“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)

UNIVERSITY OF EXETER

History of DNA and modern approaches to sequencing

Konrad Paszkiewicz

January 2017

Contents

- A short history of DNA
- Review of first generation sequencing techniques
- Short-read second generation sequencing technology
  - Illumina
- Third generation single molecule sequencing
  - PacBio
  - Oxford Nanopore

“DNA is a stupid molecule”

Max Delbruck

“Never under-estimate the power of … stupidity”

Robert Heinlein

“It was believed that DNA was a stupid substance, a tetranucleotide which couldn't do anything specific”

Max Delbruck

The first person to isolate DNA

- Friedrich Miescher
  - Born with poor hearing
  - Father was a doctor and refused to allow Friedrich 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

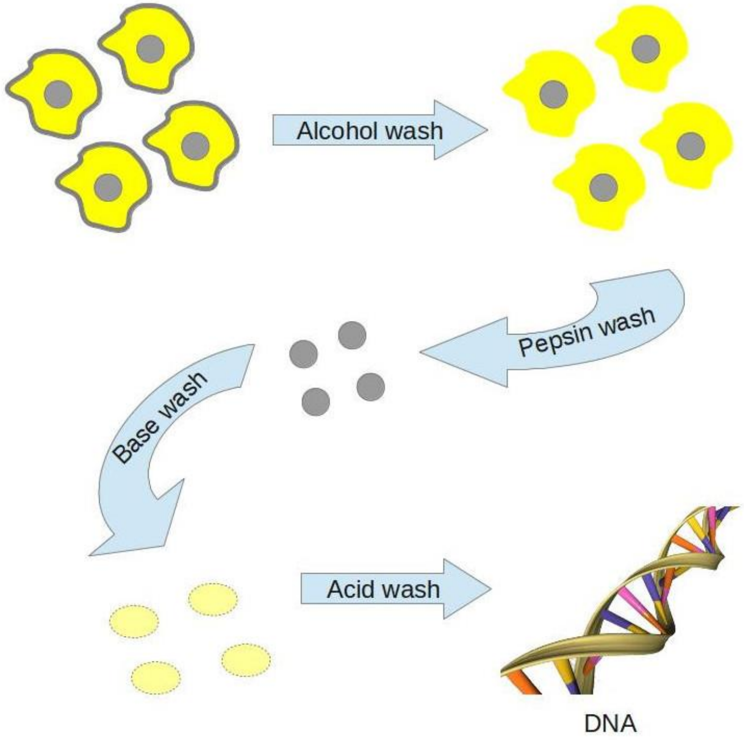
- Went to work in Felix Hoppe-Seyler's laboratory in Tubingen, Germany
  - The founding father of biochemistry and focussed on the study of protein
  - The lab was one of the first to crystallise haemoglobin and describe the interaction between haemoglobin and oxygen using spectroscopy
  - Also played host Paul Ehrlich who later went on to develop gram staining and immunological advances
- Freidrich's work on DNA was regarded as a side-project
- Freidrich extracted 'nuclein' on cold winter nights
  - Initially from human leukocytes extracted from bandage pus from the local hospital filled with soldiers from the Austro-Prussian war
  - 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

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.“

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

Friedrich Miescher

1874 - First hints to composition

- By 1874 Meischer had determined that nuclein was
  - A 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”
  - Later discarded the idea because he thought it unlikely that nuclein could encode sufficient information
- He returned to working on his former supervisor’s haemoglobin work and made the discovery that carbon dioxide rather than oxygen regulated breathing

Friedrich Miescher

1881 - Discovering the composition of nuclein

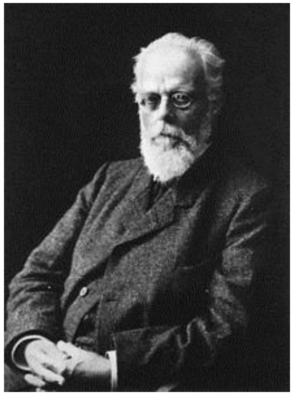
- Kossel worked in the same lab as Freidrich Miescher
- Wanted to relate chemical composition to biological function
- Discovered fundamental building blocks of nuclein
  - Adenine, Cytosine, Guanine, Thymine, and Uracil
  - Identified histone proteins and that nuclein was bound to histone in the nucleus
  - By observing cell division inferred that nuclein was not used for energy storage but was linked to cell growth

Albrecht Kossel

1890s – Hints at the molecular basis of heredity

- How are characteristics transmitted between generations?

- 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

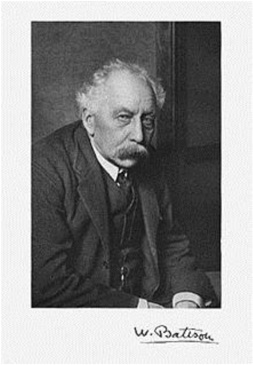
19001-1905: Re-discovery of Mendel’s laws and the birth of genetics

- Concept of genes as independent particles of information
- No ‘blending of traits’
  - Almost simultaneously rediscovered by de Vries, Correns and Von Tschermak
- Bateson coins the word ‘genetics’ from the Greek ‘genno’ – to give birth
- Bateson became known as ‘Mendel’s bulldog’ and popularised Mendel’s work

Hugo de Vries

Carl Correns

Erich Tschermak



William Bateson

1902 – Are chromosomes involved in heritability?

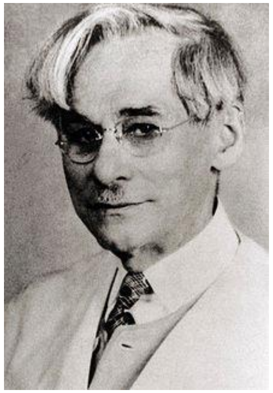
- Walter Sutton using grasshopper gametes
- Theodore Boveri using sea urchins

Walter Sutton and Theodore Boveri

"…the association of paternal and maternal chromosomes in pairs and their subsequent separation during [cell] division …may constitute the physical basis of the Mendelian law of heredity."

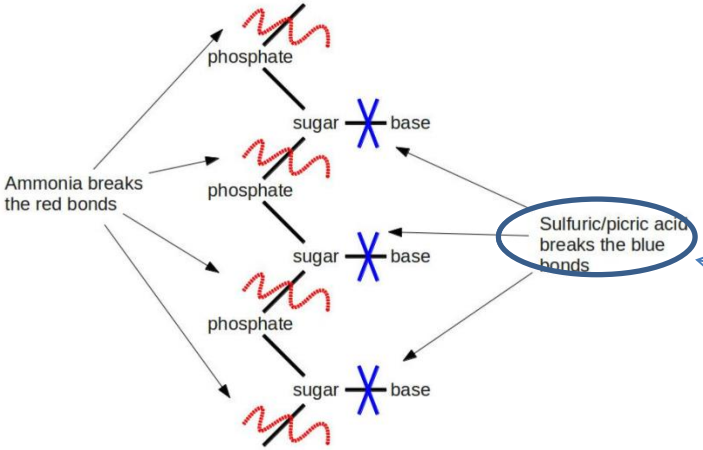
- Theodore Boveri *Sutton, W. S. 1903. The chromosomes in heredity. Biological Bulletin, 4:231-251.*

1910s - More on the composition of DNA



Phoebus Levene

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


phosphate
sugar
base
Ammonia breaks the red bonds
phosphate
sugar
base
Sulfuric/picric acid breaks the blue bonds
phosphate
sugar
base

Close relative of TNT explosive

- Enabled the discovery of DNA and RNA bases
- Unfortunately, this method can destroy bases and biases 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

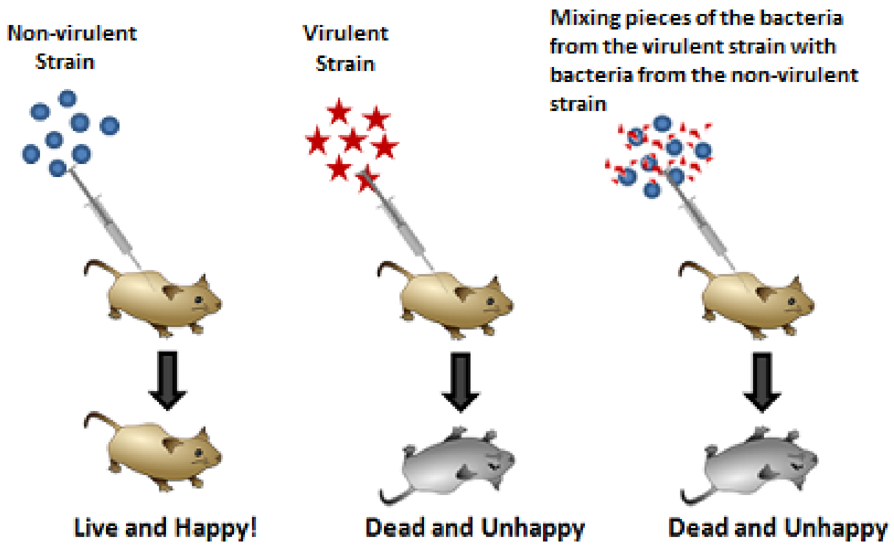
- Chromosome as a unit of heritability confirmed by Thomas Morgan by 1915
- Alfred Sturtevant creates the first genetic linkage map
- Genetic recombination shown to be caused by physical recombination of chromosomes by McClintock & Creighton

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


Non-virulent Strain
Virulent Strain
Mixing pieces of the bacteria from the virulent strain with bacteria from the non-virulent strain
Live and Happy!
Dead and Unhappy
Dead and Unhappy

- 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

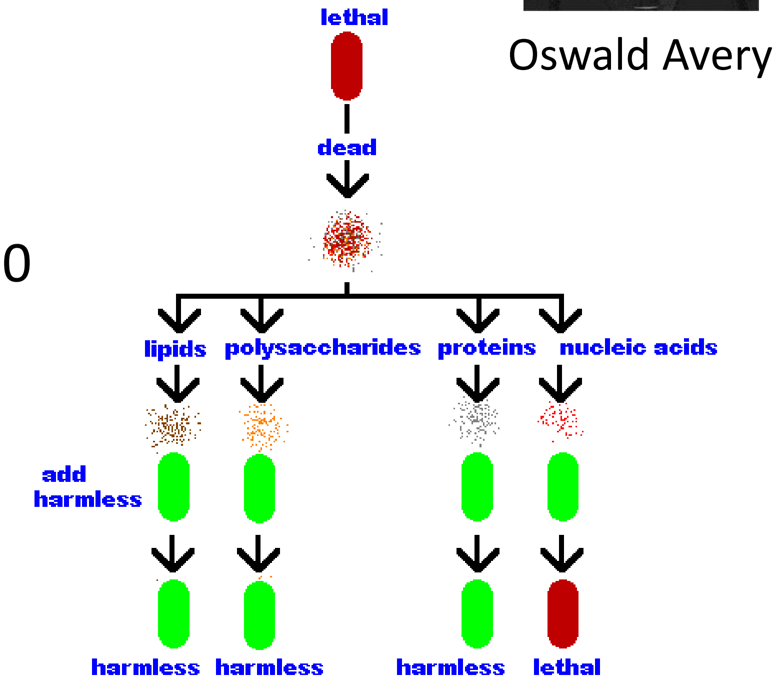
- Served as inspiration for Watson & Crick

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

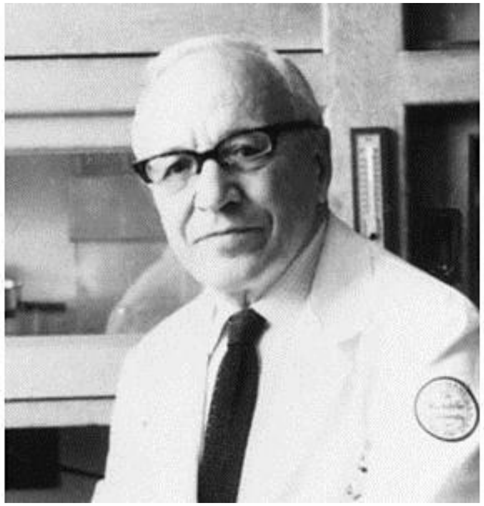
Oswald Avery



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

1945 – 1952 Critique

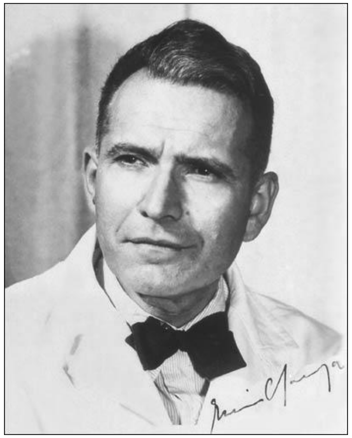
- Alfred Mirksy was a pioneer of molecular biology
- Isolated chromatin from a wide variety of cells
- He was concerned that Avery's results could be the result of protein contamination
- Convinced the Nobel panel not to award a prize to Avery
- Later, Mirsky would actually demonstrate the 'constancy' of DNA throughout somatic cells



Alfred Mirsky

1950 – Base composition between organisms

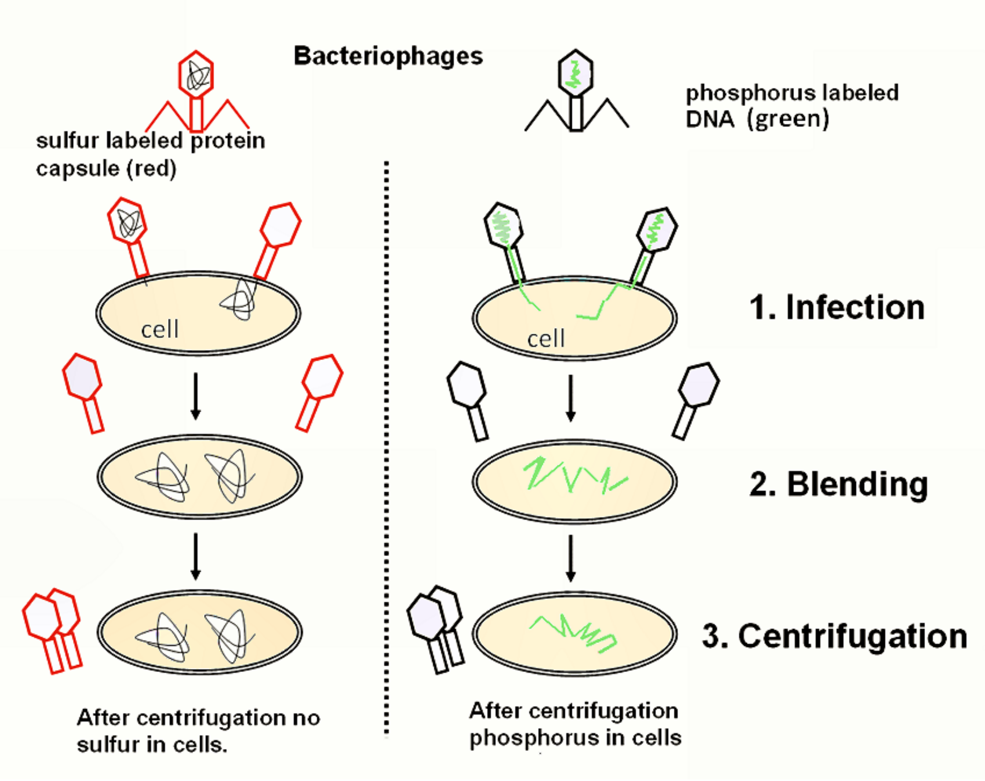
- Erwin Chargaff hit back at Mirsky and developed the base complementarity hypothesis with Masson Gulland
- Determined that the molar ratio of A:T and G:C were always very close to 1
- Relative proportions of bases varied between species but was the same within species
- Refuted Levene's 30 year-old tetranucleotide hypothesis



Erwin Chargaff

1952- Confirmation of Avery's experiment

- Grow bacteriophage using radioactive substrates
  - Protein with radioactive sulphur
  - DNA with radioactive phosphorous
- Bacteriophages infected bacteria by injecting DNA, not protein
- Indicated that protein could not be the heritable genetic material
- Yet there was still the possibility of small amounts of protein contamination which led some to have doubts about the role of DNA


Bacteriophages
sulfur labeled protein capsule (red)
phosphorus labeled DNA (green)
cell
cell
1. Infection
2. Blending
3. Centrifugation
After centrifugation no sulfur in cells.
After centrifugation phosphorus in cells

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

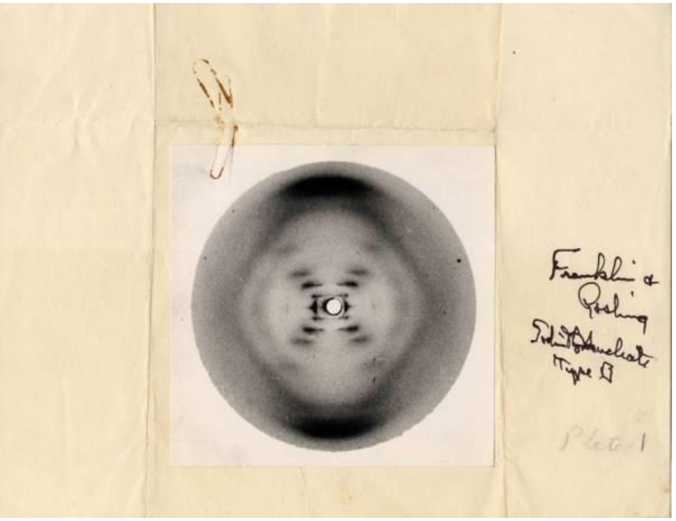
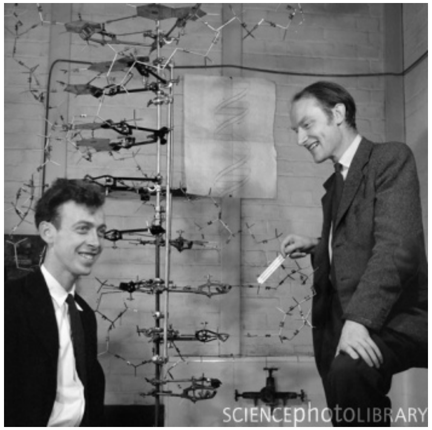


Photo Number 51

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 or wet-form 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 and role of DNA did not occur until around 1960

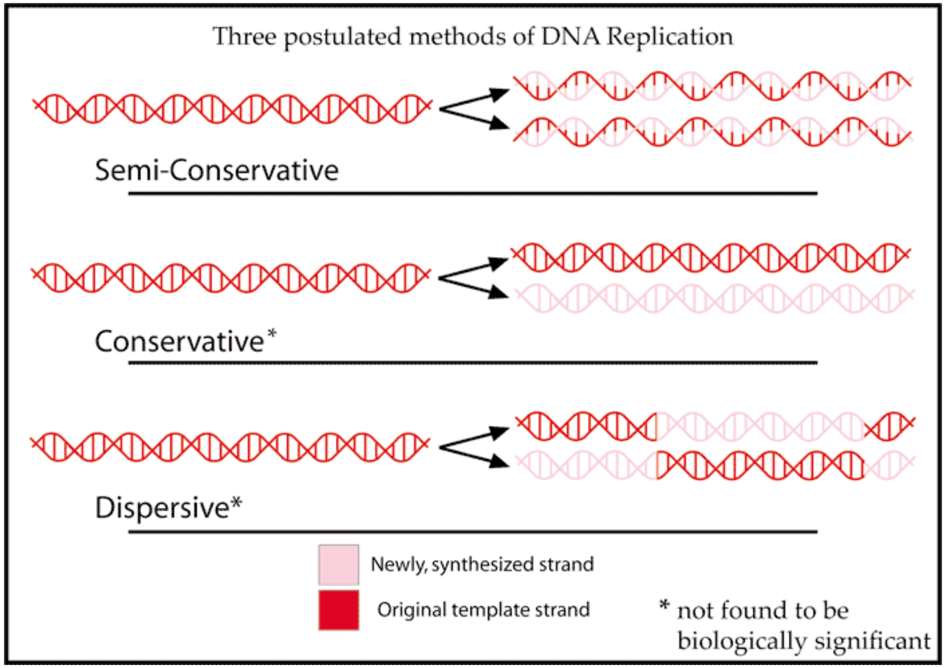


Francis Crick & James Watson

1958 – Evidence for the mechanism of DNA replication

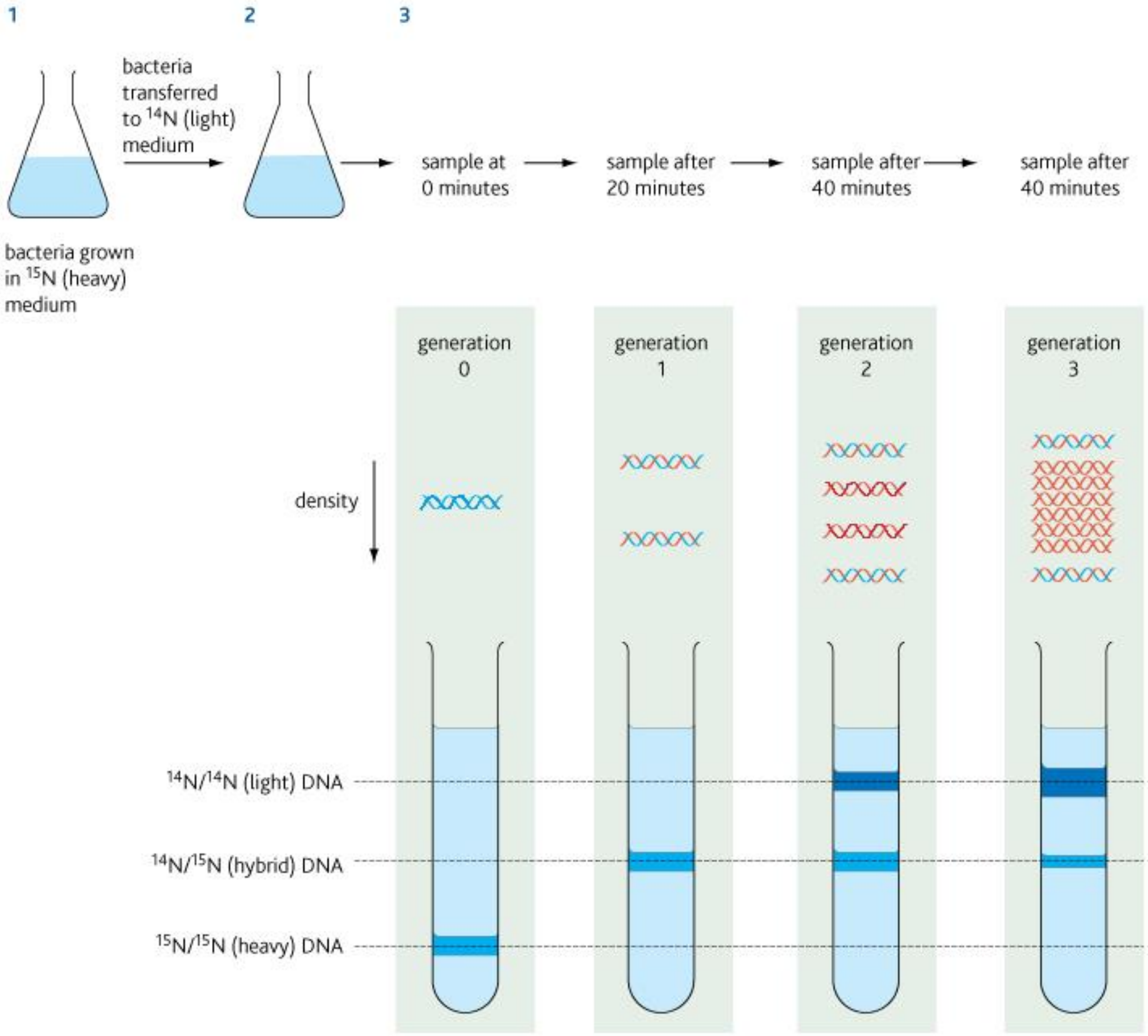
- Meselson & Stahl

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


Three postulated methods of DNA Replication
Semi-Conservative
Conservative*
Dispersive*
Newly, synthesized strand
Original template strand
* not found to be biologically significant

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

1958 – Evidence for the semi-conservative 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

- Set the stage for…

1961 - Deciphering the genetic code

- How did DNA code for proteins?

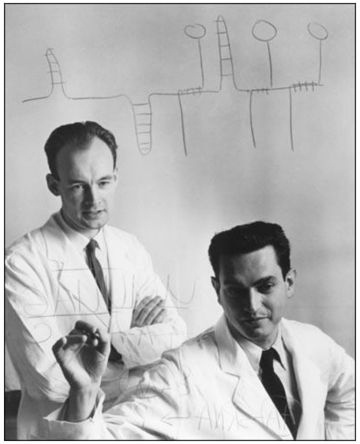
- Nirenberg and Matthaei

- 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

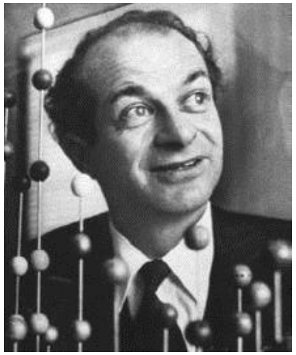
- Completed in mid 1960s by Har Gobind Khohrana



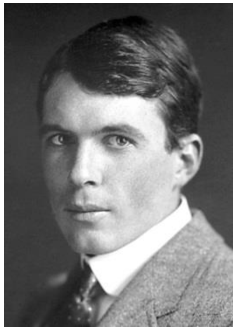
Other key figures

- 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…

Oswald Avery

- Avery died in 1955
- It is unknown whether he learned of Watson & Crick’s structure of DNA
- However his 1944 paper is cited around 40 times a year and has cited over 2000 times

Oswald Avery

Further reading

- Eighth day of creation – Horace Freeland Judson

- Life's Greatest Secret – Matthew Cobb

- Oswald Avery, DNA, and the transformation of biology. Cobb, M. Current Biology. Volume 24, Issue 2, 20 January 2014, Pages R55–R60

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

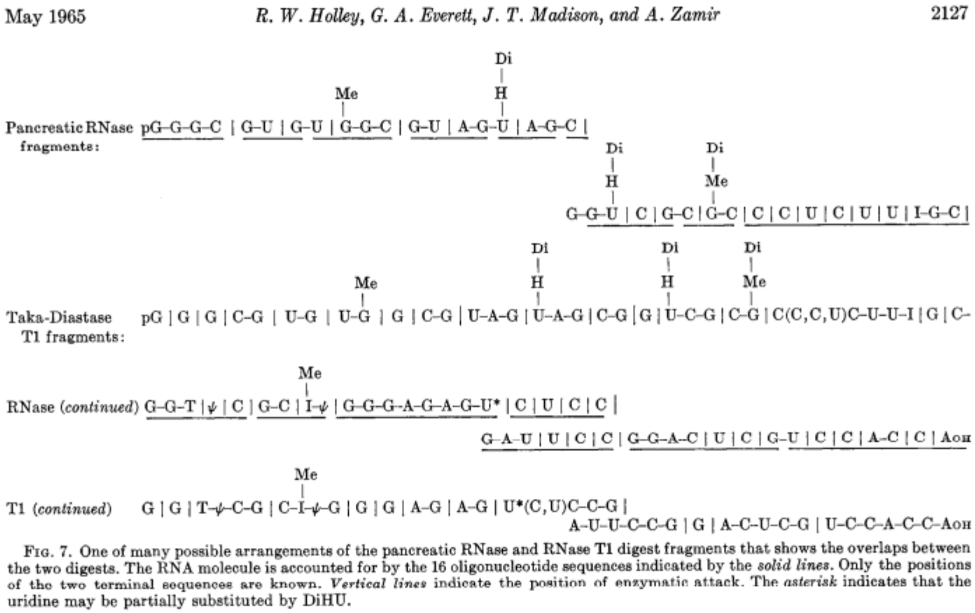
- Sequenced bovine insulin
- Developed a method to label N-terminal amino acids
  - Enabled him to count four polypeptide chains
- Used hydrolysis and chromatography to identify fragments

