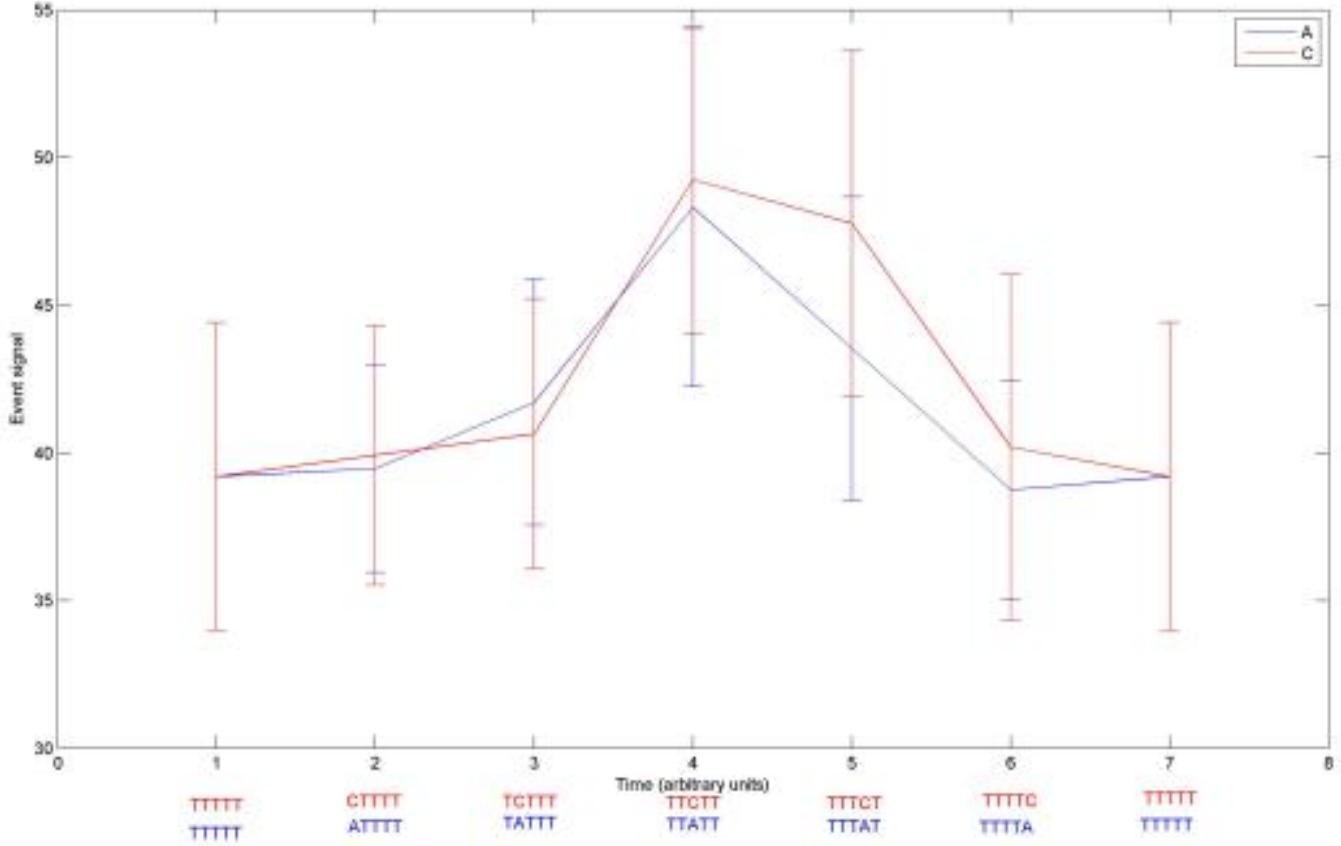
Preliminary Oxford Nanopore R6 results

- Oxford Nanopore have two chemistries – R6 and R7
- R7 is more advanced but testing is only just beginning