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



May 1965 *R. W. Holley, G. A. Everett, J. T. Madison, and A. Zamir* 2127

```
                                                           Di
                                                           |
                                    Me                     H
                                    |                      |
Pancreatic RNase  pG-G-G-C | G-U | G-U | G-G-C | G-U | A-G-U | A-G-C |
  fragments:
                                                          Di           Di
                                                          |            |
                                                          H            Me
                                                          |            |
                                                      G-G-U | C | G-C | G-C | C | C | U | C | U | U | I-G-C |

                                                           Di             Di          Di
                                                           |              |           |
                                    Me                     H              H           Me
                                    |                      |              |           |
Taka-Diastase     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-
  T1 fragments:

                                    Me
                                    |
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

                                    Me
                                    |
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.

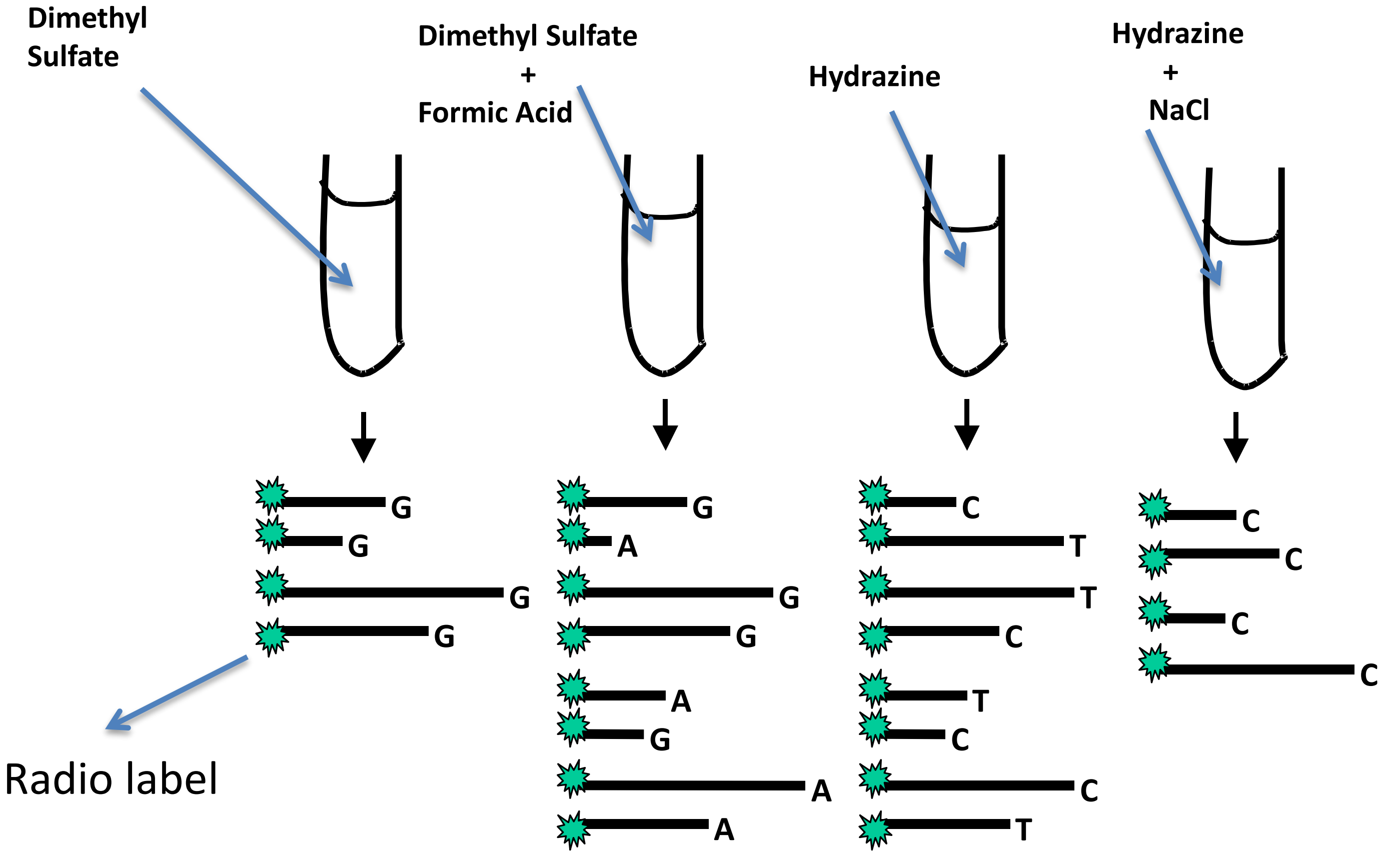
http://www.sciencemag.org/content/147/3664/1462

1975 - The dawn DNA sequencing

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

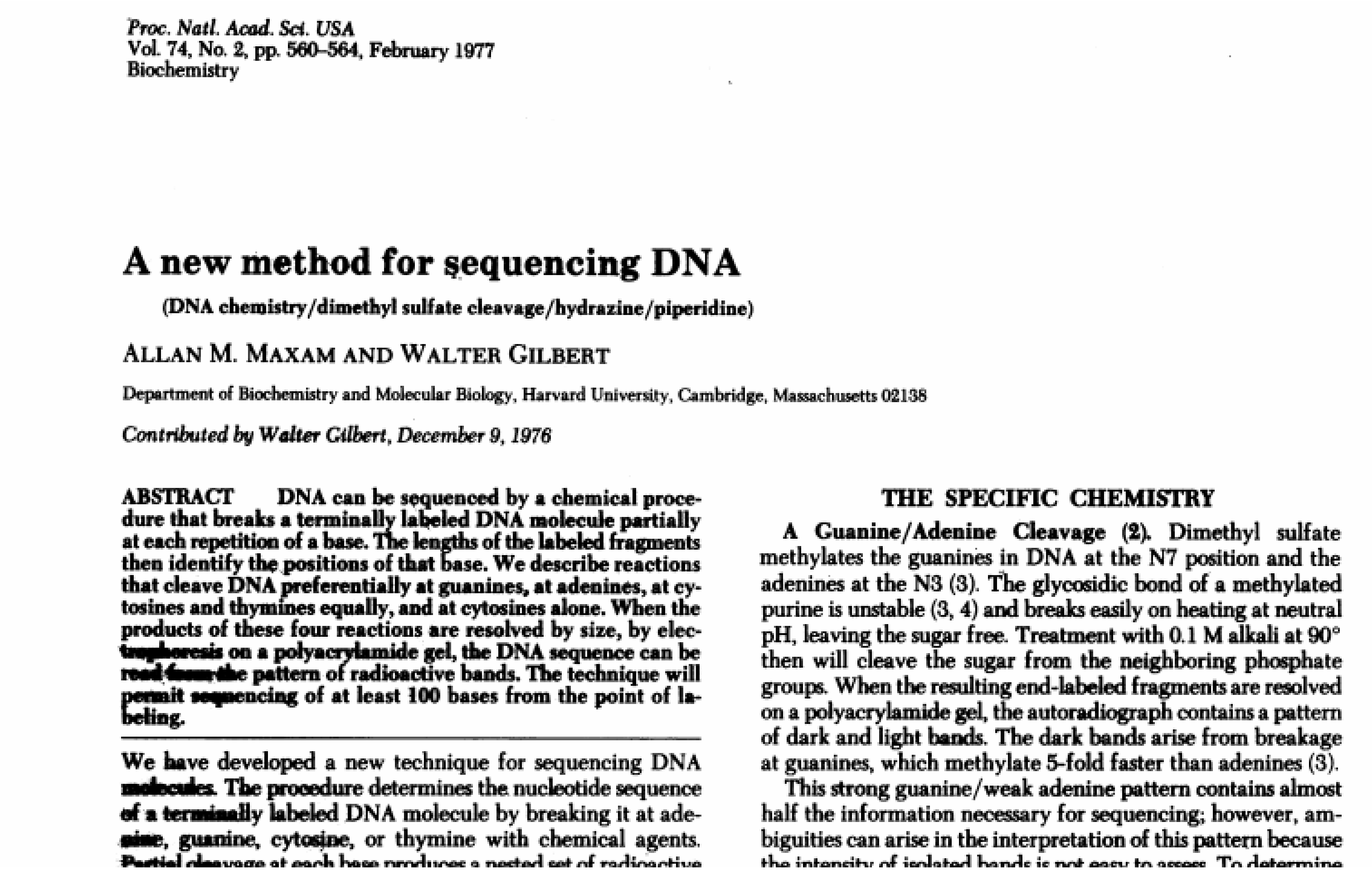
- Fred Sanger’s Plus/Minus method
- Maxam-Gilbert
- Fred Sanger’s chain termination method

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.

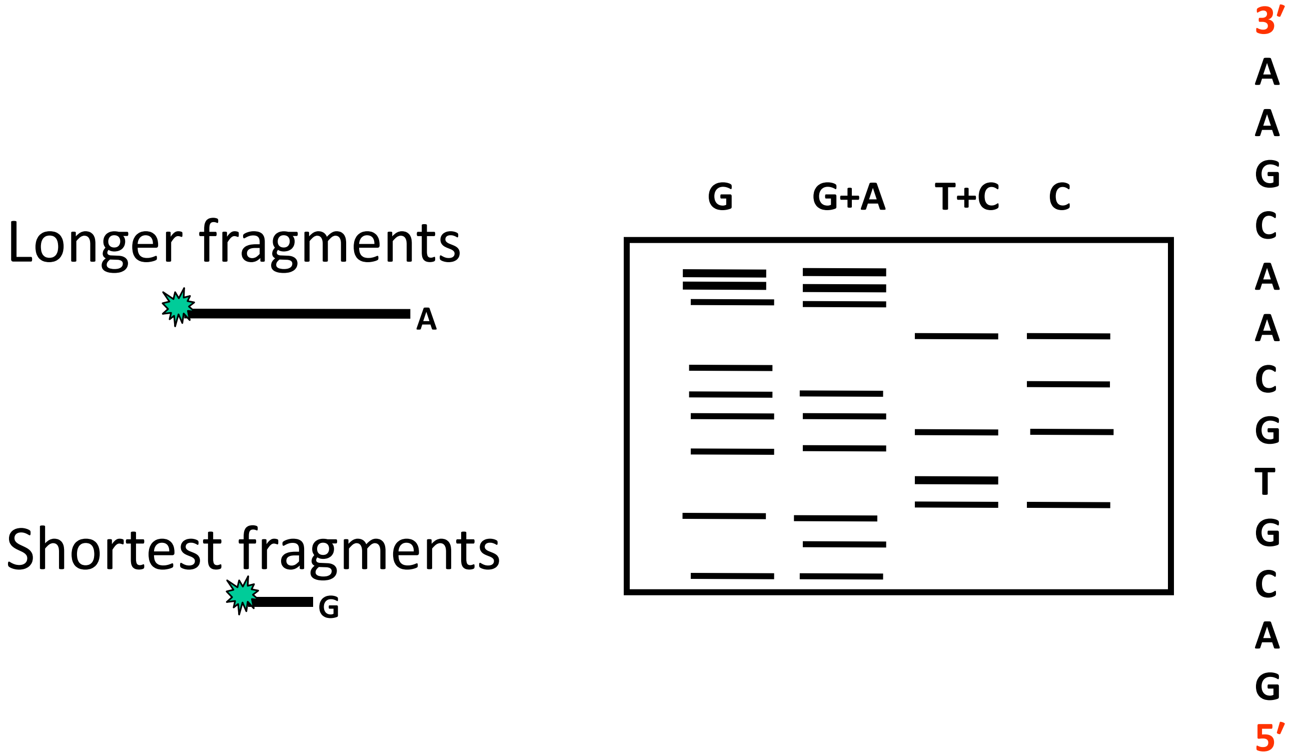
We have developed a new technique for sequencing DNA molecules. The procedure determines the nucleotide sequence of a terminally labeled DNA molecule by breaking it at adenine, guanine, cytosine, or thymine with chemical agents. Partial cleavage at each base produces a nested set of radioactive

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



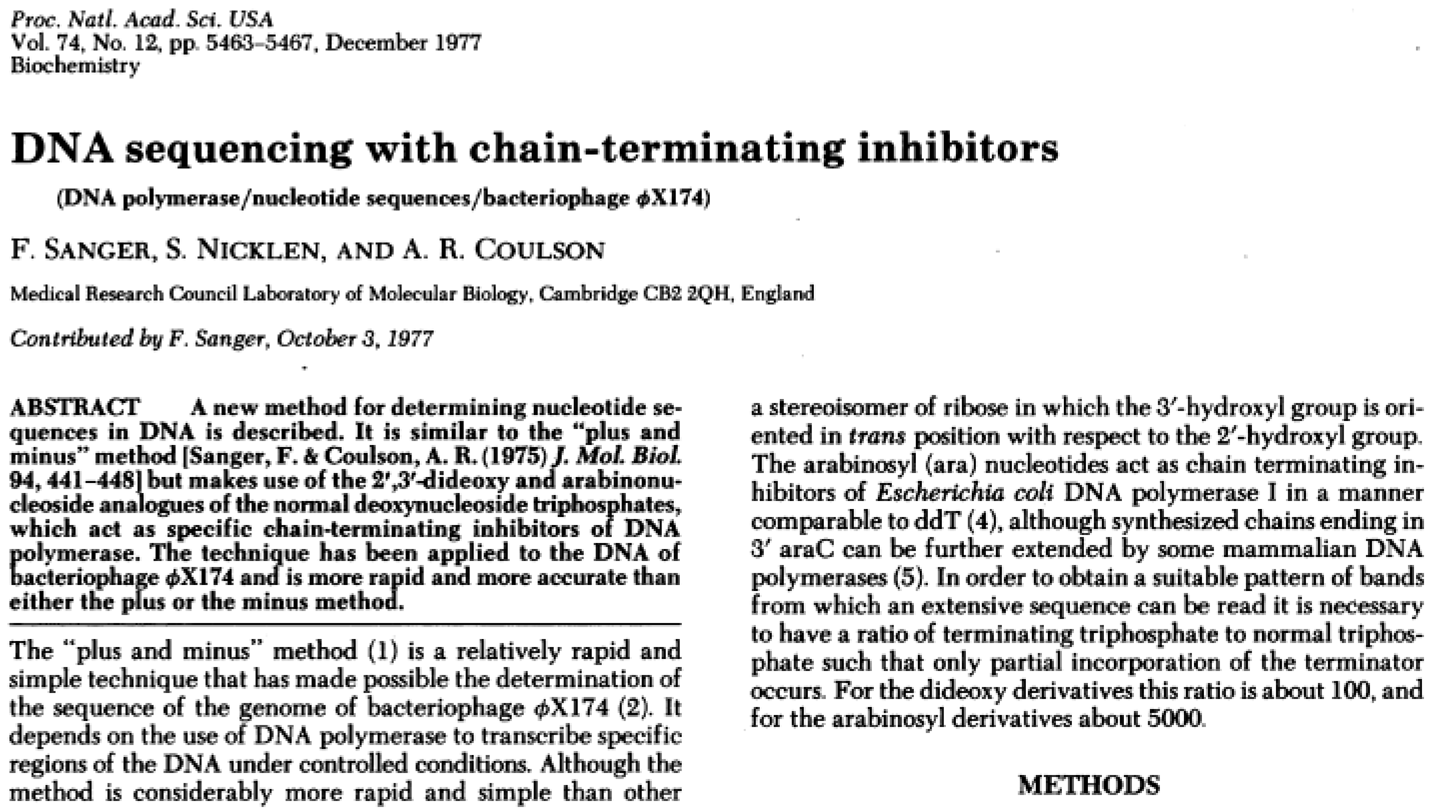
Longer fragments

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.

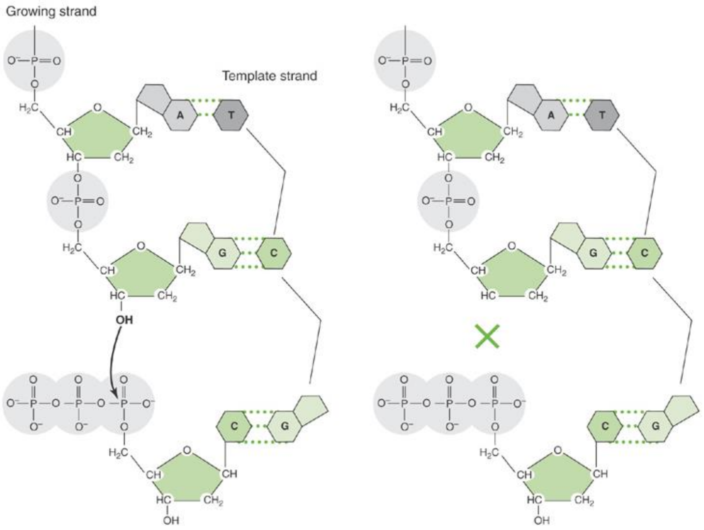
The "plus and minus" method (1) is a relatively rapid and simple technique that has made possible the determination of the sequence of the genome of bacteriophage ϕX174 (2). It depends on the use of DNA polymerase to transcribe specific regions of the DNA under controlled conditions. Although the method 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 DNA Sequencing

- Uses two classes of de-oxy nucleocide tri-phosphate
  - Regular de-oxy NTP nucleotides (i.e dATP, dGTP, dCTP and dTTP)
  - di-deoxy NTP molecules which are radio-labelled and lack a 3’ hydroxyl group
- The lack of 3’ hydroxyl bond prevents extension of growing strand
- 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



Copyright © 2007 F.A. Davis Company www.fadavis.com

Sanger sequencing

AGCTGCCCG

Possible fragment lengths

A

ddATP +
four dNTPs

ddA
dAdGdCdTdGdCdCdCdG

1 or 9bp

C

ddCTP +
four dNTPs

dAdGddC
dAdGdCdTdGddC
dAdGdCdTdGdCddC
dAdGdCdTdGdCdCddC

3, 6, 7 or 8bp

G

ddGTP +
four dNTPs

dAddG
dAdGdCdTddG
dAdGdCdTdGdCdCdCddG

2, 5, or 9bp

T

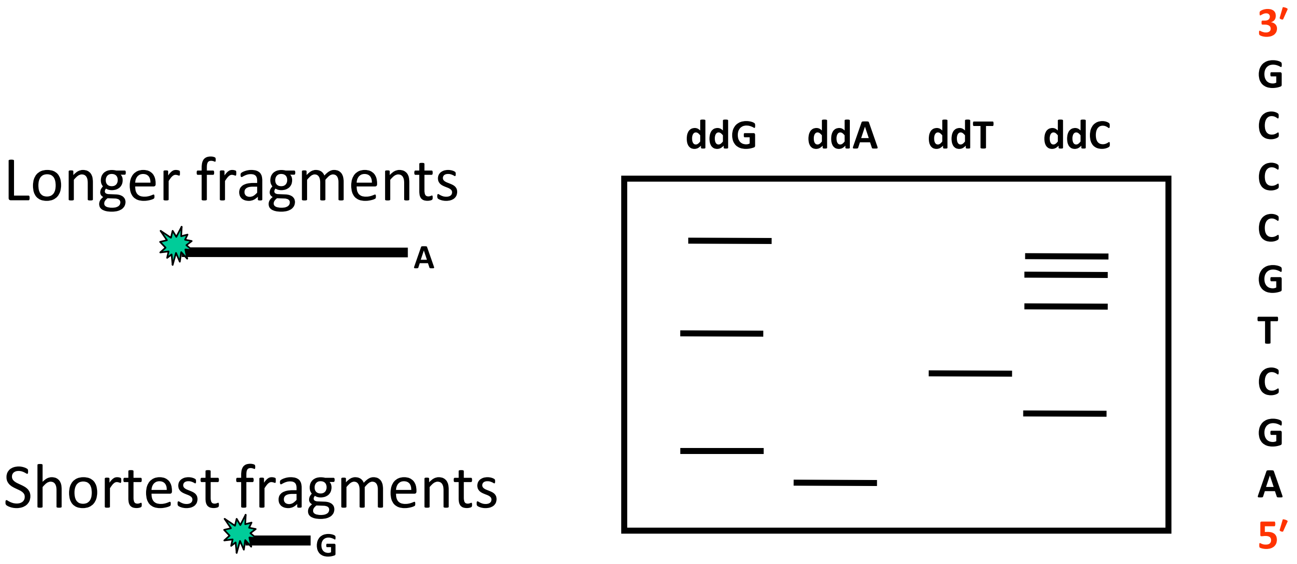
ddTTP +
four dNTPs

dAdGdCddT
dAdGdCdTdGdCdCdCdG

4 or 9bp

Sanger di-deoxy method

5’ AGCTGCCCG 3’



Longer fragments

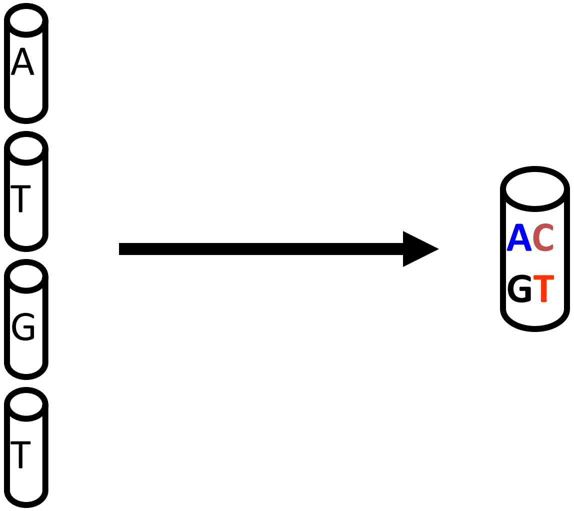
Shortest fragments

1985: Automating Sanger Sequencing

- Disadvantages of manual Sanger sequencing
  - Labour intensive
  - Used radioactive labels
  - Interpretation/analysis was subjective
- Difficult to scale up
- Leroy Hood, Michael Hunkapiller developed an automated method utilising:
  - Fluorescent labels instead of radioactivity
  - Utilise computerised algorithms to analyse data
  - Robotics
- Development of PCR by Kary Mullis (NGS would be impossible without it)

Dye Terminator Sequencing

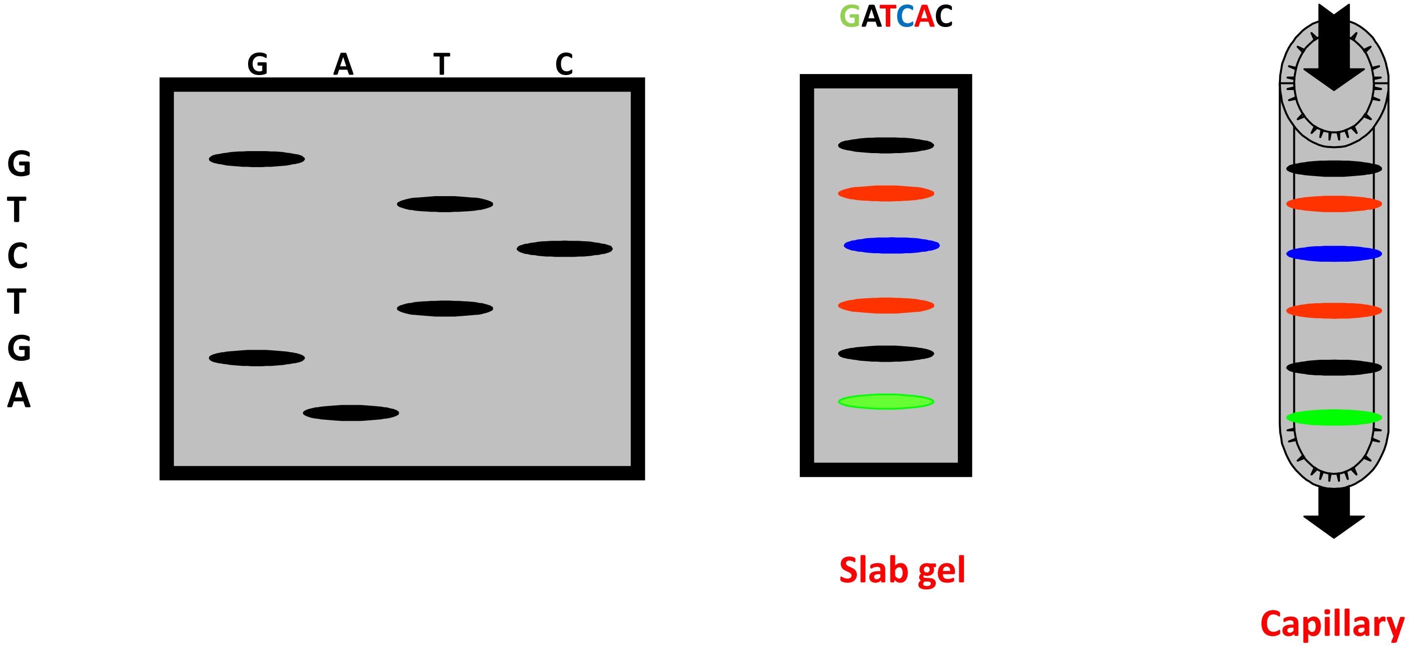
- A distinct dye or “color” is used for each of the four ddNTP.
- Since the terminating nucleotides can be distinguished by color, all four reactions can be performed in a single tube.



The fragments are distinguished by size and “color.”

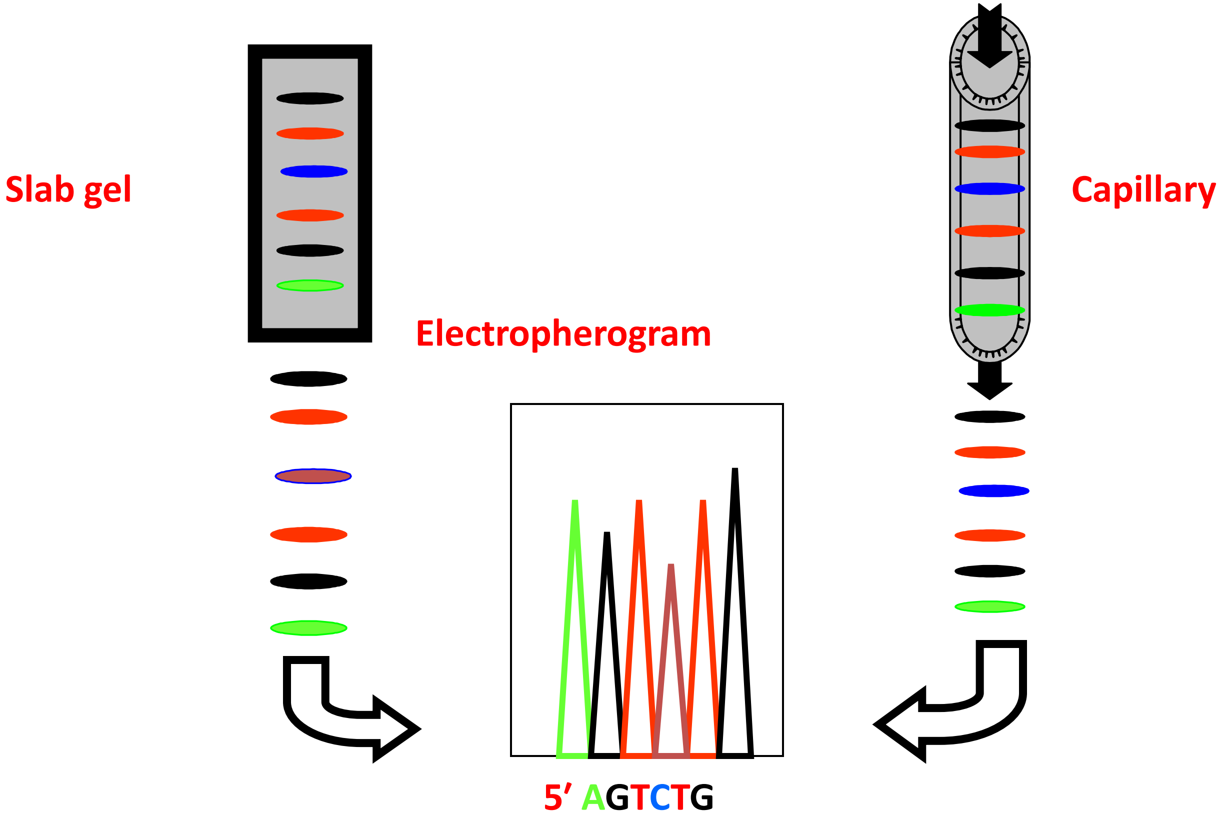
Dye Terminator Sequencing

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



Dye Terminator Sequencing

- The DNA ladder is read on an electropherogram.



Slab gel

Electropherogram

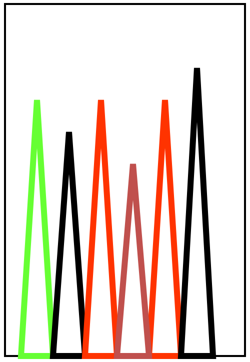
Capillary

5′ AGTCTG

Automated Sequencing

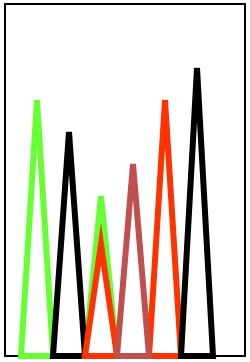
- Dye primer or dye terminator sequencing on capillary instruments.
- Sequence analysis software provides analyzed sequence in text and electropherogram form.
- Peak patterns reflect mutations or sequence changes.

T/T



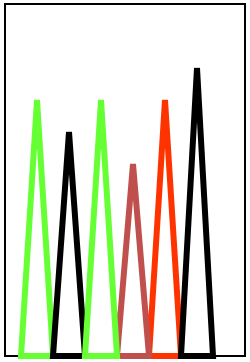
5’ A G T C T G

T/A



5’ AG(T/A)CTG

A/A



5’ A G A C T G

Sanger Sequencing
Useful videos

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

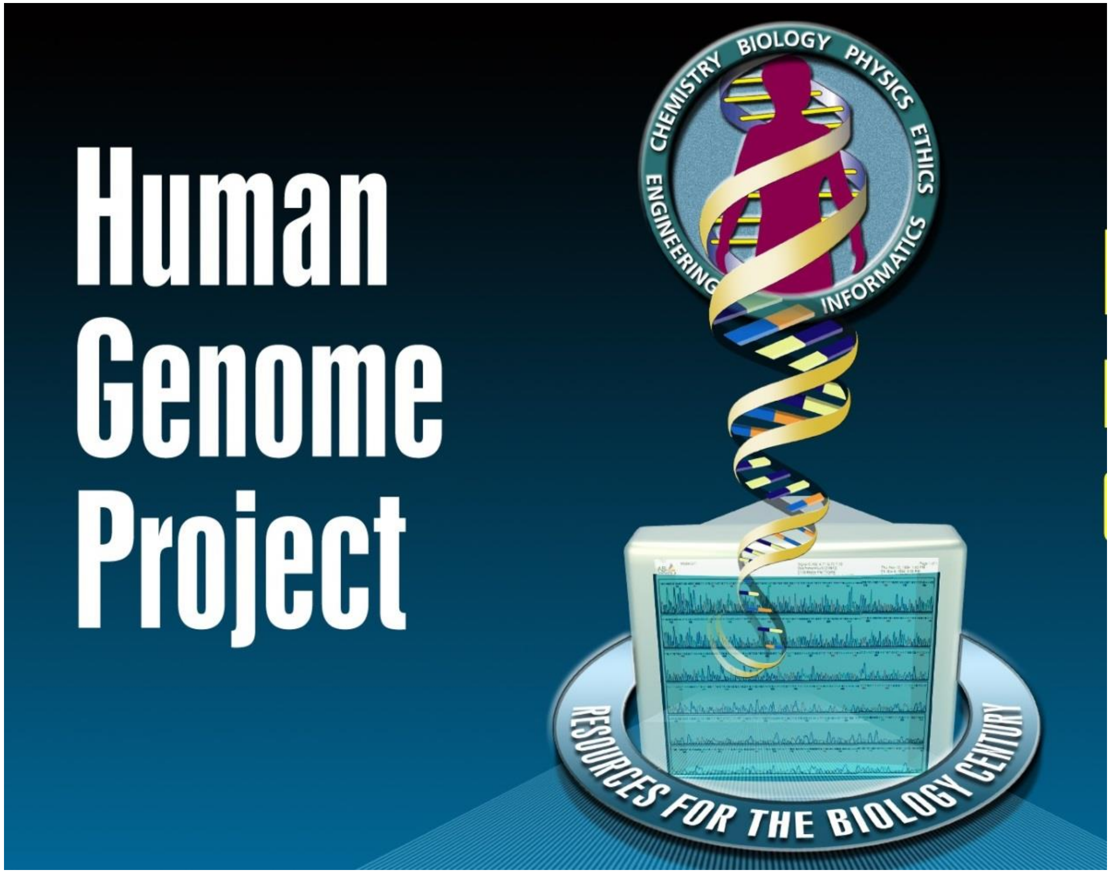
Features of Sanger Sequencing

- 96-384 sequences per run
- 500bp-1kb read lengths
- $100 per megabase
- Accuracy decreases with length (99.999% at 500bp down to 99% at 900bp)
- Still the most accurate technique for sequencing

Limitations of Sanger Sequencing

- Cloning/Sub-cloning
  - DNA must be compatible with biological machinery of host cells and can introduce bias
  - Labour and/or machines to prepare clones requires significant capital
- Difficult to distinguish allele frequency
  - Usually 10% is the limit of detection for clinical variants
- Cost
  - $10,000,000 to sequence a 1Gbase genome to 10x coverage

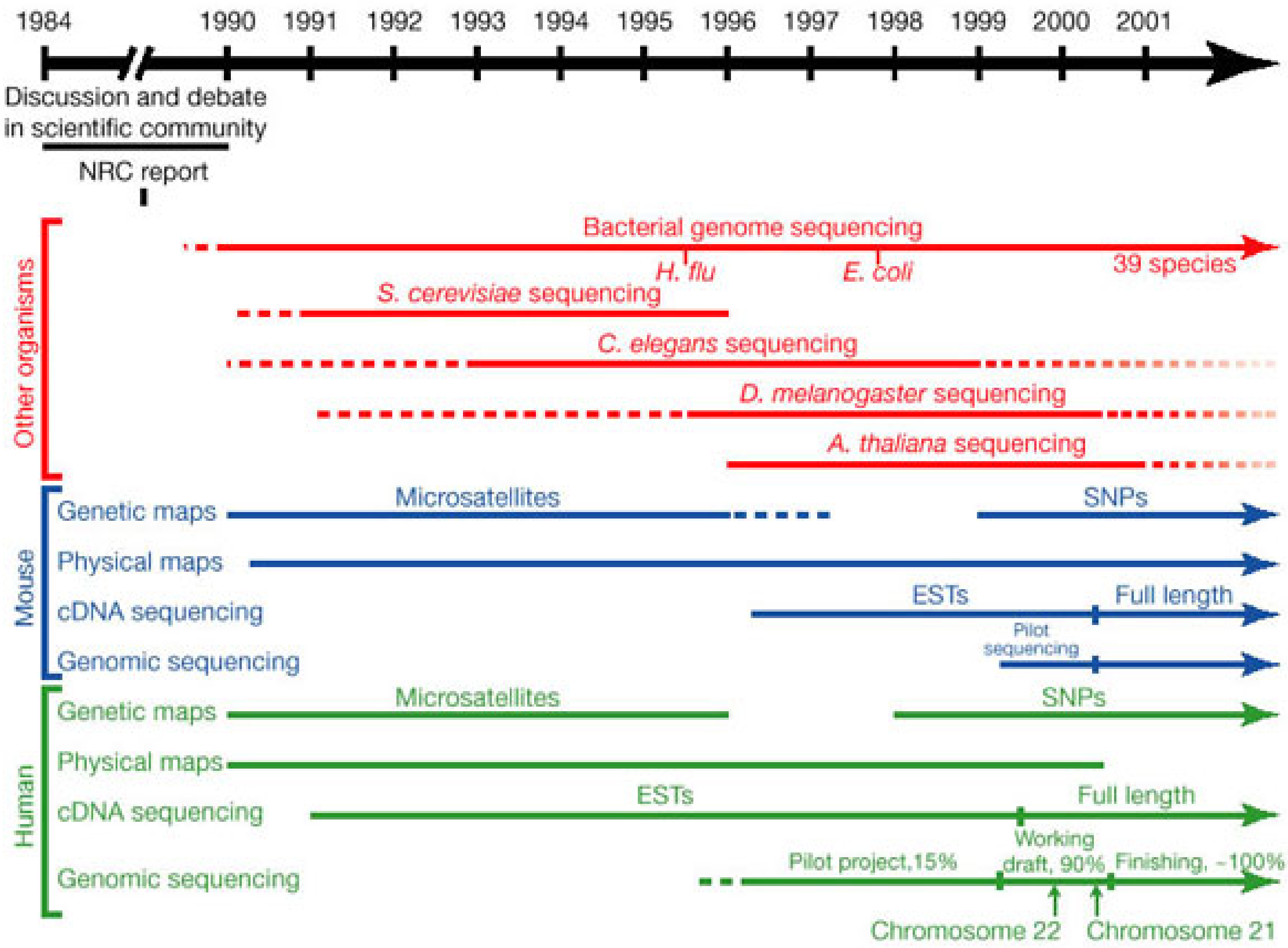
Human genome project



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

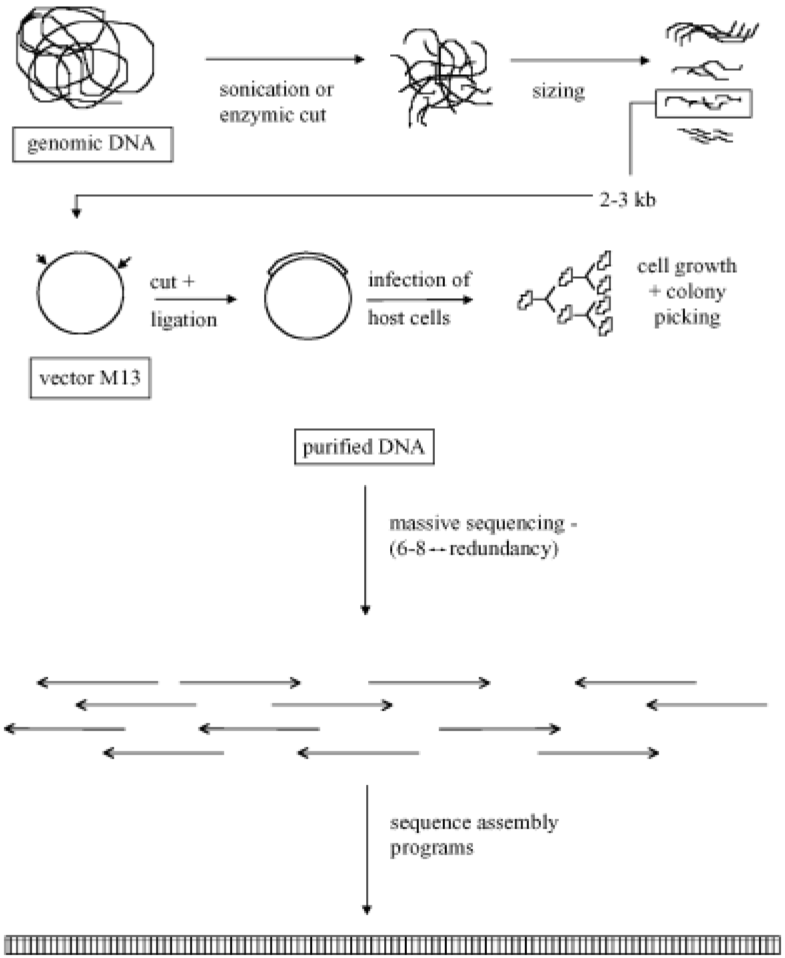


http://bit.ly/q3Qsd5

How it was Accomplished

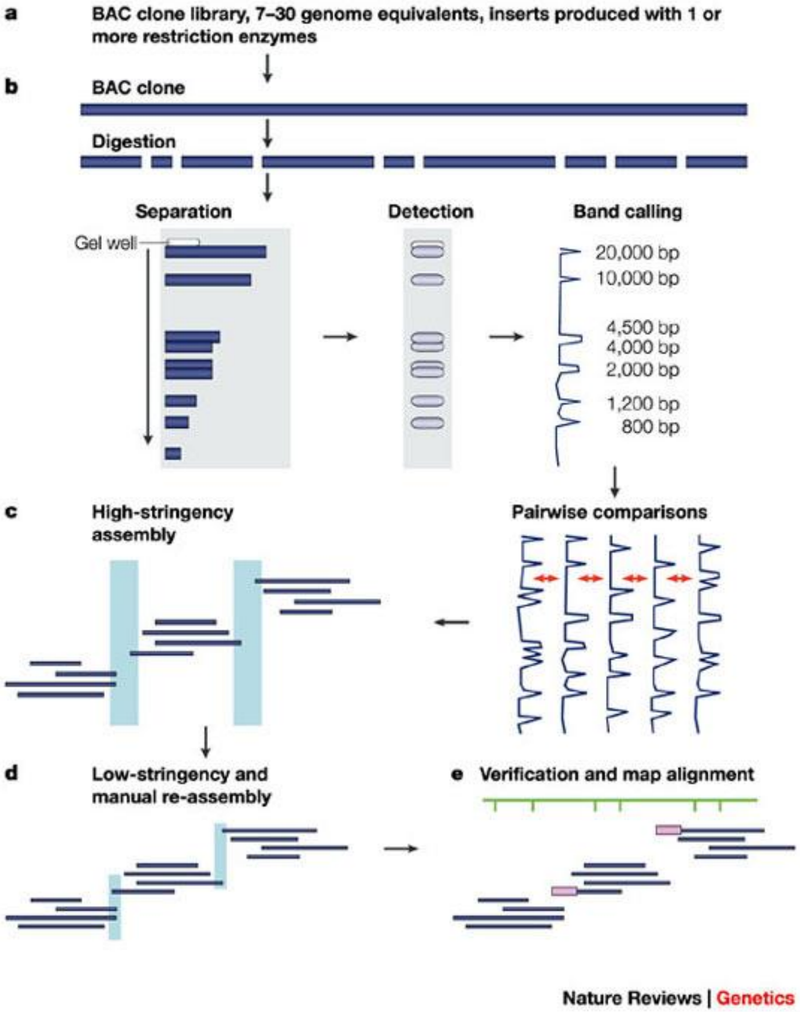
- Public Project
  - Hierarchical shotgun approach
  - Large segments of DNA were cloned via BACs and located along the chromosome
  - These BACs where shotgun sequenced
- Celera
  - Pure shotgun sequencing
  - Used public data (released daily) to help with assembly

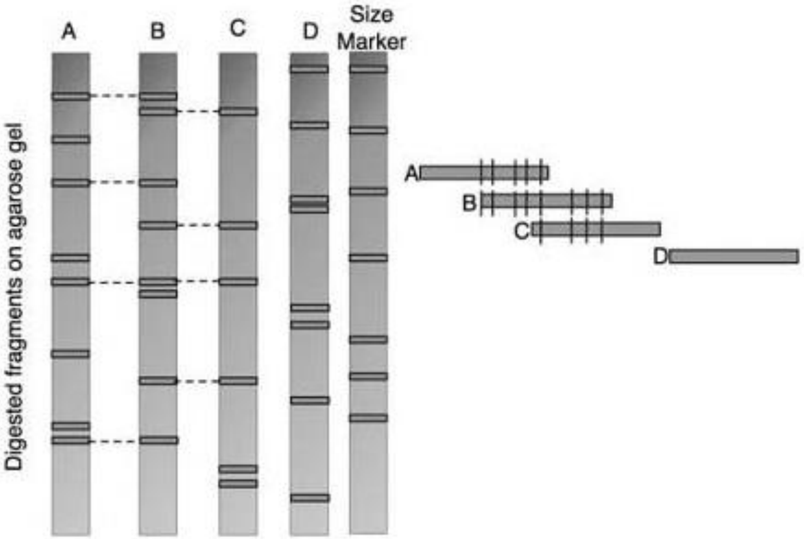
Method 1: Hierarchical Sequencing



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

Using Bacterial artificial chromosomes (BACs) to aid assembly

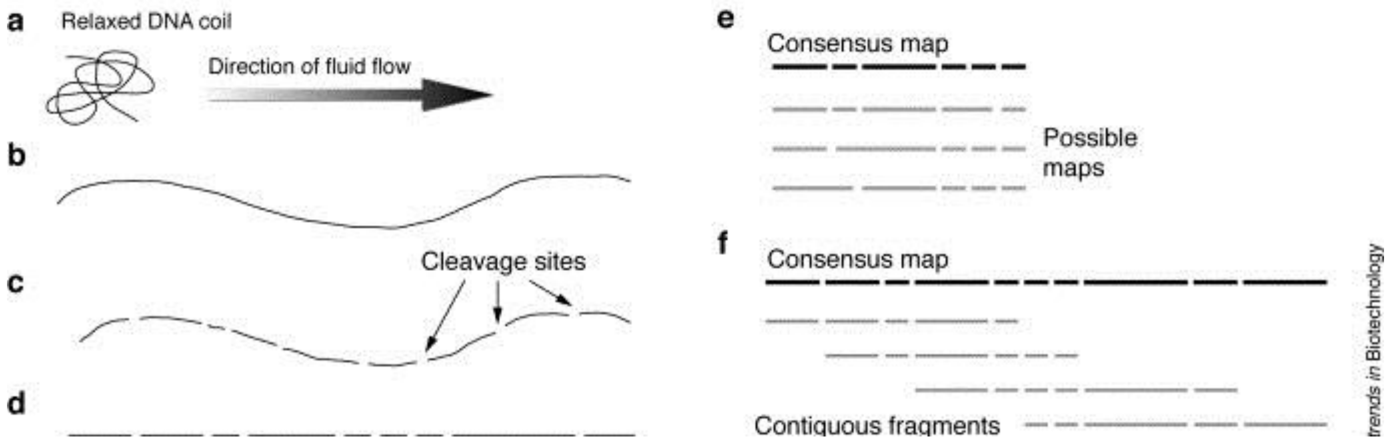

a
BAC clone library, 7–30 genome equivalents, inserts produced with 1 or more restriction enzymes
b
BAC clone
Digestion
Separation
Gel well
Detection
Band calling
20,000 bp
10,000 bp
4,500 bp
4,000 bp
2,000 bp
1,200 bp
800 bp
c
High-stringency assembly
Pairwise comparisons
d
Low-stringency and manual re-assembly
e
Verification and map alignment
Nature Reviews | Genetics

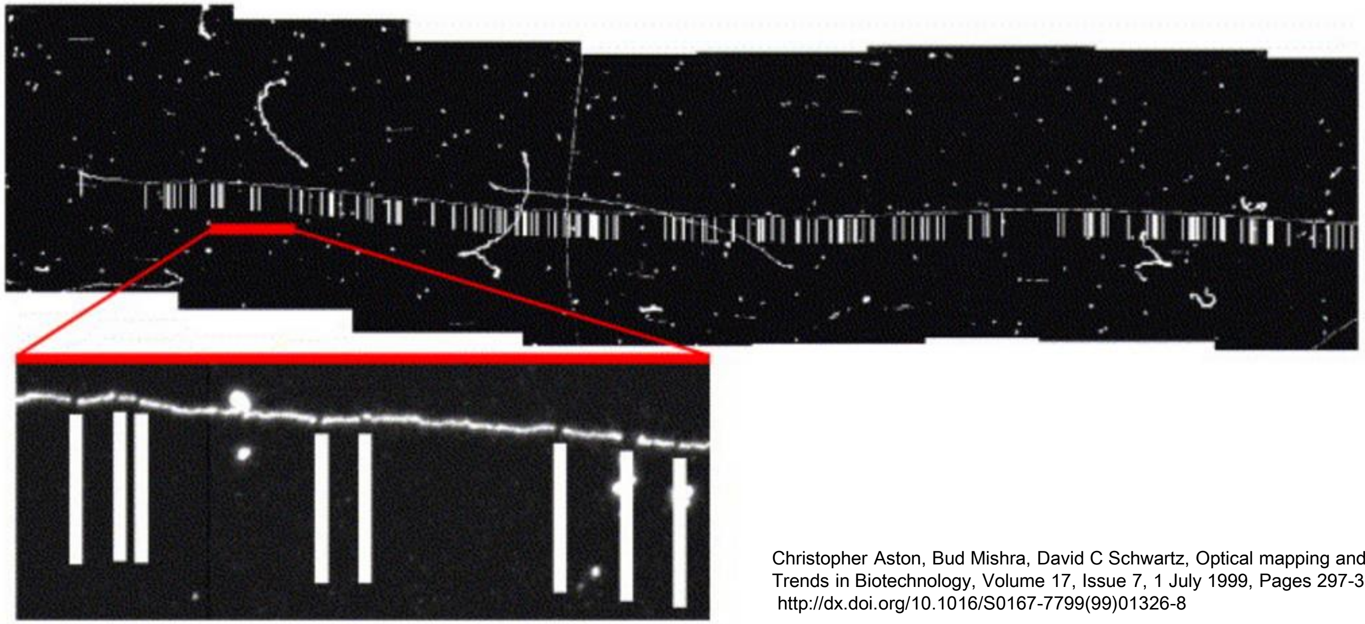


Nature Reviews Genetics **5**, 578-588 (August 2004) | doi:10.1038/nrg1404

Genetics, Genomics and Breeding of Oilseed Brassicas, Page 134

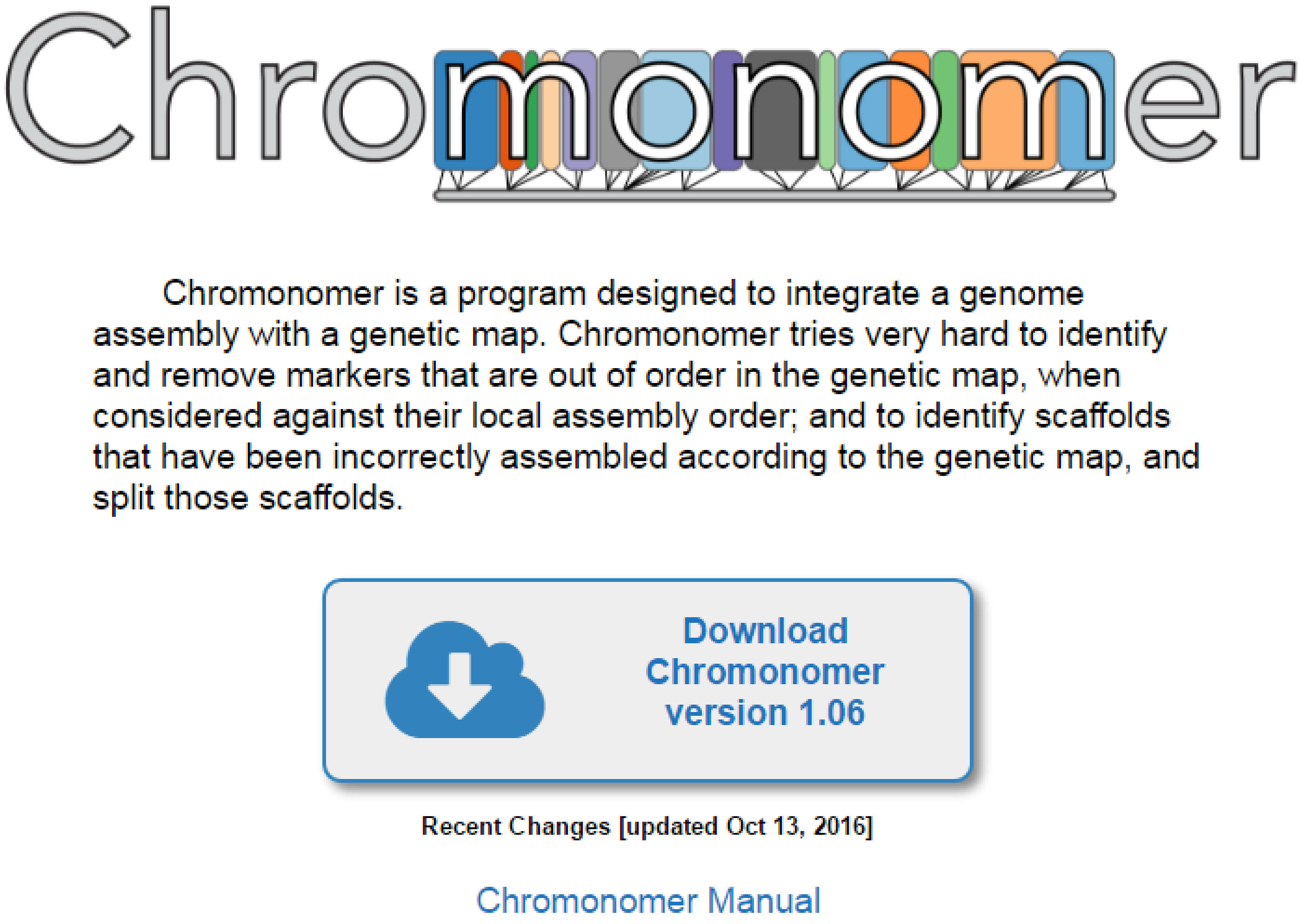
Using optical mapping approaches to aid assembly





Christopher Aston, Bud Mishra, David C Schwartz, Optical mapping and its potential for large-scale sequencing projects, Trends in Biotechnology, Volume 17, Issue 7, 1 July 1999, Pages 297-302, ISSN 0167-7799, http://dx.doi.org/10.1016/S0167-7799(99)01326-8

Using genetic maps to improve



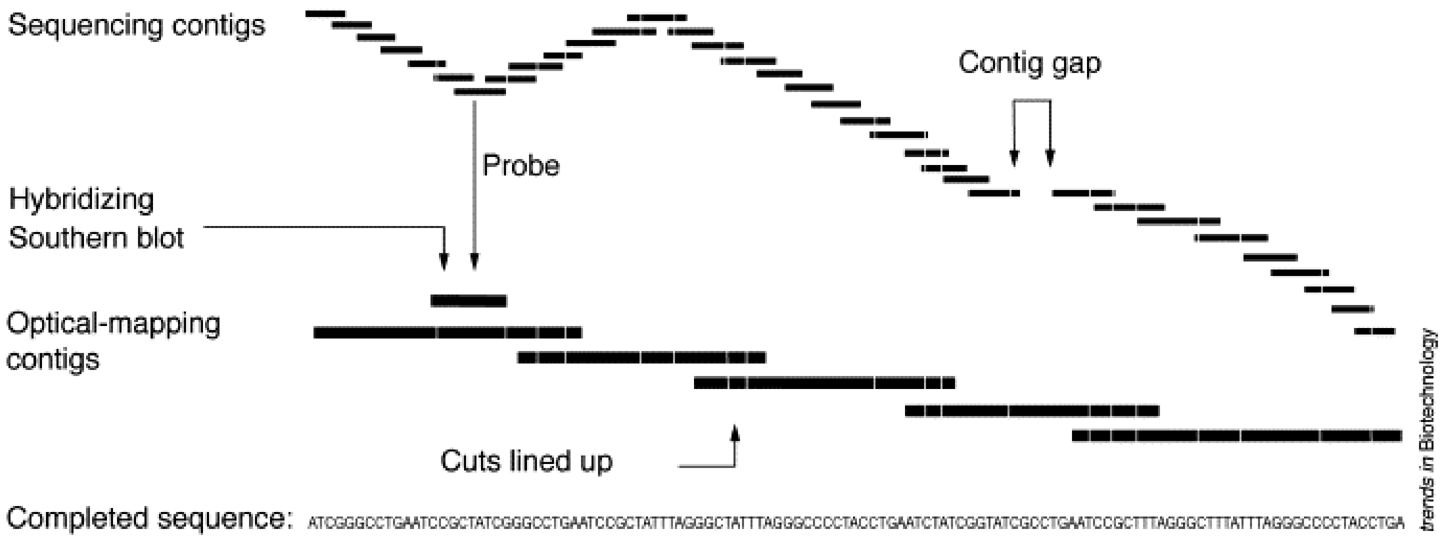
Chromonomer is a program designed to integrate a genome assembly with a genetic map. Chromonomer tries very hard to identify and remove markers that are out of order in the genetic map, when considered against their local assembly order; and to identify scaffolds that have been incorrectly assembled according to the genetic map, and split those scaffolds.

Recent Changes [updated Oct 13, 2016]

Chromonomer Manual

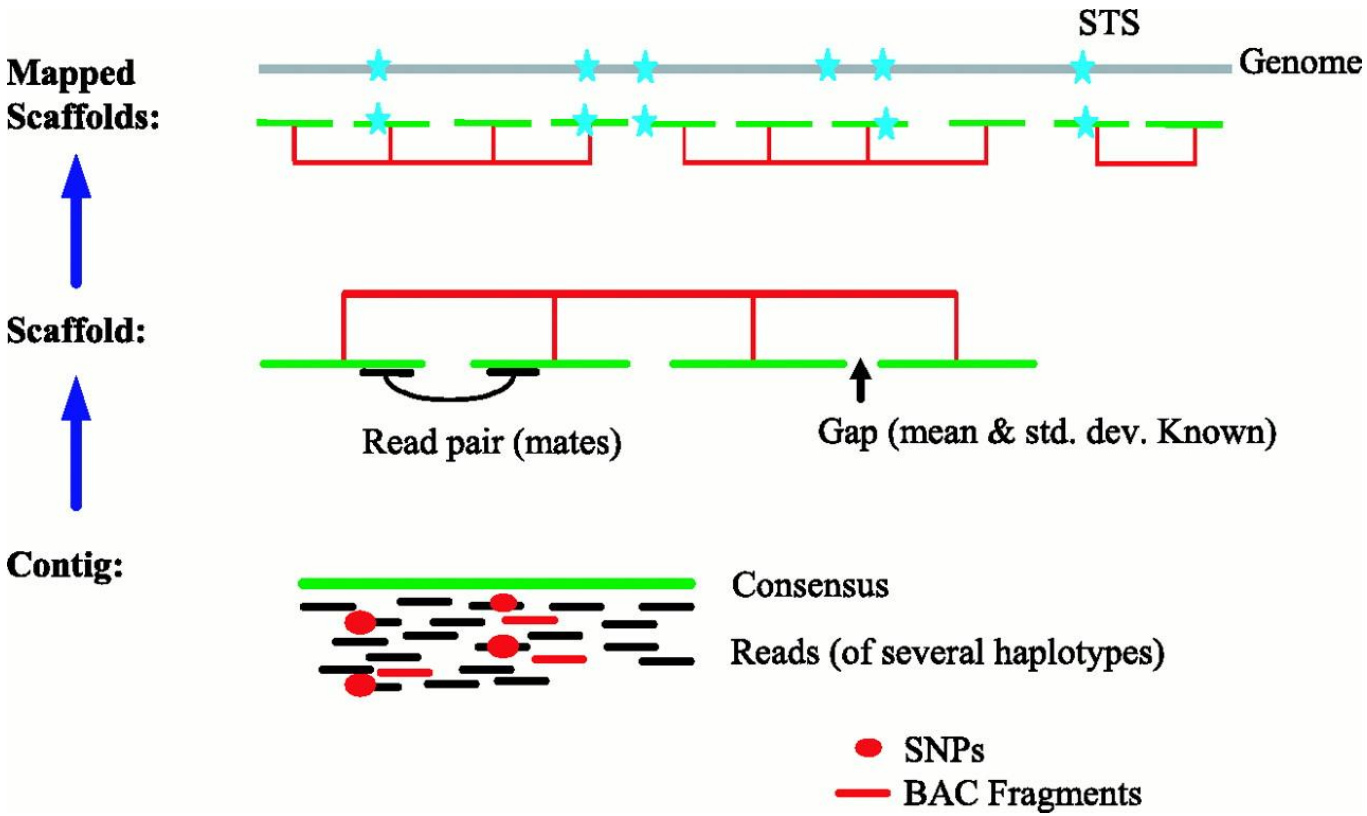
http://catchenlab.life.illinois.edu/chromonomer/

Using hybridisation probes to aid assembly


Sequencing contigs
Contig gap
Probe
Hybridizing
Southern blot
Optical-mapping
contigs
Cuts lined up
Completed sequence: ATCGGGCCTGAATCCGCTATCGGGCCTGAATCCGCTATTTAGGGCTATTTAGGGCCCCTACCTGAATCTATCGGTATCGCCTGAATCCGCTTTAGGGCTTTATTTAGGGCCCCTACCTGA

trends in Biotechnology

Method 2: Celera Shotgun Sequencing



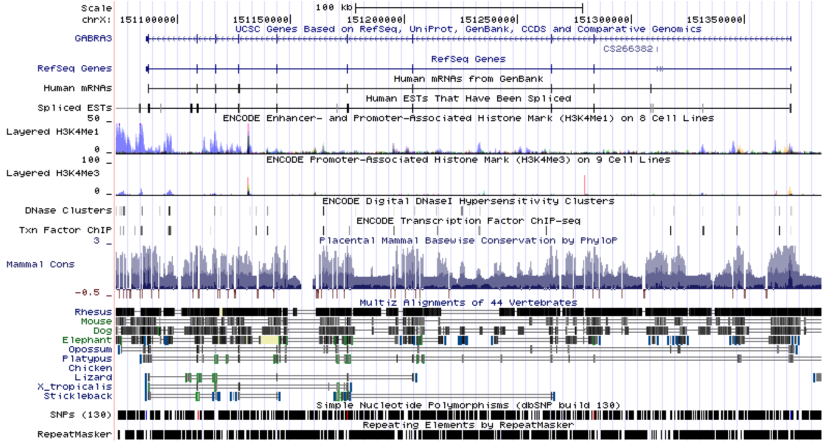
- Used paired-end strategy with variable insert size: 2, 10, and 50kbp

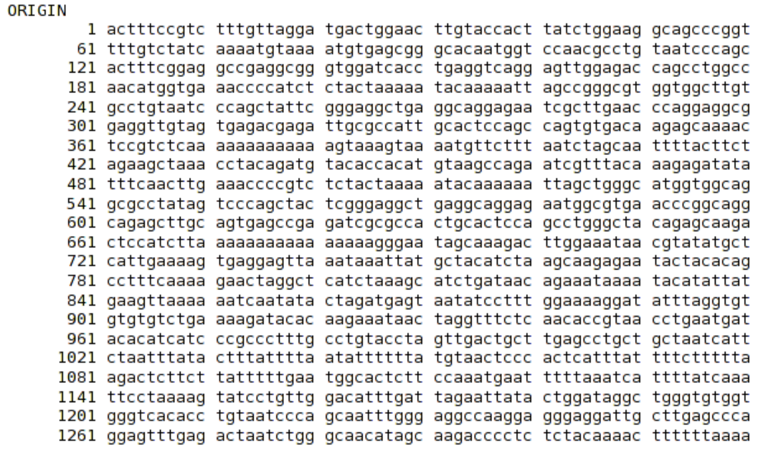
Outcome of the HGP

- 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

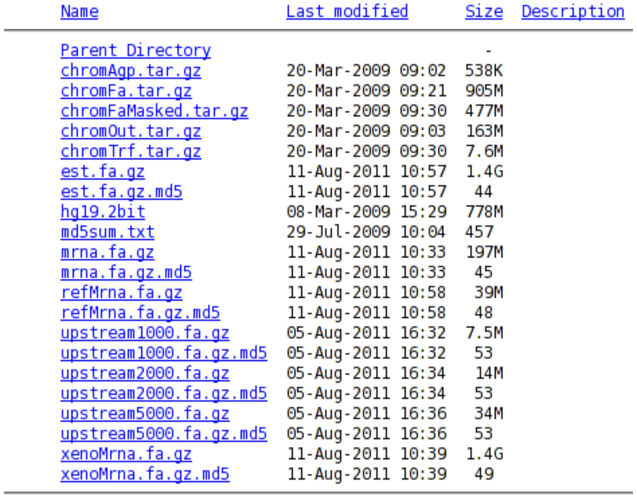
20110813 http://www.ncbi.nlm.nih.gov/sites/genome

HGP Data Access

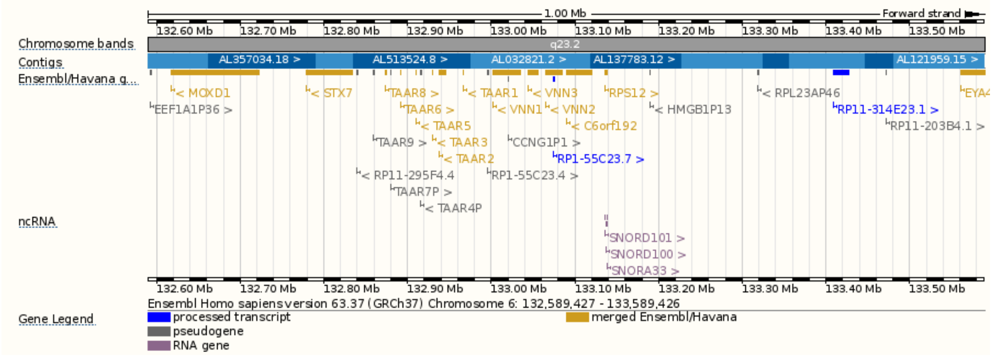

Scale
chrX:
100 kb
151100000
151150000
151200000
151250000
151300000
151350000
UCSC Genes Based on RefSeq, UniProt, GenBank, CCDS and Comparative Genomics
GABRA3
CS266382
RefSeq Genes
RefSeq Genes
Human mRNAs from GenBank
Human mRNAs
Human ESTs That Have Been Spliced
Spliced ESTs
50 _
ENCODE Enhancer- 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
```



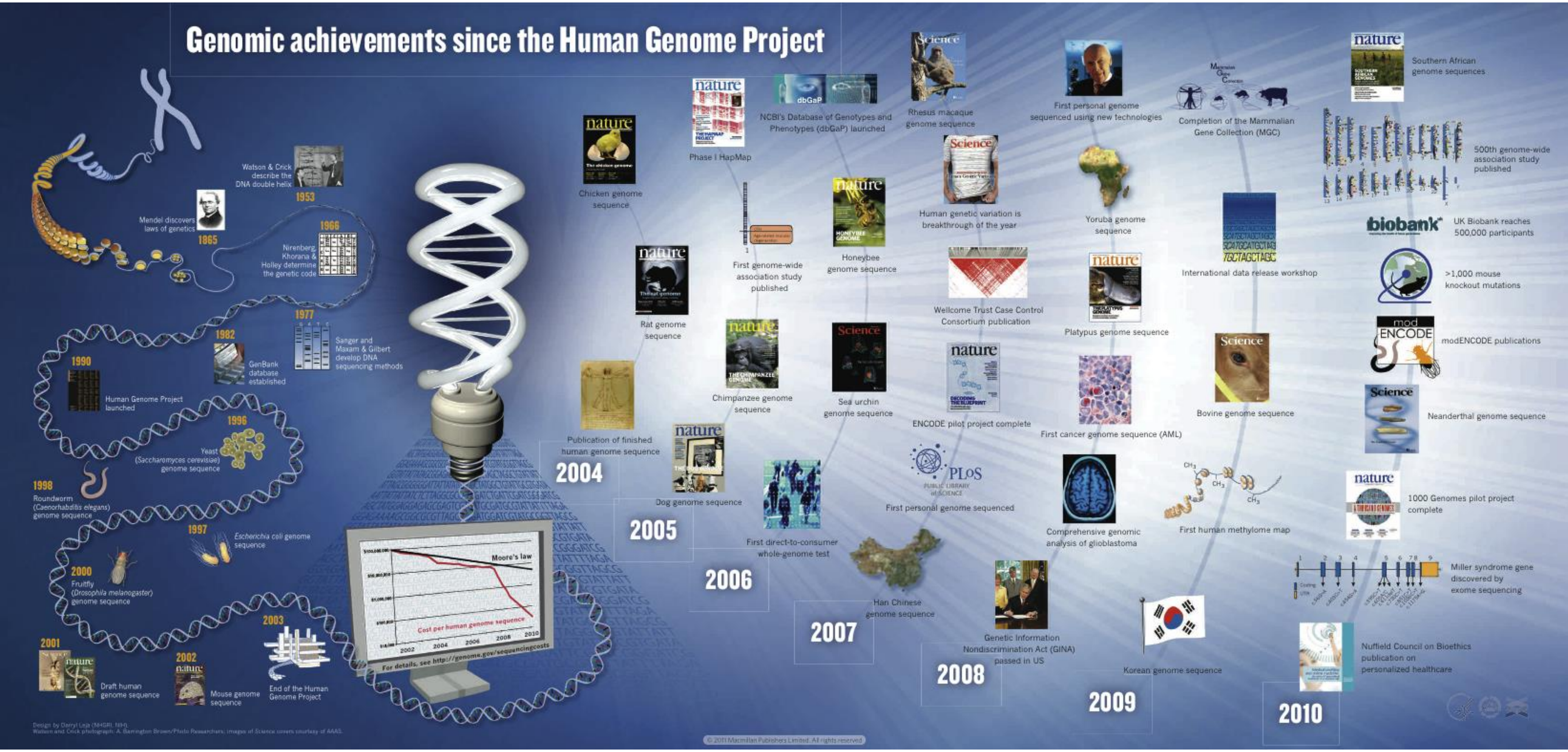
Name	Last modified	Size	Description
Parent Directory		-	
chromAgp.tar.gz	20-Mar-2009 09:02	538K	
chromFa.tar.gz	20-Mar-2009 09:21	905M	
chromFaMasked.tar.gz	20-Mar-2009 09:30	477M	
chromOut.tar.gz	20-Mar-2009 09:03	163M	
chromTrf.tar.gz	20-Mar-2009 09:30	7.6M	
est.fa.gz	11-Aug-2011 10:57	1.4G	
est.fa.gz.md5	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

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
1997
Escherichia coli genome sequence
1998
Roundworm (Caenorhabditis elegans) genome sequence
2000
Fruitfly (Drosophila melanogaster) genome sequence
2001
Draft human genome sequence
2002
Mouse genome sequence
2003
End of the Human Genome Project
Moore's law
Cost per human genome sequence
$100,000,000
$10,000,000
$1,000,000
$100,000
$10,000
2002
2004
2006
2008
2010
For details, see http://genome.gov/sequencingcosts
2004
Publication of finished human genome sequence
Chicken genome sequence
Rat genome sequence
2005
Phase I HapMap
Dog genome sequence
Chimpanzee genome sequence
2006
First genome-wide association study published
Honeybee genome sequence
Sea urchin genome sequence
NCBI's Database of Genotypes and Phenotypes (dbGaP) launched
dbGaP
Rhesus macaque genome sequence
First direct-to-consumer whole-genome test
2007
Human genetic variation is breakthrough of the year
Wellcome Trust Case Control Consortium publication
ENCODE pilot project complete
First personal genome sequenced
PLoS PUBLIC LIBRARY of SCIENCE
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
CH3
Korean genome sequence
2010
Southern African genome sequences
500th genome-wide association study published
biobank
UK Biobank reaches 500,000 participants
>1,000 mouse knockout mutations
mod ENCODE
modENCODE publications
Neanderthal genome sequence
1000 Genomes pilot project complete
Miller syndrome gene discovered by exome sequencing
Coding
UTR
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 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 $100,000s to $10s of millions using capillary sequencers
- Labour intensive methodology

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

Second generation short read technologies

Common features

- Generation and sequencing of monoclonal populations of DNA molecules
- Rely on polymerases to re-synthesise complementary strands of DNA
- Typically rely on either fluorescently-labelled nucleotides or monitoring of hydrogen release upon incorporation

Sequencing – 1990s-2007

PRODUCTION

Rooms of equipment

Subcloning > picking > prepping

35 FTEs

3-4 weeks

SEQUENCING

74x Capillary Sequencers

10 FTEs

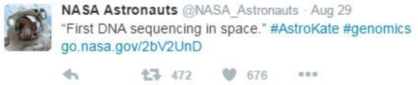
15-40 runs per day

1-2Mb per instrument per day

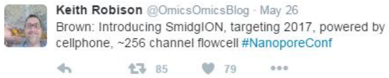
120Mb total capacity per day

Sequencing today

Sequencing today?


NASA Astronauts @NASA_Astronauts · Aug 29
"First DNA sequencing in space." #AstroKate #genomics
go.nasa.gov/2bV2UnD

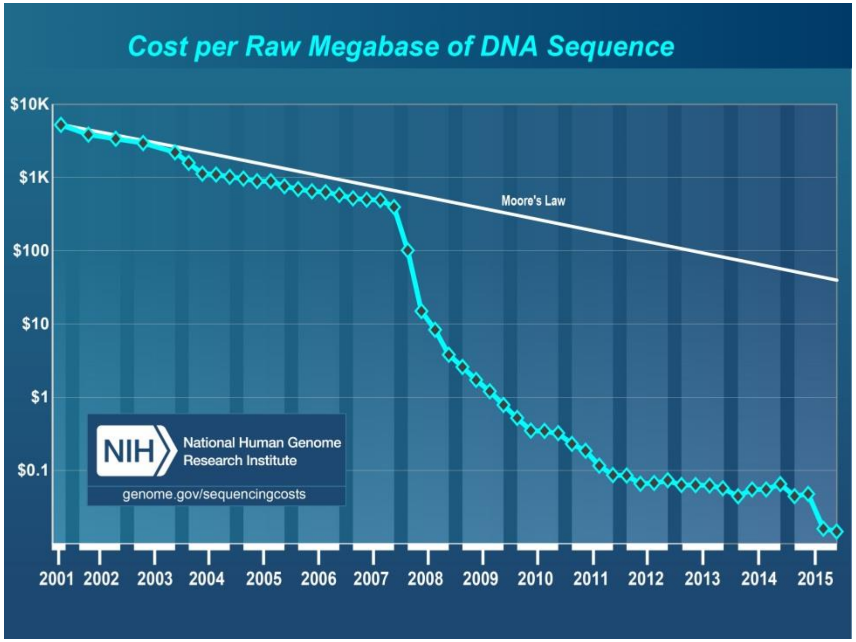
472 676


Keith Robison @OmicsOmicsBlog · May 26
Brown: Introducing SmidgION, targeting 2017, powered by cellphone, ~256 channel flowcell #NanoporeConf

85 79

Key advantages over Sanger Sequencing

- Hugely reduced labour requirements
  - No need to perform cloning
- Reduced cost per sequence
- Reduced time to result
- Decentralisation
- Enabling new techniques (e.g. gene expression profiling)



Platform comparison

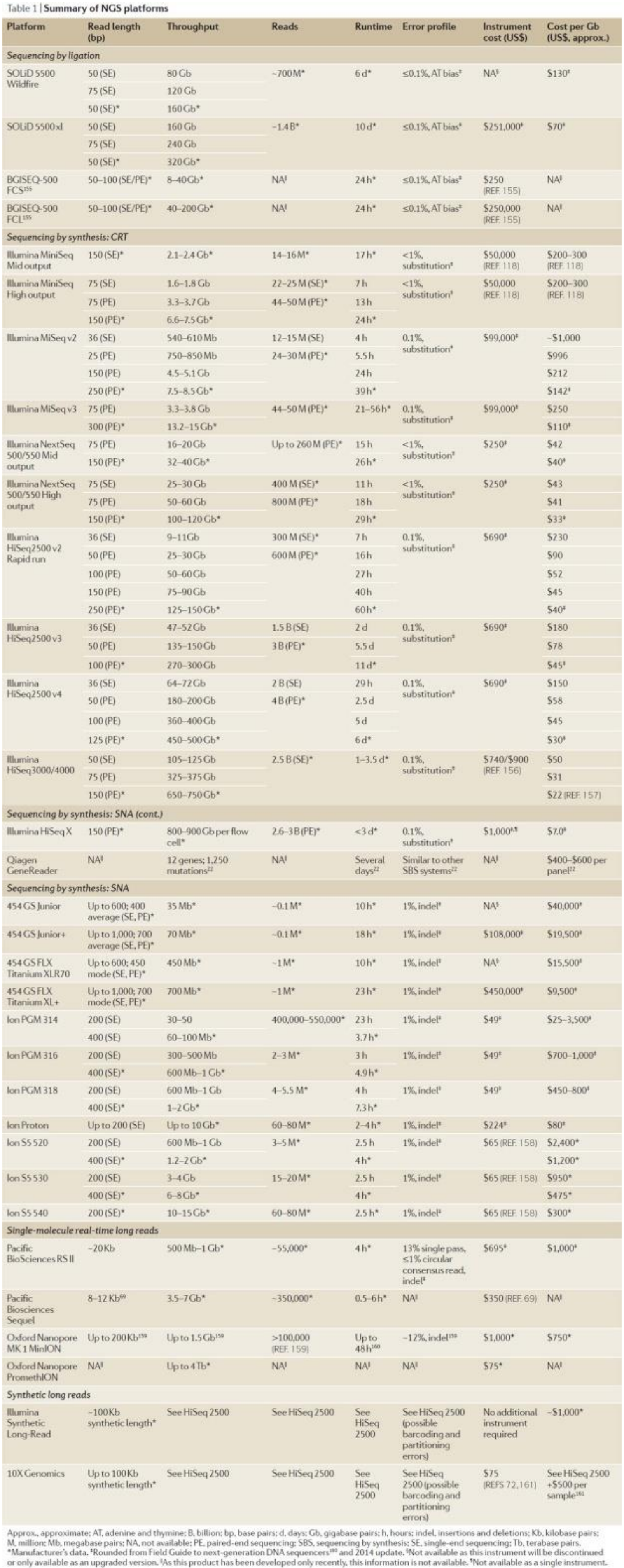


Table 1 | **Summary of NGS platforms**

Platform	Read length (bp)	Throughput	Reads	Runtime	Error profile	Instrument cost (US$)	Cost per Gb (US$, approx.)
Sequencing by ligation							
SOLiD 5500 Wildfire	50 (SE)	80 Gb	~700 M*	6 d*	≤0.1%, AT bias[‡]	NA[§]	$130[‡]
	75 (SE)	120 Gb					
	50 (SE)*	160 Gb*					
SOLiD 5500xl	50 (SE)	160 Gb	~1.4 B*	10 d*	≤0.1%, AT bias[‡]	$251,000[‡]	$70[‡]
	75 (SE)	240 Gb					
	50 (SE)*	320 Gb*					
BGISEQ-500 FCS[155]	50–100 (SE/PE)*	8–40 Gb*	NA[||]	24 h*	≤0.1%, AT bias[‡]	$250 (REF. 155)	NA[||]
BGISEQ-500 FCL[155]	50–100 (SE/PE)*	40–200 Gb*	NA[||]	24 h*	≤0.1%, AT bias[‡]	$250,000 (REF. 155)	NA[||]
Sequencing by synthesis: CRT							
Illumina MiniSeq Mid output	150 (SE)*	2.1–2.4 Gb*	14–16 M*	17 h*	<1%, substitution[‡]	$50,000 (REF. 118)	$200–300 (REF. 118)
Illumina MiniSeq High output	75 (SE)	1.6–1.8 Gb	22–25 M (SE)*	7 h	<1%, substitution[‡]	$50,000 (REF. 118)	$200–300 (REF. 118)
	75 (PE)	3.3–3.7 Gb	44–50 M (PE)*	13 h			
	150 (PE)*	6.6–7.5 Gb*		24 h*			
Illumina MiSeq v2	36 (SE)	540–610 Mb	12–15 M (SE)	4 h	0.1%, substitution[‡]	$99,000[‡]	~$1,000
	25 (PE)	750–850 Mb	24–30 M (PE)*	5.5 h			$996
	150 (PE)	4.5–5.1 Gb		24 h			$212
	250 (PE)*	7.5–8.5 Gb*		39 h*			$142[‡]
Illumina MiSeq v3	75 (PE)	3.3–3.8 Gb	44–50 M (PE)*	21–56 h*	0.1%, substitution[‡]	$99,000[‡]	$250
	300 (PE)*	13.2–15 Gb*					$110[‡]
Illumina NextSeq 500/550 Mid output	75 (PE)	16–20 Gb	Up to 260 M (PE)*	15 h	<1%, substitution[‡]	$250[‡]	$42
	150 (PE)*	32–40 Gb*		26 h*			$40[‡]
Illumina NextSeq 500/550 High output	75 (SE)	25–30 Gb	400 M (SE)*	11 h	<1%, substitution[‡]	$250[‡]	$43
	75 (PE)	50–60 Gb	800 M (PE)*	18 h			$41
	150 (PE)*	100–120 Gb*		29 h*			$33[‡]
Illumina HiSeq2500 v2 Rapid run	36 (SE)	9–11Gb	300 M (SE)*	7 h	0.1%, substitution[‡]	$690[‡]	$230
	50 (PE)	25–30 Gb	600 M (PE)*	16 h			$90
	100 (PE)	50–60 Gb		27 h			$52
	150 (PE)	75–90 Gb		40 h			$45
	250 (PE)*	125–150 Gb*		60 h*			$40[‡]
Illumina HiSeq2500 v3	36 (SE)	47–52 Gb	1.5 B (SE)	2 d	0.1%, substitution[‡]	$690[‡]	$180
	50 (PE)	135–150 Gb	3 B (PE)*	5.5 d			$78
	100 (PE)*	270–300 Gb		11 d*			$45[‡]
Illumina HiSeq2500 v4	36 (SE)	64–72 Gb	2 B (SE)	29 h	0.1%, substitution[‡]	$690[‡]	$150
	50 (PE)	180–200 Gb	4 B (PE)*	2.5 d			$58
	100 (PE)	360–400 Gb		5 d			$45
	125 (PE)*	450–500 Gb*		6 d*			$30[‡]
Illumina HiSeq3000/4000	50 (SE)	105–125 Gb	2.5 B (SE)*	1–3.5 d*	0.1%, substitution[‡]	$740/$900 (REF. 156)	$50
	75 (PE)	325–375 Gb					$31
	150 (PE)*	650–750 Gb*					$22 (REF. 157)
Sequencing by synthesis: SNA (cont.)							
Illumina HiSeq X	150 (PE)*	800–900 Gb per flow cell*	2.6–3 B (PE)*	<3 d*	0.1%, substitution[‡]	$1,000[‡,¶]	$7.0[‡]
Qiagen GeneReader	NA[||]	12 genes; 1,250 mutations[22]	NA[||]	Several days[22]	Similar to other SBS systems[22]	NA[||]	$400–$600 per panel[22]
Sequencing by synthesis: SNA							
454 GS Junior	Up to 600; 400 average (SE, PE)*	35 Mb*	~0.1 M*	10 h*	1%, indel[‡]	NA[§]	$40,000[‡]
454 GS Junior+	Up to 1,000; 700 average (SE, PE)*	70 Mb*	~0.1 M*	18 h*	1%, indel[‡]	$108,000[‡]	$19,500[‡]
454 GS FLX Titanium XLR70	Up to 600; 450 mode (SE, PE)*	450 Mb*	~1 M*	10 h*	1%, indel[‡]	NA[§]	$15,500[‡]
454 GS FLX Titanium XL+	Up to 1,000; 700 mode (SE, PE)*	700 Mb*	~1 M*	23 h*	1%, indel[‡]	$450,000[‡]	$9,500[‡]
Ion PGM 314	200 (SE)	30–50	400,000–550,000*	23 h	1%, indel[‡]	$49[‡]	$25–3,500[‡]
	400 (SE)	60–100 Mb*		3.7 h*			
Ion PGM 316	200 (SE)	300–500 Mb	2–3 M*	3 h	1%, indel[‡]	$49[‡]	$700–1,000[‡]
	400 (SE)*	600 Mb–1 Gb*		4.9 h*			
Ion PGM 318	200 (SE)	600 Mb–1 Gb	4–5.5 M*	4 h	1%, indel[‡]	$49[‡]	$450–800[‡]
	400 (SE)*	1–2 Gb*		7.3 h*			
Ion Proton	Up to 200 (SE)	Up to 10 Gb*	60–80 M*	2–4 h*	1%, indel[‡]	$224[‡]	$80[‡]
Ion S5 520	200 (SE)	600 Mb–1 Gb	3–5 M*	2.5 h	1%, indel[‡]	$65 (REF. 158)	$2,400*
	400 (SE)*	1.2–2 Gb*		4 h*			$1,200*
Ion S5 530	200 (SE)	3–4 Gb	15–20 M*	2.5 h	1%, indel[‡]	$65 (REF. 158)	$950*
	400 (SE)*	6–8 Gb*		4 h*			$475*
Ion S5 540	200 (SE)*	10–15 Gb*	60–80 M*	2.5 h*	1%, indel[‡]	$65 (REF. 158)	$300*
Single-molecule real-time long reads							
Pacific BioSciences RS II	~20 Kb	500 Mb–1 Gb*	~55,000*	4 h*	13% single pass, ≤1% circular consensus read, indel[‡]	$695[‡]	$1,000[‡]
Pacific Biosciences Sequel	8–12 Kb[69]	3.5–7 Gb*	~350,000*	0.5–6 h*	NA[||]	$350 (REF. 69)	NA[||]
Oxford Nanopore MK 1 MinION	Up to 200 Kb[159]	Up to 1.5 Gb[159]	>100,000 (REF. 159)	Up to 48 h[160]	~12%, indel[159]	$1,000*	$750*
Oxford Nanopore PromethION	NA[||]	Up to 4 Tb*	NA[||]	NA[||]	NA[||]	$75*	NA[||]
Synthetic long reads							
Illumina Synthetic Long-Read	~100 Kb synthetic length*	See HiSeq 2500	See HiSeq 2500	See HiSeq 2500	See HiSeq 2500 (possible barcoding and partitioning errors)	No additional instrument required	~$1,000*
10X Genomics	Up to 100 Kb synthetic length*	See HiSeq 2500	See HiSeq 2500	See HiSeq 2500	See HiSeq 2500 (possible barcoding and partitioning errors)	$75 (REFS 72,161)	See HiSeq 2500 +$500 per sample[161]

Approx., approximate; AT, adenine and thymine; B, billion; bp, base pairs; d, days; Gb, gigabase pairs; h, hours; indel, insertions and deletions; Kb, kilobase pairs; M, million; Mb, megabase pairs; NA, not available; PE, paired-end sequencing; SBS, sequencing by synthesis; SE, single-end sequencing; Tb, terabase pairs. *Manufacturer's data. [‡]Rounded from Field Guide to next-generation DNA sequencers[180] and 2014 update. [§]Not available as this instrument will be discontinued or only available as an upgraded version. [||]As this product has been developed only recently, this information is not available. [¶]Not available as a single instrument.

http://www.nature.com/nrg/journal/v17/n6/fig_tab/nrg.2016.49_T1.html

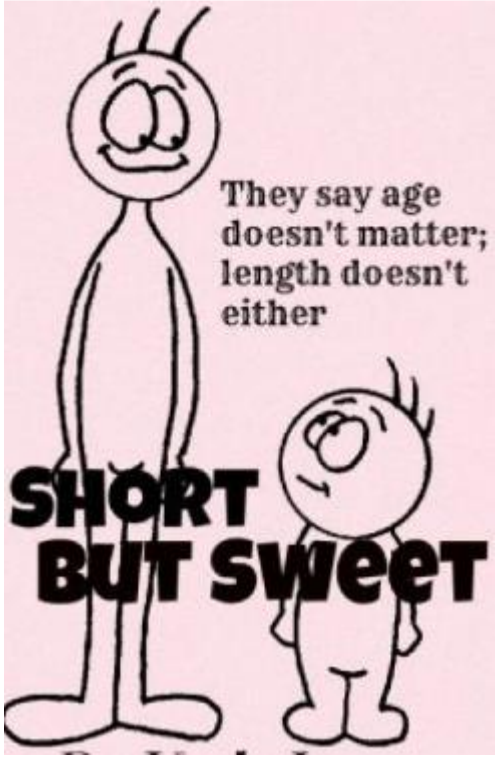
Illumina Sequencing By Synthesis

Illumina HiSeq

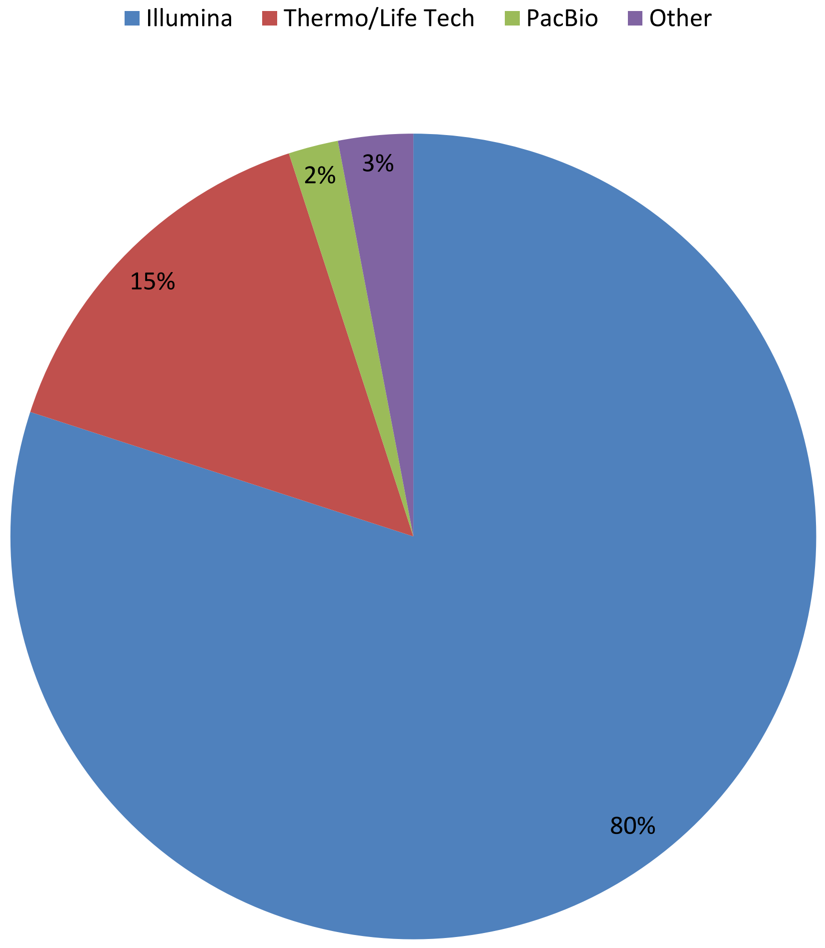
Illumina NextSeq

Illumina MiSeq

The workhorse of modern genomics



Approximate Market Share 2015



http://www.gosreports.com/global-sequencing-instrument-industry-2015-market-research-report/

Fun fact

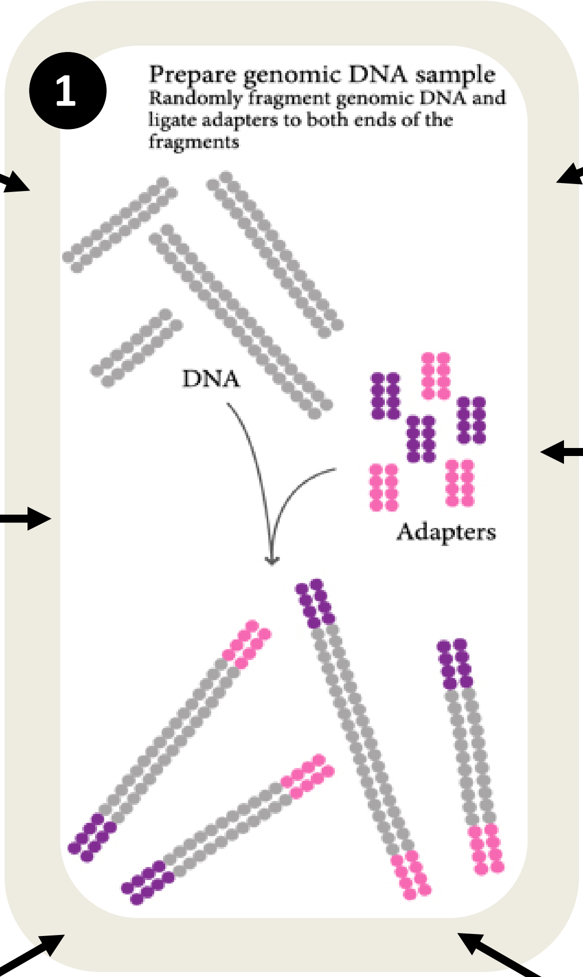
- Clive Brown
- Formerly director of Computational Biology at Solexa (Illumina)
- Chief Scientific Officer at Oxford Nanopore

Step 1: Sample Preparation

Active Chromatin

Genomic DNA

min 1ng



1 Prepare genomic DNA sample

Randomly fragment genomic DNA and ligate adapters to both ends of the fragments

mRNA

0.5ug total RNA

Small RNA

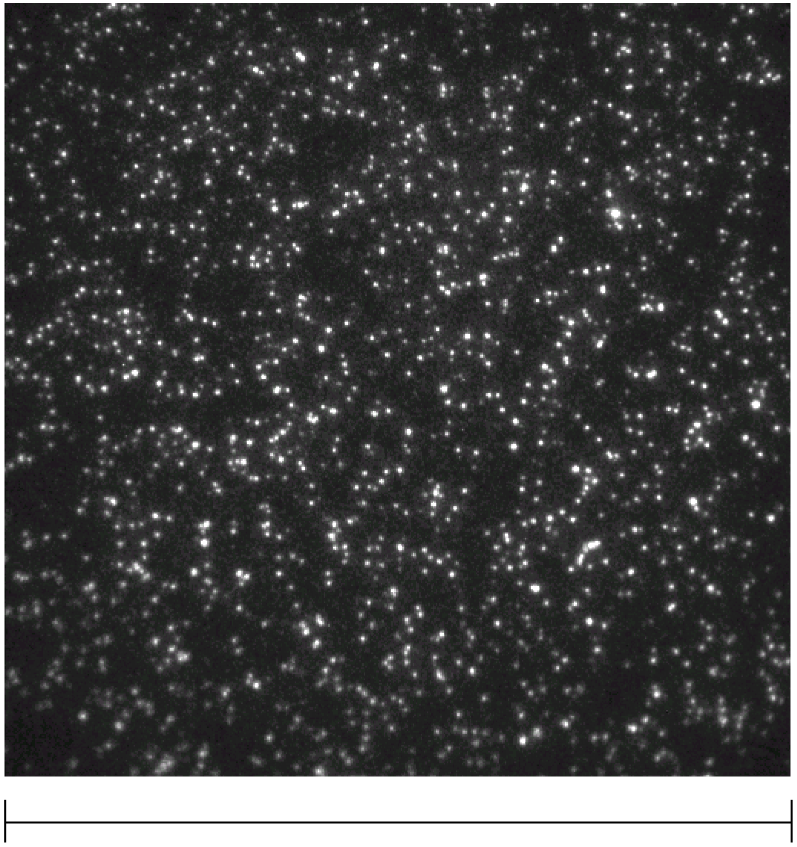
1 ug total RNA

10 ng

ChIP-Sequencing

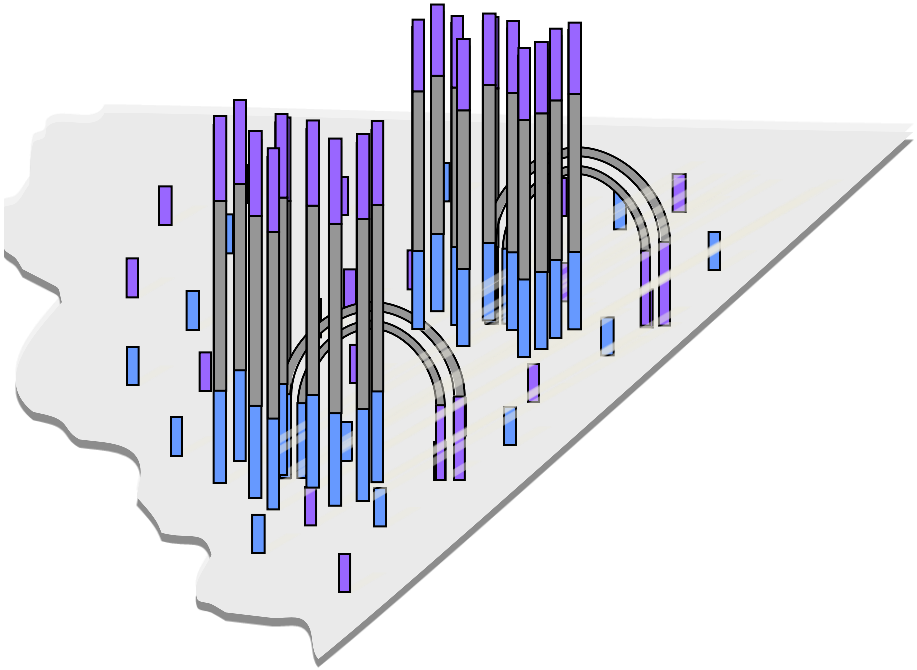
Other Apps

Step 2: Clonal Single Molecule Arrays



Random array of clusters

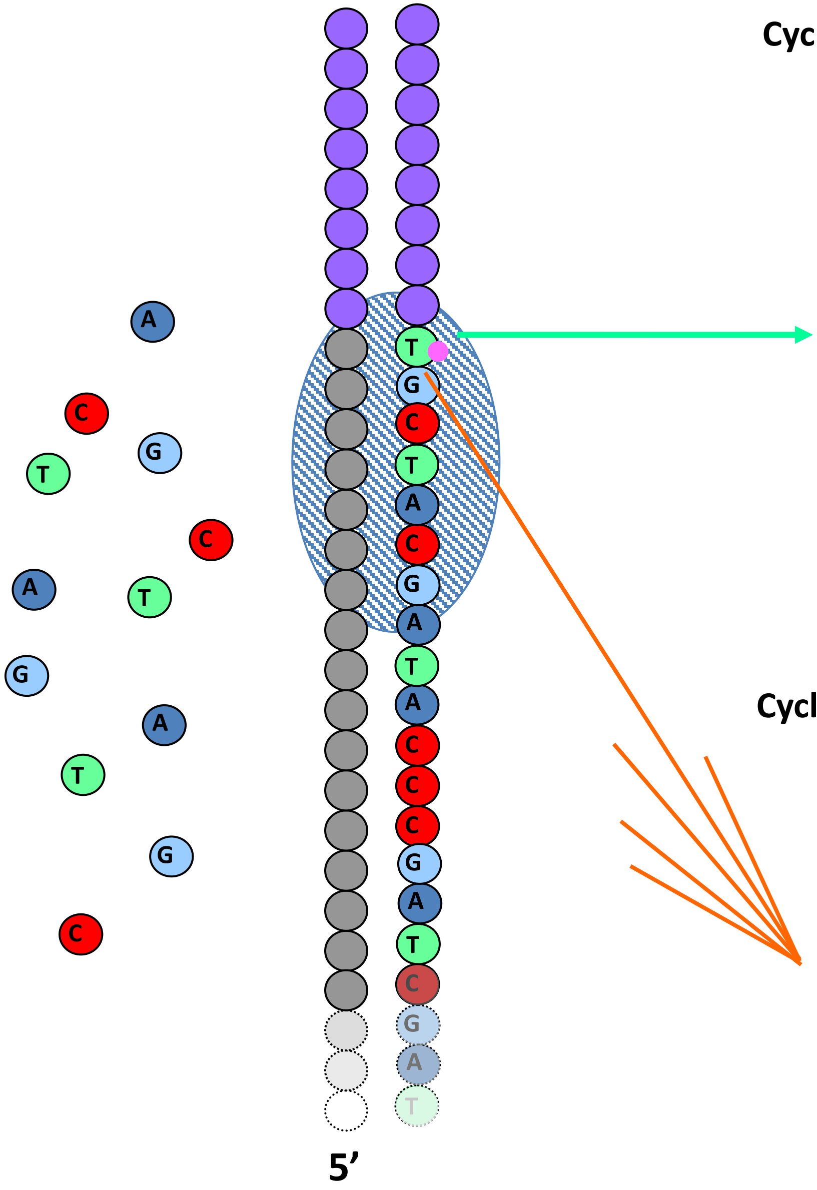
Attach single molecules to surface
Amplify to form clusters



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

1 cluster = 1 sequence

Step 3: Sequencing By Synthesis (SBS)



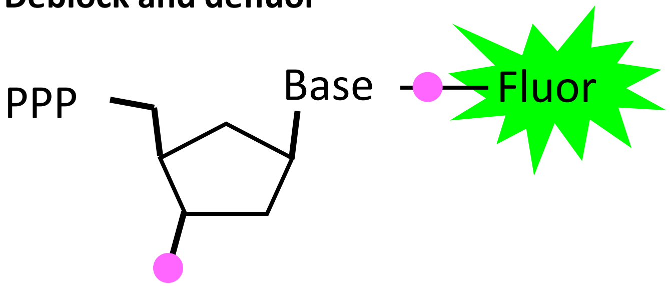
Cycle 1: **Add sequencing reagents**

First base incorporated

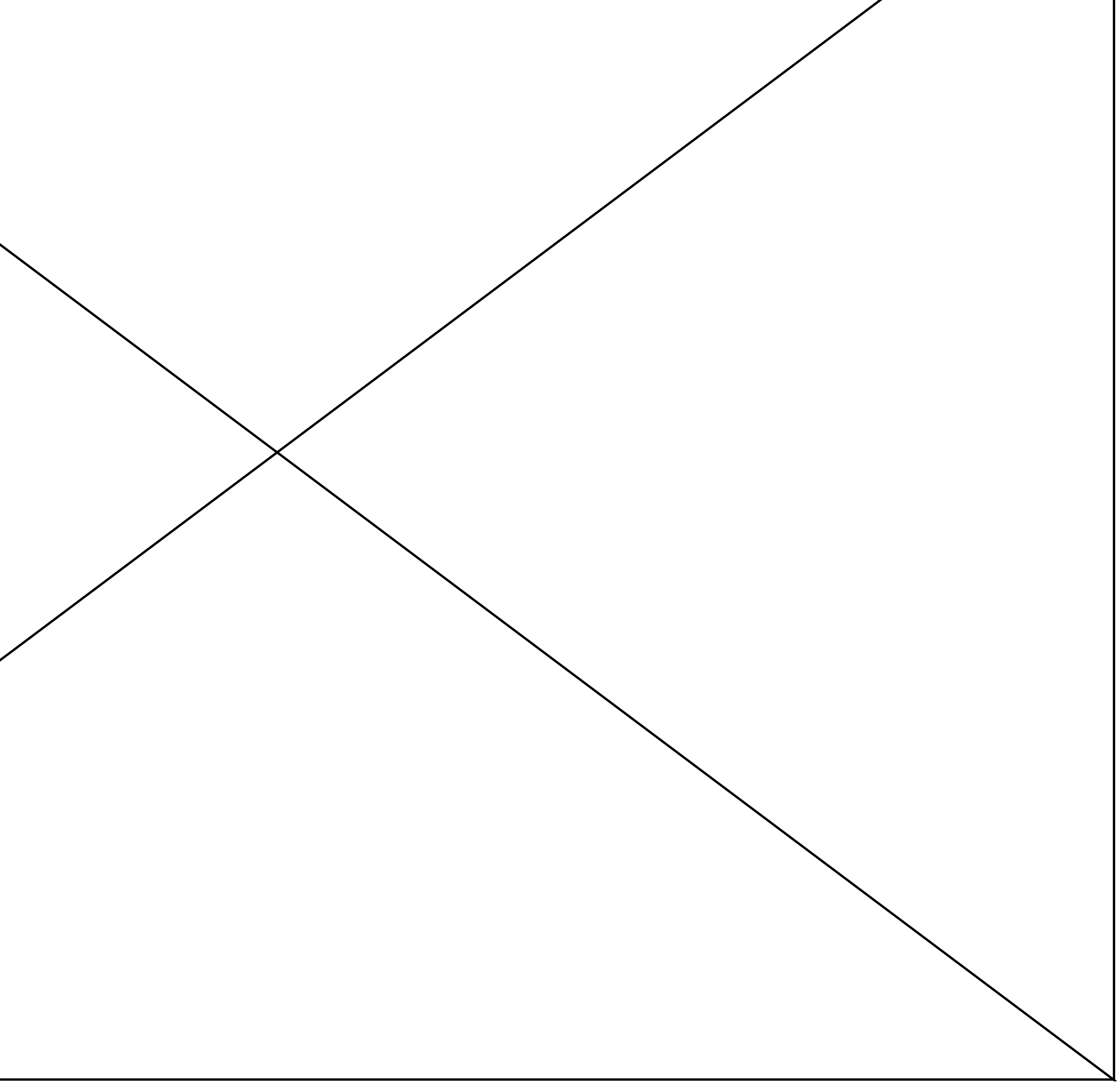
Remove unincorporated bases

Detect signal

Deblock and defluor

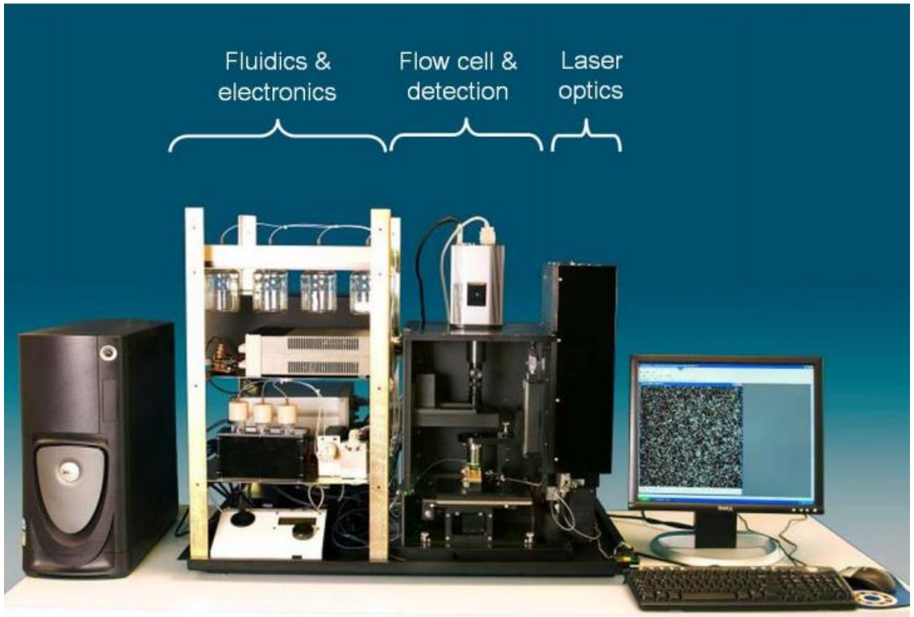


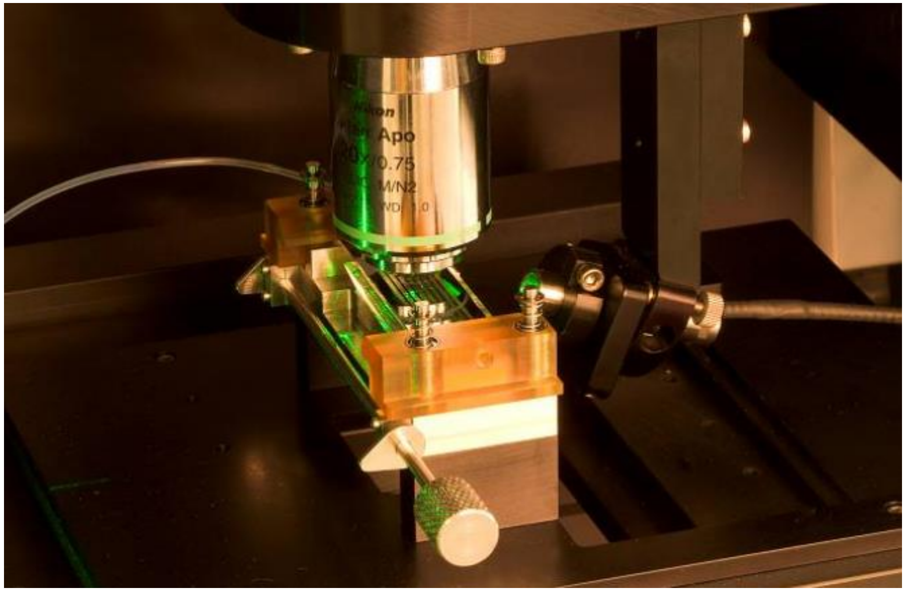
Cycle 2: Add sequencing reagents and repeat



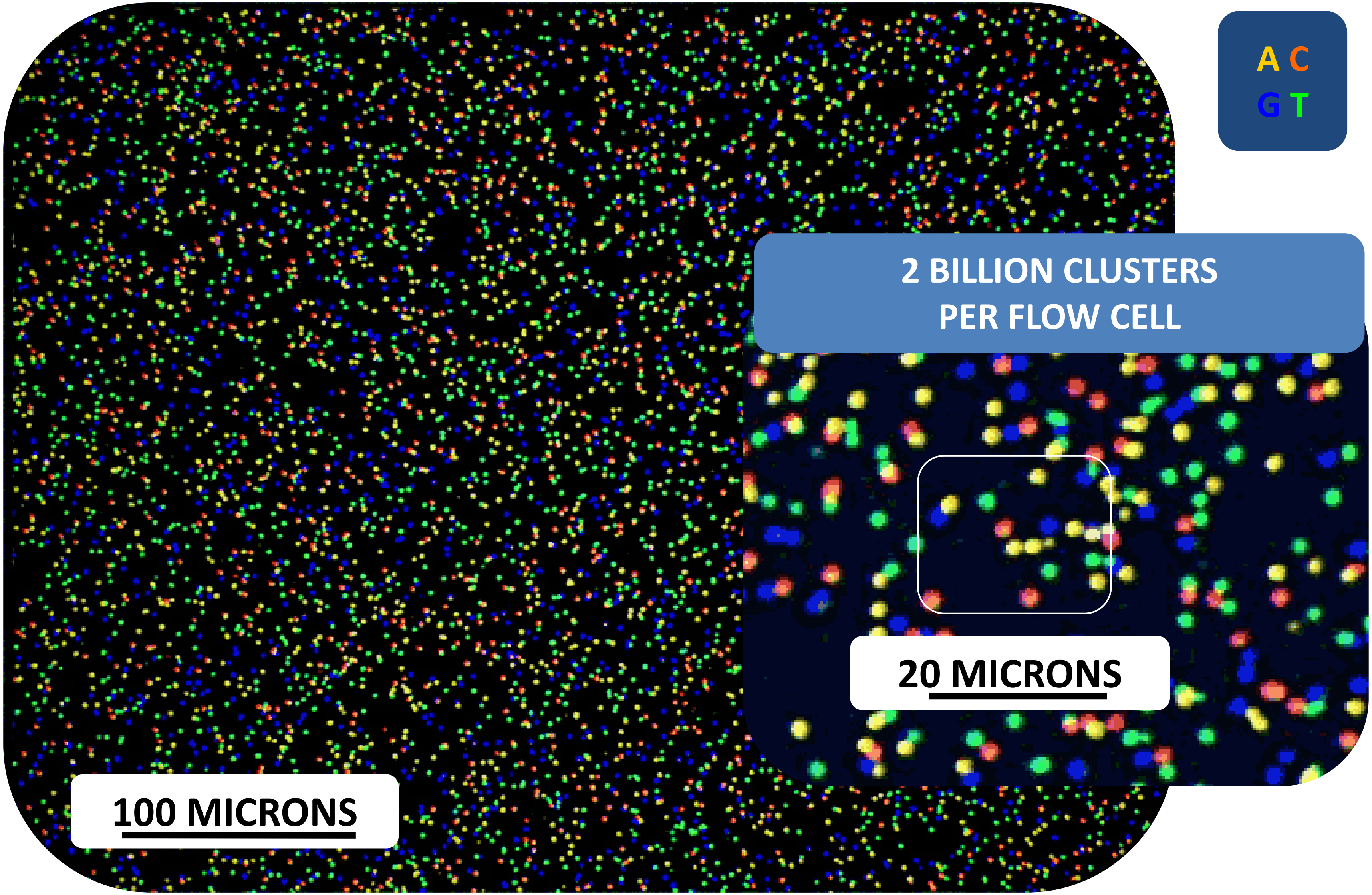
https://www.youtube.com/v/HMyCqWhwB8E

Under the hood:





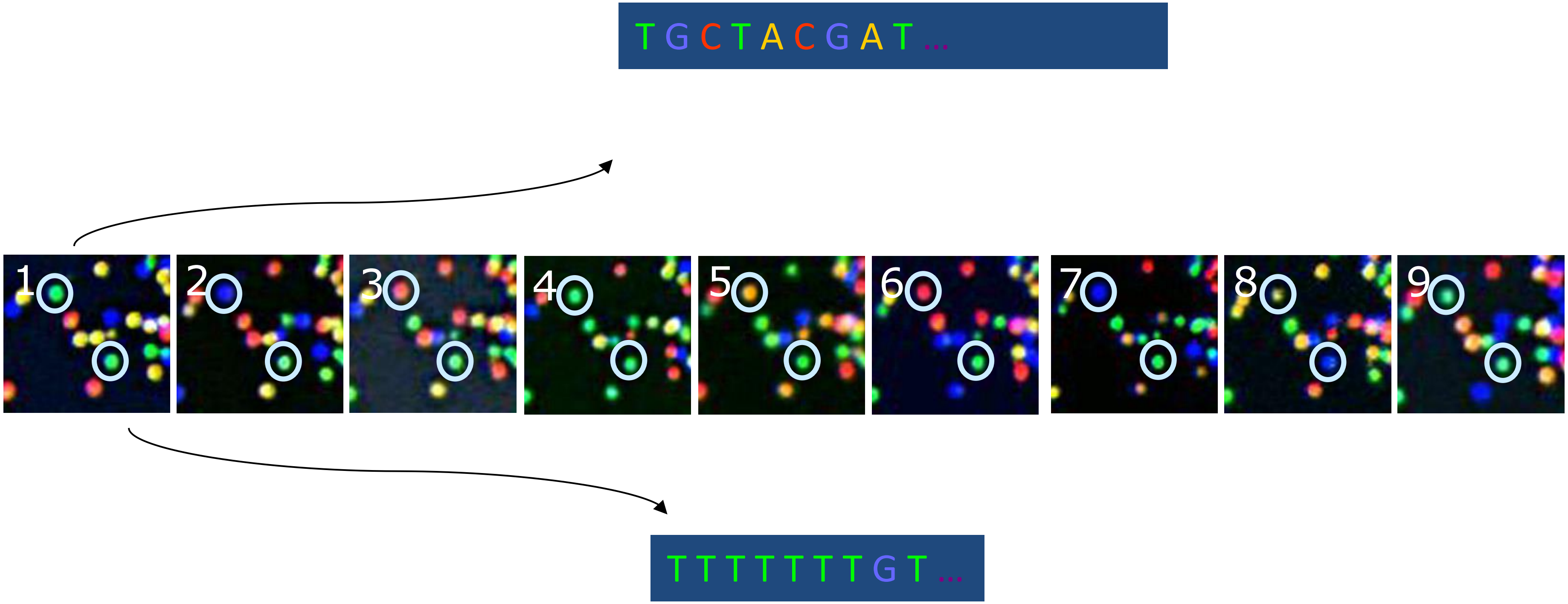
Illumina Sequencing : How it looks



2 BILLION CLUSTERS
PER FLOW CELL

A C
G T

Base calling from raw data



TGCTACGAT...