- R8 should arrive in the coming months

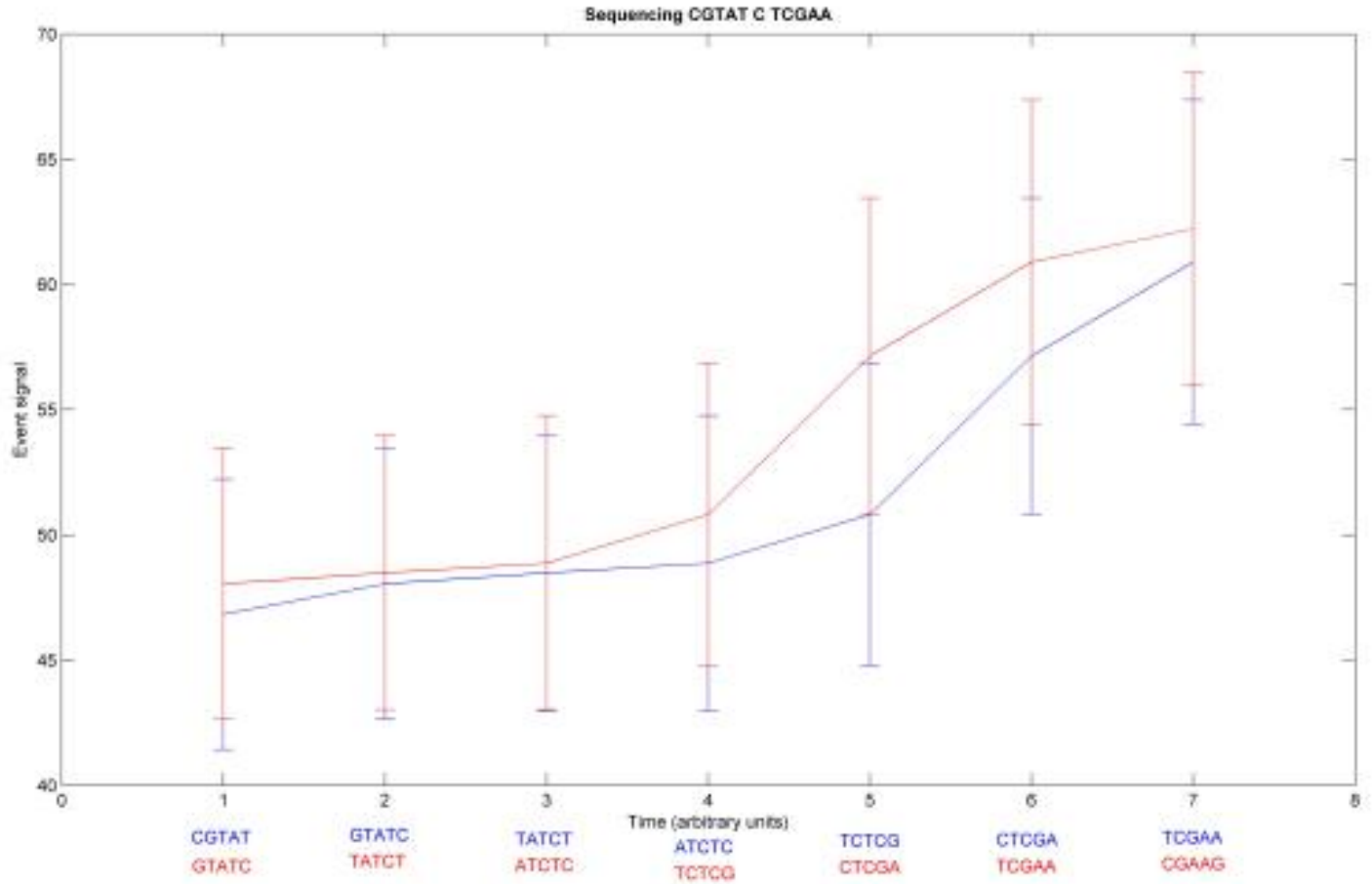
Challenge of 5-mer basecalling

- TTTTTATTTTT vs TTTTTCTTTTT



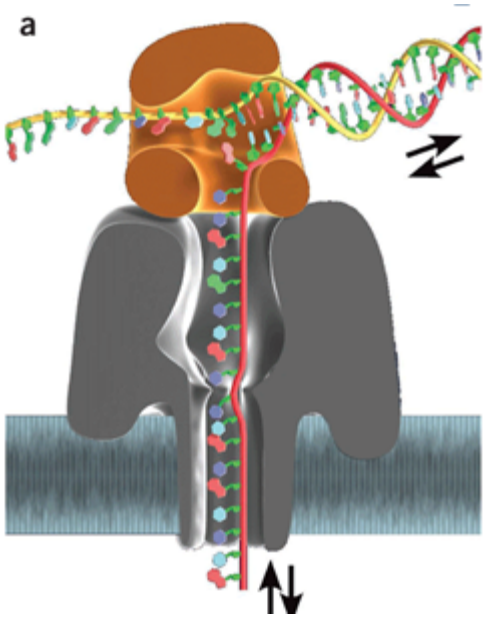
Challenge of 5-mer basecalling

- CGTATTCGAA vs CGTATCTCGAA

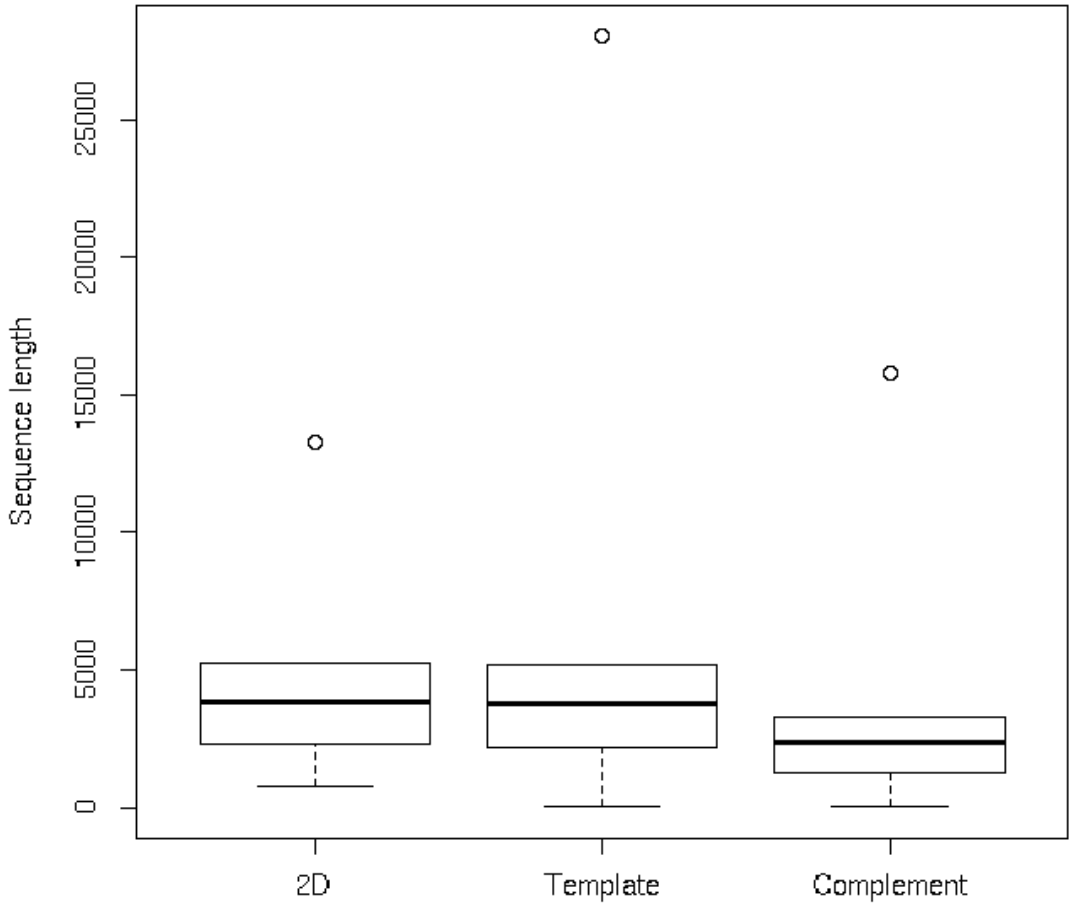


Basecalling 1D vs 2D reads

- Both the template and complementary strand can be sequenced
- This doesn't always work
- If it does, the base-calling can be improved
- Different kmers at the same locus can improve basecalling
- Up to 40% of reads using latest chemistry are 2D
- Attempts are being made to modify the library preparation to increase the proportion of 2D reads

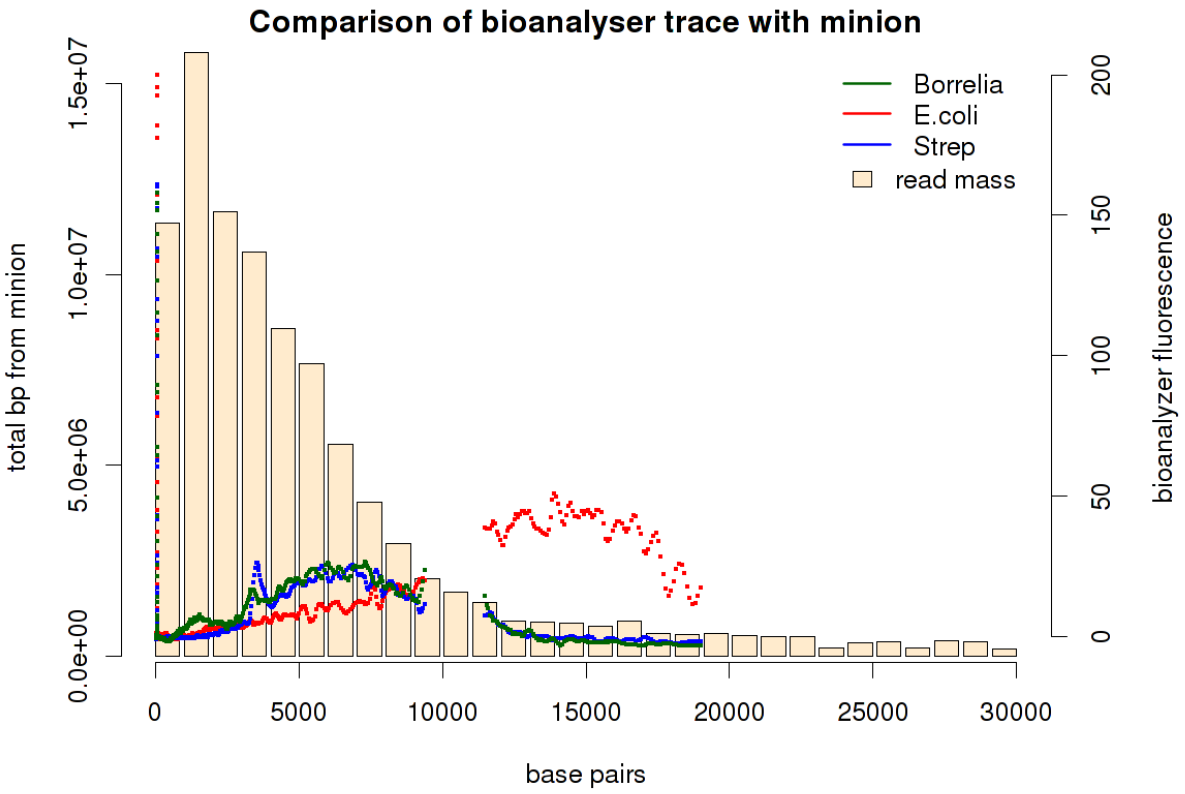


Lengths of 2D vs 1D reads

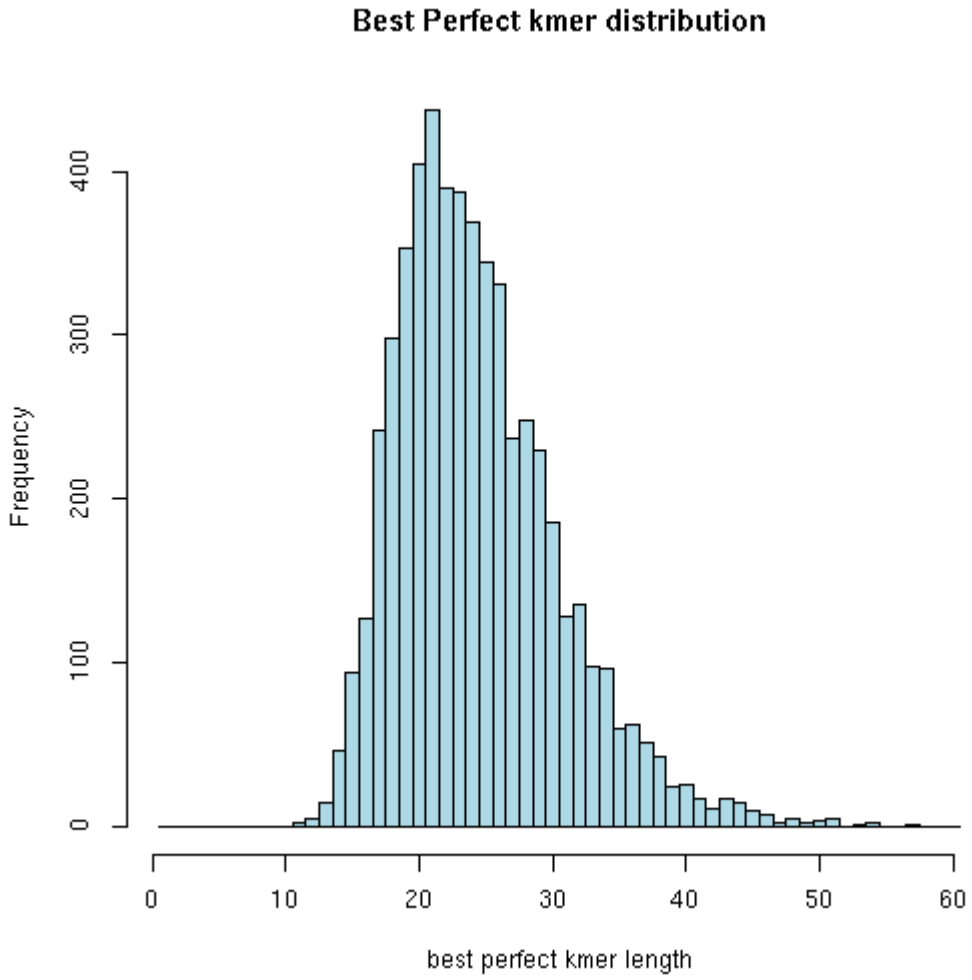


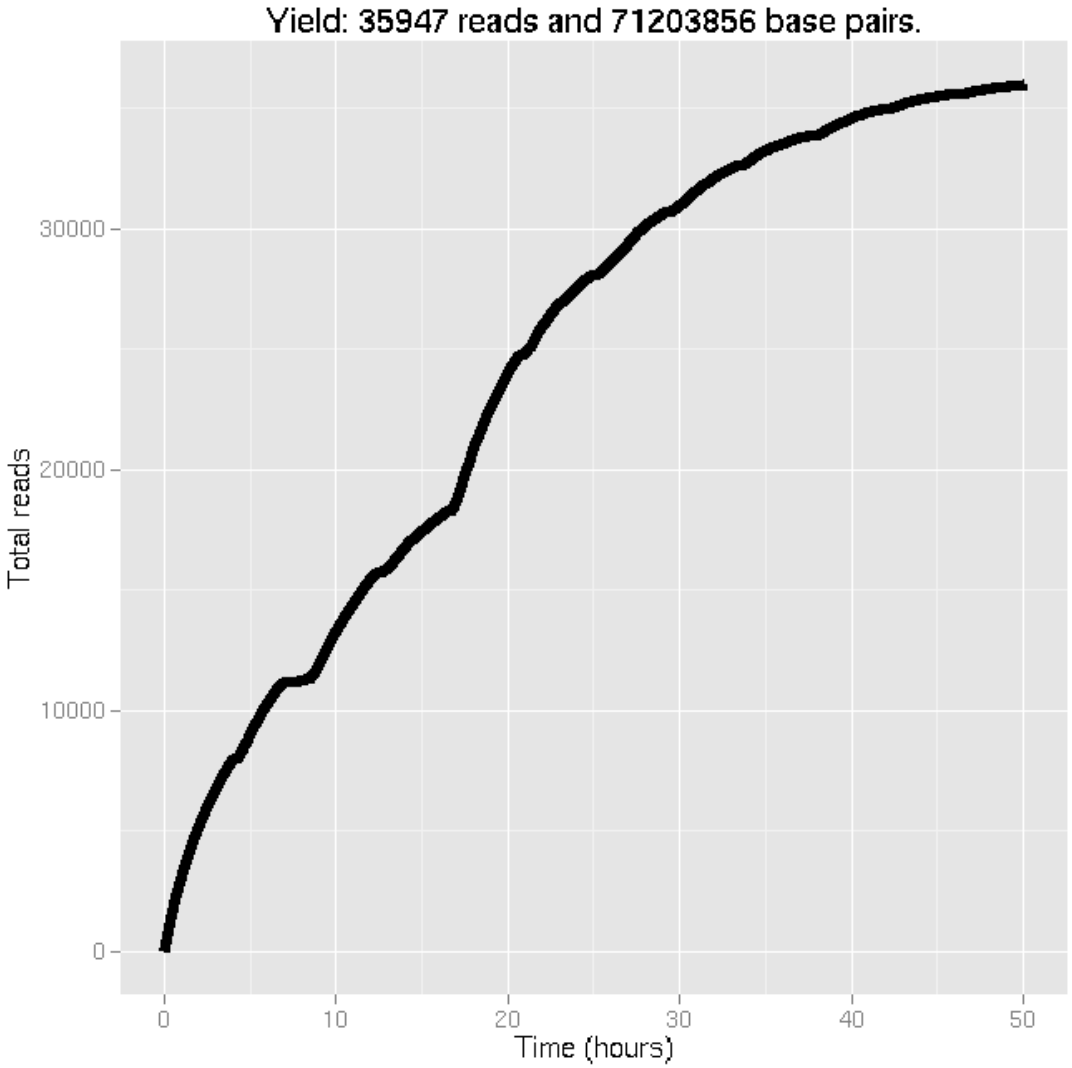