TTTTTTTGT...

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

Illumina platforms

Illumina HiSeq

- 500Gbase/flowcell
- 8 human genomes
- 6 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
- £650,000 incl 3 year servicing

Illumina NextSeq 500

- 90 - 120Gbase/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

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

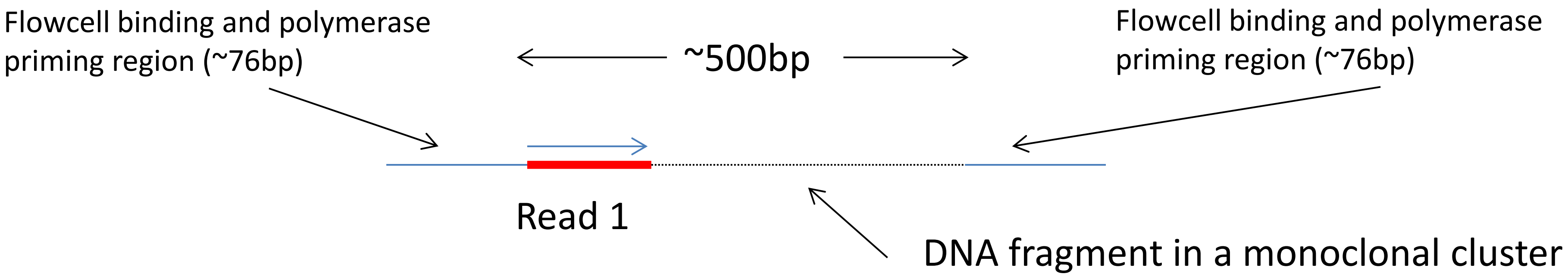
- 4 colour chemsitry

- £90,000 incl 3 year servicing

Types of Illumina reads

- Single-end
- Paired-end
- Barcode/index
- Mate-pair
- Long synthetic reads
- 10X Genomics Linked Reads

Single-end reads



Flowcell binding and polymerase priming region (~76bp)

~500bp

Flowcell binding and polymerase priming region (~76bp)

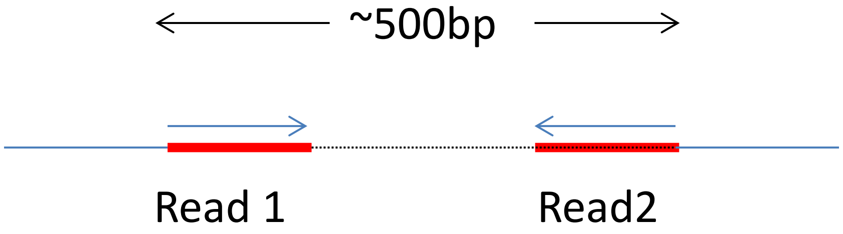
Read 1

DNA fragment in a monoclonal cluster

Read 1 ACTGATTCTTATTATCACTATTGGTAGCTGGTATTGGGTAT…..

- This would be a 16bp single-end read from a 500bp fragment
- Most common Illumina read lengths are 50, 75bp, 100, 125 or 150bp
- Usually cheapest but may not always be available for small projects
- Useful for counting applications (e.g. gene expression profiling in bacteria)

Paired-end reads

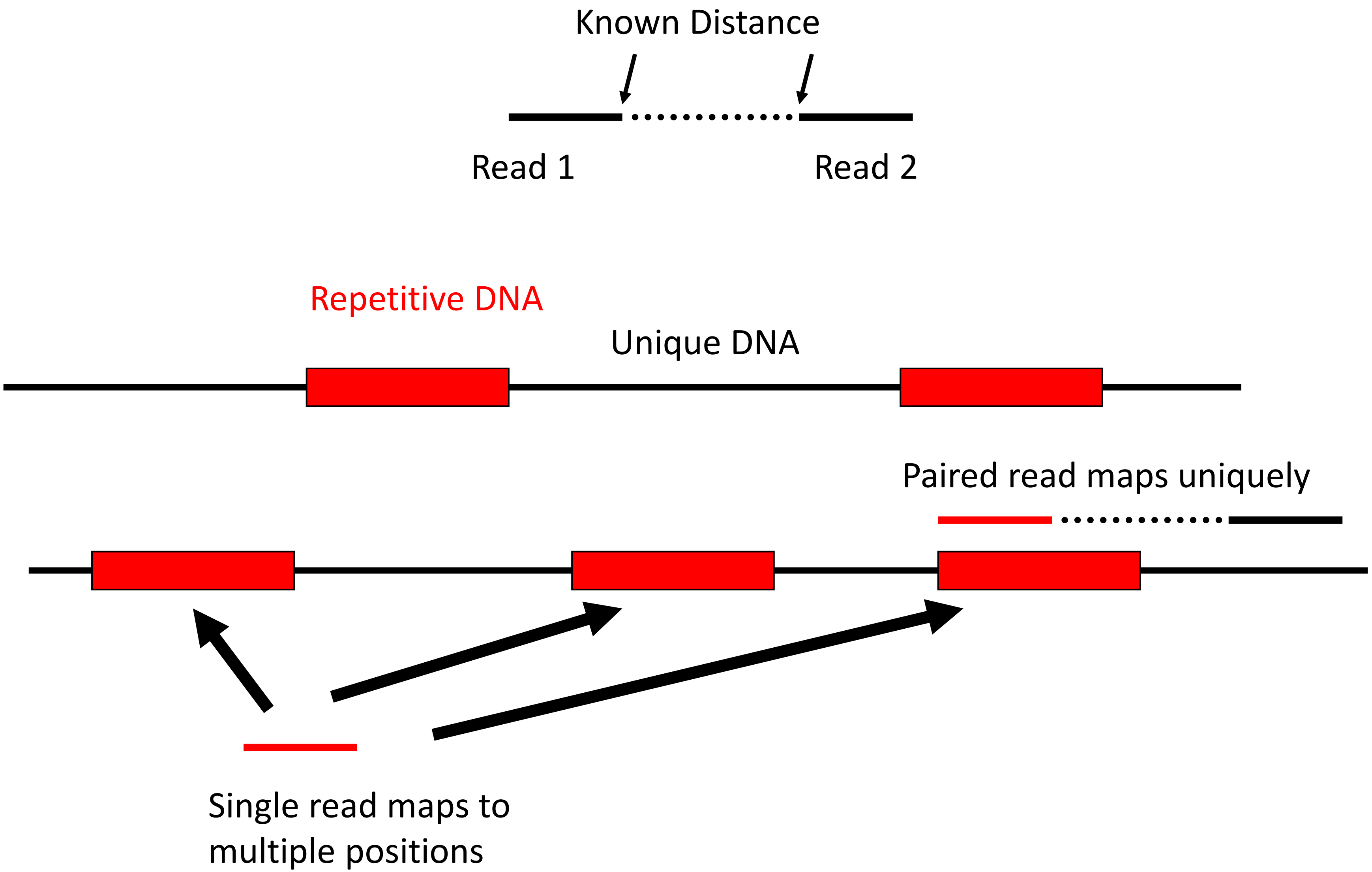


Read 1 ACTGATTCTTATTATCACTATTGGTAGCTGGTATTGGGTAT…..

Read 2 GCCTATCATCTGTATCGTCTATATGTGAGGTCGTAGCCCTA…..

- This would be a 16bp *paired-end* read from a 500bp fragment
- Most common Illumina read lengths are 50, 75bp, 100, 125 or 150bp
- Some facilities will mostly run paired-end reads
- Often a requirement for de-novo assembly or isoform quantification
- For some applications its desirable to have read 1 and read 2 overlap to increase accuracy (e.g. 16S amplicon sequencing)
  - This is achieved by careful design of the amplicon or size selection to ensure it is shorter than read 1 + read 2

Paired-end reads are important



Known Distance

Read 1

Read 2

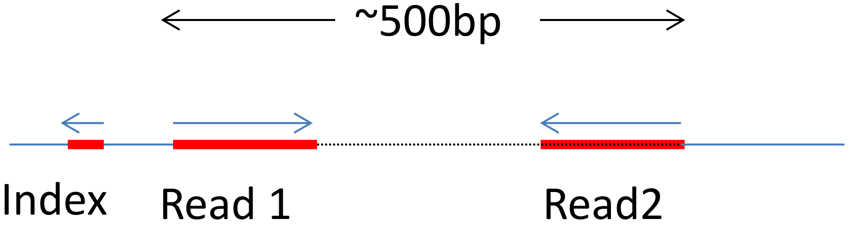
Repetitive DNA

Unique DNA

Paired read maps uniquely

Single read maps to
multiple positions

Barcodes/index reads

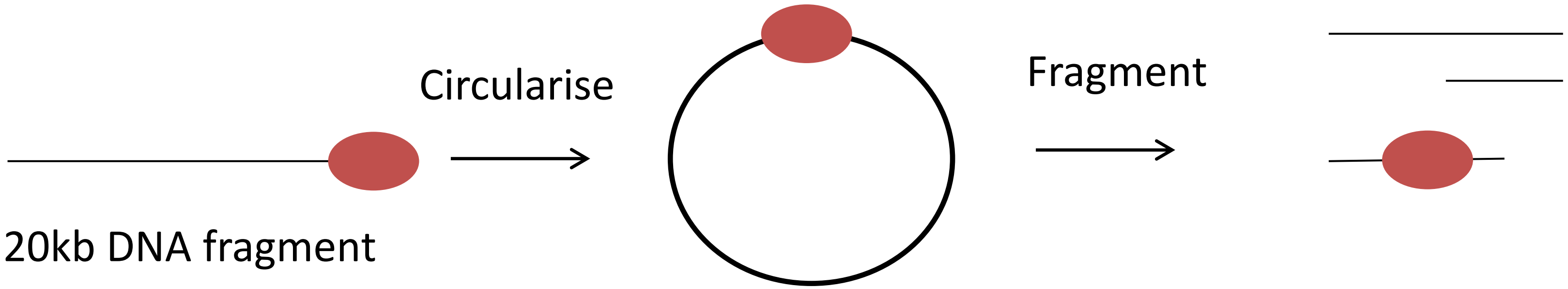


Read 1	ACTGATTCTTATTATCACTATTGGTAGCTGGTATTGGGTAT.....
Index read	ACTGTGTA
Read 2	GCCTATCATCTGTATCGTCTATATGTGAGGTCGTAGCCCTA.....

- This would be a 16bp *paired-end* read with an 8bp index from a 500bp fragment
- Achieved by adding one or more priming sites for the polymerase
- Enables multiplexing (mixing) of samples on the same physical space on the flowcell
- Virtually all libraries are indexed today even if they are
- Barcodes can also be introduced in-line as part of read 1 or read 2 (e.g. some RAD libraries, amplicons)

Mate pair libraries

- Maximum DNA fragment size for Illumina is ~800bp



Circularise

Fragment

20kb DNA fragment

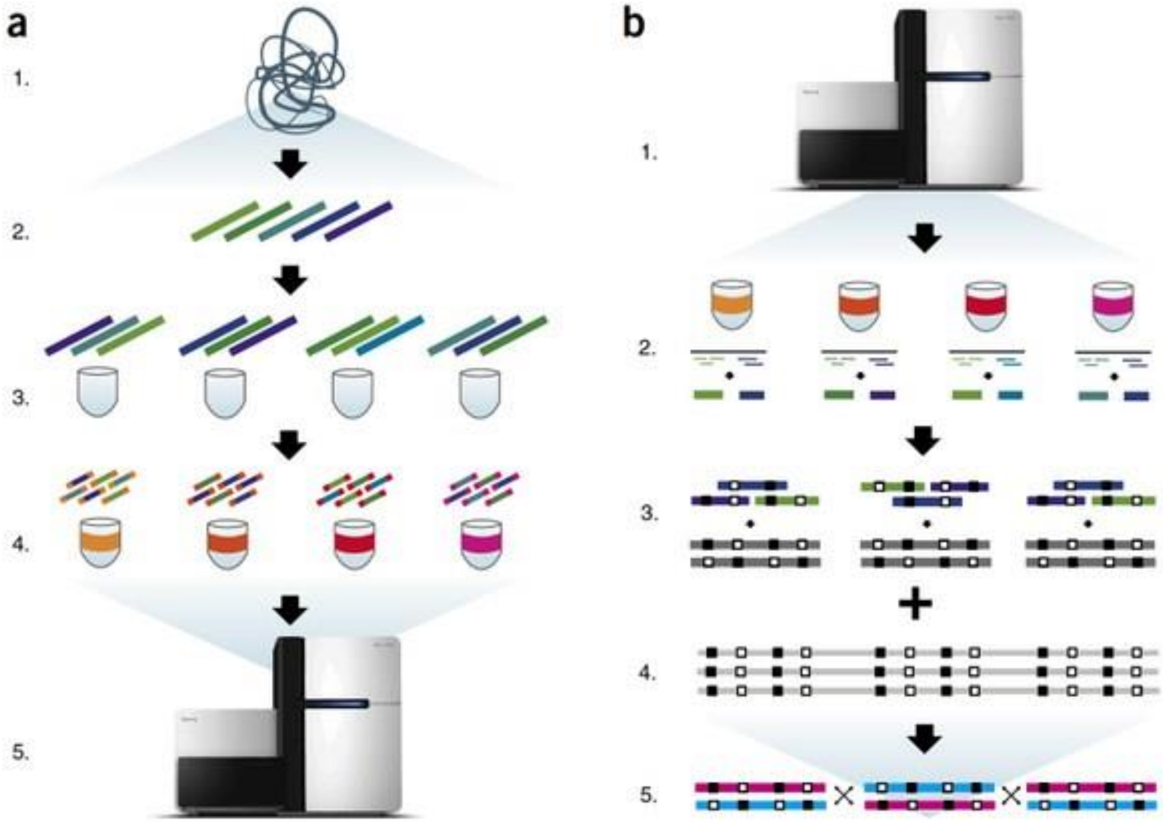
- Purify fragments containing biotin moiety using Streptavidin beads
- Create a standard Illumina library and sequence using paired-end reads
- Physical fragment size is 500bp
- Genomic distance between read 1 and read 2 is 19.5kb
- Valuable for de-novo assembly

● = Biotin moiety

Long synthetic read libraries

- Generate 2-10kb reads by partitioning DNA
- Dilute bulk DNA into wells
- ~3000 molecules/well
- Create individual barcoded libraries for the molecules in each well
- Assemble reads from each well separately
- Reduces complexity and enables assembly of long synthetic reads from standard paired-end reads

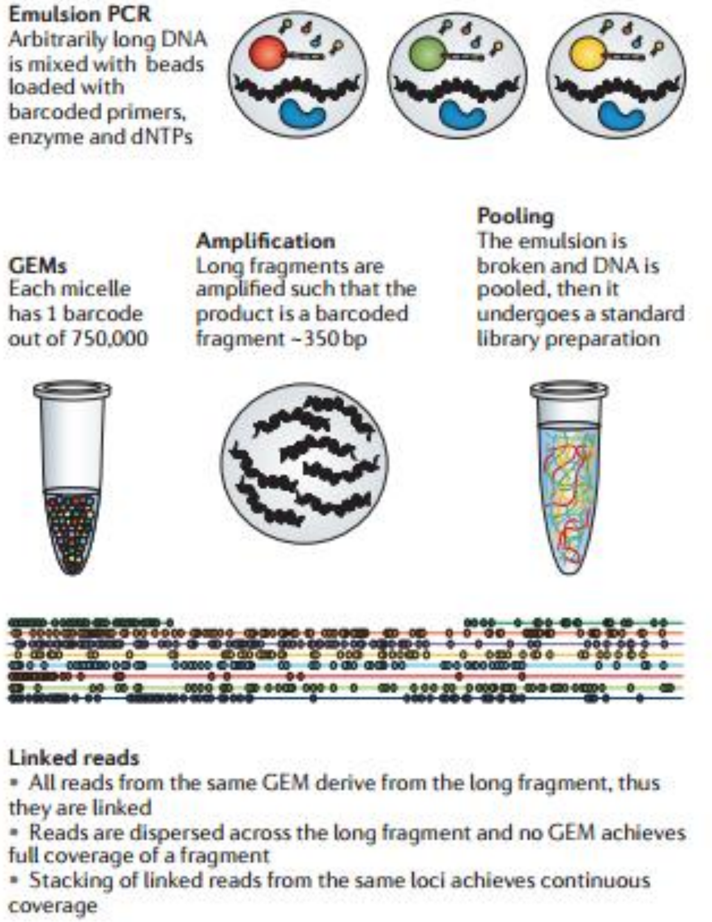
- Useful for denovo assembly, haplotyping and phasing



http://i2.wp.com/nextgenseek.com/wp-content/uploads/2014/06/Moleculo-TruSeq-Synthetic-longreadkit.jpg

10X Genomics

- A third-party system which provides additional capabilities
  - ‘Linked-reads’ to enable the formation of haplotypes and improve genome assemblies
  - Single-cell gene expression profiling
- £150k purchase price and approx £500 per library



Emulsion PCR
Arbitrarily long DNA is mixed with beads loaded with barcoded primers, enzyme and dNTPs

GEMs
Each micelle has 1 barcode out of 750,000

Amplification
Long fragments are amplified such that the product is a barcoded fragment ~350 bp

Pooling
The emulsion is broken and DNA is pooled, then it undergoes a standard library preparation

Linked reads
- All reads from the same GEM derive from the long fragment, thus they are linked
- Reads are dispersed across the long fragment and no GEM achieves full coverage of a fragment
- Stacking of linked reads from the same loci achieves continuous coverage

http://www.nature.com/nrg/journal/v17/n6/pdf/nrg.2016.49.pdf

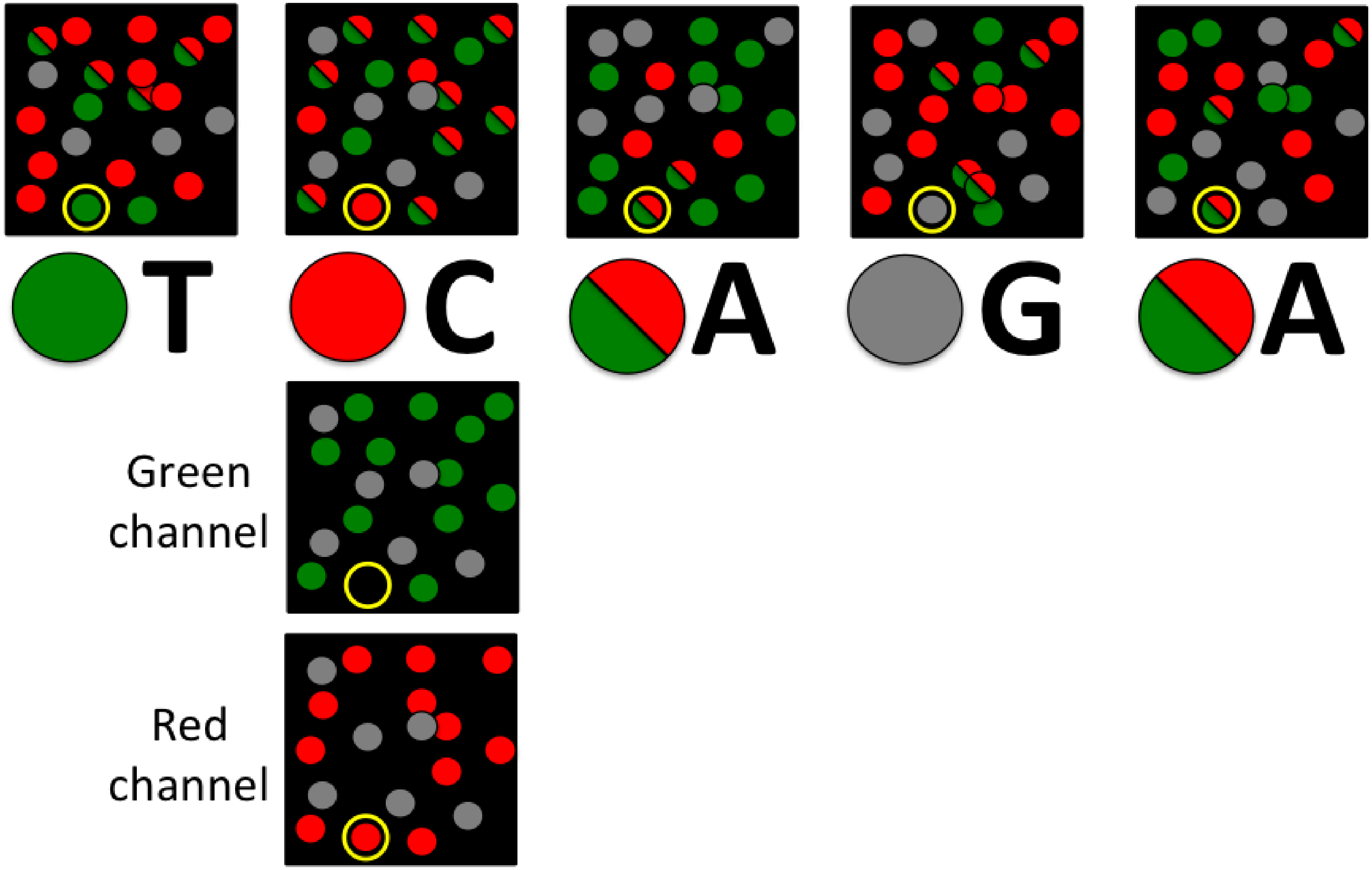
Illumina technological developments

- NextSeq
  - Utilises 2-colour instead of 4-colour chemistry to reduce sequencing time
- HiSeq 2500, NextSeq and MiSeq
  - Clusters formed randomly on the surface of the flowcell
- HiSeq 3000, 4000, X series
  - Clusters only form within nanowells
  - Patterned flowcells

2-colour chemistry

- Instead of using 4 different dyes for each nucleotide, use 2
- Label T as green and C as red
- Label A as green and red
- Label G with no dye
- Rely on cluster position to call G bases

2-colour chemistry

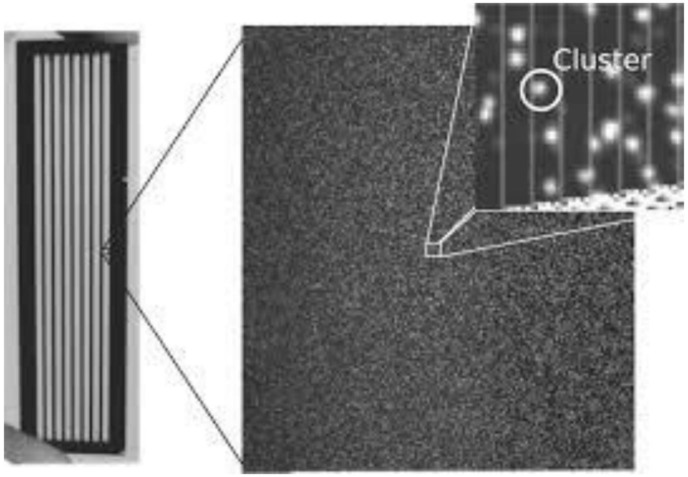


http://core-genomics.blogspot.cz/2014/01/nextseq-500s-new-chemistry-described.html

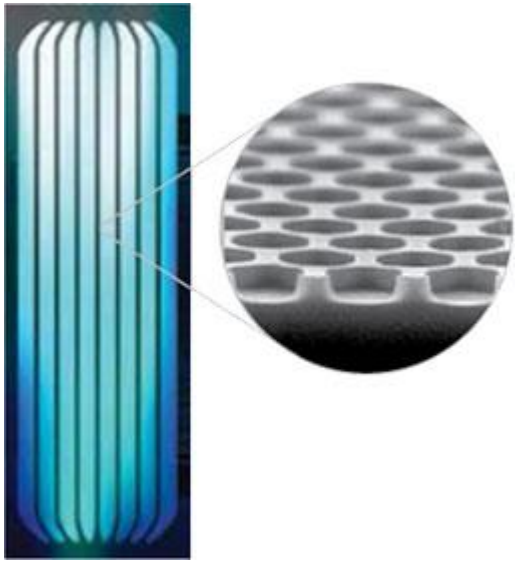
Advantages/disadvantages

- Advantages
  - Speed
  - Only two pictures need to be taken each cycle instead of four
- Disadvantages
  - Higher likelihood of errors
  - Difficult to calibrate guanine quality scores
  - With fragments shorter than read length, tendency is to call G with high quality scores
  - Note recommended for low-diversity samples (e.g. 16S amplicons)

Patterned flowcells

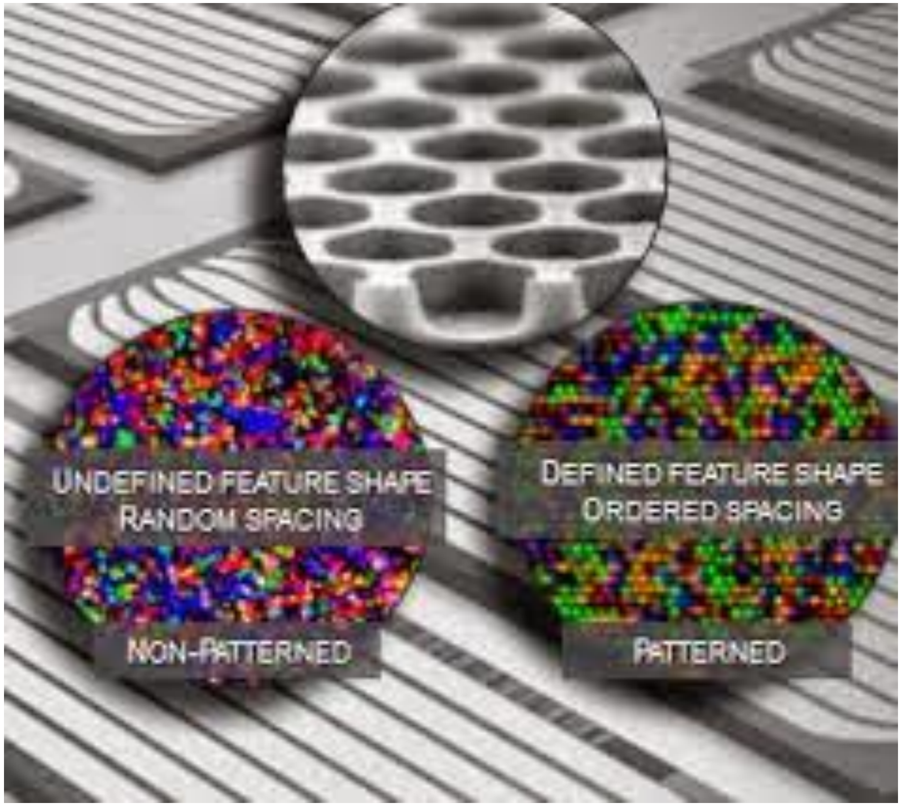


Randomly clustered flowcell
(2500)



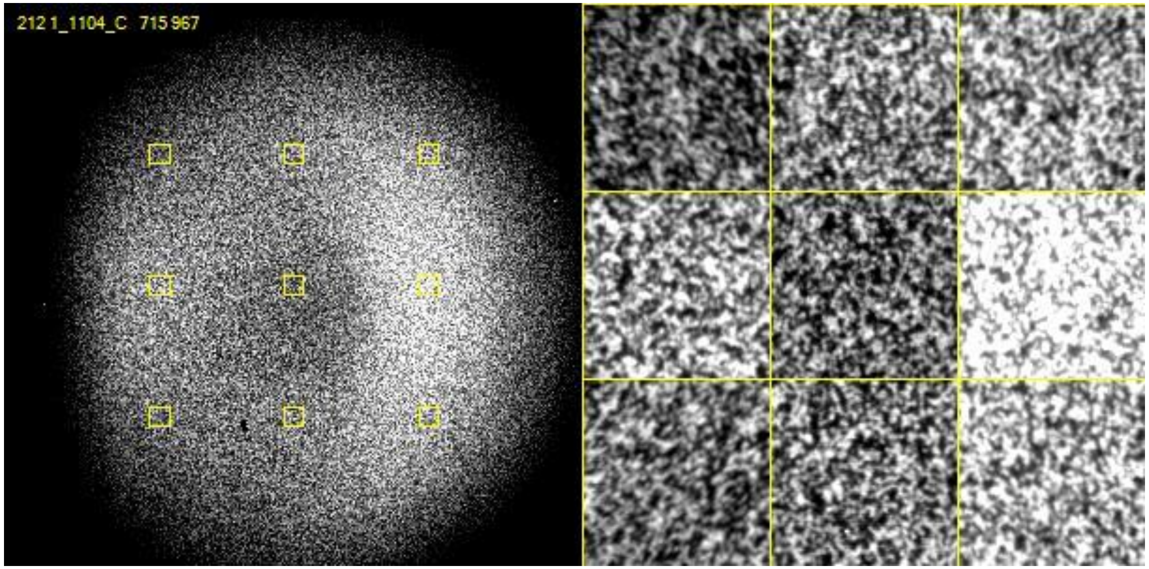
Patterned flowcell
(3000/4000)

Comparison


UNDEFINED FEATURE SHAPE
RANDOM SPACING
NON-PATTERNED
DEFINED FEATURE SHAPE
ORDERED SPACING
PATTERNED

Advantages

- Removes the need to detect cluster location during first 4 cycles of sequencing
- Lower sensitivity to over-clustering



Advantages

- Allows for exclusion amplification to reduce the number of polyclonal clusters
- Utilises an electric field to transport labelled dNTPs to wells faster than amplicons can diffuse between wells
- 1 sequence per well
- Whichever sequence starts replicating first within a well will rapidly out-compete other sequences
- Removes upper poisson-limit on random flowcell clustering (~37%)

http://www.google.com/patents/WO2013188582A1?cl=en

Disadvantages

- Possibility to obtain large number of duplicated sequences across wells
  - Caused by seeding of adjacent wells
  - Can be caused by under-clustering
  - Still important to load correct concentrations
- Limits on DNA fragment length

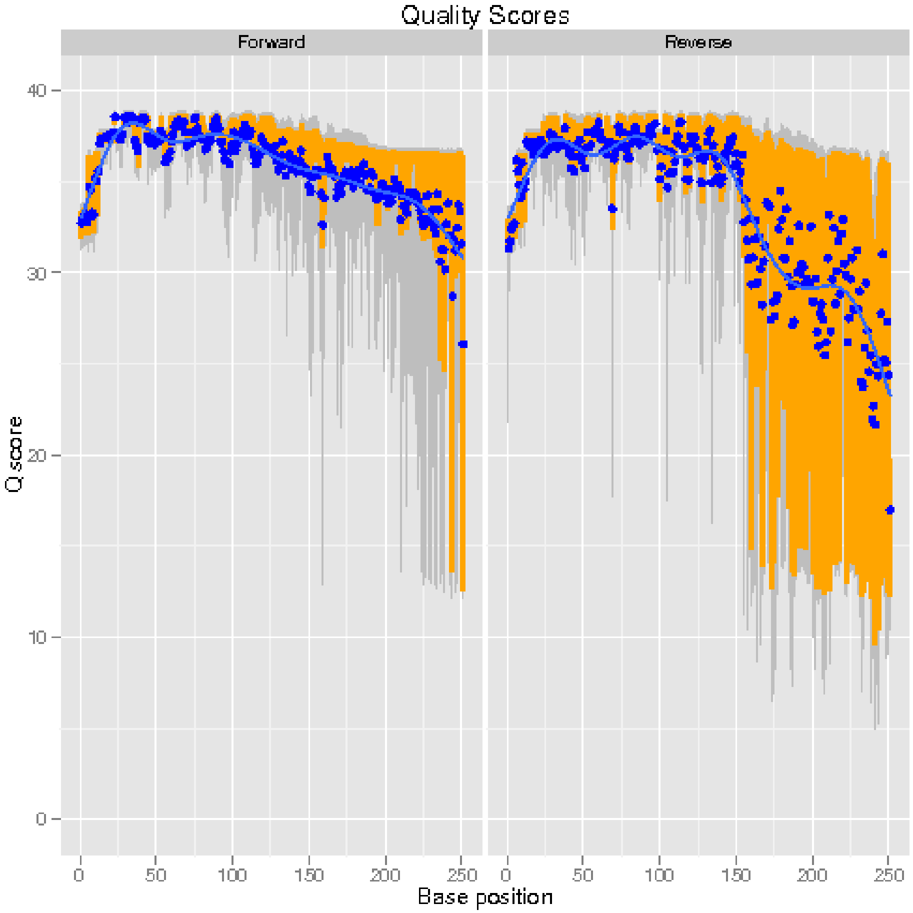
Potential issues with Illumina sequencing

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

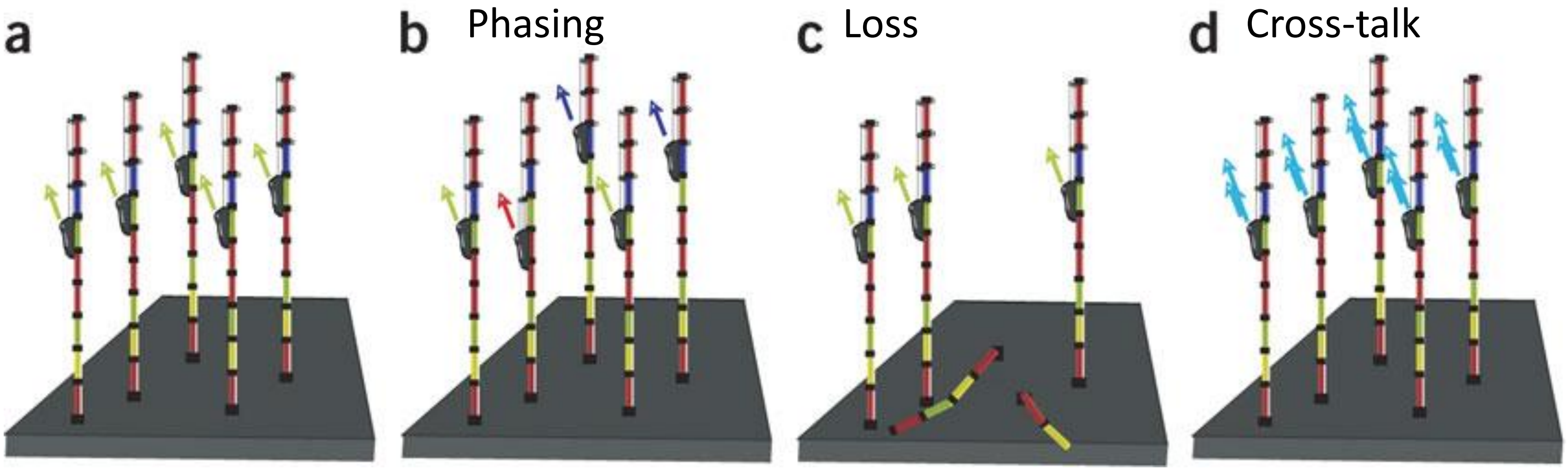
- Now mostly resolved
  - Low diversity sequences
    - 16S/amplicon sequences
    - Custom adaptors with barcodes at 5’ end
    - Now a much reduced problem thanks to software updates
  - GC/AT bias
    - GC clusters are smaller than AT
    - GC bias during amplification is still a problem

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)

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
    - Places limits on denovo assembly

Features of Illumina Sequencing

- 1 – 300 million sequences per run/lane (depending on platform and configuration)
- 36-300bp read lengths
- $0.01 - $0.1 per megabase
- Accuracy decreases along read length but ~0.1-0.3%

Other second-generation technologies

- 454
- Life Tech/Ion Torrent
- BGI-Seq 500
- Complete Genomics (primarily a service for human genomics)
- Qiagen Genereader (gene panels)

Third generation sequencers

Third generation sequencers

- Single-molecule DNA sequencing
- Some rely on detecting the incorporation of complementary bases to single-stranded DNA (PacBio)
- Some rely on changes in electrical current as DNA passes through a sensor (Oxford Nanopore)

Pacific Biosciences RS II and Sequel

Key features

- 10-15x more expensive than Illumina per base
- Sequences single molecules of DNA
- Read lengths 500-90,000bp
- Does not require amplification
- Can directly detect base modifications (DNA methylation)
- For many applications it requires high quality DNA and high amounts of DNA (>20ug)
  - Where amplification is involved prior to sequencing this requirement is relaxed somewhat

Illumina Read 100bp

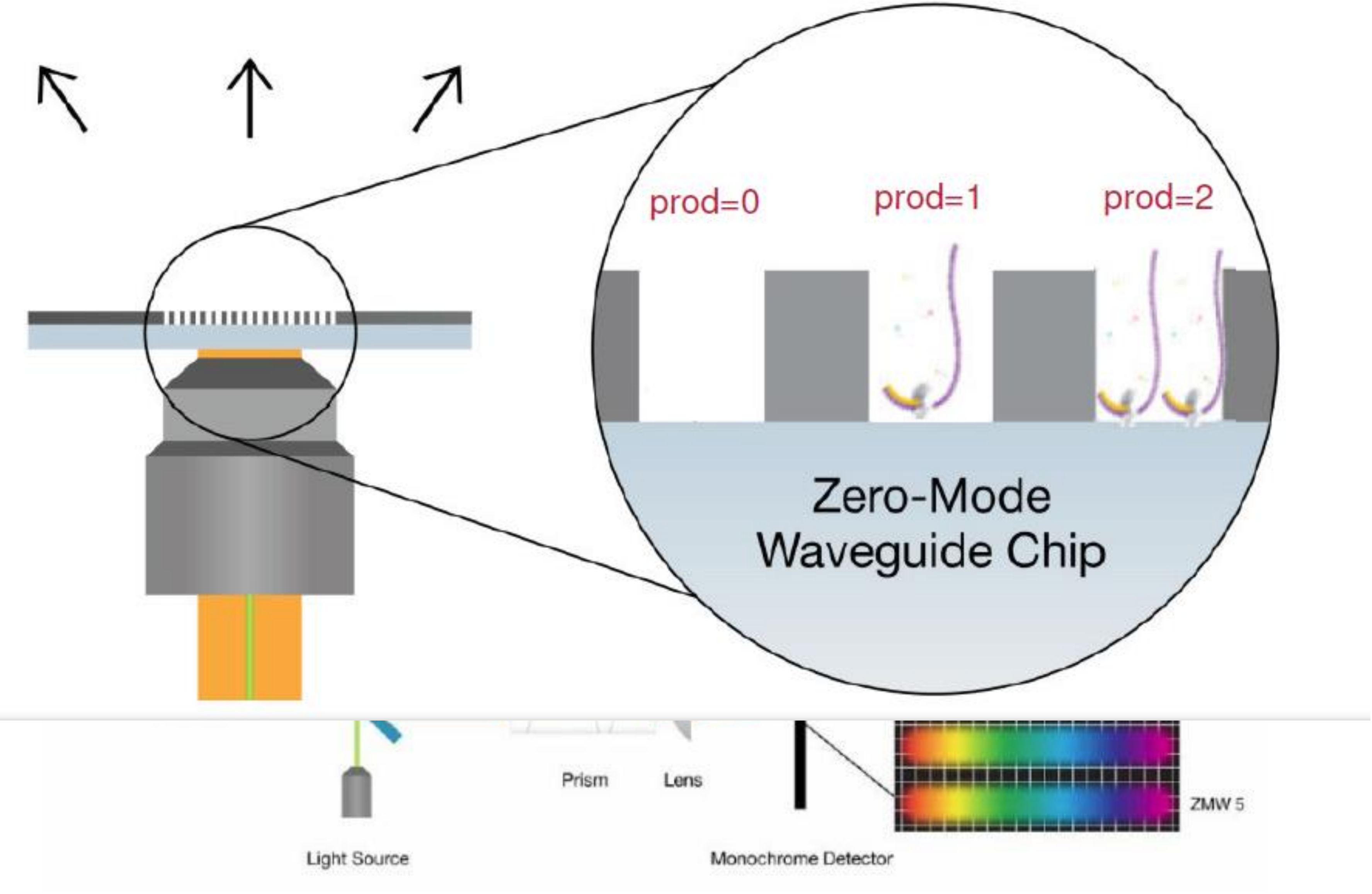
```
AAGAAACTGATCAGGGATAGCGGTCAGGTGTTTTACAACCACTAAACCCACAGTACCAATGATCCCATGCAATGAGAGTTGTTCCGTTGTGGGGAAAGTT
```