Read lengths

- Highly dependent on input DNA length
- Difficult to preserve DNA lengths

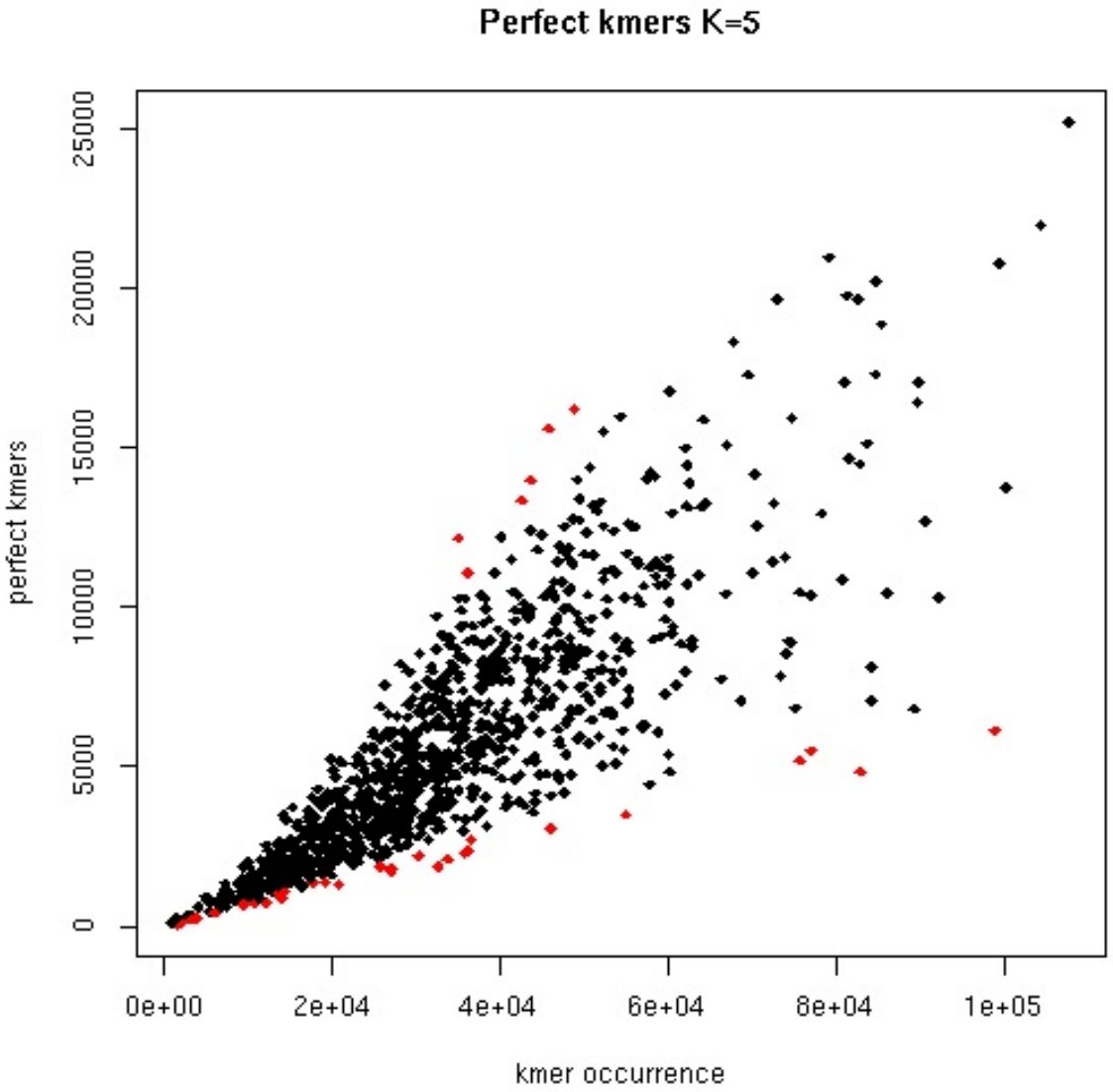


Longest perfect stretches




Yield: 35947 reads and 71203856 base pairs.
Total reads
0
10000
20000
30000
0
10
20
30
40
50
Time (hours)