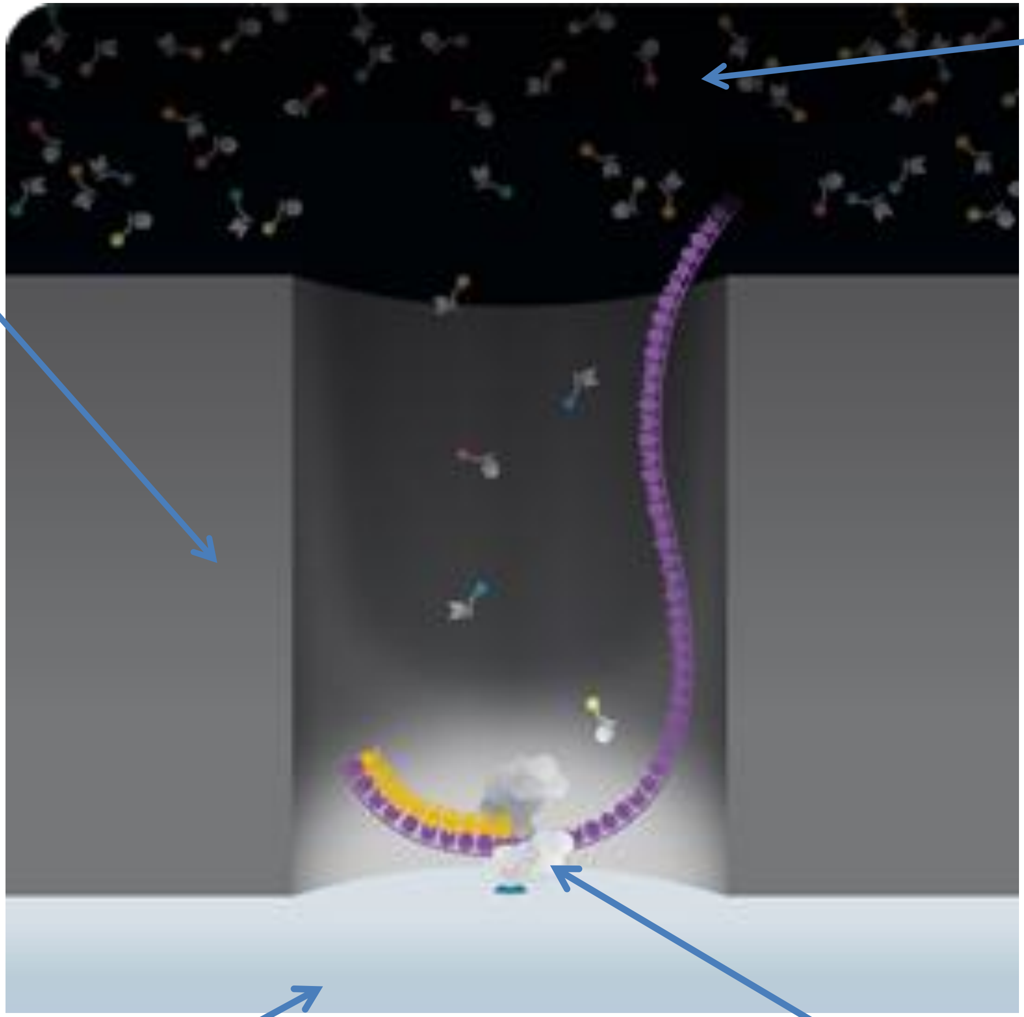
PacBio read 9700bp

```
CCTAGCAAACTCGGAAGATTTTTTCAGAGGGATCTAGAATATGATGAAAGATAGAAAATTACGACGCTTATCGGAAGTGACGAATACTTTTATATGAGGAGGGCTGTTTTTACAAAATCCGGTAGTAACTTGCTAACCAATTCCTAGGCAGGTCATTGGCAACAGTGGCATGCACCG
AGAAGGACGTTTGTAATGTCCGCTCCGGCACATAGCAGTCCTAGGGACAGTGGCGTACAGTCATAGATGGTCGTGGGAGGTGGTACAATTCTCTCATGCAAAAATATGTAAAACGGTAGCAACTGGAAATCATTCAACACCCGCACTATCGGAAGTTCACCAGCCAGCCGCAGCAC
GTTCCTGCATACGACGTGTCTGCGGCTCTACATATCTCCTATGAGCAACGTGTTAGCAGAGCCAAGCACAACTCTAATTTTAATACATAATGATGATAATATAATATTAAAAATTTCCTGTGTACTAATTTTACTATGGTTTCTGATAAGAATCATTGCAAAGATCAACAACTTGTATTA
CATTGACAGTTAAGCAGTTAATTTTATCACCTCTAAAATATATCAGCATCTAGCATGCAACTATCAAAATGGAGAGTTTTATGACTAAAAACCATGGGAAAGAAGACTTAAAGATTTATCGCACTTGCTCAAATGCTGCATTGATACATATTTTGACCCTGAATTATTTCGCTTGAATTT
GAATCAATTCCTCCAAACCGCAAGAACAGTAACATTTATTATTCAAAAAAACAAAAAACAAGATTATAGGATATGACATTTGGTATAAAATAATGTTATTGAAAAATGGAAAAATGATCATTAATGGCTTGGGCTAAAAATTCTCGGCAATAGCGATAGAAAAACAAGGCGATTTAG
AAATGTATAGCGAGGCAAAGGCTACTCTTATTTCATCTTACATTGAAGAAAATGACATTGAGTTTATTACAAATGAAAGTATGTTAAACATTGGTATAAAAAGTTAGTCAGACTTGCACAAAGAAATTACCTTCATATTTAACTGAATCATCTATTATTAAATCAGAAAGACGATGGGT
CGCTAATACGCTAAAAGATTACGAATTATTACATGCTTAGCTATAATTATGGCAGAATGTATACTGCTGTAACTCTCTTGGCATACAAAACAATCCAATGGGTGACGATGTGATTTCGCCAACATCATTCGACTCTTTATTTGATGAAGCCAGGAGAATAACTTATTTAAATTAAAAGA
TTACTCCATAAGCAAATTGTCATTTAGCATGATACAATATGACAATAAAATAATTCCTGAAGATATTAAAGAGCGTCTAAAACTGGTAGATAAGCCTAAAAATATCACTTCGACAGAAGAGTTAGTTGACTATACAGCCAAGCTTGCAGAAACGACTTTTTTAAAGGACGGTTATCAC
ATTCAAACATTAATTTTTTATGATAAACAATTCCATCCAATTGATTTAATCAATACAACATTTGAAGATCAAGCAGATAAATTATTTTTTGGCGTTATGCAGCTGACAGAGCCAAAATAACAAATGCCTATGGCTTCATTTGGATATCAGAGCTATGGCTCAGAAAAGCAAGCATCTACT
CCAATAAACCAATACATACAATGCCAATTATAGATGAAAGACTTCAGGTAATTGGAATTGATTCAAATAATAATCAAAAATGTATTCATGGAAAATAGTTAGAGAAAACGAAGAAAAAACCGACTTTAGAAATATCAACAGCAGACTCAAAATGACGAAAACCATATTTCATGCGTT
CAGTCTTAAAAGCAATTGGCGGTGATGTAACACTATGAACATTGAGTCATAGAACTTCCATTATTCTCCTGAAGATAATAATGCGCCAATAAACAATACTCAGCTTTACAATATACTAACTAACCGCAGAACGTTATTTCATACAACGTTTTGGGGCATATCACAAAACGATTACTCCA
TAACAGGGACAGCAGGCCACTCAATATCAGGTGCAGTTGATGTATCACACGGTTCAGCAACACCCGATACTTTTCCAGGCTTCCAGCAACGAGGTTTCTTCTTCGTTGCAATTTCCAGATCTGCAGCATCCTGAAGCGGCGCAATATGCTCACTGGCTACCTGCATCAGGCTTTTTTTG
TTTCTTCCGCCTCCGGATCCGGAACAGTTTTTCTGCTTCCGTATCCTTCACCCAGGCTGTGCGTTCCACTTCTGATATTCCCTCCGGCGATAACCAGGTAAAATTTTCCGGTAACGGACCGAGTTCAGAATAAATAACGCGTCGCCGGAAGCCACGTCATAGACGGTTTTACCCCGATG
GTCTTCAACGAGATGCACGATGCCATCACTGTTGAAAACAGCCACAAGCCAGCCGGAATATCTGGCGGTGCAATATCGGTACTGTTTGCAGGCAGACCGGTATGAGGCGGAATATATGCGTCACTTCACCAATAAATTCATTAGTTCCGGCCAGCAGATTATAATTTTTATGGTCGT
GGTTGTTCACTCATTCTGAATGCCATTATGCAAGCCTCACATATAGTTAAATGCATGTTTTTGACGGTGTTTTCCGCGTTACCGCAGCGTTAACGGTGATGGTGTGTCCGTGTGAACCAATACTGAAAGAATGGGCATGAGCACCGATAACAACCGGATGCTGGTGCGCACCAATACA
ACTGTATGCGCATGTGCACCGGCACTCACGGCTGTACCGGACAATGAGTGACTGTGGCTGCCCTGACTGTCCGTTTCGATAAATAAGCAATACCTGTGTGCTGGTTCCTTTAACTGTGGATAAACTTCCTGTAATGGTTGCTGTTCATACTGACTCCAGCAGAACTGTTCATCCTTAA
ACCACTTGTGTGGGCATGAGCACCCGCGGCCCTGTTGAACCGCTCAGACTGTGAGCATGAGCCCCGTGTTATTCGTCGATTTGGTGCGTAATCGAAACTGCTGTTGTTTTCGTCCCGTAATCAAACGACGATGTGGTTTTCGTCCAAATCCGTACCGGATGCACTGGCACTGTGGGT
GTGCGACTTAATTCCATCCTGTTCCTGAGACAATACAGCACGACCGCTGGCGGGTTTCCCCTTGATTGTCCAGCCTCGCATATCAGGAAGCACCCGATGGATACGCGACAGCAAGTTTTGGGTAGGCTGATTTGTCAAACGCCTGCCCCTGCATCAGGACGTAGCCAGACGGAACGA
TATCTGATGGCCACGGGATCGGCGCACCTGCCGGAAAGGCCGAATTCTCACCGGCCCAAGGTATTCAAGACATCTGCAACGGAATTTTTTGCCAGAATATCCTGCCAACCTGAGTCAGTTCAGTCAGGCTGGCGGCATCATTTTCCGCAAAATACGGTAATTTATTTTTCGCCGTGGA
AAGCCTGCCAGCGCCGTCAGTGTCGCATTCTTCGGTTGTTTACCCGCAAGCGCGTTAGTCATGGTGGTAGCAAAATCTGGATCATTCCCGAGCGCTGCGGCCAGTTCATTCAGCGTATTCAGTGCGTCAGGTGACGCGTCGATAACATCTGCAATCGCGGCCAGTAAAAAGCGGTGT
TCGCAATCTGGGTATTGTTTGTTCCCCTGAGCGCGGTTGGTGCTGTTGGCGTTCCGGTCAGTGCCGGACTGTCCAGTGGCTTTTTTTCTGTTCGTTTCATCCATTACACCTTAACCGCCTTTGGCGTTGCAGCAAGCGTTTCAGACGTGCGTTGGTTGCACTGCTGAGCTGCACTATCCC
TTTCTCGTTGTGTCGCGCATCCTCAAGCGCGACAGCTGAAGCTATATCTTCTGCACGTTTTGCCGAATTTTTTGCACGTATTGCCGCCGTTCTGCCGCACTTTTGCTCTGCGATGCTGATACCGCACTTCCCGCAGCTCTGTCGCCTTCGTGGATGCCGTTGACGCACTCCCCGCCGCCG
CTGTTTTTGCGTCTGCCGCGGCAGAGGCGCTCCGTTCCGCTGCTGTTTCAGATGACTGGCATTCGTCTCGGACGTTTTTGCCGCCCTGGCAGAATTTTCTGCCGCCGTTGCCGAGGAAGCTGCACGACCGGCACTTGATGATGCGTTCGTTTCTGATGATTTTGCTGCCTCTTTGAGGC
CACCGCATCTCGTGCTGAAGTGGCGGCCTCTGACGCTTTCGTGGCCGCGGTGGAGGCAGACGTGGCGGCTGATTGTTGTGACGCTGCAGCATTCGTTTCTGACGTTTTCGCCGCACCGGCACTGGTGGCCGCCGCGTTTTTTGAGGACTCTGCGGCTGCGGCACTTTTTTCGCGCTT
CAGTGGCCTTTGCTGATGCGCTTCTGCGCCGGAGGACGCTTCCTGAGCTGACGATGCAGCCTGTCCGGCGGACGTGCTGGCGGCGTGCTGAGTCAGTTGCATCAGTCACAAGGGCGCGACCTGAGCAGCTGATGCACTGGCATCGCCGGCTGATTTCTTCGCGTCTGCCGTACTCT
GTGCCACCACGGACGCGTTACGCGCCACCTCTTCCACATCAGTTCAAGACGACGCAGCACCTCCGGCCGGGCATCATCCTCCGTCATGGCACAGAGAAAATCATTCAGCGTCCCGGGTTGTGAATCTTCATACACGGTGATGGTCCCGGGCGTGCGATGGTGGAAAACCGTCAACC
TGCAGGATGACACTGTACTGACCGTACTCACATCCATGCTGTAACGCCCGGCTTCATCCGGATTCTCTGAGCCCCGTGTTCACCACCACCGTGGTGCTGTTACGTCTGGCTTTCAGCTGAATGGTGGCAGTTCTGTACCGGTTTTCCTGTGCCGTCTTTCAGGACTCCTGAAAATCTTTA
CTGCCATATTCACCCCACAAAAAGCCGGTTCCGGCGGGCTGTCATAACACTGTGTTACCTGGCTAATCAGAATTTATAACCGACCCAACGATGAATCCGTCAGTACGCCAGTCGCCACTGCCGGAGCCTTCATAAGCAATATCAACAACGACGGACGCTGCCGGATTAATCTGTATA
CCTGCACTCCACGCCACTGAGGTATGCCGCATTGCACTTTCGTCCCTGGCAGTGGTCGTCTCTTTCATATACCGGGGAGTGATTTCCGTCTTACGGTAATCCATTGTACTGCCGGACCACGACTGTGAGCCACTCCGGCCATGGCGTACGCACTGACCTGCTTACTGATTTGTAAAACC
GGTCCGGCCATCACGCTCACATAACGTCACGCAGGCTCTCATAGTGAAACGTATCCTCCCCGGTCATCACTGTGCTGCTCTTTTTCGACGCGGCGAACCCCAGGGAAGCCATCACCCCACACTGTCCGTCAGCTCATAACGGTACTTCACGTTAATCCCTTTCAGATGACTCACACC
GGTATCCCCGCCGACAACGACGGCAATGTACCGGTTTCACTTGAAAATAGCCCACGTAAACGTACATGTCCACCTTCCGCACGGGCCGGAGTGACTGTCACCGCAAGTGCGGCAAAGACAGCAACGGCAATACACACATTACGCATCGTTCACCTCTCACTGTTTTATAATAAAACG
CCCGTTCCCGGACGAACCTCTGTAACACACTCAGACCACGCTGATGCCCAGCGCCTGTTTCTTAATCACATAACCTGCACATCGCTGGCAAACGTATACGGCGGAATATCTGCCGAATGCCGTGTGGACGTAAGCGTGAACGTCAGGATCACGTTTCCCCGACCCGCTGGCATGTCA
ACAATACGGGAGAACCTGTACCGCTCGTTCGCCGCGCCATCATAAATCACCGCACGTTCATCCAGTACTTTCAGATAACACATCGAATACGTTGTCCTGCGCTGACAGTACGCTTACTTCCGCGAAACGTCAGCGAAGCACCACTATCTGGCGATCAAAGGATGGTCATCGGTCACG
GTGACAGTAGGGTACTGACGGCCAGTCACACTGCTTTCACGCTGGCGCGGAAAAGCCGCGCTCGCCGCCTTTACAATGTCCCGACGATTTTTTCCGCCCTCAGCGTACCGTTTATCGTACAGTTTTCAGCTATCGTCACATTACTGAGCGTCCCGAGTTCGCATTCACACTGCACTGAT
ATCGCATTTTTAGCGGTCAGCTTTCCGTCCGGTGTCAGGGAAAAGGCGGAGGATTGCCGCCGTGGTAATGGTGGGGGCCGTCAGGCGCTTCAGGAACACGTCGTTCATGAATATCTGGTTGCCCTGCGCCACAAACATCGGCGTTTCATTCCCGTTTGCCGGGTCAATAAATGCG
ATACGATTGGCGGCAACCAGAACTGGCTCAGTTTGCCTTCCTCCGTGTCCTCCATGCTGAGGCCAATACCCGCGACATAATGTTTGCCGTCTTTGGTCTGCTCAATTTTGACAGCCCACATGGCATTCCACTTATCACTGGCATCCTTCCACTCTTTCGAAAACTCCTCCAGTCTGCTGG
CGTTATCCTCCGTCAGCTCGACTTTTCCCAGCAGCTCTTTGCCGAGATGGGATTCGGTTATTTGCTTTGAAAAATCCAGGTAACTTCCGCATCATCGCTCGCCCGACGACGGCCTCCACGAATGCGATTTTGCCAACGGTGTTCAACTGCGGATATAAAAGTAATAATCATGGCCCGGT
TTGATATTGATACTGGCGGCTATCCAGTACAGCGCCGTACCAAGATAACGCGTGCTGGTTTCAACCTGTCTGATATCCGCAATCTGCTTTTCCGAGAACAGAACTCAAACTGTACCGTCGGGTCATAAACGGCAAGATGCGGCGTGGCGGTTATCTGAAAATAGCCCGGCGTCAGCT
CAATCCTCGACGGTGCTGCCGGTGCGGCAATCGGAACGATACCGACGCGGATCGCCCTGCTGCCCCACGCATTTACCGCCCGGACTGTCAGCCTGTAGTTCCCCAGCGCCAGTTGCGTGAAGCGGTATGTGGTTTCCGTCGTCCGGGCCGTGCTGACCAGCCGCTCACTGCGTCGT
CCGCTGTTACGGTCAGACGGAGCAGGAACTCACGCCTTCACCACCTTCGGTGTGTCCATCGCGCCAGCACCTGATATTCCGCTGTCTGCAGTGACTTCTGCGGTCAGGTGCTGCACGCTGGCGGCGTGACACCATTCACCGTGCCACTCTGTTCGCCGTCAAAGTGCGCCCGTTATCC
ACGATGGCTCTTTTCCGGCACATGCTGCACGGCGGTGATGGCATACGTGCCGTCGTCGTTCTCACGGATACTCACGCAGCGGAACAGTCGCTGGCGCAGCGTCGGCAGCTTCAGCTCCCATACGCTGTATTCAGCAACACCGTCAGGAACACGGCTCACTTTTACCTTCACGCCGTC
GGTGACGGACTGAACCTCCACGCTGACCGGATTGCCACTTCCGTCAACAGGCTTATCAGCGCGGTACGGAGGATGGCAGCGTGATTTCACGGTCGAGCGTCAGCGTCCGGGTCTGGCTGTTCACCGCCAGCACACGACCACCGGTGCTGATACCGGCATAGTCATCATCGAGATTT
CAATAACATCGCCCGGTACATGGCGAAGCCTTCTGCGCCGACGCTGAAATCCACGGTCTGCGTTTCCAGCAGTTCTGTTTTAATCAGCCACAGCCCGGCGCGGTGTGCCTGCCCCCGGCTGGTACAGCCAAAGGCATCATCTTCGTAACATTACGACGTAACGGGCAATGGCCTGCG
TATCTTCAACAAGCTCTGTCGCCGTCTCCCAGCCGTTGTTCGGGTCAATCAGTTCACCTCAACGGCATTTATGGGCGGTCTTCAGGGCGCTGAGCTGTAGCGGAACGGCGCGCCATCATCCGGCATCACACATTACTGCGGTTATAGGTCACGTCTTATCCGACGGTCGGTCCTGCAC
GAACGTCAGCGTCTGCCGTTCATACCGGCATACAGCGCATCGCCGAGCAGAAATCGCTGAGCACTACGCCTTACGCTGTGTGTCAGGTACGCATTACAGGTGATGCGCGCTCCGTGCGCCAAAGCGTCGGCACTGACTGGTCGCAGTACTGGCCGATGACATACAGCGCCCATTTA
TCCACATCCGCCGCACCAAGACGTTTCCCCATGCCGTAGCGCGGATGGGTCAGCATATCCCACAGACACCAGGCCATGTTGTTGCTGTATGCCGGTTTAAACGTTCCGTCCCAGATACGCGCTGTATTGCCGCGTCTGCGGGTTATAGTTCGACGGCACCTGCAGAATACGCCCGCG
CAGATGATAATTACGGCTCACCTGCTGGCTGCCGAACTGCTCCGAGTCCCCTGCACGCCGACCAGTGCCGTGTTCGGGTAGCACTGTTTCACATCGATGATTTCAGTGTATGACGACAGAGCGTTTTGTTCTGCAGCTGGTCTGTGGTGCTGTCCGGCGTCATCCTGCGCATCGGATA
TTAAACGGGCGCGGCGGCAGGTTACCCATCACCACCGAGGCAGATACTGCGAGGTGGTTTTGCCTTATGGTGATGTCTTTTTCCGTCACCCAGCACCGTTACGTTGTATCTGAACCAGCAGGCGACTTCCGACGGATTCCTGTCACCCTTTGAGGTGGTTTCCACAGTGCTGTACACC
GAAGGTAAAGCGCAGACGGTCGATGTTTGCAGACGTAATGGTGCGGGTGATCGGCGTGTCATATTTCACTTCCGTACCAGCACCGTCTCGGAGCCGGAGGATTCAAATCCTCCGGCGGAGTCTGCTCCTGCTCACCGCCCGGAACACCACGTGACACCGGATATGTTGGTATTCCC
CTCAGTGTCCAGCACCGGCGTACTGTTCAGCAGCACGCTTTTTAAGCCATCCACCGGACCTTCATCGGCCTTCGCTGATGGCATCGATCACACTCAGCAACTGCGTGGACTTCAGGTTGTCCTTCGCTTCGCGCGGGGTATGCCTTACTGCTTCCTTTACCCATTCTCACGCTCCATAAA
TGACAAACGCCGCAGGCGGTTTCACATAAAACATTTTGCATCAGCGACAAATCACCCAACCTGACCACGTCCCTTCGTCTGCGTGCTGATCTCCTGAGAAACACGCGTGACCACGCGCATTTCCGTACAGAACAGGCAGAAATTGCCCTGGGCAACCATGTTATCCAGTGAGGAGA
AATAGGTGTTCTGCTTACGTTATCCGTTGTCTGTATACGGGAGTTCTGGCTTTCGGTGCCAGCATCTGCGCACACCACCGAGCACCATACTGGCACGAGAGAAAACAGGATGCCGGTCATACCACCGGCCCAATGGCTGCCCCATGCTGCAAGGGTGGCCCGGCGGTAAAGATGAT
CCGGCAATGGCGGCAGCCCCCAGGACATCTGGAATACGCACTGACTTGGCCCGGGCGACTCTGGGAACAATATGAATTACAGCGCATCAGGCAGAGTCTCATGTAACTGCGCCGTTAACCCGGACGTGCTGACGTCCGCGCAATCCGTACCTGATACCAGCCGTCGCAGTTTCTGA
CGAAAACGCCGGGAGCTGTGTGGCCAGTGCCGGATGGCTTCAGCCCC
```

https://www.youtube.com/watch?v=NHCJ8PtYCFc

SMRT Cell


prod=0
prod=1
prod=2
Zero-Mode Waveguide Chip
Prism
Lens
ZMW 5
Light Source
Monochrome Detector



Zero mode waveguide

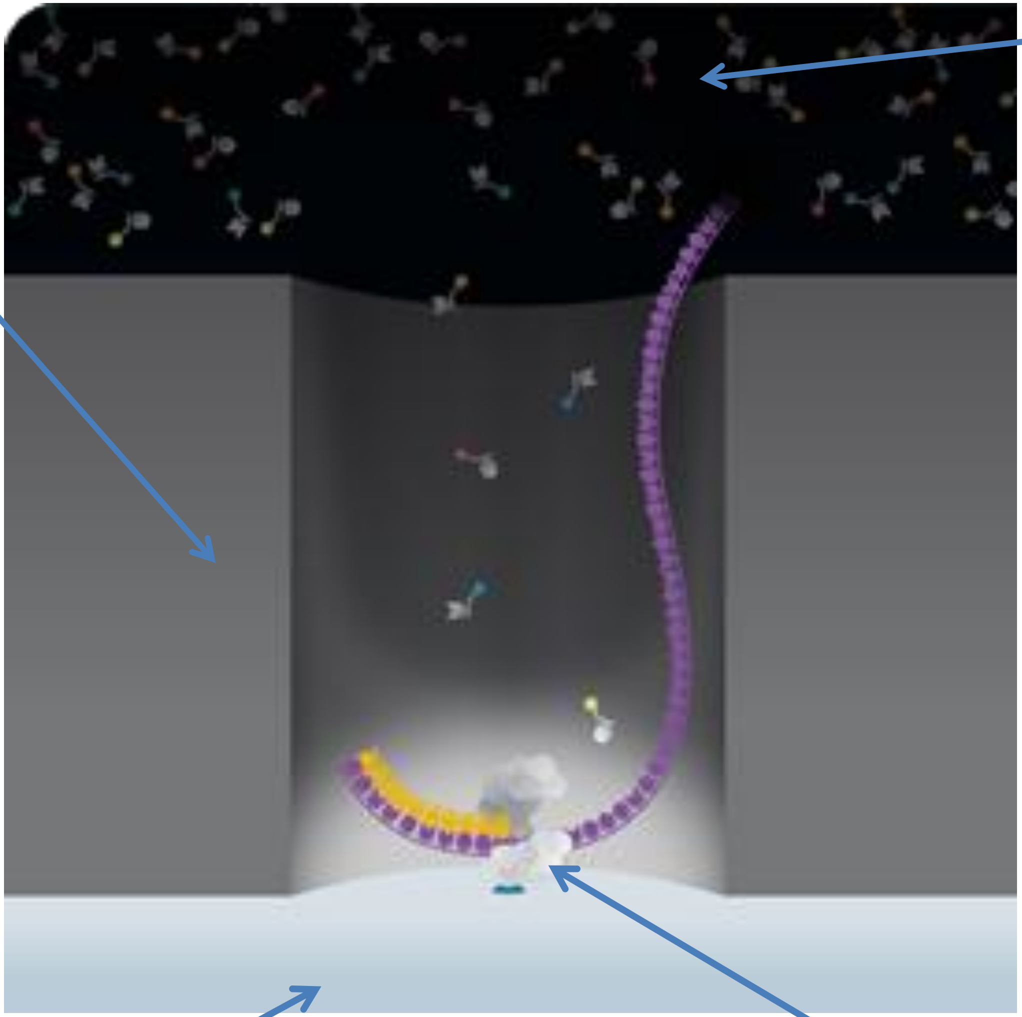
Free nucleotides

Immobilised DNA polymerase

Laser and detector

Reducing noise in the Zero Mode Waveguide (ZMW)

- Sequencing takes place in the ZMW
- Sequencing can only take place if one and only one DNA/polymerase complex is present in each ZMW
- Each ZMW is just 70nm wide
- Wavelength of laser light used to illuminate ZMW ~500nm
- Therefore light incident on ZMW will act as an *evanescent wave* and only penetrate the first ~30nm
- This reduces the amount of noise from fluorescence of non-incorporated fluorophores



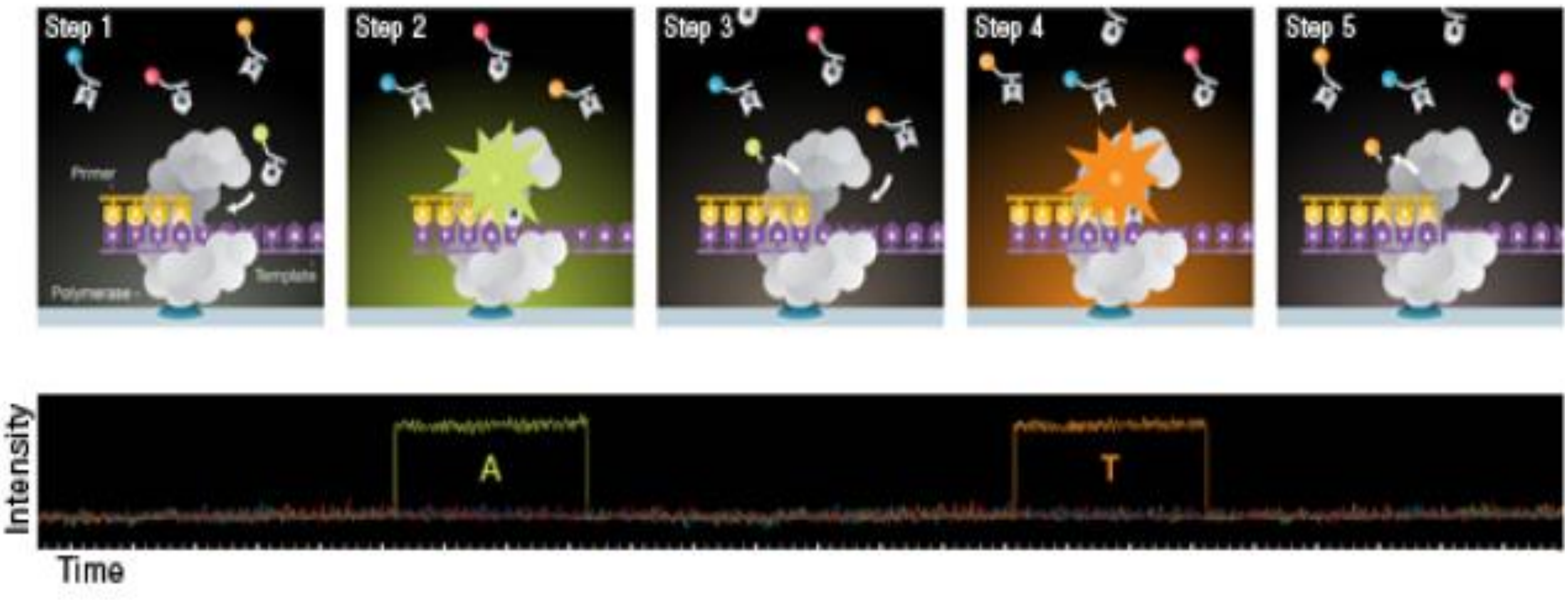
Zero mode waveguide

Free nucleotides

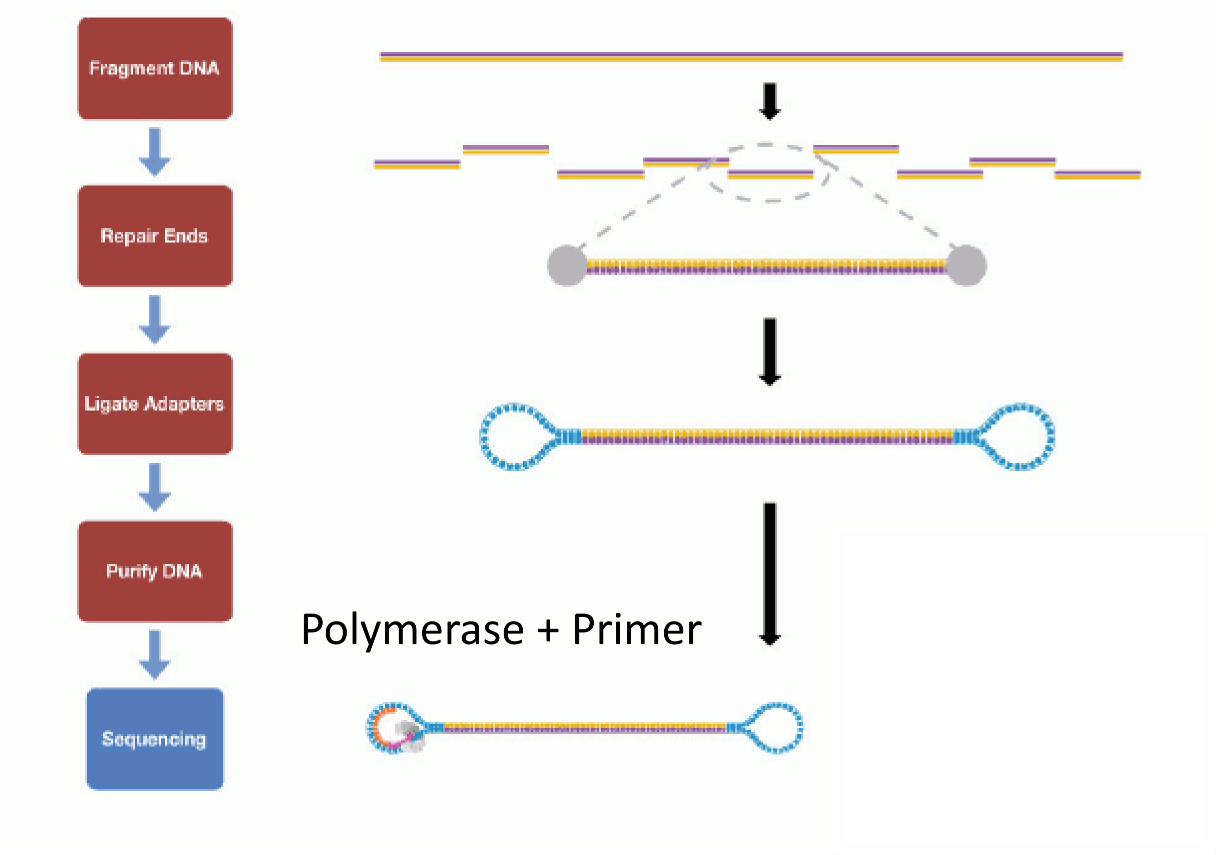
Immobilised DNA polymerase

Laser and detector

Observing a single polymerase



Library preparation steps



Fragment DNA

Repair Ends

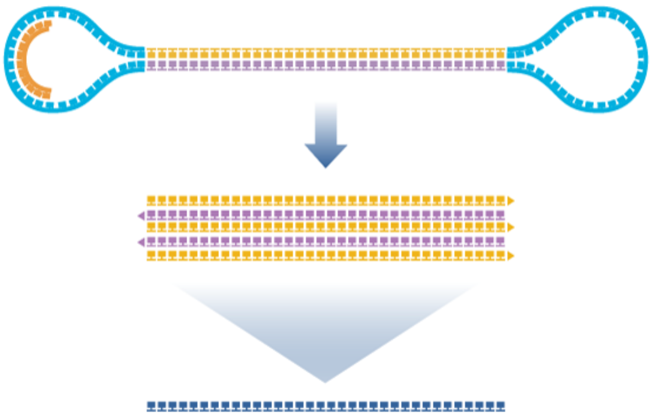
Ligate Adapters

Purify DNA

Sequencing

Polymerase + Primer

Circular consensus sequencing

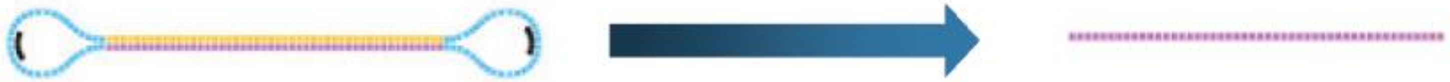


PacBio nomenclature

- Polymerase read length
  - Lifetime of the polymerase (how many kilobases it sequences)
  - Directly impacts both quality and length of reads
- Read of insert length
  - Length of the DNA fragment between adaptor sequences
  - Ultimate limit on read lengths
  - Short fragments will tend to load preferentially so size selection is important to remove these
- Subread
  - The data from a single pass of a polymerase across the DNA fragment between the adaptors
- Circular consensus
  - The consensus of multiple subreads from a single piece of DNA
- P0, P1 and P2
  - Occupancy of Zero-Mode-Waveguides – we want to maximise P1 (one DNA/polymerase complex)

Circular consensus sequencing

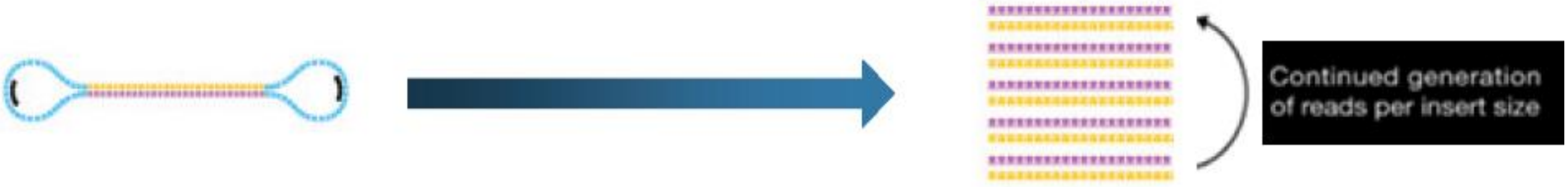
Standard Sequencing for Continuous Long Reads (CLR)



Large Insert Sizes

Generates one pass on each molecule sequenced

Circular Consensus Sequencing (CCS)

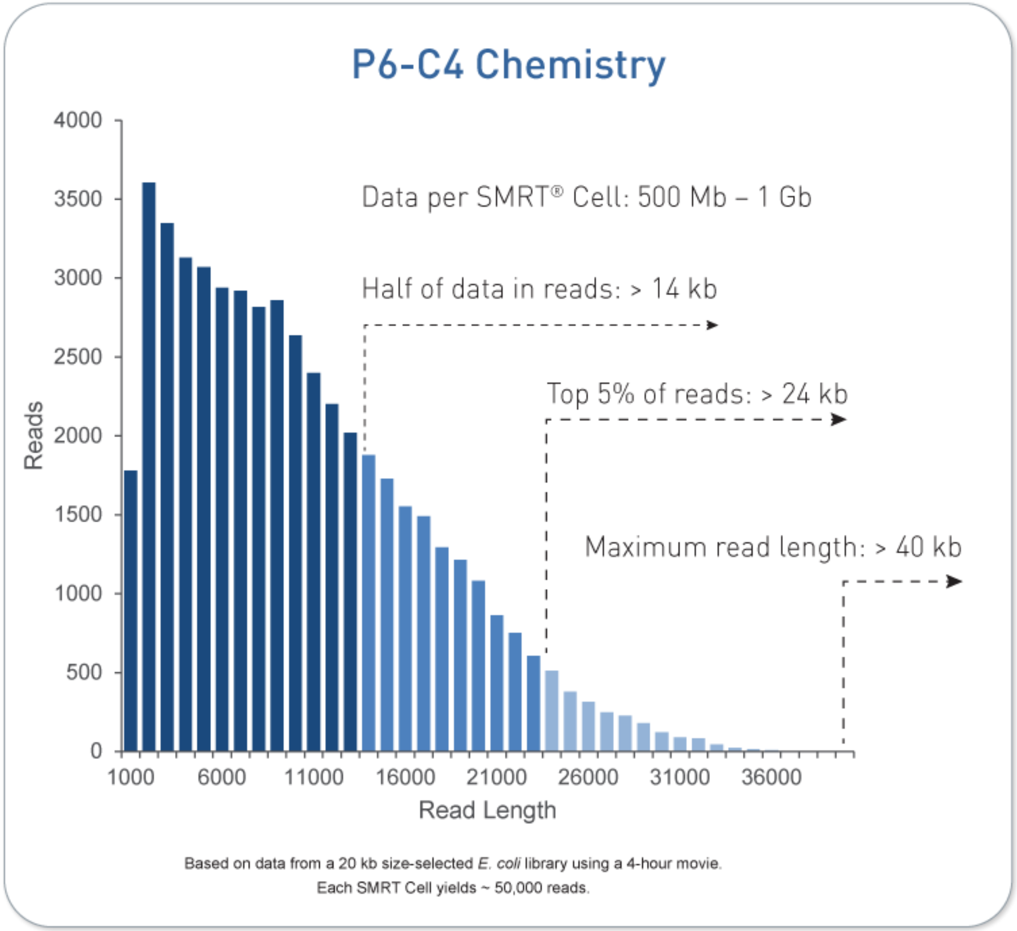


Continued generation of reads per insert size

Small Insert Sizes

Generates multiple passes on each molecule sequenced

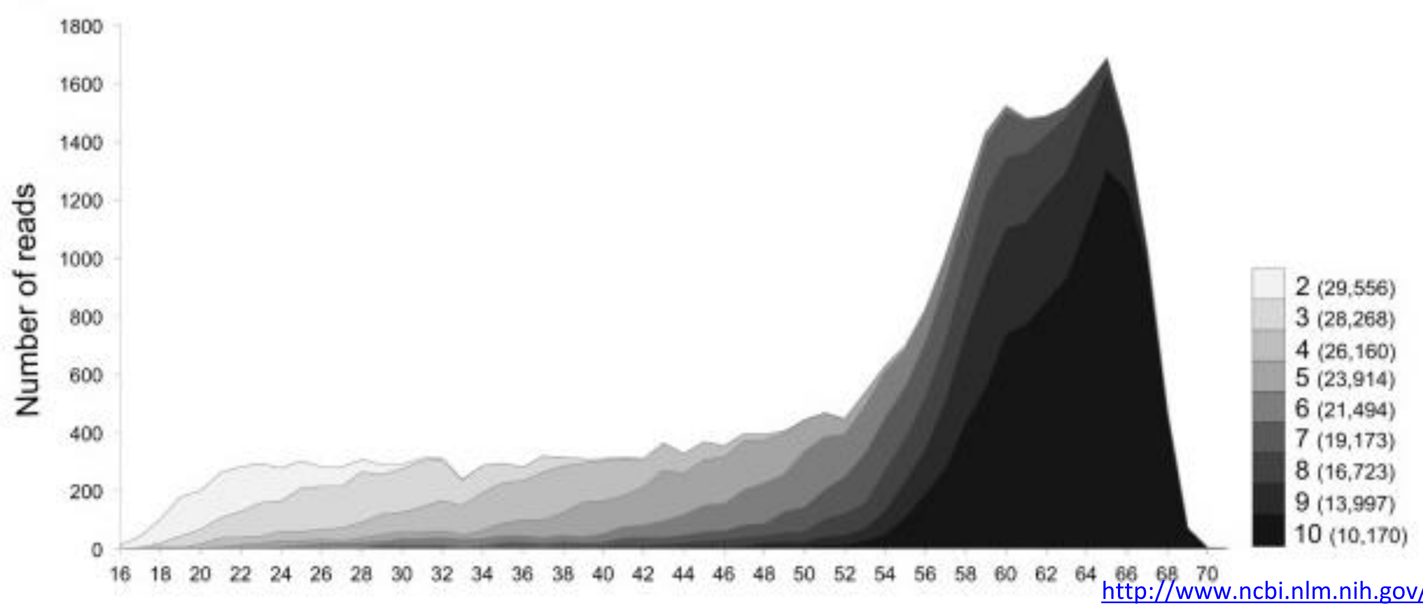
Read lengths



Based on data from a 20 kb size-selected *E. coli* library using a 4-hour movie.
Each SMRT Cell yields ~ 50,000 reads.