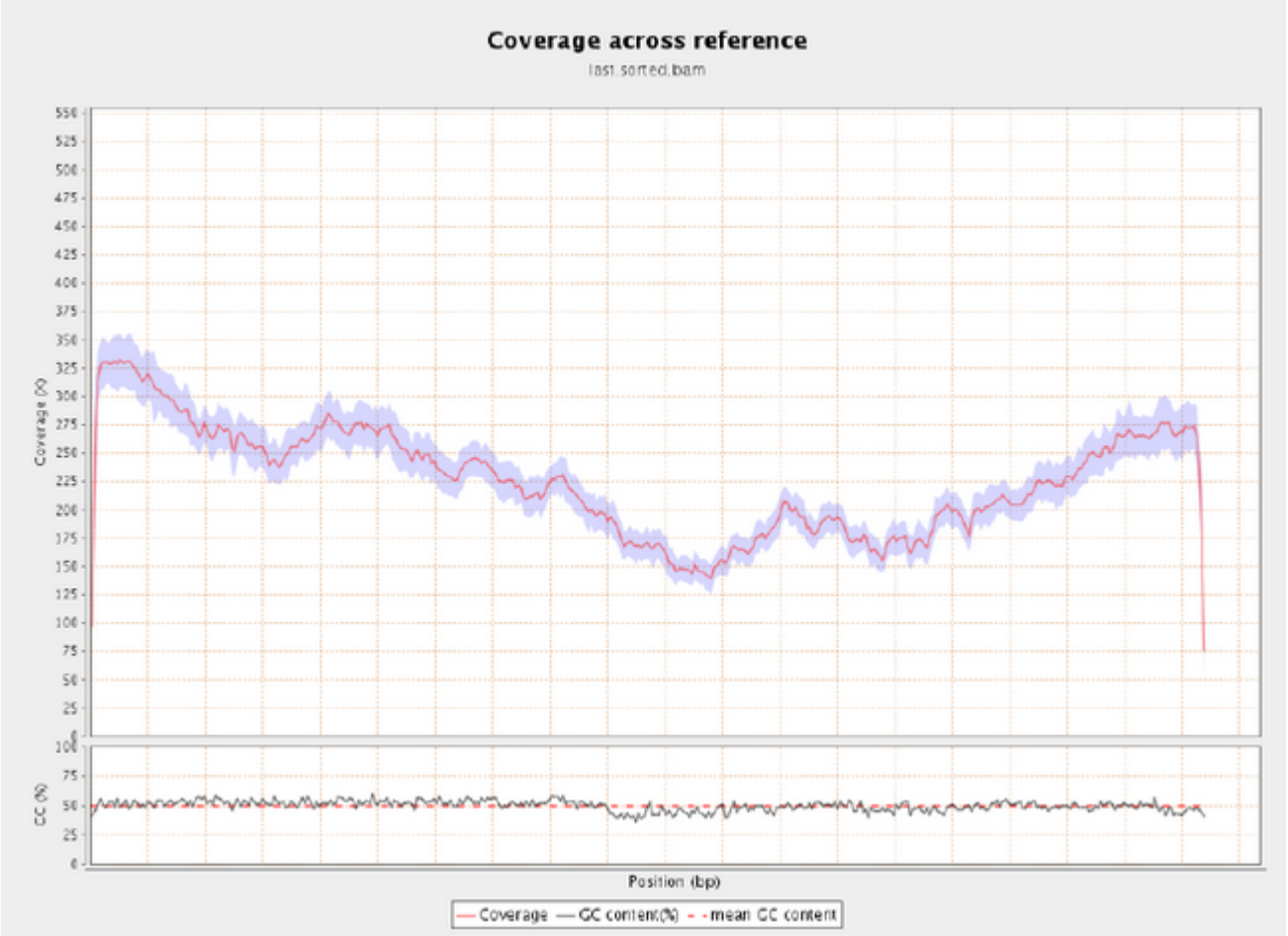
Hard to read motifs



Hard Kmers
AAAAA
ACCTA
ACCTC
AGCGC
AGCTA
AGGTC

Easy Kmers
AAGAA
ACGAA
CGTTC
CTTTC
GAACG

Coverage of an E.coli genome




12 kb
511 bp

Borrelia burgdorferi
Assembly stats

Assembly	N50	Longest contig	Percentage of chromosome recovered in one contig	Misassemblies (Quast)	Mismatch errors/Indel errors
MiSeq only	288kb	434kb	~48%	3	161 / 46
MiSeq & MinION	910kb	910kb	~100%	6	662 / 107