Note on read length and accuracy

- Longer reads are more error prone because they are read fewer times
- Error rate of a single read has a phred quality score of approximately 9 (~12% error rate)
- Therefore to obtain a phred quality above 30 we need to have at least 2-3 passes of the molecule
- With a median polymerase read length of 12kb, this is achievable with for 3kb fragments



http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4152597/

Note on read length and accuracy

- In summary:
  - Shorter reads will tend to be higher quality
  - Longer reads will tend to be lower quality

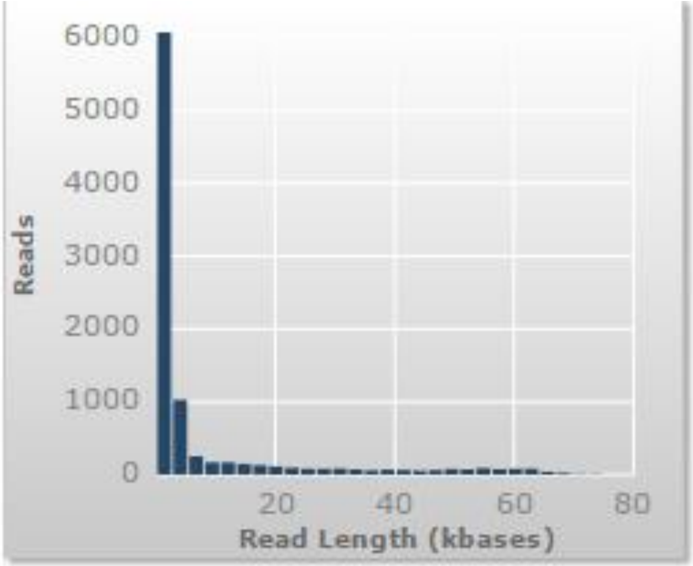
- Depending on the application this may or may not be a problem
  - Long, error prone reads are good for scaffolding genome assemblies
  - May not be suitable for long amplicons

Clean and high quality DNA is CRUCIAL

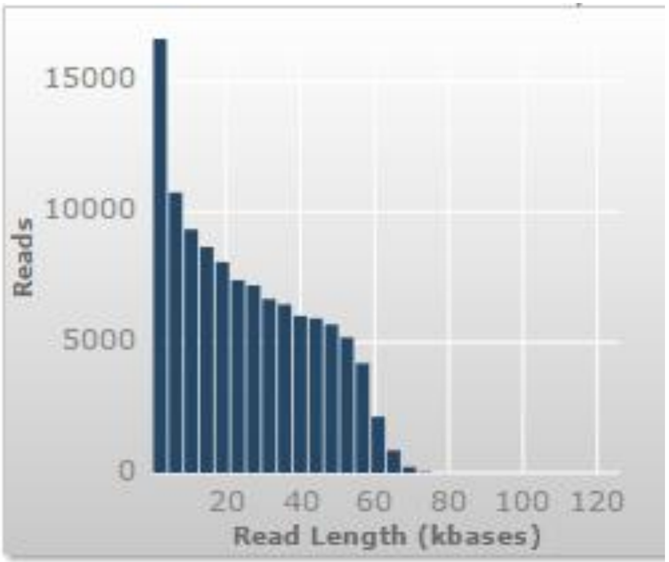
- Optimising a DNA extraction to achieve obtain clean, high quality DNA is the most time consuming step for non-model organisms
- **Avoid:**
  - **DNA damage**: alkylation, oxydation, UV-crosslinks, AP-sites, intercalating agents
  - **DNA binders**: polyphenols, secondary metabolites , pigments, polysaccharides
  - **Polymerase inhibitors**: salts (EDTA), phenol, alcohols
  - **DNA Fragmentation:** If your DNA is fragmented during extraction then you will not obtain long reads

Example: Effect of contaminants

Sample with contaminants

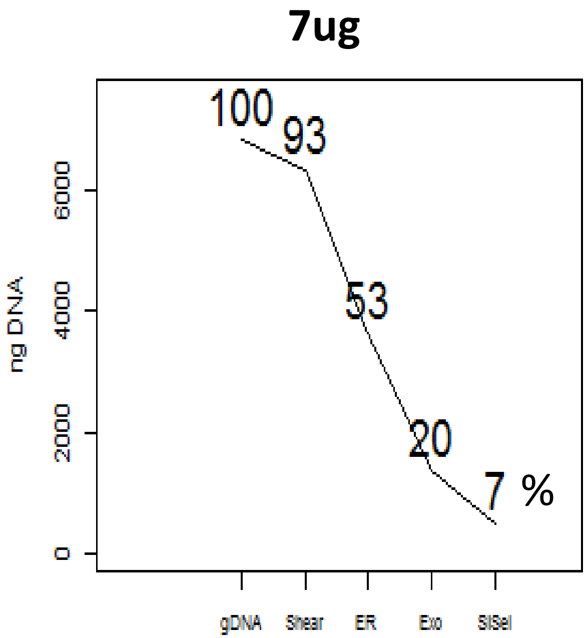


Same sample without contamination

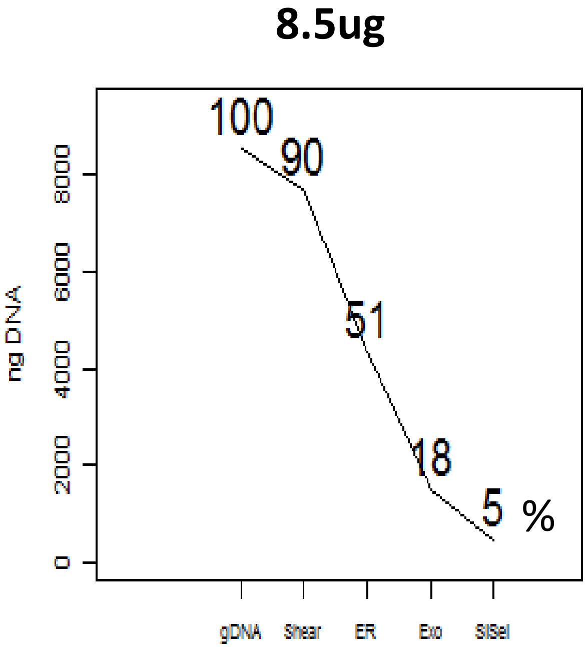


This is *native* DNA sequencing as opposed to *PCR-cleaned* sequencing

You need a lot of DNA

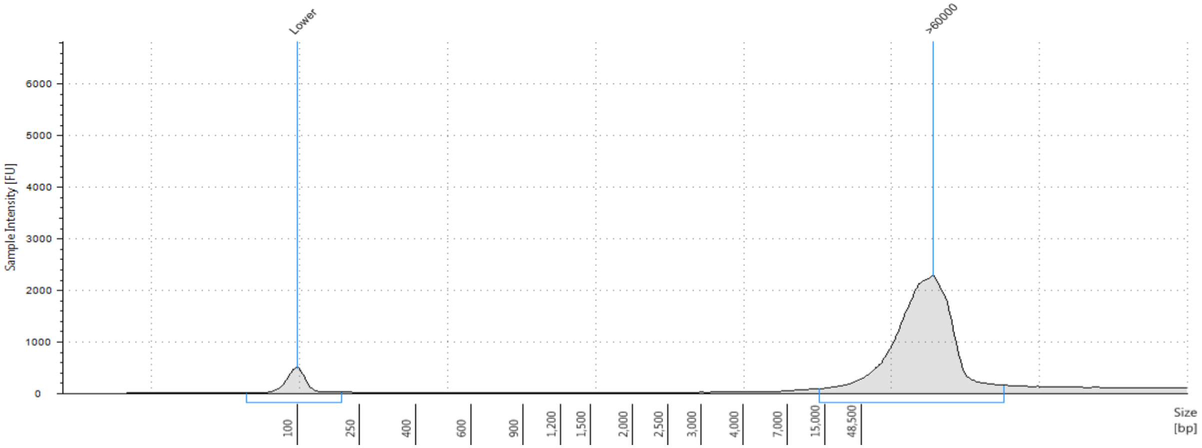


Template = 0.475 ug

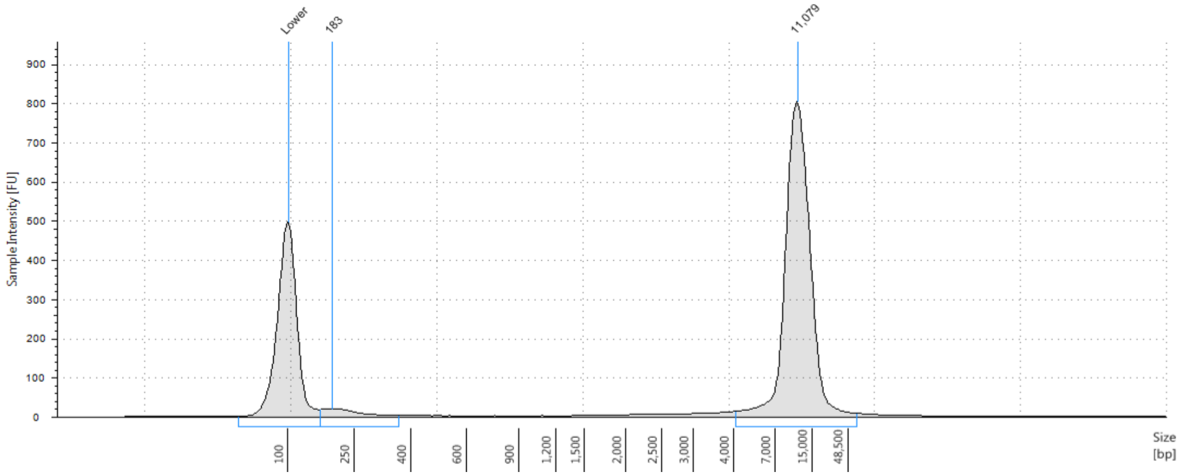


Template = 0.46 ug

Fragment DNA



Shearing: hydro-shear, g-tube, needle, sonication etc.



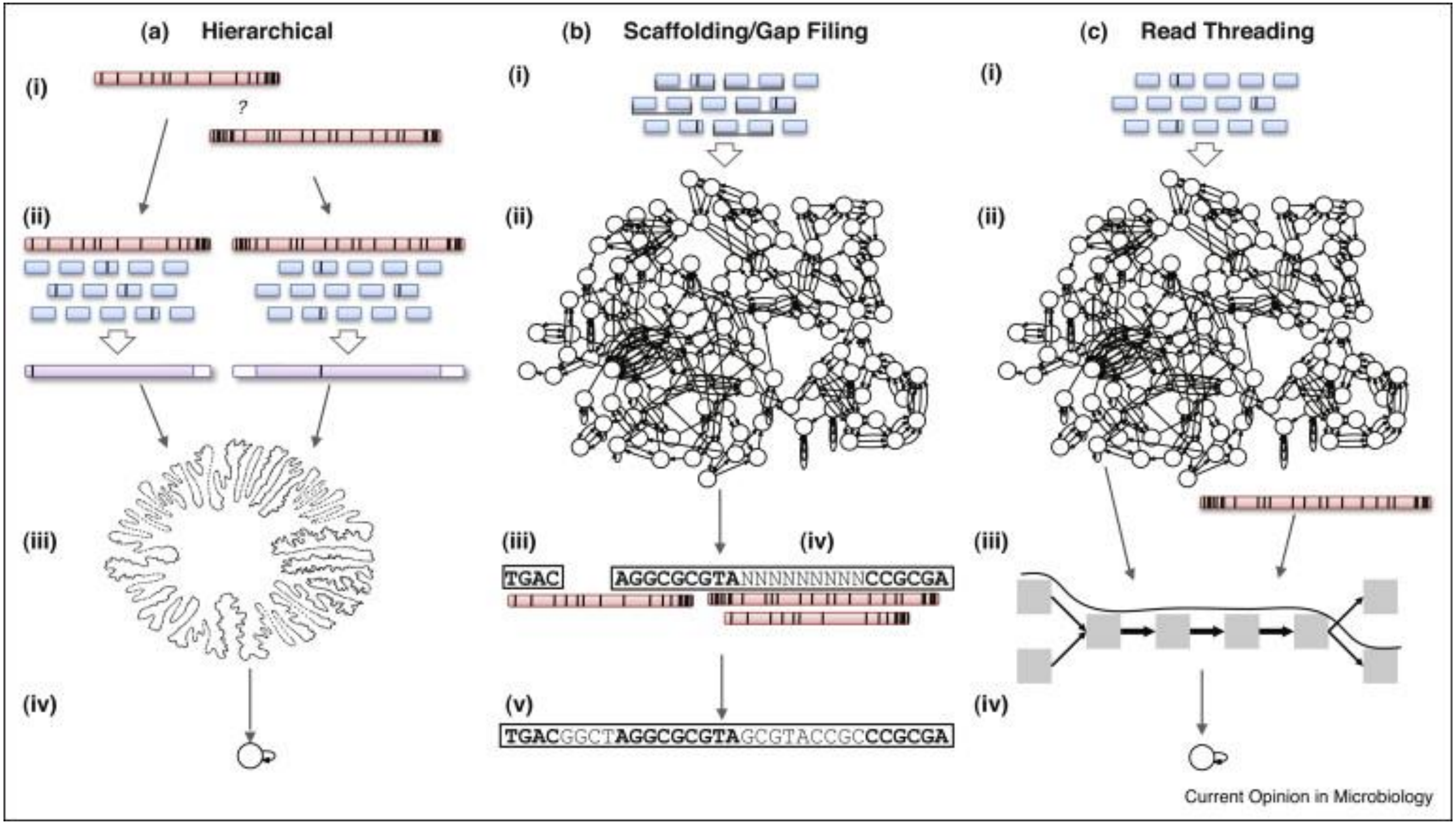
Types of libraries

- Non-size selected
  - Sequencer will tend to load shorter fragments so best avoided unless material is of a uniform length
- 10, 20, 30kb libraries
  - Increasingly difficult to size select sufficient material beyond that
  - Molar concentration of long DNA fragments tends to be much lower than shorter fragments
- IsoSeq libraries
  - Based on clontech polyA libraries
- Barcoding/multiplexing is possible

PacBio applications

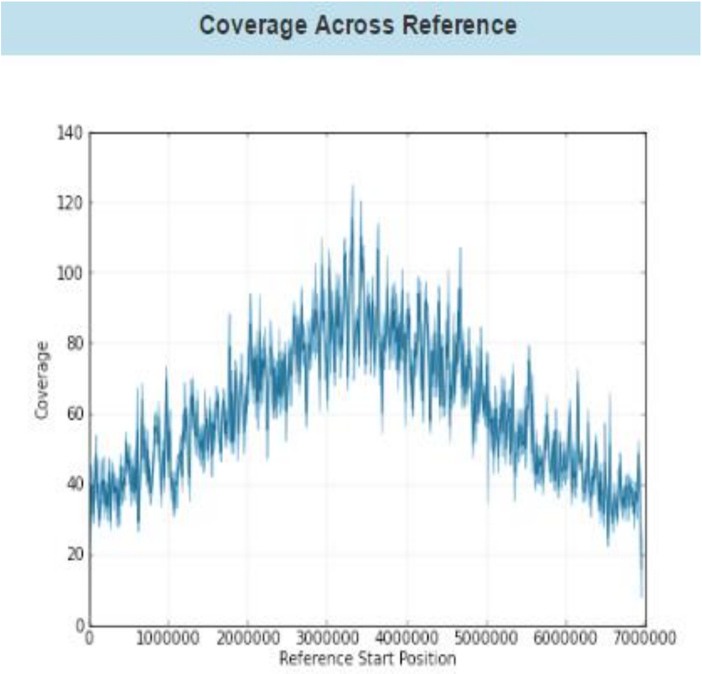
- Denovo assembly
- Hybrid assembly
- Complex regions and structural variation
- PCR Amplicon
- Iso-seq
- Targeted capture of regions
- Haplotype phasing
- DNA modifications

Genome assembly methods

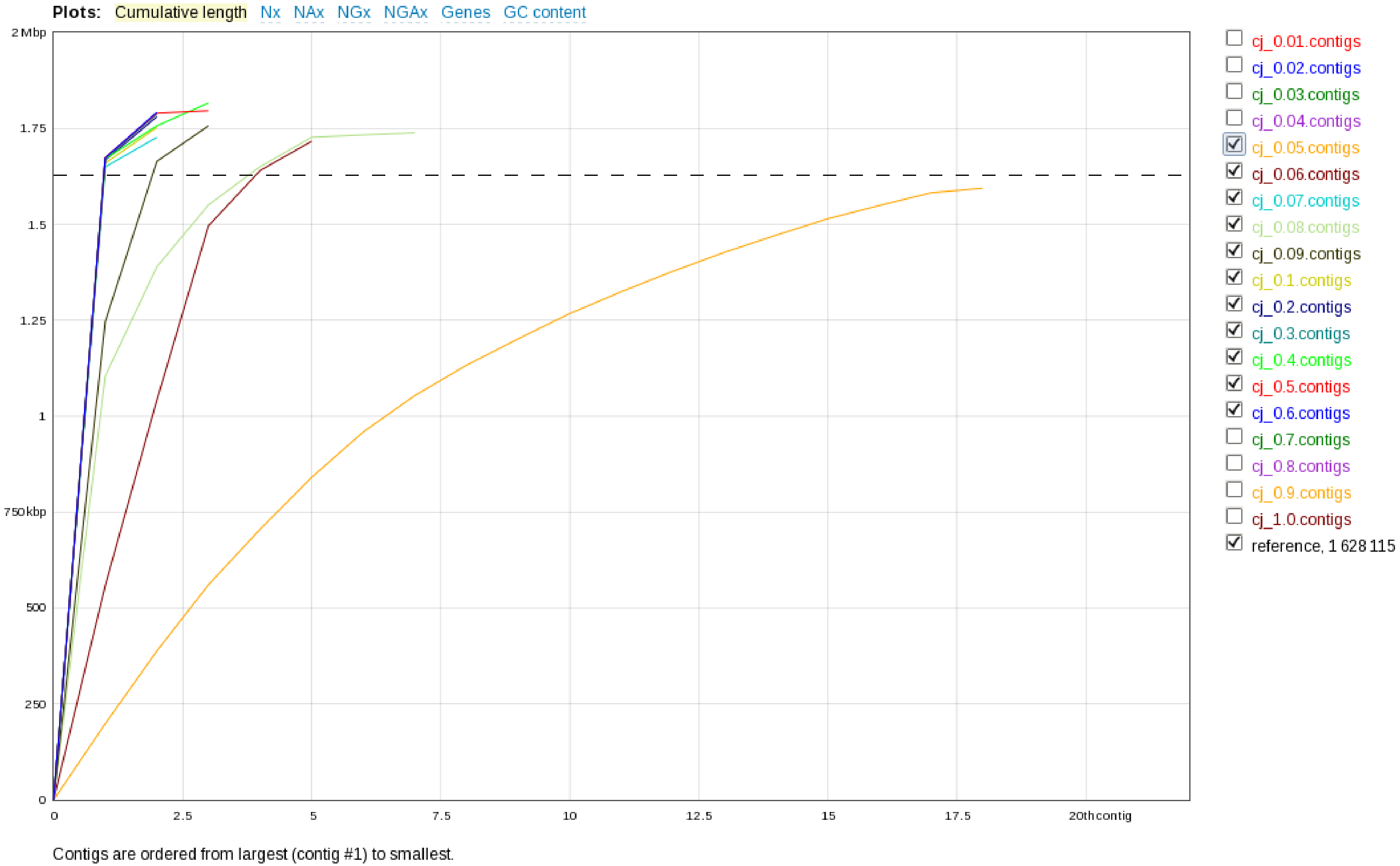

(a) Hierarchical
(b) Scaffolding/Gap Filing
(c) Read Threading
(i)
?
(ii)
(iii)
(iv)
(v)
TGAC
AGGCGCGTANNNNNNNNNCCGCGA
TGACGGCTAGGCGCGTAGCGTACCGCCCGCGA
Current Opinion in Microbiology

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

Sequencing of Pseudomonas aeruginosa (6.3Mb)
1 contig



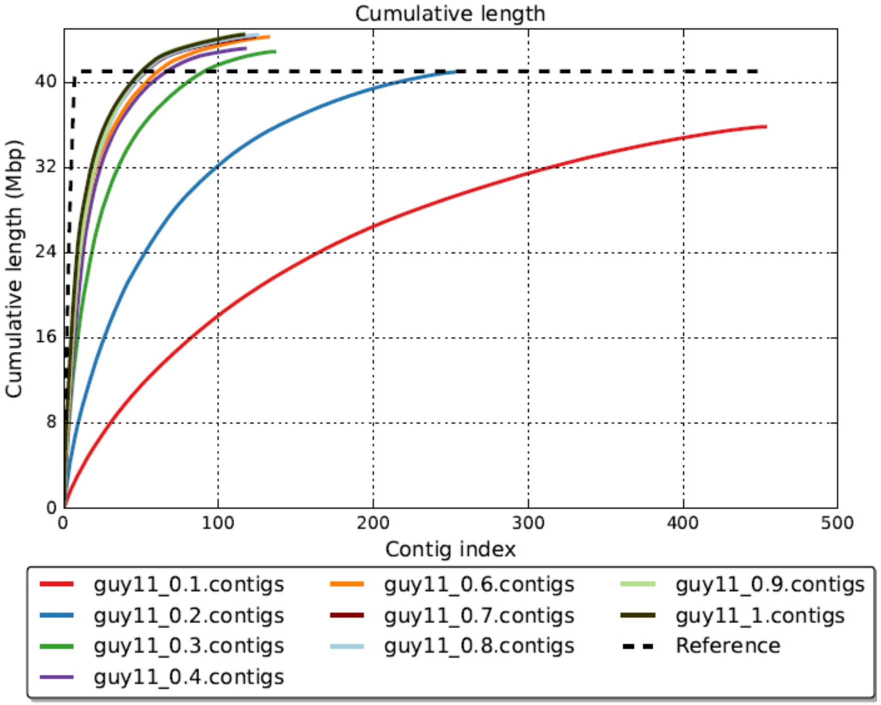
How much data do we need from a 20kb library?



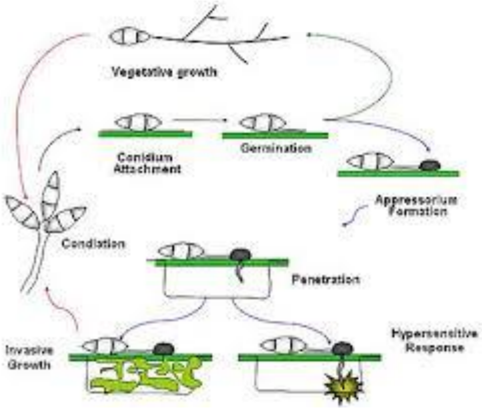
Contigs are ordered from largest (contig #1) to smallest.

Fungal de-novo assemblies

- Magnaporthe oryzae
- ~40 Mbase genome

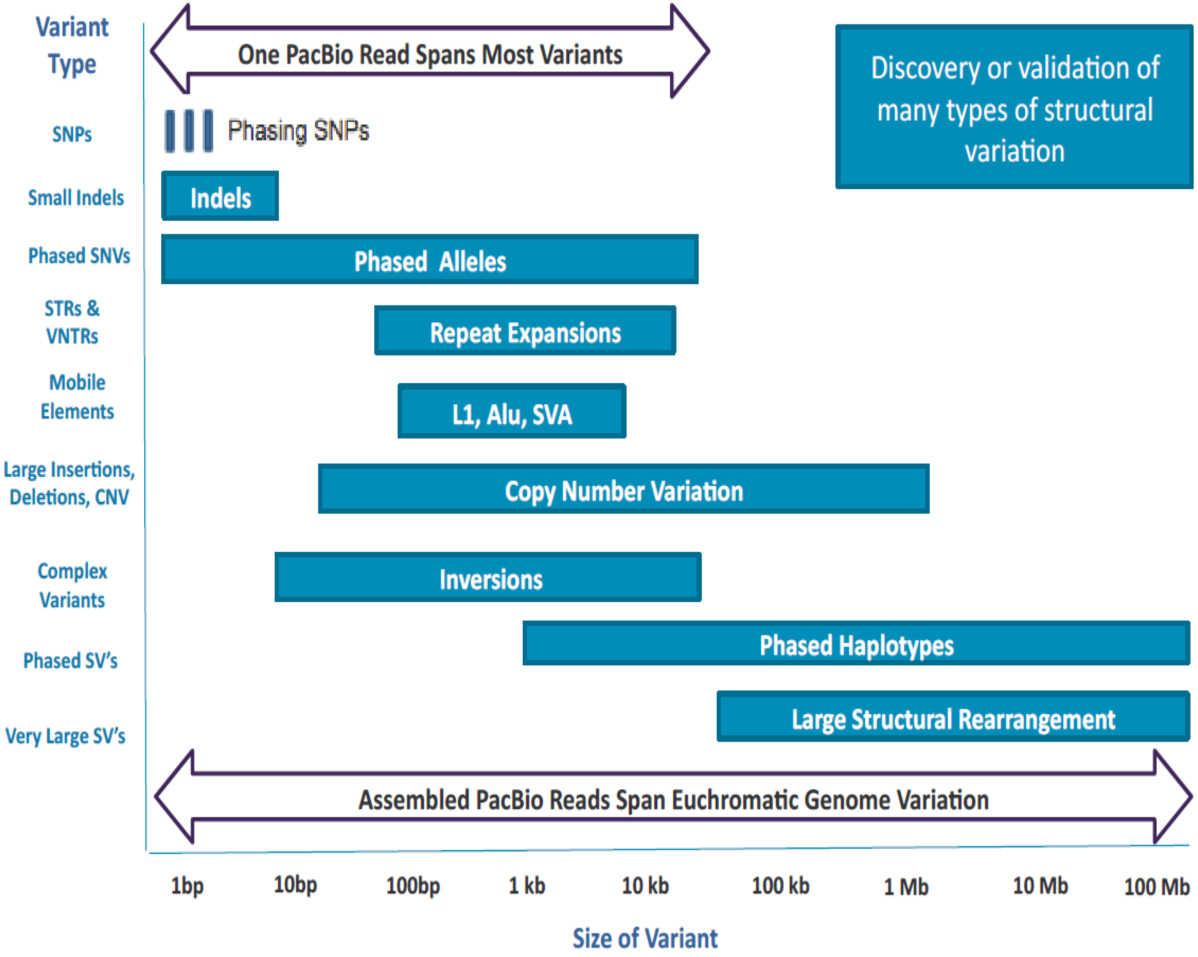


Nature Reviews | Microbiology

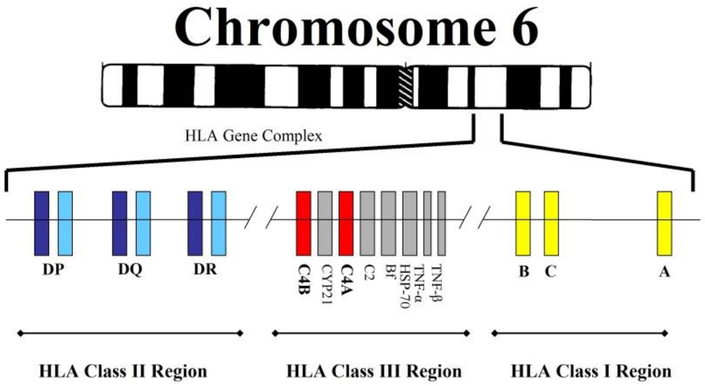


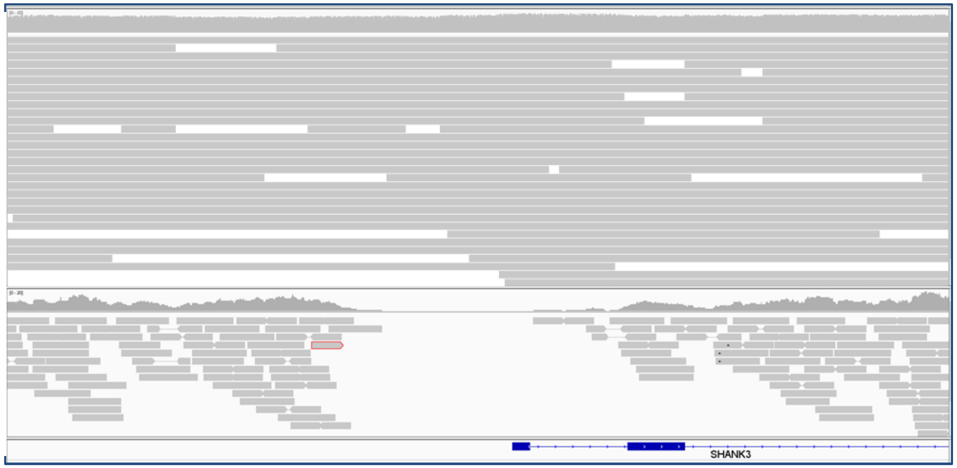
http://www.nature.com/nrmicro/journal/v7/n3/abs/nrmicro2032.html

Structural variation



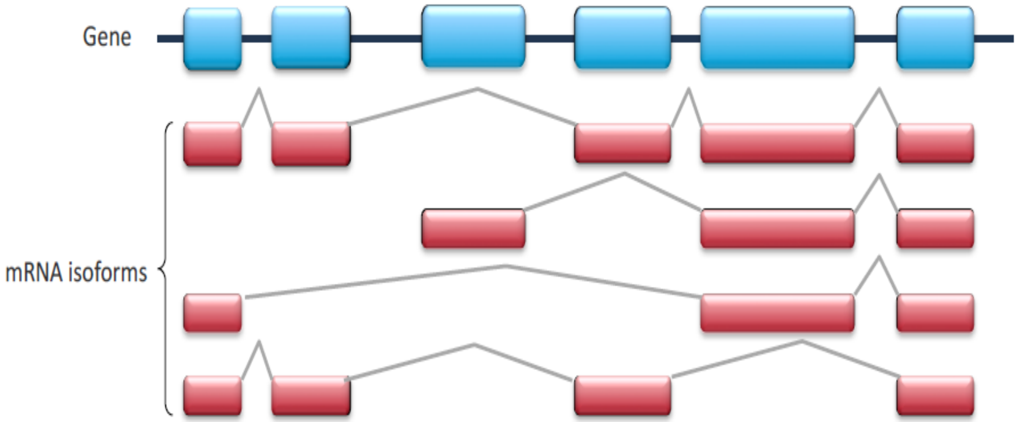
Sequencing of complex regions
Major histocompatibility complex





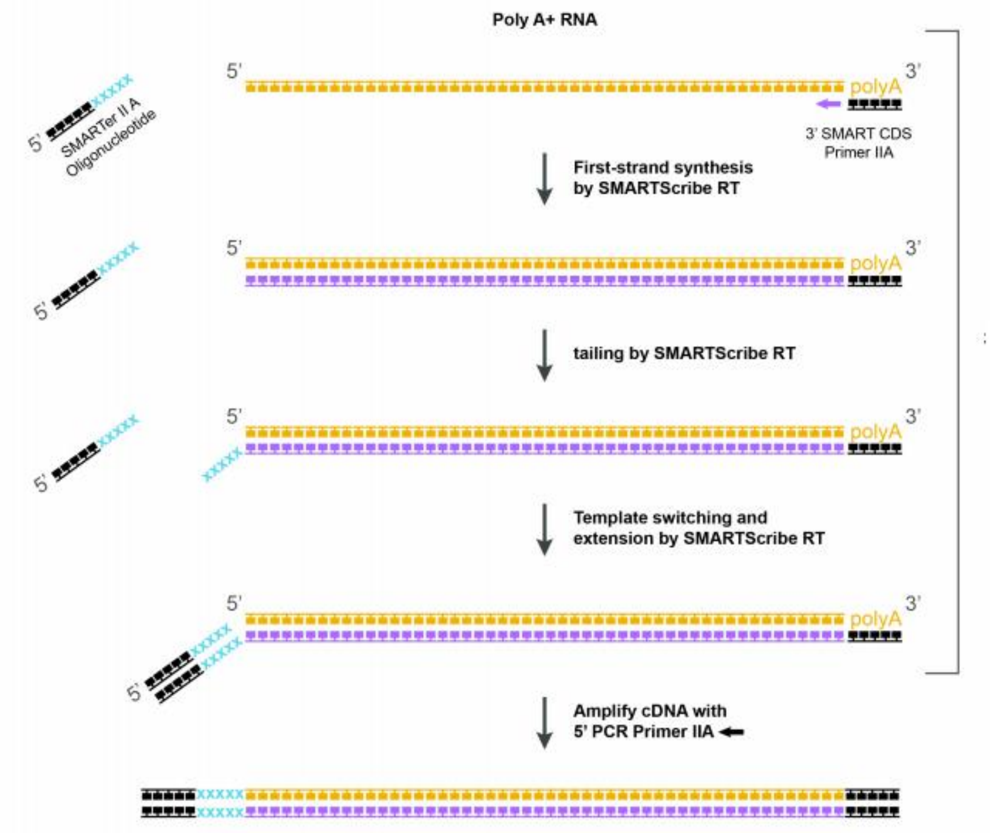
```
CCCCGTCACAGCCCCCAGACCCCCGCCCCGTGGCTCGGCCCCCGCCCTCCGCACACACCT
CCCGCCCCCACCCGGGACCCCGCAAGTAACCCCCCAGCACTGGCCCTGAGCCCTCCCGGCC
CCCGCCTCCGGCGCAGCCCCCTCGCCACCCCCGCTTCCCTCCCGTCTCAGGCCCCCTCCCCC
CGCCGCCCCCGCCCCCGGGGAAGGCAGGCGCCGAGCTGAGCCGGGGCCGATGCAGCTG
AGCCGCGCCGCCGCCGCCGCCGCCGCCGCCCCTGCGGAGCCCCCGGAGCCGCTGTCCCCC
GCGCCGGCCCCGGCCCCGGCCCCCCCCGGCCCCCTCCCGCGCAGCGCGGCCGACGGGGC
TCCGGCGGGGGGGAAGGGGGGGCCGGGGCGCCGCGCGCGGAGTCCCCGGGCGCTCCG
TTCCCCGGCGCGAGCGGCCCCGGCCCGGGCCCCGGCGCGGGGATGGACGGCCCCGGGG
CCAGCGCCGTGGTCGTGCGCGTCGGCATCCCGGACCTGCAGCAGACGGTGAGCCCCG
```

Isoform sequencing

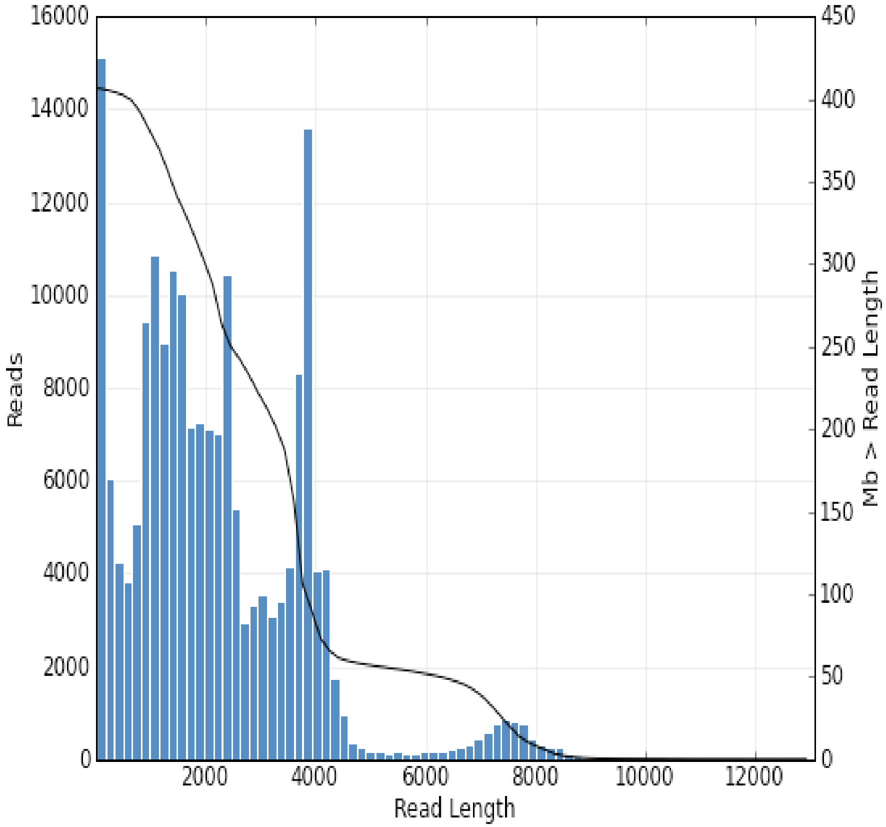


Full length transcript sequencing (IsoSeq)

- Requires polyA RNA
- Uses SMRTer approach
- Ability to sequence full length transcripts with no need for assembly

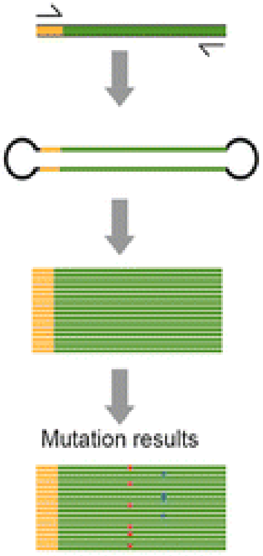

Poly A+ RNA
5'
3'
polyA
SMARTer II A Oligonucleotide
3' SMART CDS Primer IIA
First-strand synthesis by SMARTScribe RT
tailing by SMARTScribe RT
Template switching and extension by SMARTScribe RT
Amplify cDNA with 5' PCR Primer IIA
XXXXX

Iso-seq example

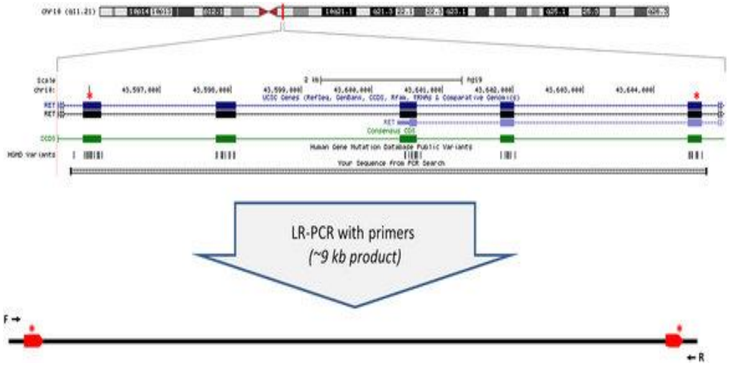


Amplicon sequencing

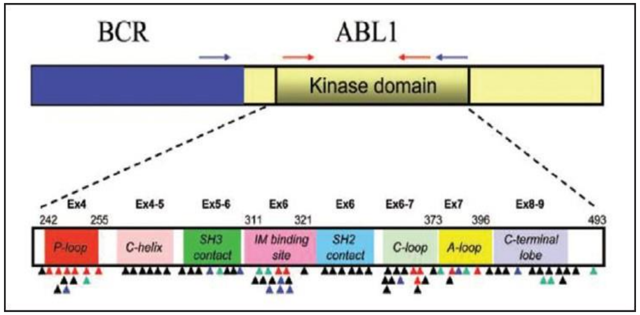
Concept



RET gene: phasing of two mutations



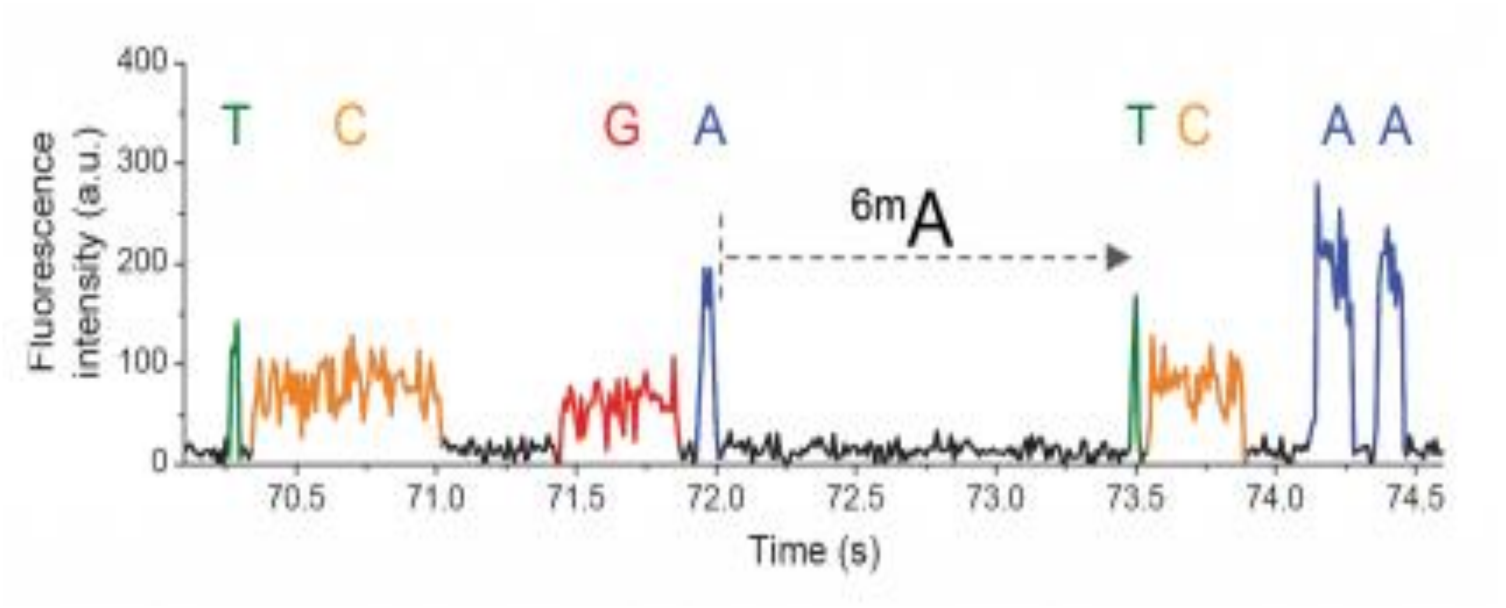
BCR-ABL : Finding mutations in a transcript



Amplicon Consensus Summary

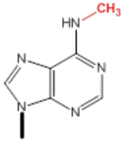
Sequence Cluster	Length (Bp)	Estimated Accuracy	Subreads Coverage
BCR-ABL	1,579	99.994%	500

Epigenetic modifications

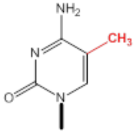


Types of modifications

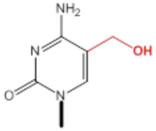
Eukaryotes:



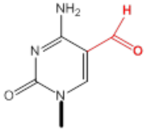
6-mA



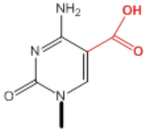
5-mC



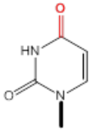
5-hmC



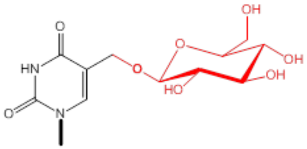
5-fC



5-caC

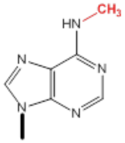


U

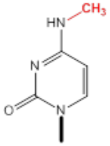


Base J

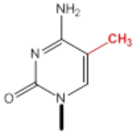
Prokaryotes:



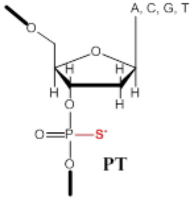
6-mA



4-mC



5-mC



PT

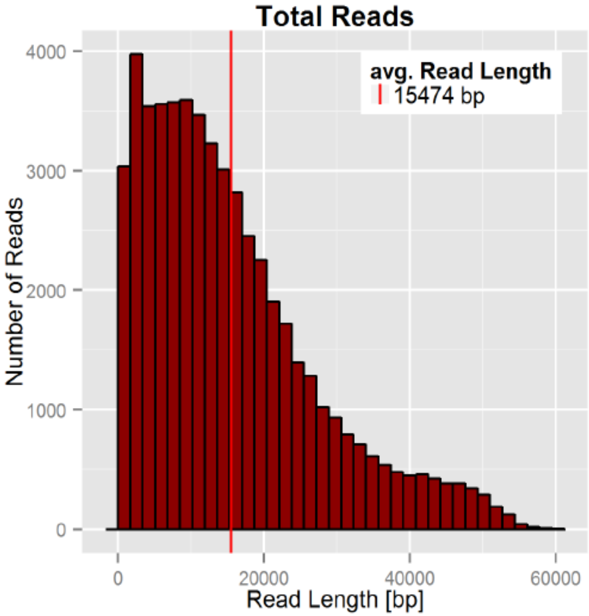
http://www.pacb.com/applications/epigenetics/

RSII vs Sequel

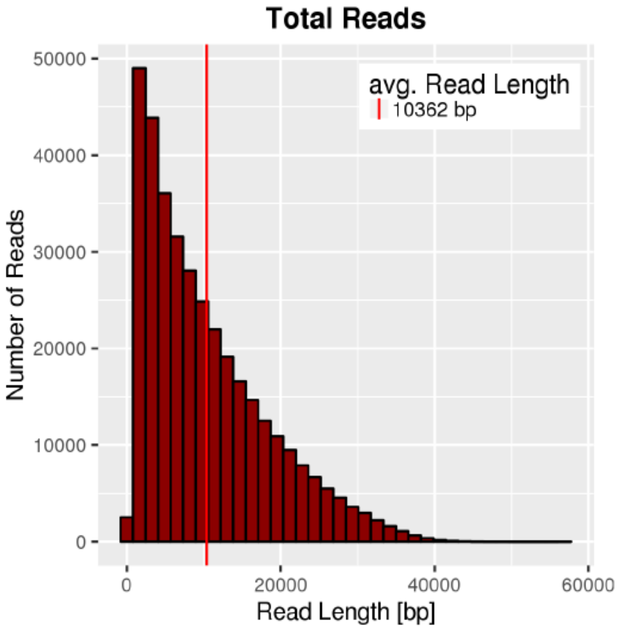
	RSII	Sequel
Cost of instrument (with service contract)	$800k	$450k
Cost per library	$300-$400	$400-500
Cost per SMRT cell	$175	$600
Median polymerase read length	12-15kb	7-8kb
Maximal polymerase read length	90-100kb	30-40kb
Number of reads/cell	50,000-100,000	300,000-600,000
Data volume	750Mbase-2Gbase	3-6Gbase
Cost per megabase (inc library) assuming lower limits of throughput (approx)	$0.67	$0.37

Consumable costs only

Initial chemistry issues with Sequel



P1	35%
Yield	820 Mb

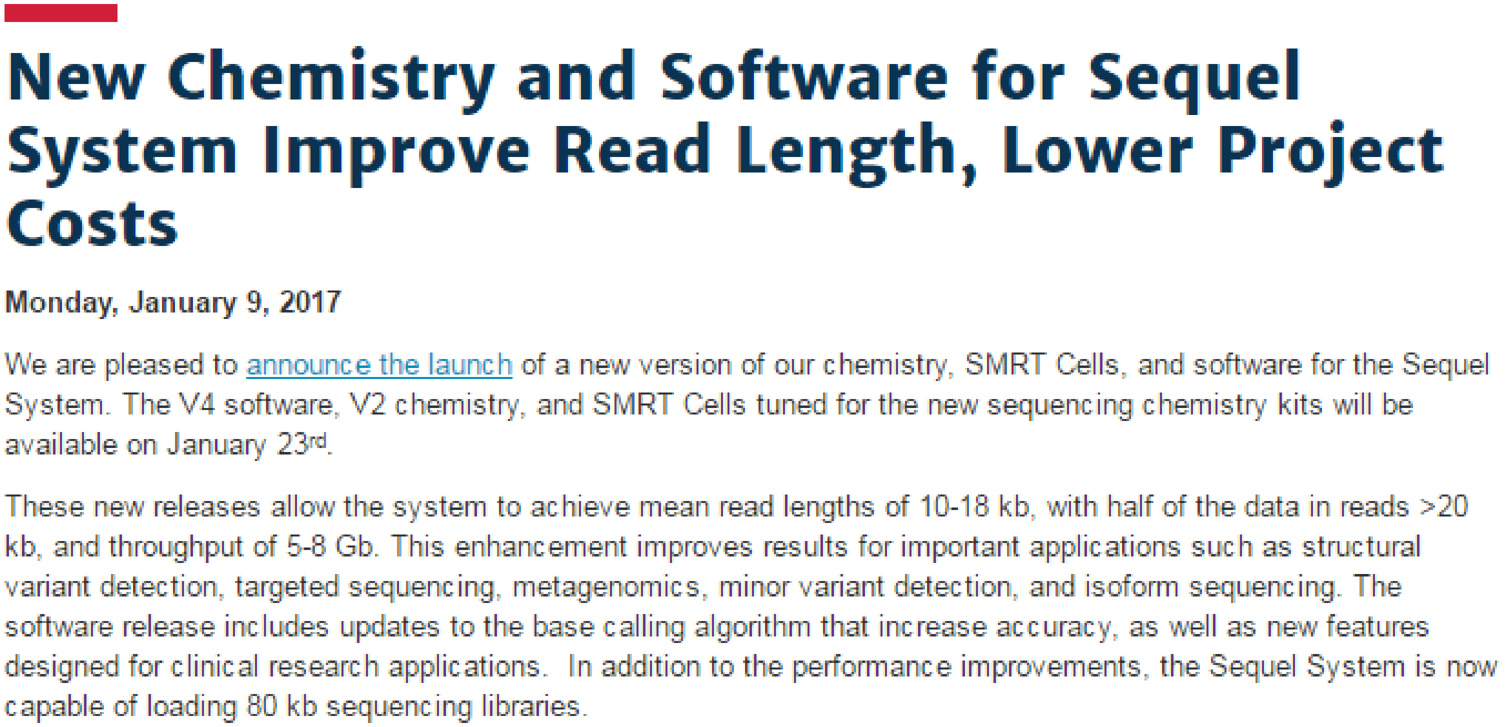


P1	36%
Yield	3716 Mb

Initial chemistry issues with Sequel

- Limited to 20kb inserts
- Limited to 6-9kb polymerase reads
- Output limited to 2-3Gbases/cell

Initial chemistry issues with Sequel



New Chemistry and Software for Sequel System Improve Read Length, Lower Project Costs

Monday, January 9, 2017

We are pleased to announce the launch of a new version of our chemistry, SMRT Cells, and software for the Sequel System. The V4 software, V2 chemistry, and SMRT Cells tuned for the new sequencing chemistry kits will be available on January 23rd.

These new releases allow the system to achieve mean read lengths of 10-18 kb, with half of the data in reads >20 kb, and throughput of 5-8 Gb. This enhancement improves results for important applications such as structural variant detection, targeted sequencing, metagenomics, minor variant detection, and isoform sequencing. The software release includes updates to the base calling algorithm that increase accuracy, as well as new features designed for clinical research applications. In addition to the performance improvements, the Sequel System is now capable of loading 80 kb sequencing libraries.

http://www.pacb.com/blog/new-chemistry-software-sequel-system-improve-read-length-lower-project-costs/

Improvements

- Loading DNA fragments up to 80kb
- PEG buffer to reduce loading bias for small fragments
  - May enable running of non size-selected IsoSeq libraries
- Median read lengths >20kb
- Output of 5-8Gbase/cell

Pacific Biosciences

- Advantages
  - Much longer reads lengths possible than second generation sequencers
  - Cost per SMRTcell is lower ($250 per SMRTcell 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)
  - Still enzyme based
  - Often need multiple cells to optimise loading so you may need to run a minimum of 2-3 cells
  - RSII: Only 50,000 reads/cell – approx 750Mb yield (can yield up to 2Gb for short fragments)
  - Sequel: 300,000-600,000/cell – approx. 3-6Gbases
  - High (12%) error rate per read (but consensus can reduce <0.01% although indels may require Illumina data to polish out)
  - $800k machine

Bioinformatics Implications

- Relatively low data and high per base cost limits widespread use
- Can obtain useful 20-40kb fragments (P4-C6 chemistry)
- Best used in conjunction with error correction algorithms utilising shorter PacBio reads (or Illumina data)– e.g. Wheat D genome
- Excellent to assist scaffolding of genomes
- Able to generate complete bacterial genomes
- Has been used to generate higher eukaryote genomes (e.g. Drosophila, Human) but cost can be 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)

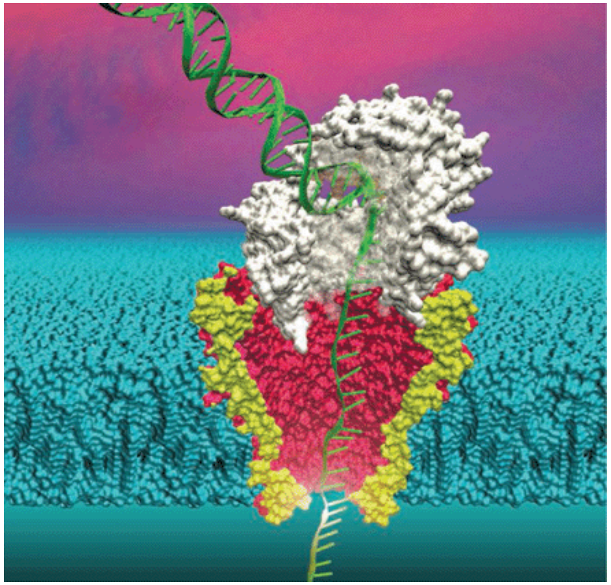
Koren, Sergey; Schatz, Michael C; Walenz, Brian P; Martin, Jeffrey; Howard, Jason T et al. (2012)
Hybrid error correction and de novo assembly of single-molecule sequencing reads
Nature biotechnology vol. 30 (7) p. 693-700

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

Useful PacBio papers

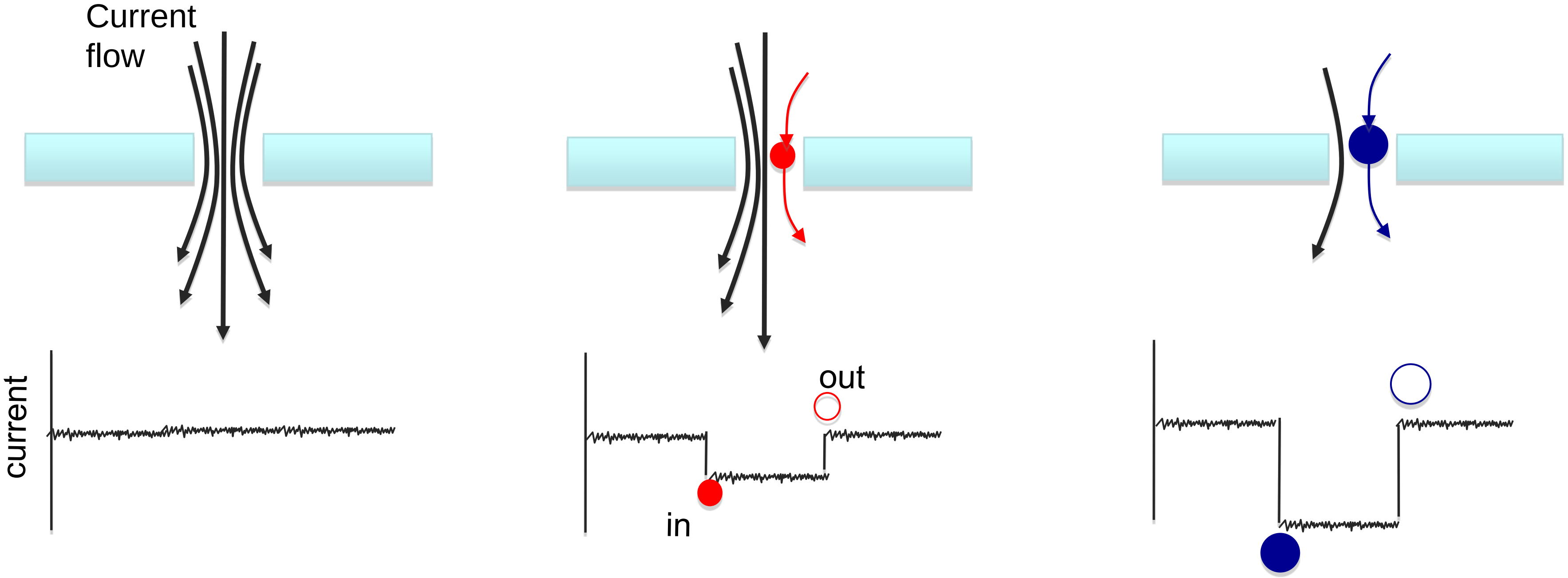
- Resolving the complexity of the human genome using single-molecule sequencing
- Defining a personal, allele-specific, and single-molecule long-read transcriptome
- Heyn, Holger et al. (2015) An adenine code for DNA: A second life for N6-methyladenine. *Cell*

Nanopore sequencing



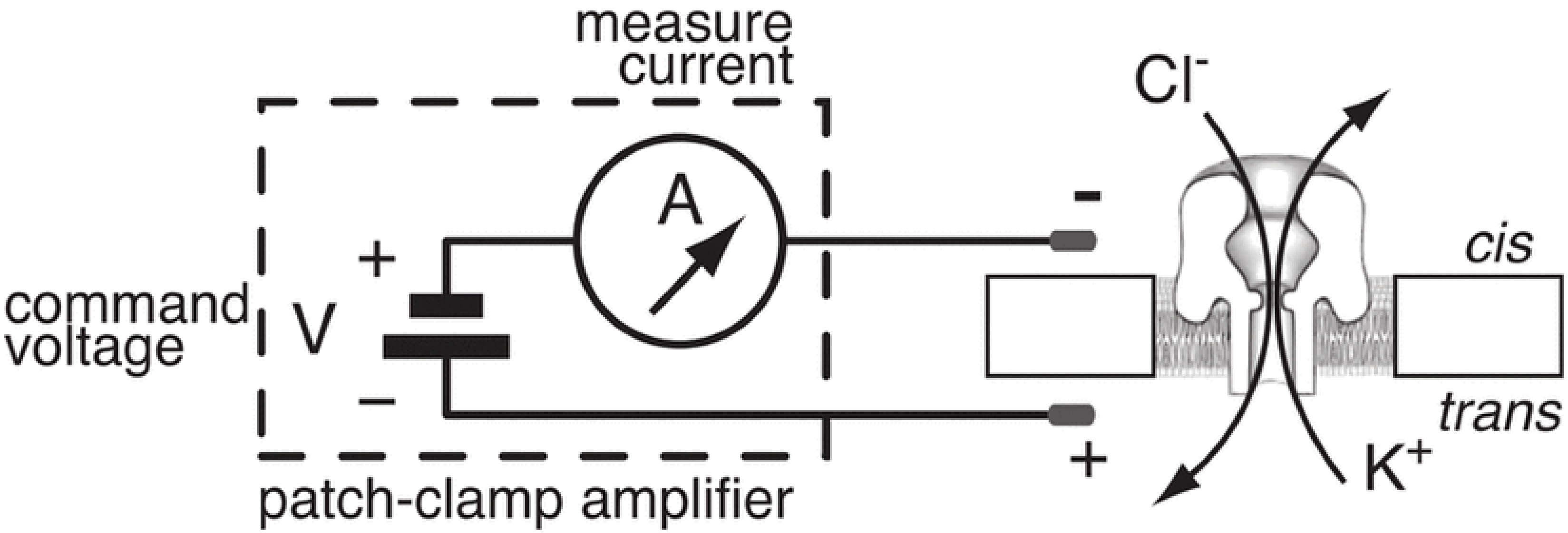
What is a nanopore?

- 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



https://nanoporetech.com/science-technology/introduction-to-nanopore-sensing/introduction-to-nanopore-sensing

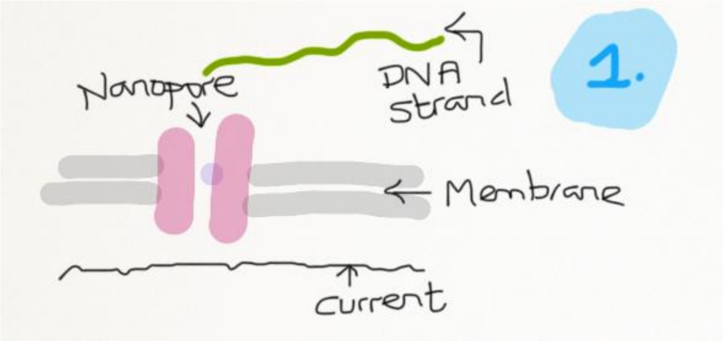
Detection

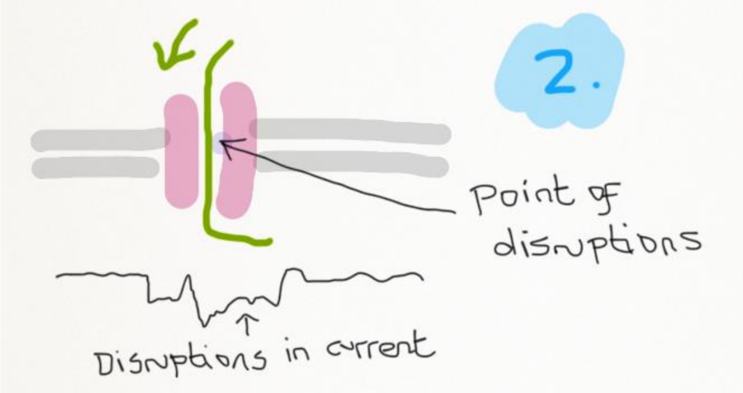


http://ieeexplore.ieee.org/xpls/icp.jsp?arnumber=5626570

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/

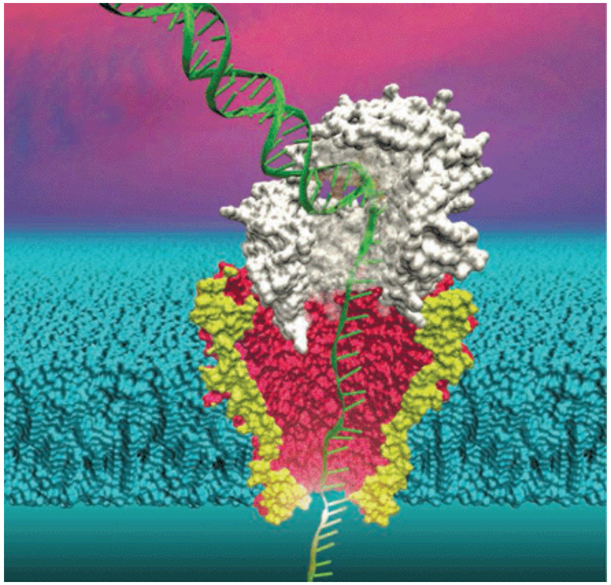
Types of pore

- Either biological or synthetic

- Biological
  - Lipid bilayers with biologically-derived 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

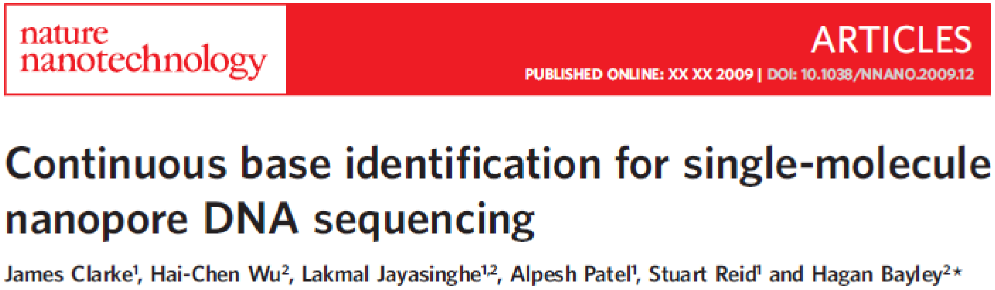
Nanopore sequencing

- 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

Nucleotide Recognition



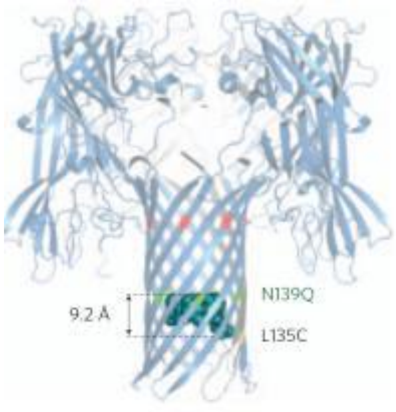
nature nanotechnology

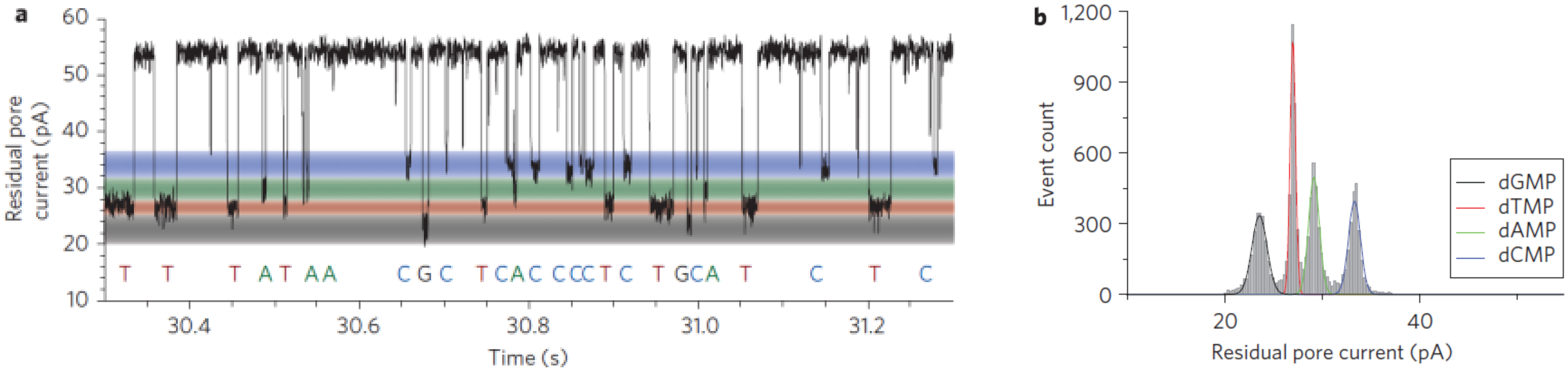
ARTICLES

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

- Oxford Nanopore is the only company with a commercialised product (MinION)

Oxford Nanopore platforms

MinION Mk 1

Up to 10Gbases/run

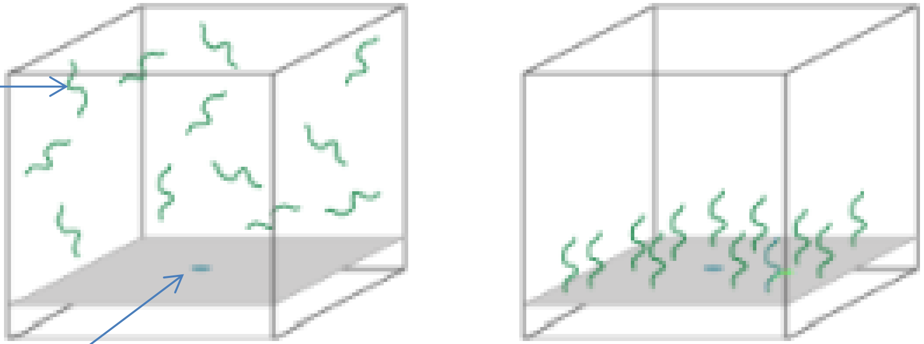
PromethION

Up to 4Tbases/run

Oxford Nanopore MinION programme

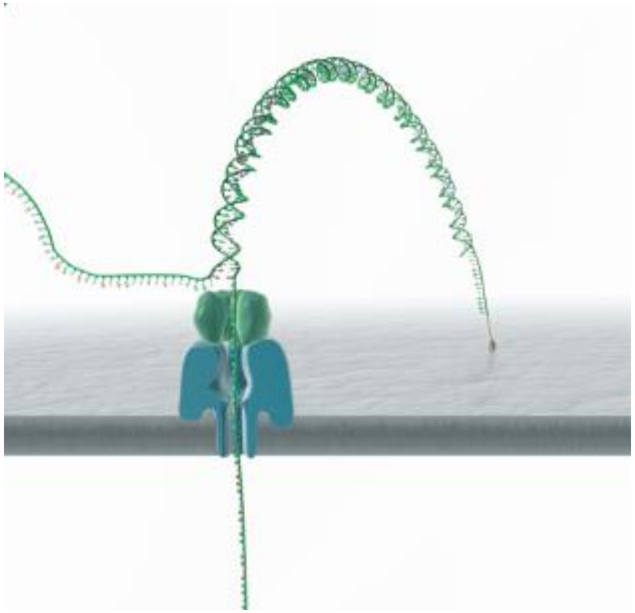
- Uses a different costing model
- Sequencer itself is provided free of charge
- $1000 buys a starter pack with 2 flowcells and basic library prep reagents and access to enhanced support for a few weeks
- Additional flowcells and library kits for $500-$900 each

DNA binding to membrane

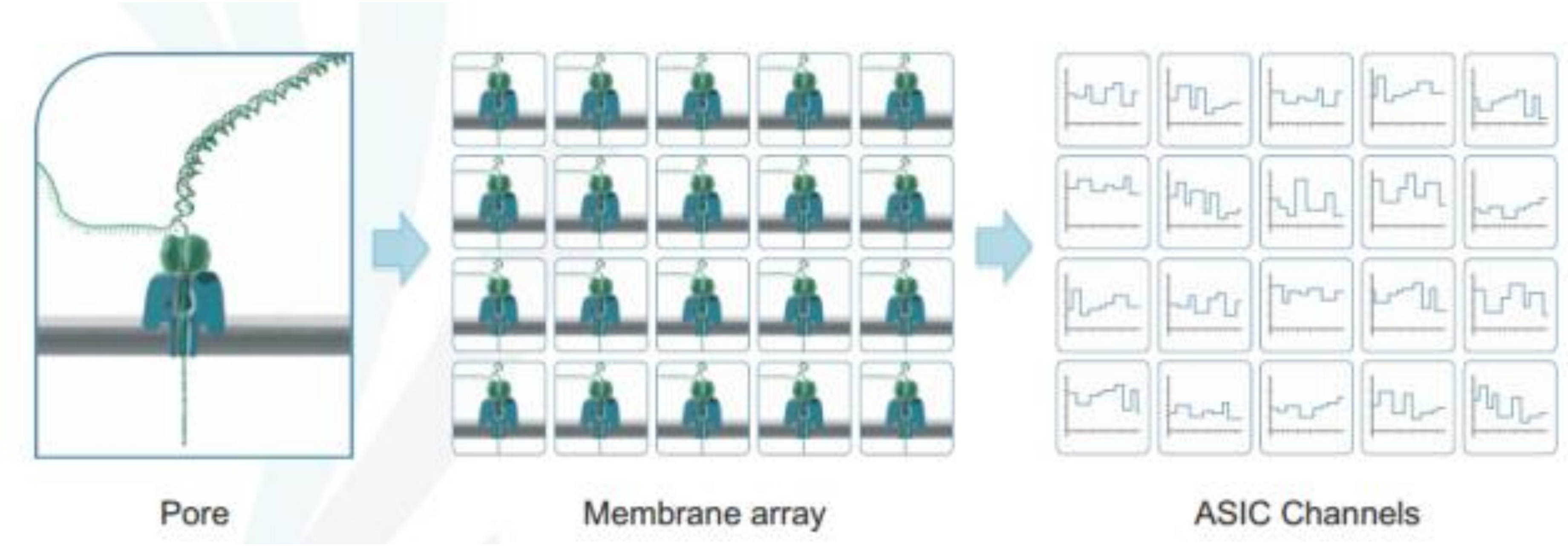


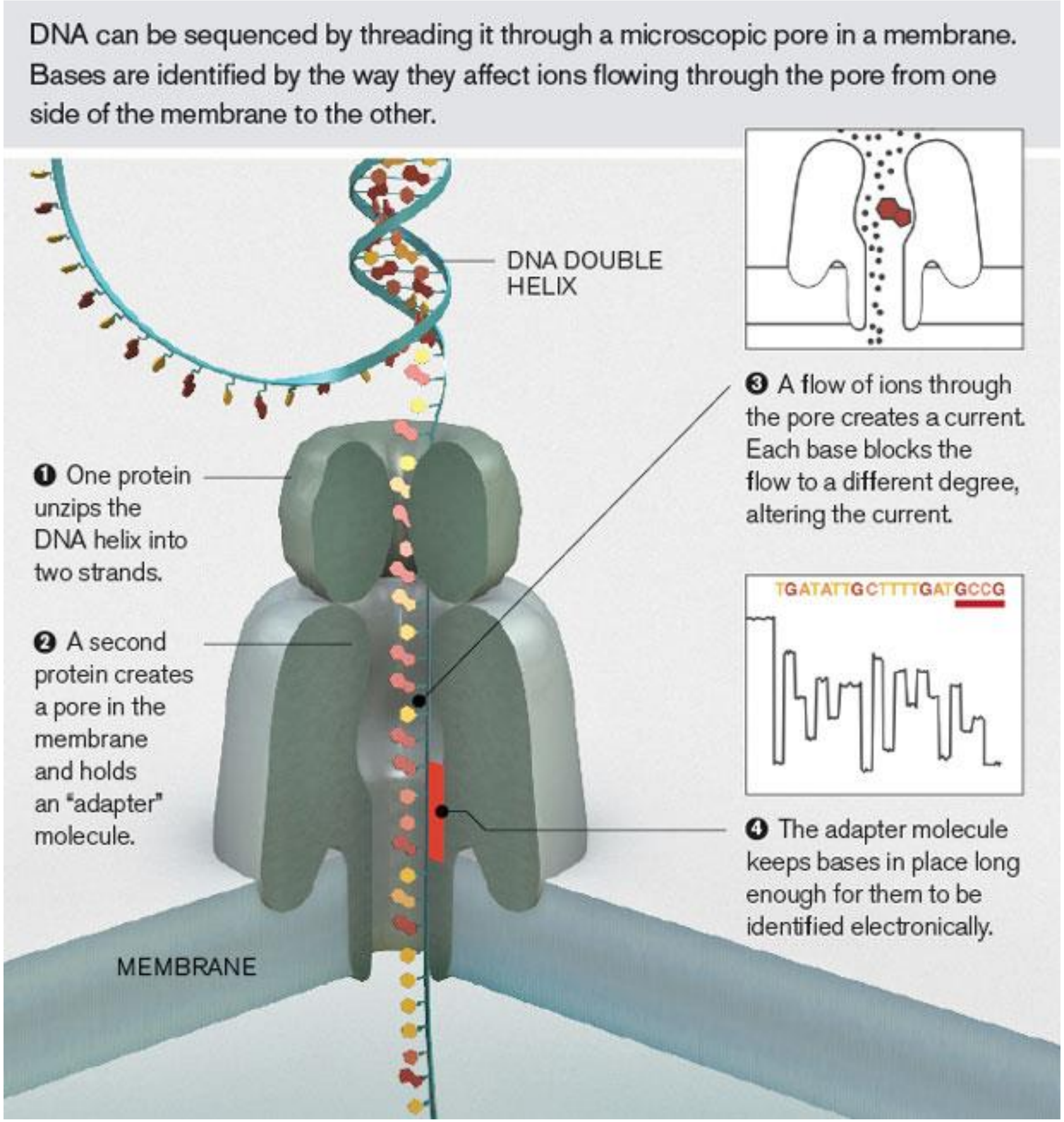
Double stranded DNA fragment

Pore



Oxford Nanopore principle





DNA can be sequenced by threading it through a microscopic pore in a membrane. Bases are identified by the way they affect ions flowing through the pore from one side of the membrane to the other.

❶ One protein unzips the DNA helix into two strands.

❷ A second protein creates a pore in the membrane and holds an "adapter" molecule.

❸ A flow of ions through the pore creates a current. Each base blocks the flow to a different degree, altering the current.

❹ The adapter molecule keeps bases in place long enough for them to be identified electronically.

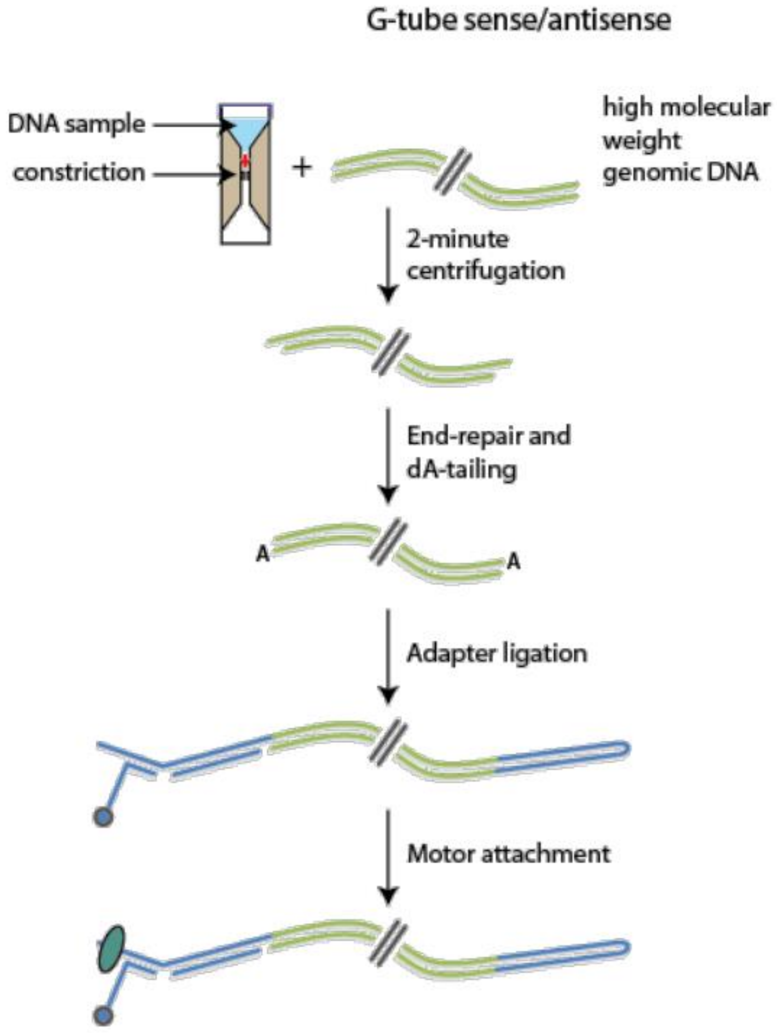
http://www2.technologyreview.com/article/427677/nanopore-sequencing/

MinIon features

- 2048 pores (512 addressable simultaneously)
- Library preparation is required
- Read lengths up to 400kb
  - Limited by input DNA
- Relatively high single pass non-random error rate (12%)
- Up to 10Gbase output

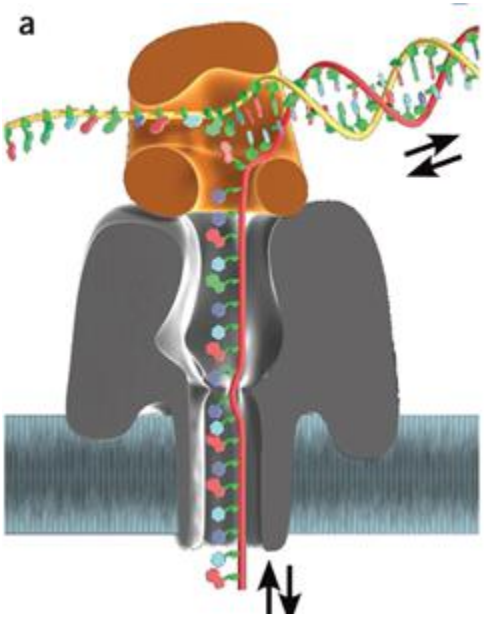
2D Library preparation

- Input requirements
  - Depends on fragment length required
  - Ideally upwards of 1ug of DNA
  - Low input option available – 20ng
- Issues – keeping long DNA fragments
- New approaches (e.g. Voltrax) attempting to create microfluidic

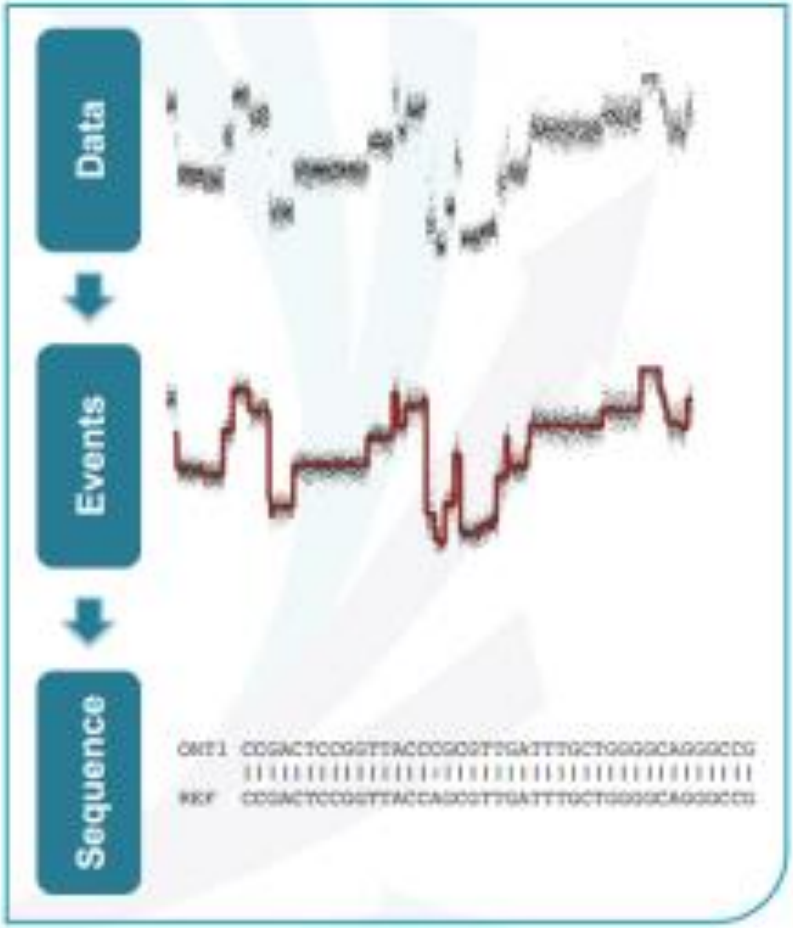


Basecalling 1D vs 2D reads

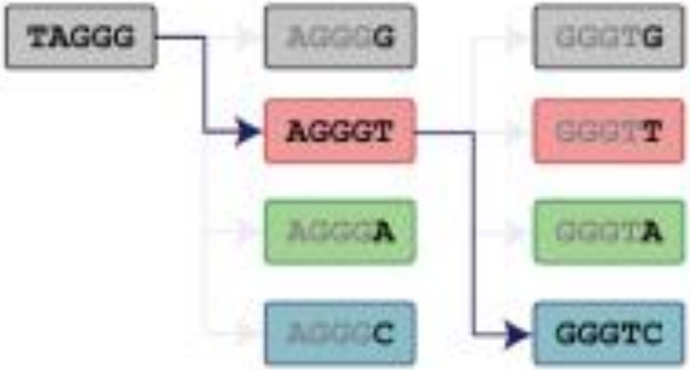
- Both the template and complementary strand can be sequenced
- This doesn’t always work
- If it does, the base-calling can be improved
- Different kmers at the same locus can improve basecalling
- The focus today is on 1D reads since library prep is easier



Challenge of basecalling



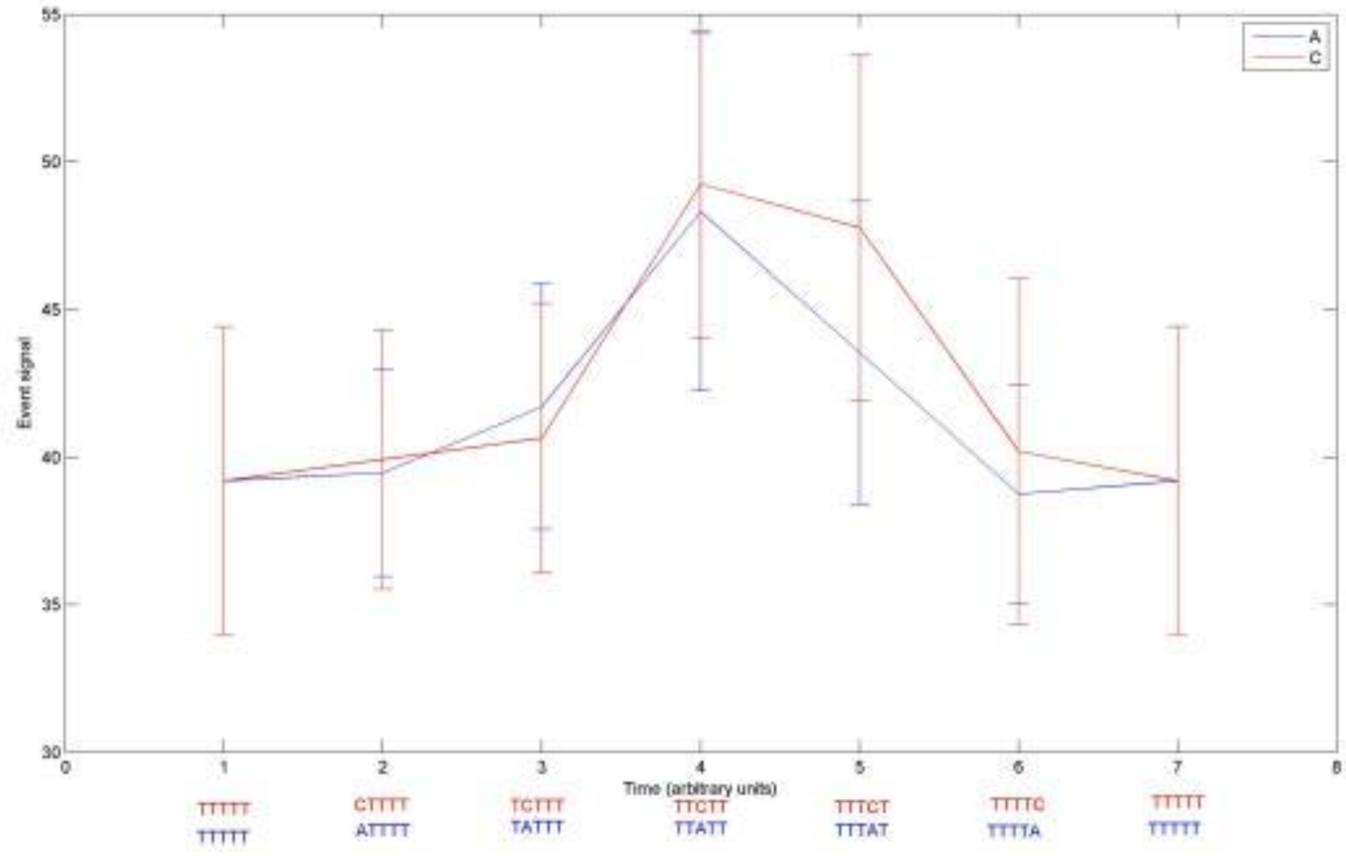
- Hidden Markov model
- Only four options per transition
- Pore type = distinct kmer length



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

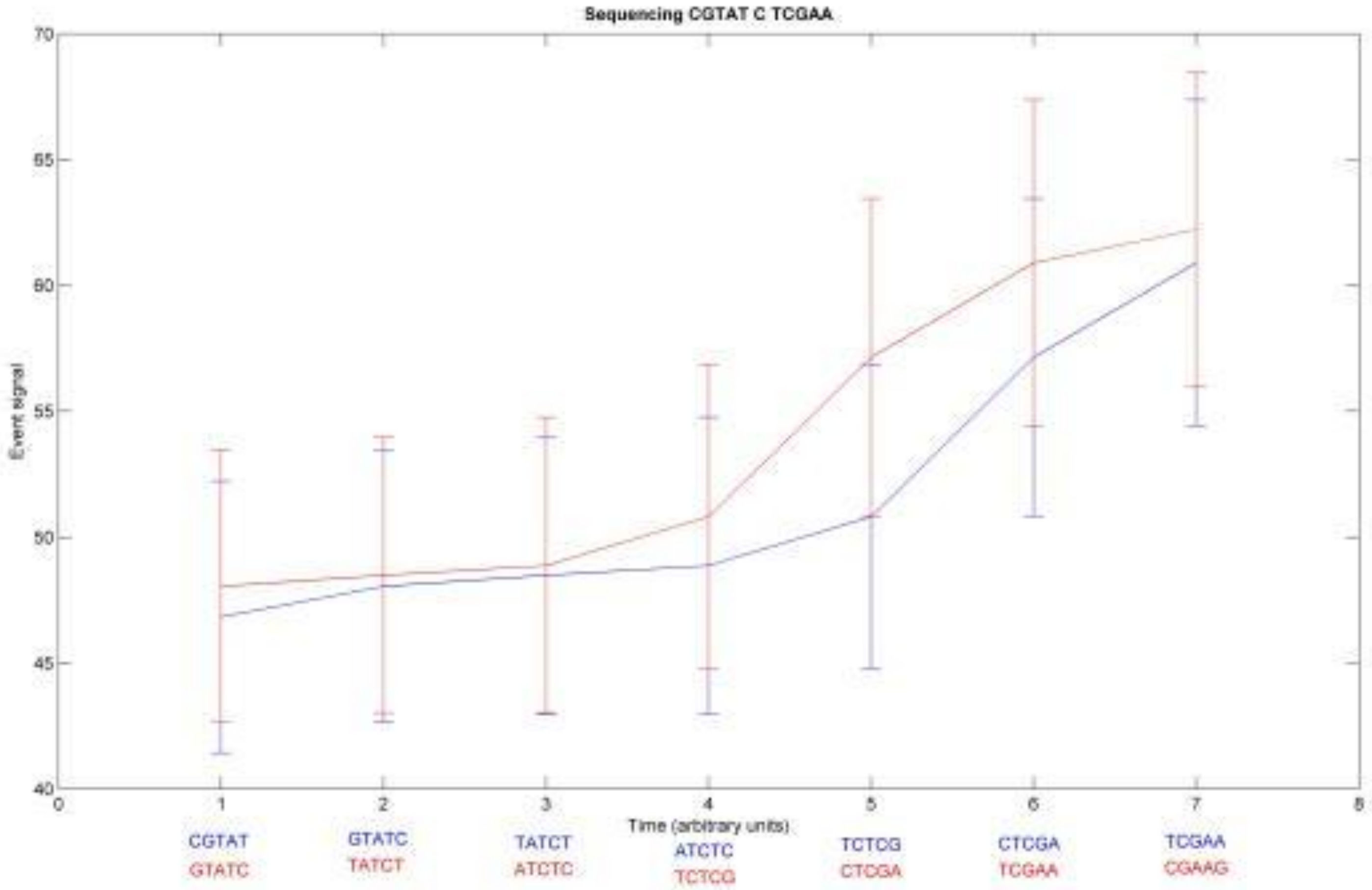
Challenge of 5-mer basecalling

- TTTTTATTTTT vs TTTTTCTTTTT



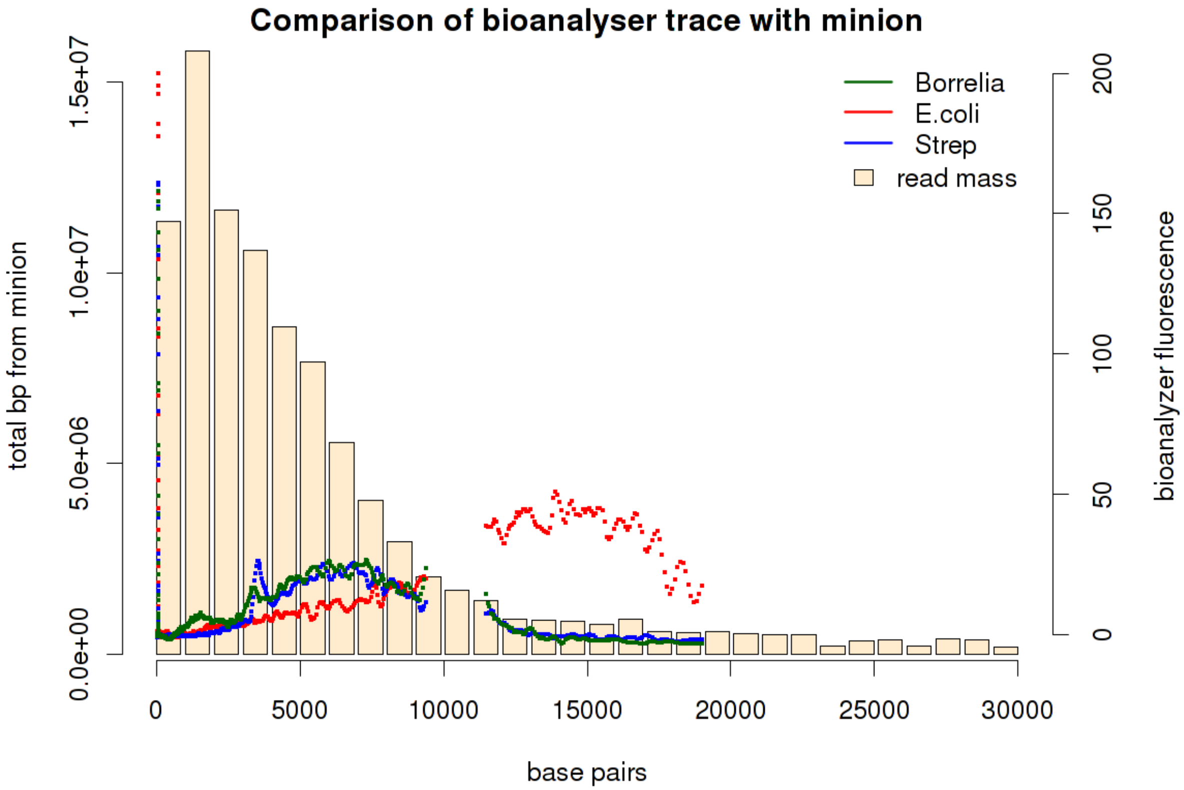
Challenge of 5-mer basecalling

- CGTATTCGAA vs CGTATCTCGAA

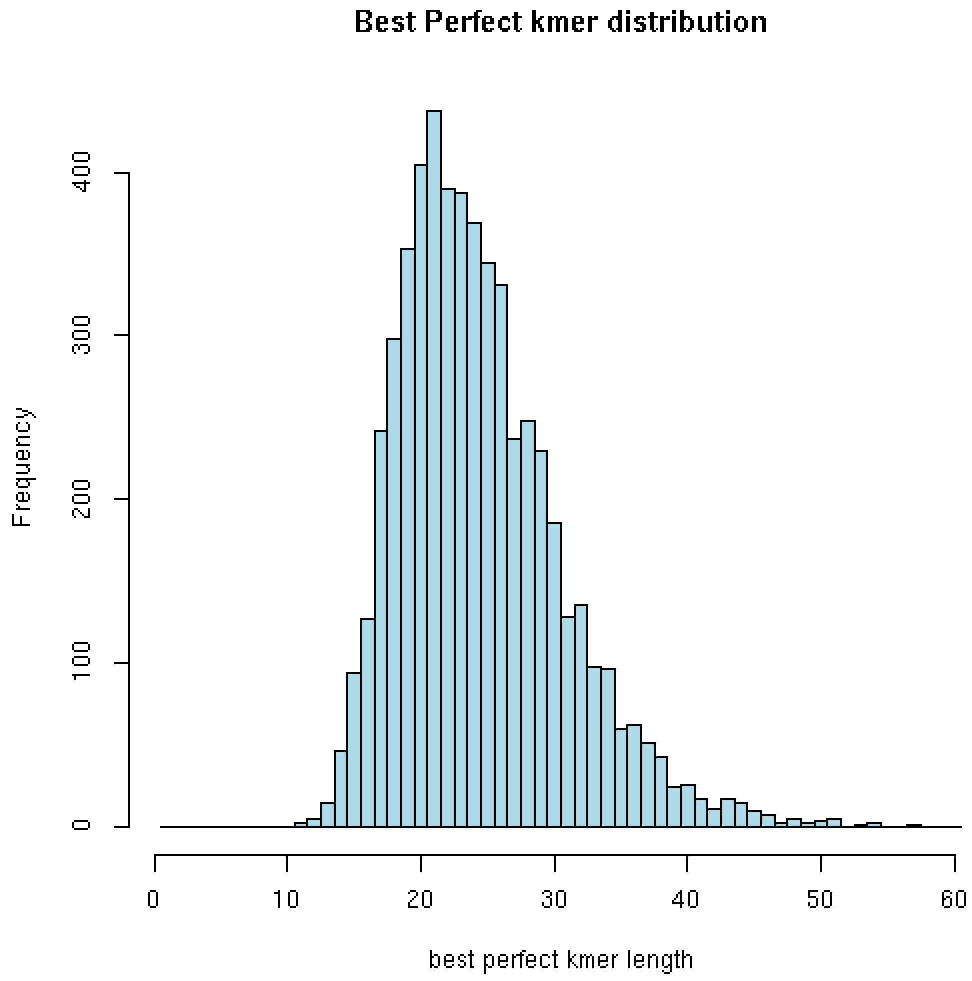


Read lengths

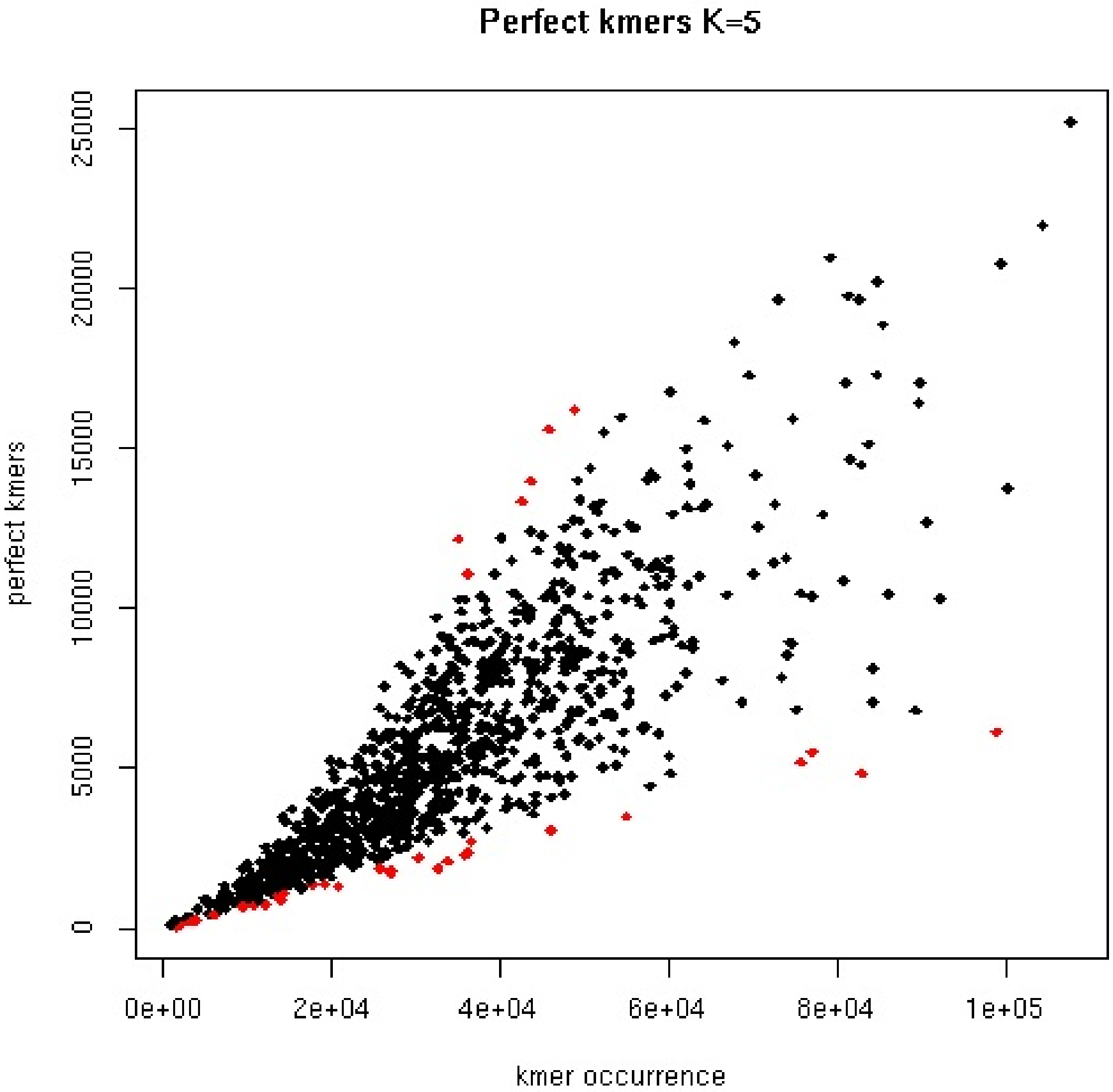
- Highly dependent on input DNA length
- Difficult to preserve DNA lengths