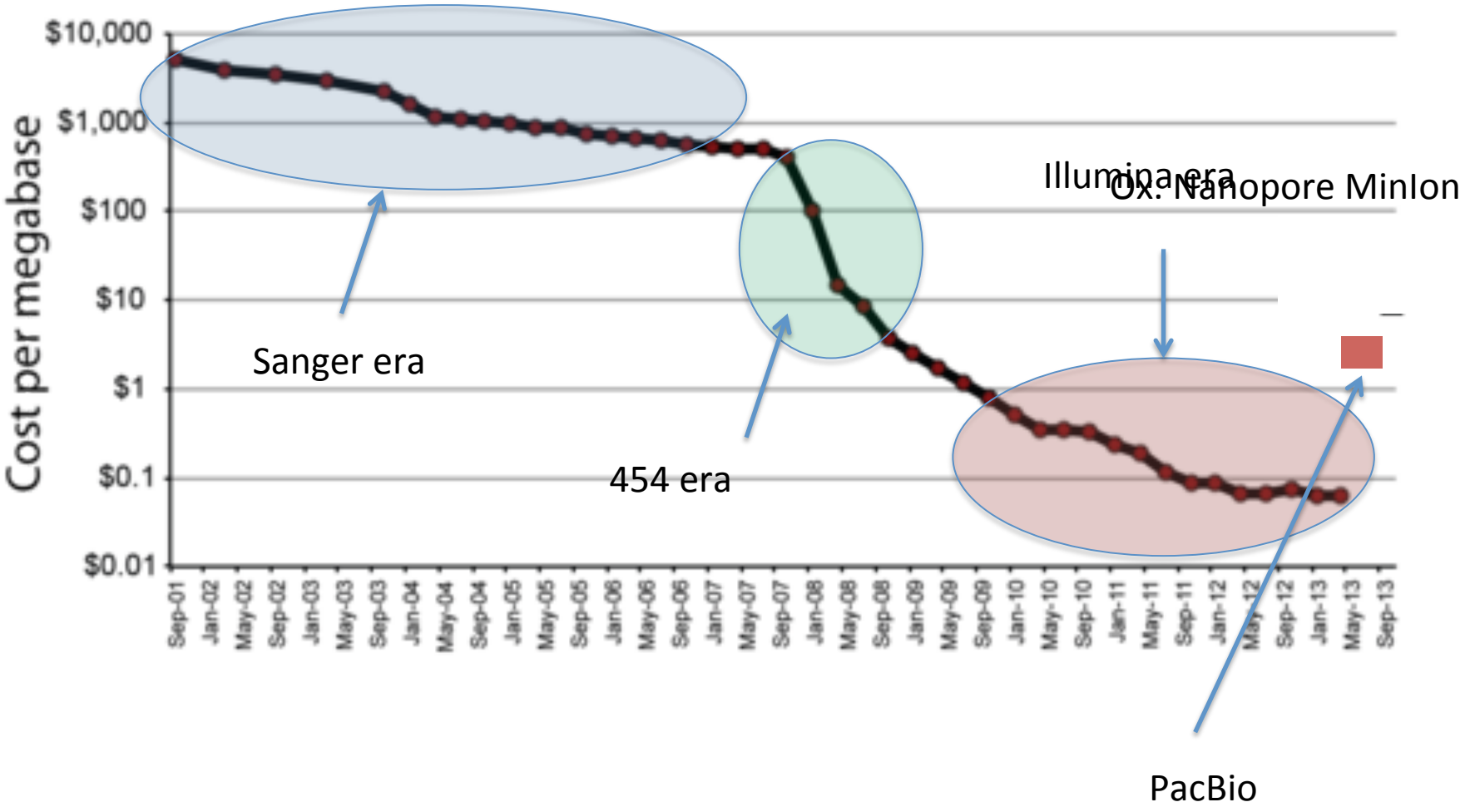
Improvements

- FASTQ – per base quality values don't make much sense
- Improvements to pore types
  - Utilise multiple pore types on a single flowcell
  - This would not make it a true single molecule sequencer since we would rely upon consensus (probably does not matter)
- Library preparation
- Flowcell reliability
- Access to basecalling API

ONT costs

- Currently ~1500 euro per flowcell
  - ONT anticipate approximate costs of 1000 euro per flowcell
- Library preparation costs are considerable
  - Hundreds of euros per sample

Cost per megabase



Software packages

- Processing ONT data
  - Poretools https://github.com/arq5x/poretools
  - poRe http://sourceforge.net/projects/rpore/
- Alignment

  Sensitive but slow aligners
  - BLAST
  - BLAT
  - LAST
  - BWA with the right parameters
- Assembly
  - At the moment data quality is too poor to attempt this with ONT data only
  - Error correction with other data types should be possible with 15-20% error rate
  - Short term fix, but may offer an alternative to PacBio or Illumina synthetic long reads

Summary

- Very encouraging but still early days
- Far from portable although the potential exists
- Error rate is ~30-40% for R6 2D reads
  - Indications that it is ~20% for R7.3 2D reads
  - Some success with completing bacterial genomes using this data
- Probably 12 months before full commercial launch

Opportunities for software development

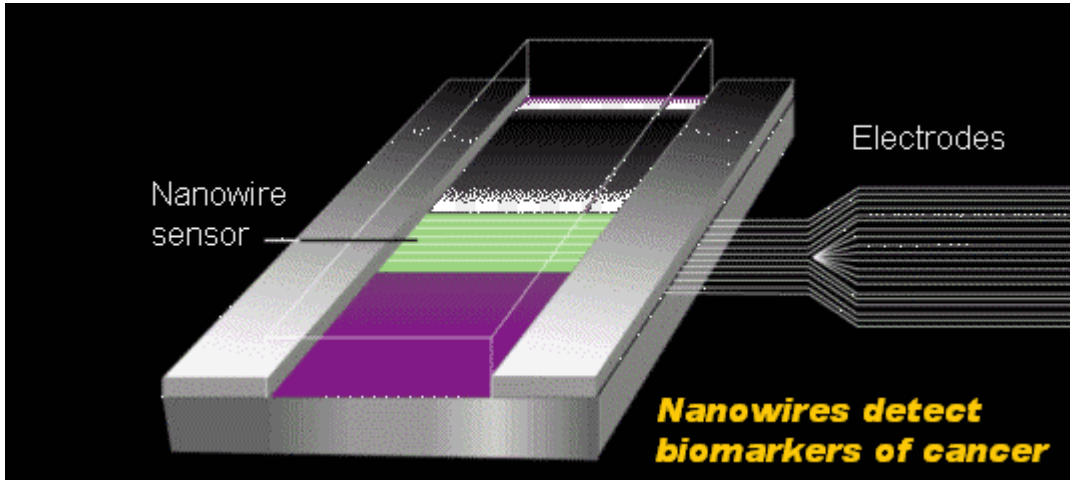
- Converting reference genomes to ‘wiggle-space’
- Developing ‘wiggle-space’ aligners
- Online ‘streaming’ bioinformatics
  - Analytics - one read at a time

General issues with nanopores

- Single base-pair resolution is not available
  - Typically 4-5 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

In conclusion

- Francis Crick's predictions for molecular biology in the year 2000:
  - Replication of DNA
  - Structure of chromosomes
  - Meaning of the nucleic acid sequence
  - Significance of repetitive sequences

Francis Crick
1971

  - "In short, the whole field is likely to be even more fascinating in the year 2000 than it is now"

Thanks to:

Karen Moore

Audrey Farbos

Paul O'Neill

Wellcome Trust

Contact me:

k.h.paszkiewicz@exeter.ac.uk

http://biosciences.exeter.ac.uk/facilities/sequencing

Supported by wellcome trust

UNIVERSITY OF EXETER