Longest perfect stretches



Hard to read motifs

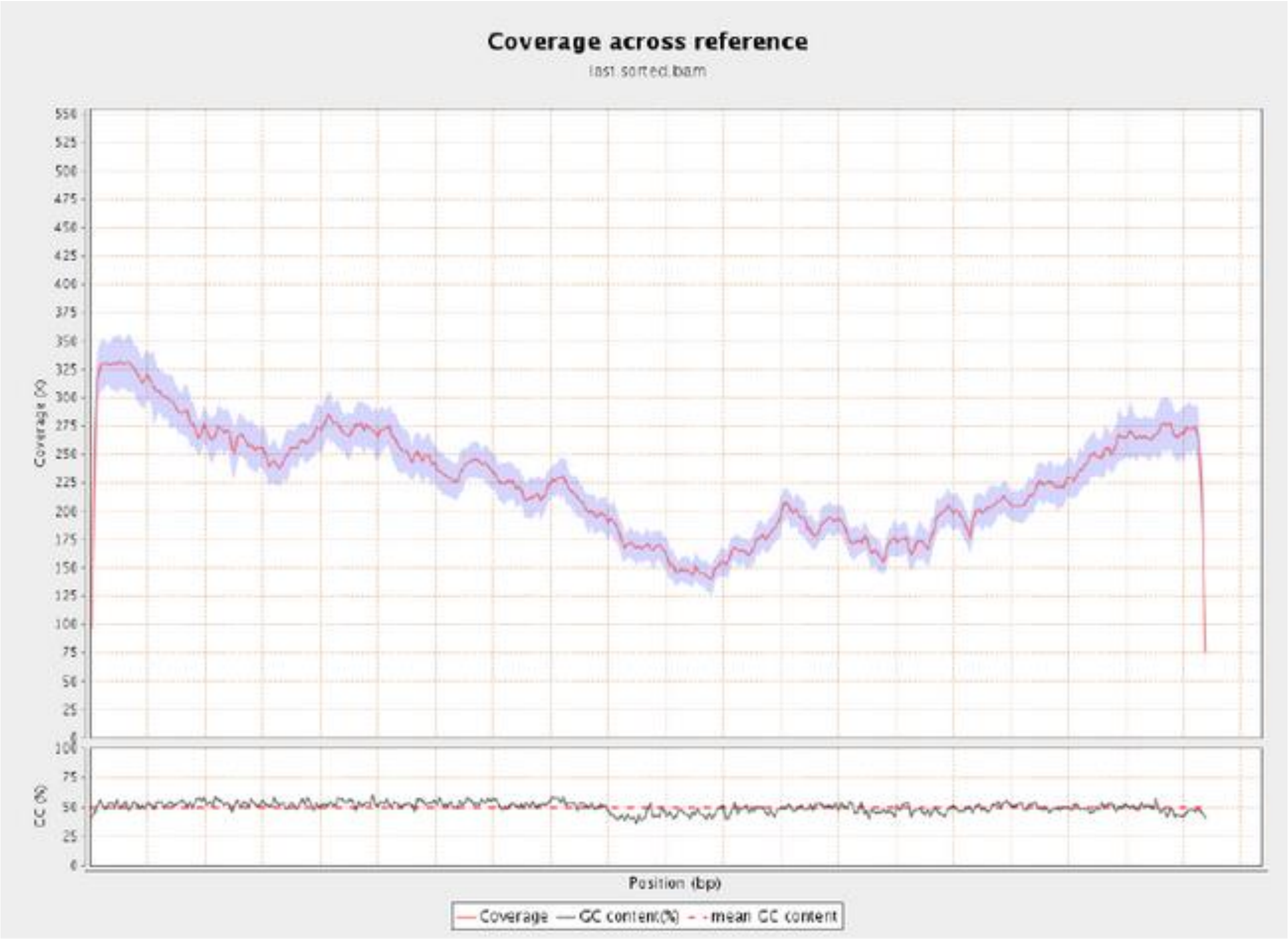


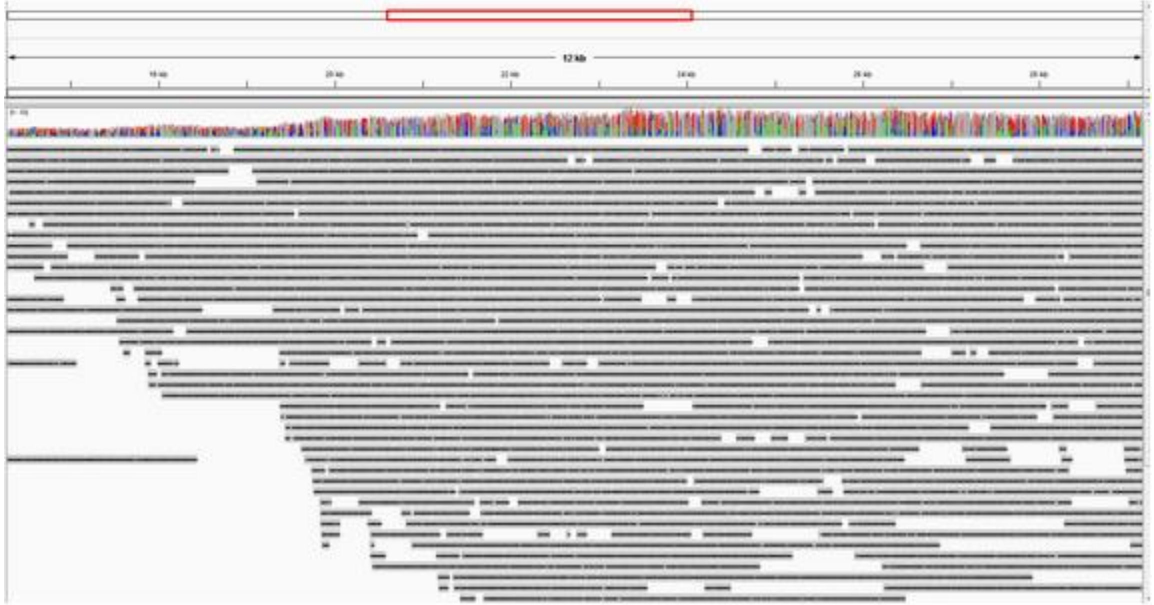
Hard Kmers
AAAAA
ACCTA
ACCTC
AGCGC
AGCTA
AGGTC

Easy Kmers
AAGAA
ACGAA
CGTTC
CTTTC
GAACG

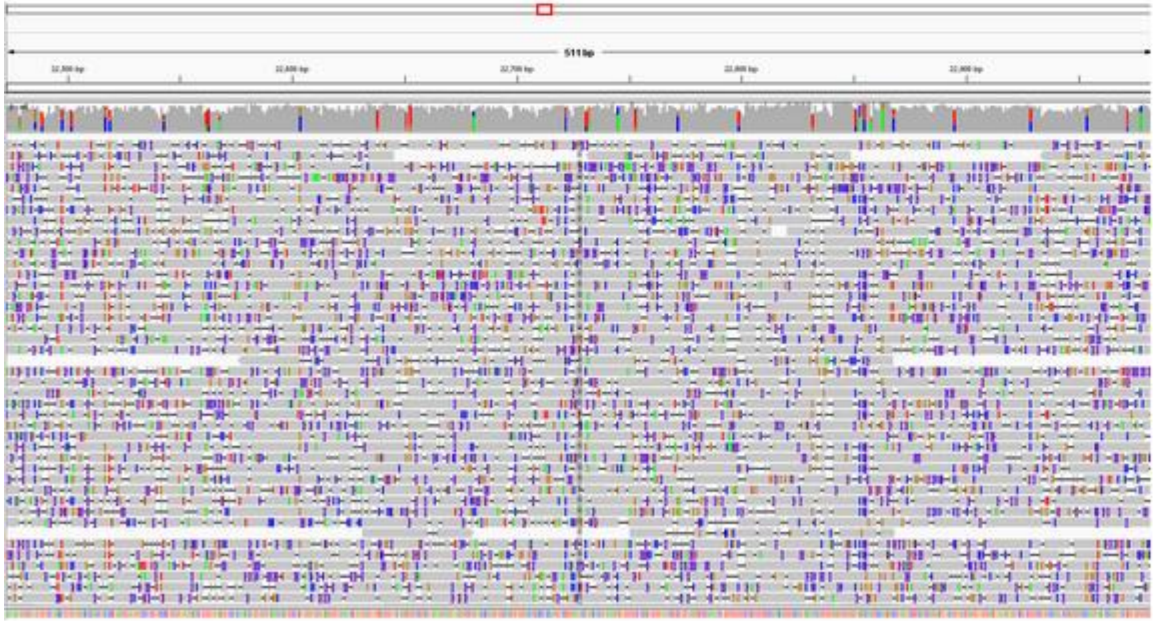
Examples and applications

Coverage of an E.coli genome





12 kb
18 kb
20 kb
22 kb
24 kb
26 kb
28 kb



511 bp
22,500 bp
22,600 bp
22,700 bp
22,800 bp
22,900 bp

Human genome data

- Are the PacBio instruments already obsolete?
  - http://www.opiniomics.org/is-the-long-read-sequencing-war-already-over/

- Initial mapping of MinION data vs PacBio data to mitochondrial genome by David Eccles:

  MinION — 2,386 reads, 25.4% mismatch
  Sequel — 2,272 reads, 8.6% mismatch

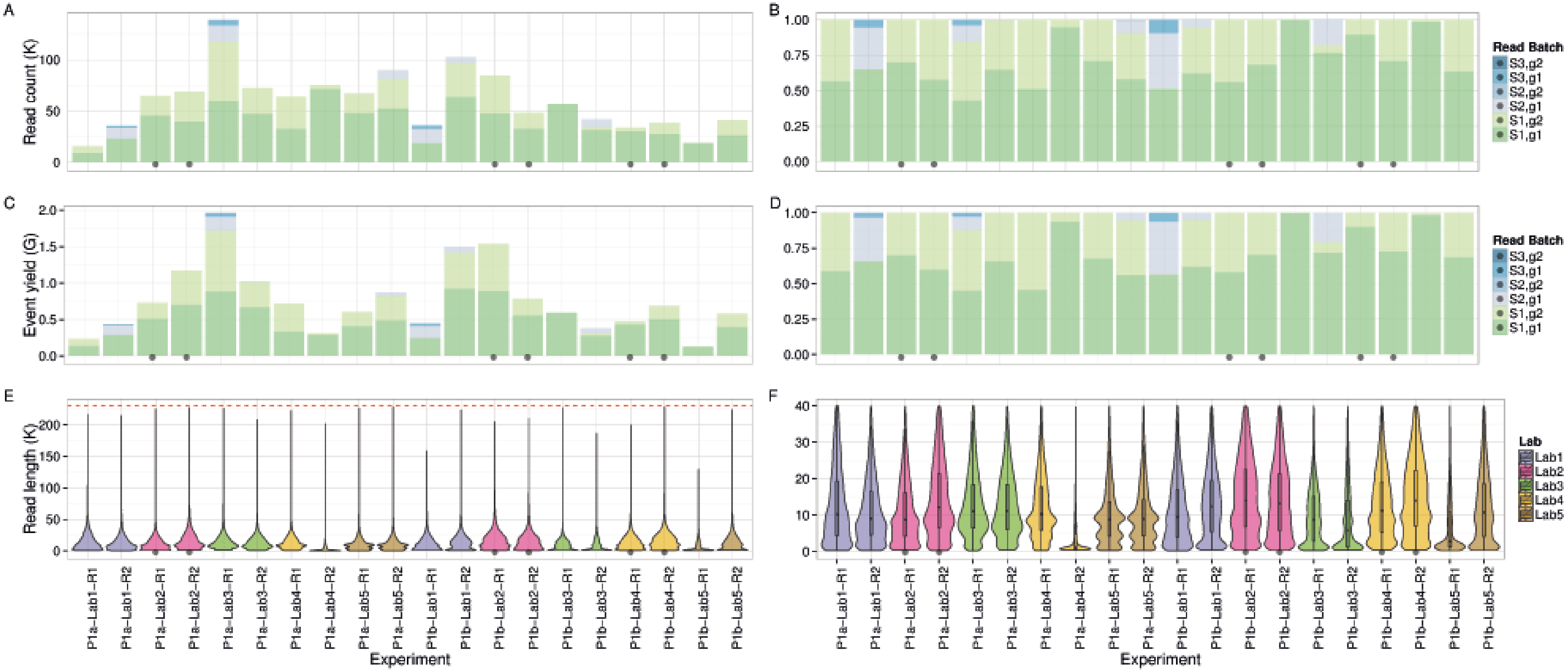
- MinION accuracy is poor but pace of nanopore change is extremely fast

- You will be able to compare MinION, RSII and Sequel datasets during the Genomics workshop

MinIon Analysis Reference Consortiom (MARC)

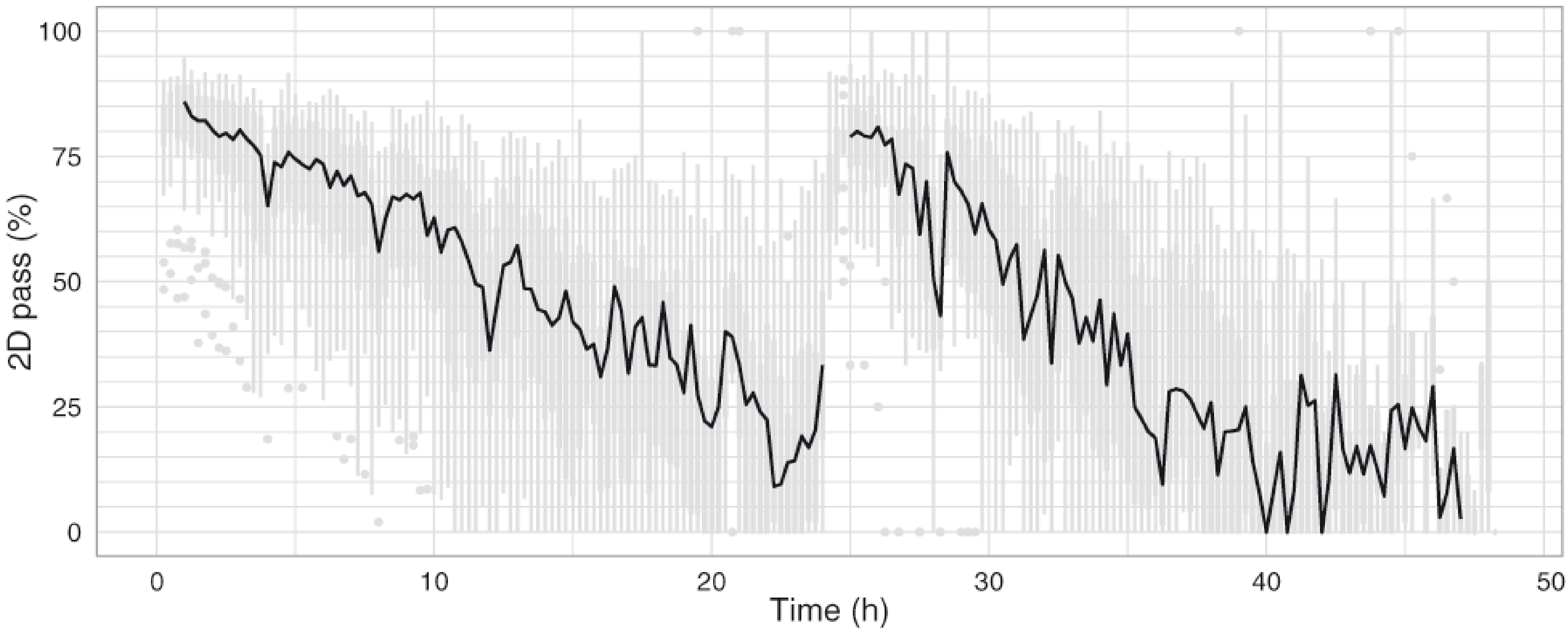
- Group of 20 labs evaluating MinIon performance using E.coli
- http://f1000research.com/articles/4-1075/v1

Yield of MinIon flowcells



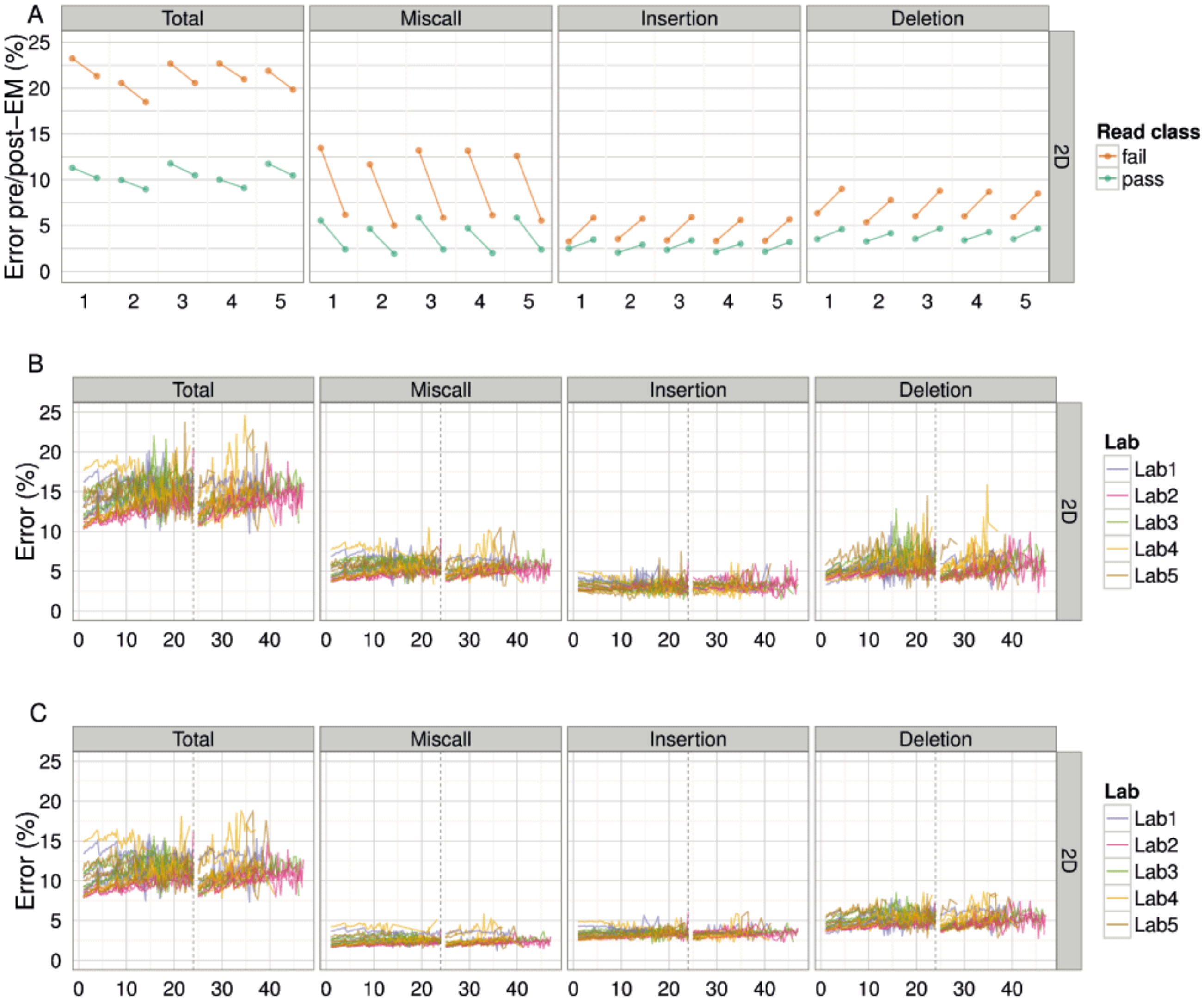
Ip CLC, Loose M, Tyson JR et al. 2015 [version 1; referees: 2 approved] F1000Research 2015, 4:1075 (doi: 10.12688/f1000research.7201.1)

Percentage of 2D pass reads produced over time.



Ip CLC, Loose M, Tyson JR et al. 2015 [version 1; referees: 2 approved] F1000Research 2015, 4:1075 (doi: 10.12688/f1000research.7201.1)

Error rates of BWA-MEM EM-corrected alignments of target 2D base-calls.

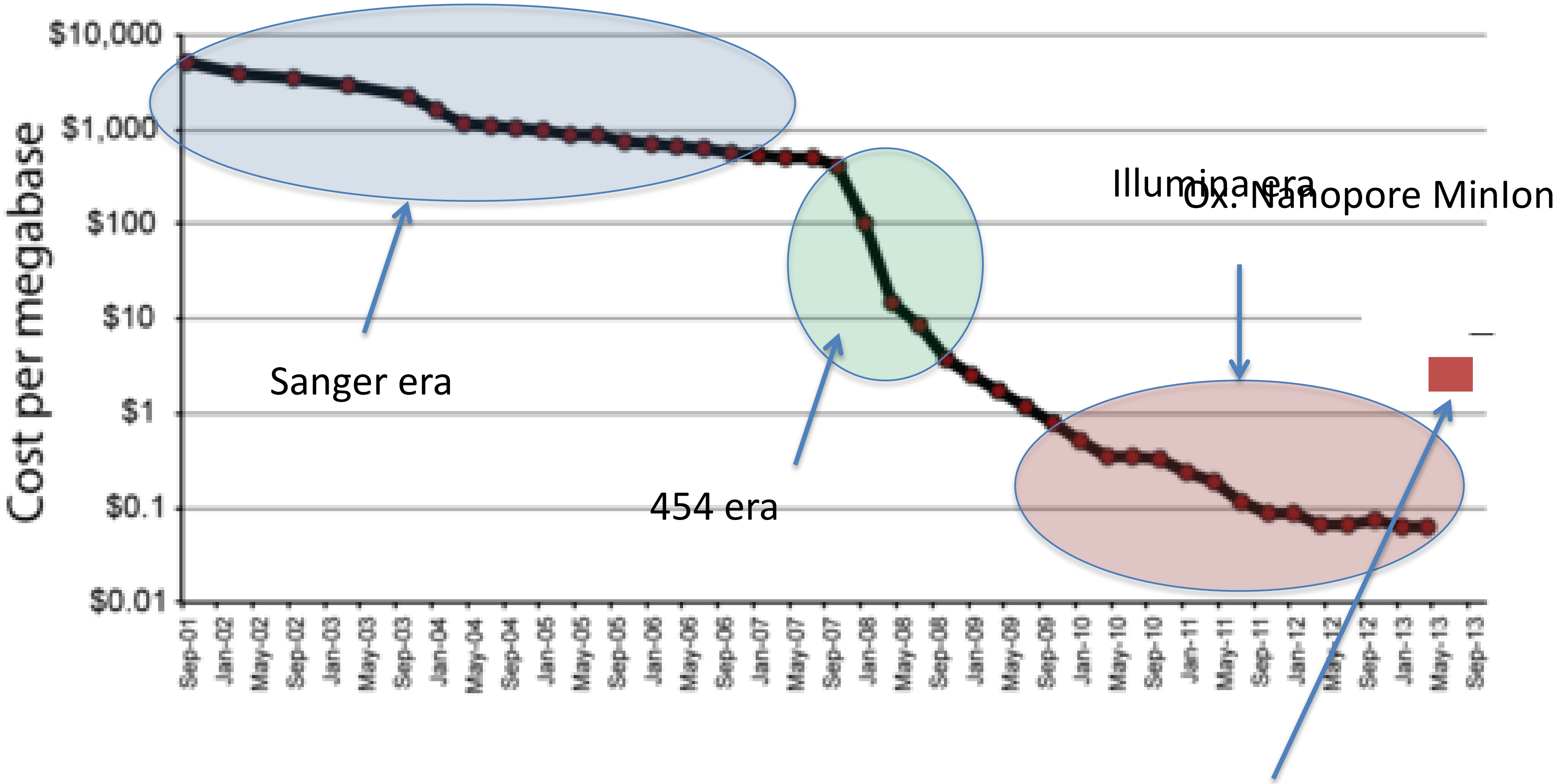


Ip CLC, Loose M, Tyson JR et al. 2015 [version 1; referees: 2 approved] F1000Research 2015, 4:1075 (doi: 10.12688/f1000research.7201.1)

Improvements

- FASTQ – per base quality values don’t make much sense
- Other types of model taking into account effects of bases sitting outside the pore
- 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)
  - Enable sequencing of RNA and perhaps enable protein sensing
- Library preparation

Cost per megabase



Illumina era

Ox. Nanopore MinIon

Sanger era

454 era

PacBio

http://wwwnc.cdc.gov/eid/article/22/2/15-1796_article

MinIon for denovo assembly and variant calling

- *De novo* sequencing and variant calling with nanopores using PoreSeq

- Oxford Nanopore sequencing, hybrid error correction, and de novo assembly of a eukaryotic genome

Novel applications

- Portable in-situ sequencing

- Nanopore sequencing as a rapidly deployable Ebola outbreak tool
- Nanopore sequencing in microgravity

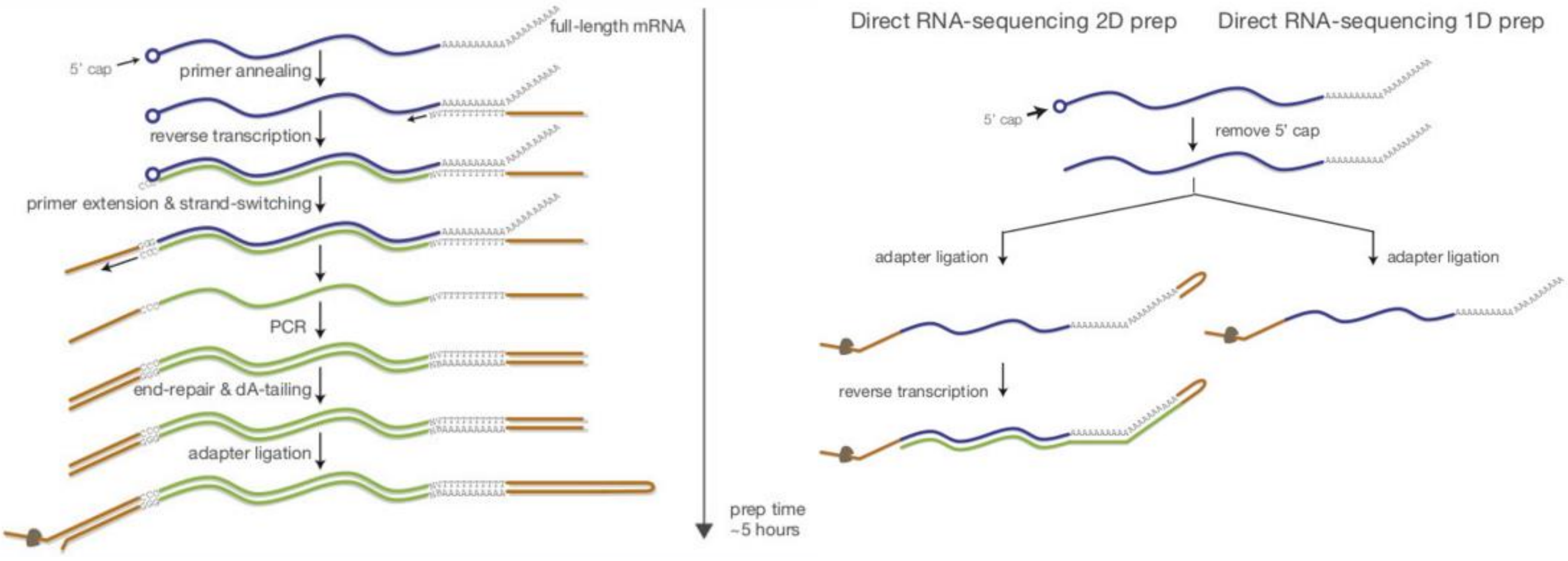
Now available... minipcr

Small, simple, accessible PCR
Portable unit you can take anywhere
Fits in the palm of your hands
Easy to use
Control via PC, Laptop or any Android device

Only
£499

Novel applications

- Direct RNA sequencing



Novel applications

- Rapid pathogen detection in clinical settings
  - Rapid identification of viral pathogens
  - Rapid draft sequencing of Salmonella during a hospital outbreak

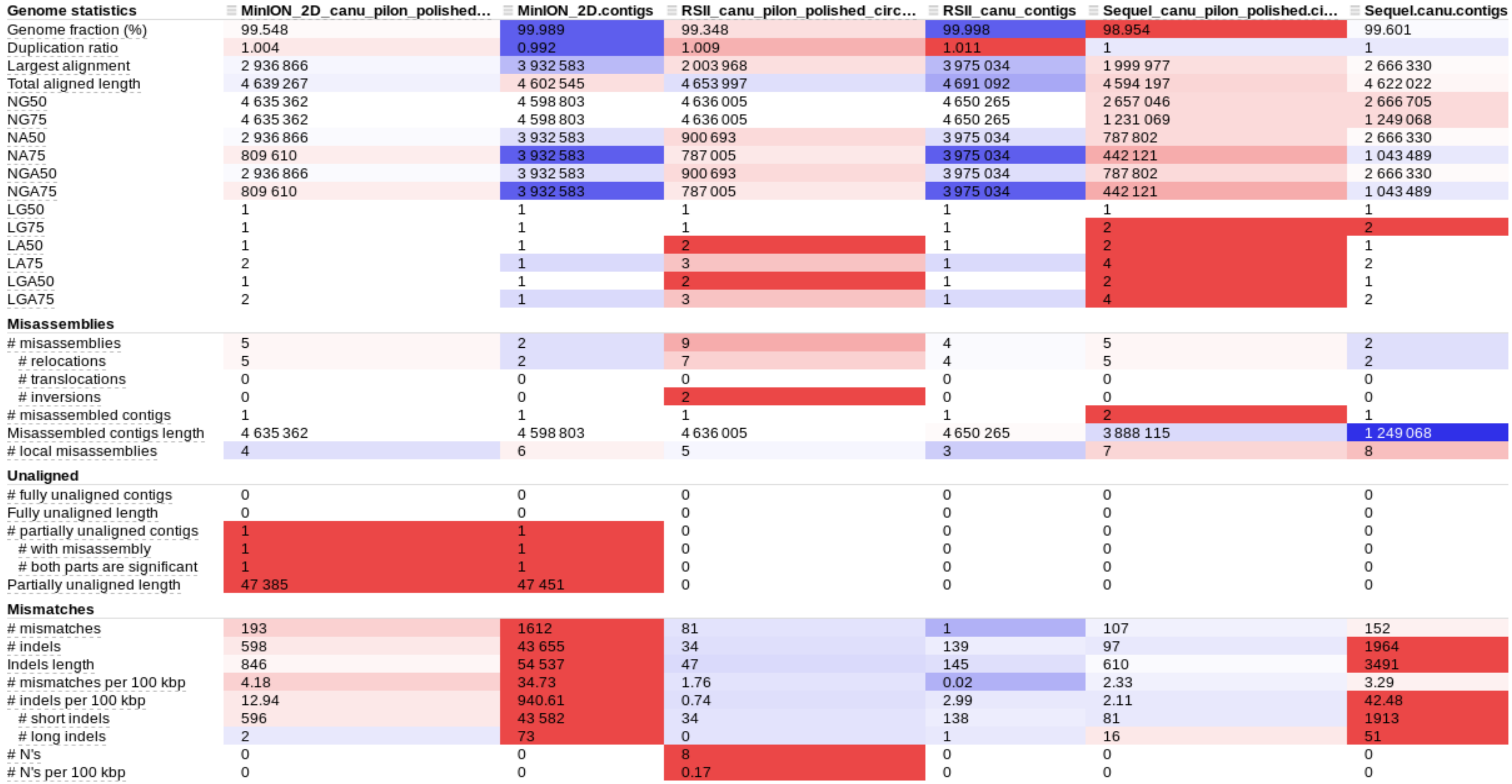
Novel applications

- Teaching aids
  - Integration of mobile sequencers into a classroom
- Sequencing as a sensor
- Portable in-situ sequencing
  - Variety of examples, but still hampered by DNA extraction

Exeter Porecamp 2016

Comparison of ONT and PacBio

- E.coli sequenced to >100x coverage on MinION, RSII and Sequel with and w/o polishing



Genome statistics	**MinION_2D_canu_pilon_polished...**	**MinION_2D.contigs**	**RSII_canu_pilon_polished_circ...**	**RSII_canu_contigs**	**Sequel_canu_pilon_polished.ci...**	**Sequel.canu.contigs**
Genome fraction (%)	99.548	99.989	99.348	99.998	98.954	99.601
Duplication ratio	1.004	0.992	1.009	1.011	1	1
Largest alignment	2 936 866	3 932 583	2 003 968	3 975 034	1 999 977	2 666 330
Total aligned length	4 639 267	4 602 545	4 653 997	4 691 092	4 594 197	4 622 022
NG50	4 635 362	4 598 803	4 636 005	4 650 265	2 657 046	2 666 705
NG75	4 635 362	4 598 803	4 636 005	4 650 265	1 231 069	1 249 068
NA50	2 936 866	3 932 583	900 693	3 975 034	787 802	2 666 330
NA75	809 610	3 932 583	787 005	3 975 034	442 121	1 043 489
NGA50	2 936 866	3 932 583	900 693	3 975 034	787 802	2 666 330
NGA75	809 610	3 932 583	787 005	3 975 034	442 121	1 043 489
LG50	1	1	1	1	1	1
LG75	1	1	1	1	2	2
LA50	1	1	2	1	2	1
LA75	2	1	3	1	4	2
LGA50	1	1	2	1	2	1
LGA75	2	1	3	1	4	2
Misassemblies						
# misassemblies	5	2	9	4	5	2
# relocations	5	2	7	4	5	2
# translocations	0	0	0	0	0	0
# inversions	0	0	2	0	0	0
# misassembled contigs	1	1	1	1	2	1
Misassembled contigs length	4 635 362	4 598 803	4 636 005	4 650 265	3 888 115	1 249 068
# local misassemblies	4	6	5	3	7	8
Unaligned						
# fully unaligned contigs	0	0	0	0	0	0
Fully unaligned length	0	0	0	0	0	0
# partially unaligned contigs	1	1	0	0	0	0
# with misassembly	1	1	0	0	0	0
# both parts are significant	1	1	0	0	0	0
Partially unaligned length	47 385	47 451	0	0	0	0
Mismatches						
# mismatches	193	1612	81	1	107	152
# indels	598	43 655	34	139	97	1964
Indels length	846	54 537	47	145	610	3491
# mismatches per 100 kbp	4.18	34.73	1.76	0.02	2.33	3.29
# indels per 100 kbp	12.94	940.61	0.74	2.99	2.11	42.48
# short indels	596	43 582	34	138	81	1913
# long indels	2	73	0	1	16	51
# N's	0	0	8	0	0	0
# N's per 100 kbp	0	0	0.17	0	0	0

Advantages of ONT vs PacBio

- Lower input amounts (ng possible)
- Less sensitivity to size distribution (although small DNA fragments will still sequence preferentially)
- Portable
- Zero capital costs
- Novel applications
- Longer read lengths (100s kb achieved)
- Higher yield (up to 10Gbases per flowcell)

Disadvantages of ONT vs PacBio

- Some material can be difficult to prepare, especially if bio-mass
- Difficult to QC libraries once made
- Higher cost per base (if capital costs are ignored)
- Non-random error profile
- Cannot read the same DNA molecule more than once

Opportunities

- Online ‘streaming’ bioinformatics
  - Analytics - one read at a time
- Developing complex sample -> sequencer ready protocols for use in the field
- Extracting/preserving long DNA fragments
- Identifying sources of bias
- Has the potential to replace established technologies and give us access to lots of long reads

Software packages

- Tracking and managing MinIon data
  - MinoTour - http://minotour.github.io/minoTour
- Processing ONT data
  - Poretools https://github.com/arq5x/poretools
  - poRe http://sourceforge.net/projects/rpore/
  - NanoCorr http://schatzlab.cshl.edu/data/nanocorr/
  - Nanopolish https://github.com/jts/nanopolish/
  - PoreSeq - http://www.nature.com/nbt/journal/v33/n10/full/nbt.3360.html
  - MarginAlign - https://github.com/benedictpaten/marginAlign
  - Lordec https://www.gatb.fr/software/lordec/
- Alignment

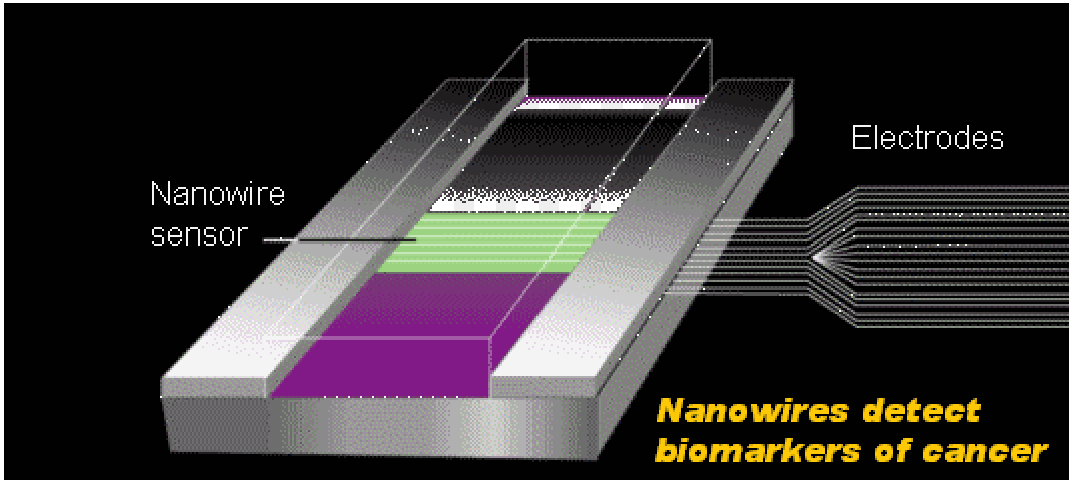
  Sensitive but slow aligners
    - BLAST
    - BLASR
    - LAST
    - BWA with the right parameters
- Assembly
  - Possible to obtain ONT-only assembly
  - Error correction with other data types is also possible
  - Offer an appealing and affordable alternative to PacBio or Illumina synthetic long reads

Beyond single nanopores?

- Single base-pair resolution is not available
  - Typically 4-5 nucleotides have to be measured simultaneously
- Only one detector per DNA strand
- Fast translocation of DNA through pore
- Small signal and high noise
- Bilayer stability

Nanowire alternatives

- QuantumDx QSEQ



Others in development

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

In conclusion

- We are mastering reading DNA (at least some of it)
- Now we are in a position to precisely edit and engineer biological systems

OUTLOOK
Vaccines

nature

THE INTERNATIONAL WEEKLY JOURNAL OF SCIENCE

SEEK AND DESTROY

PAM sequences recruit and activate Cas9 endonuclease in CRISPR/Cas9 genome-editing tool PAGE 62

CHEMISTRY
TAKING CHARGE
Energy storage beyond the lithium-ion battery
PAGE 26

COSMOLOGY
STILL IN THE DARK
Where next in the search for dark matter?
PAGE 29

DEVICE PHYSICS
SEE RADIO IN A NEW LIGHT
Laser signals from weak radio waves
PAGES 45 & 81

NATURE.COM/NATURE
6 March 2014 £10
Vol. 507, No. 7490

Thanks to:

Karen Moore

Jeremie Poschmann

Audrey Farbos

Paul O’Neill

Wellcome Trust

Contact me:

k.h.paszkiewicz@exeter.ac.uk

http://sequencing.exeter.ac.uk

Supported by wellcome trust

UNIVERSITY OF EXETER