edinburgh genomics.

Introduction to Next-Generation Sequencing Technologies

Javier Santoyo-Lopez

javier.santoyo@ed.ac.uk

22nd October 2015

Outline

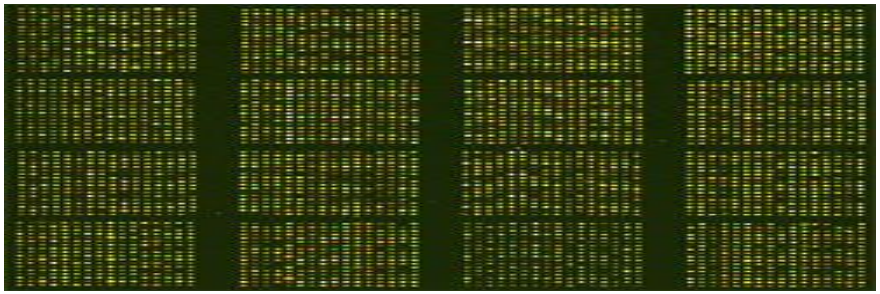
- Introduction to high-throughput technologies
- Description of the NGS technologies
- Anatomy of an NGS library
- NGS for variant calling

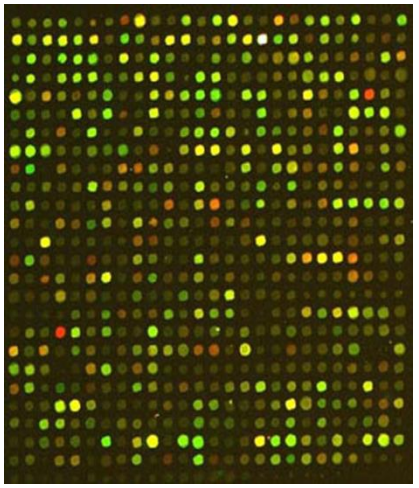
Introduction to high-throughput technologies

First HT Technology

edinburgh genomics.

- 1988 arrayed DNAs were used
- 1991 oligonucleotides are synthesized on a glass slide through photolithography (Affymax Research Institute)
- 1995 DNA Microarrays
- 1997 Genome wide Yeast Microarray

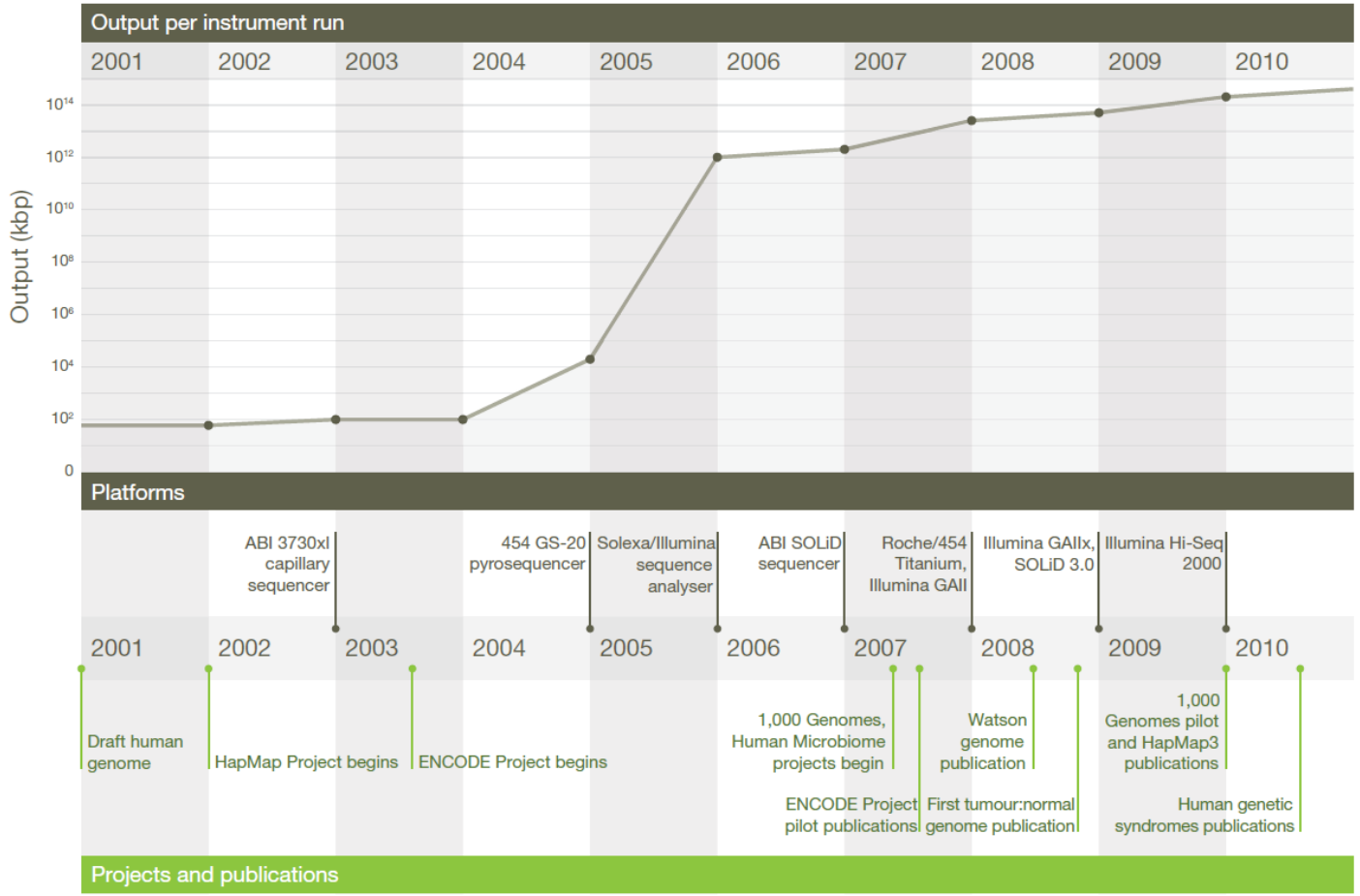




Milestone of DNA Technologies

Projects & sequence output

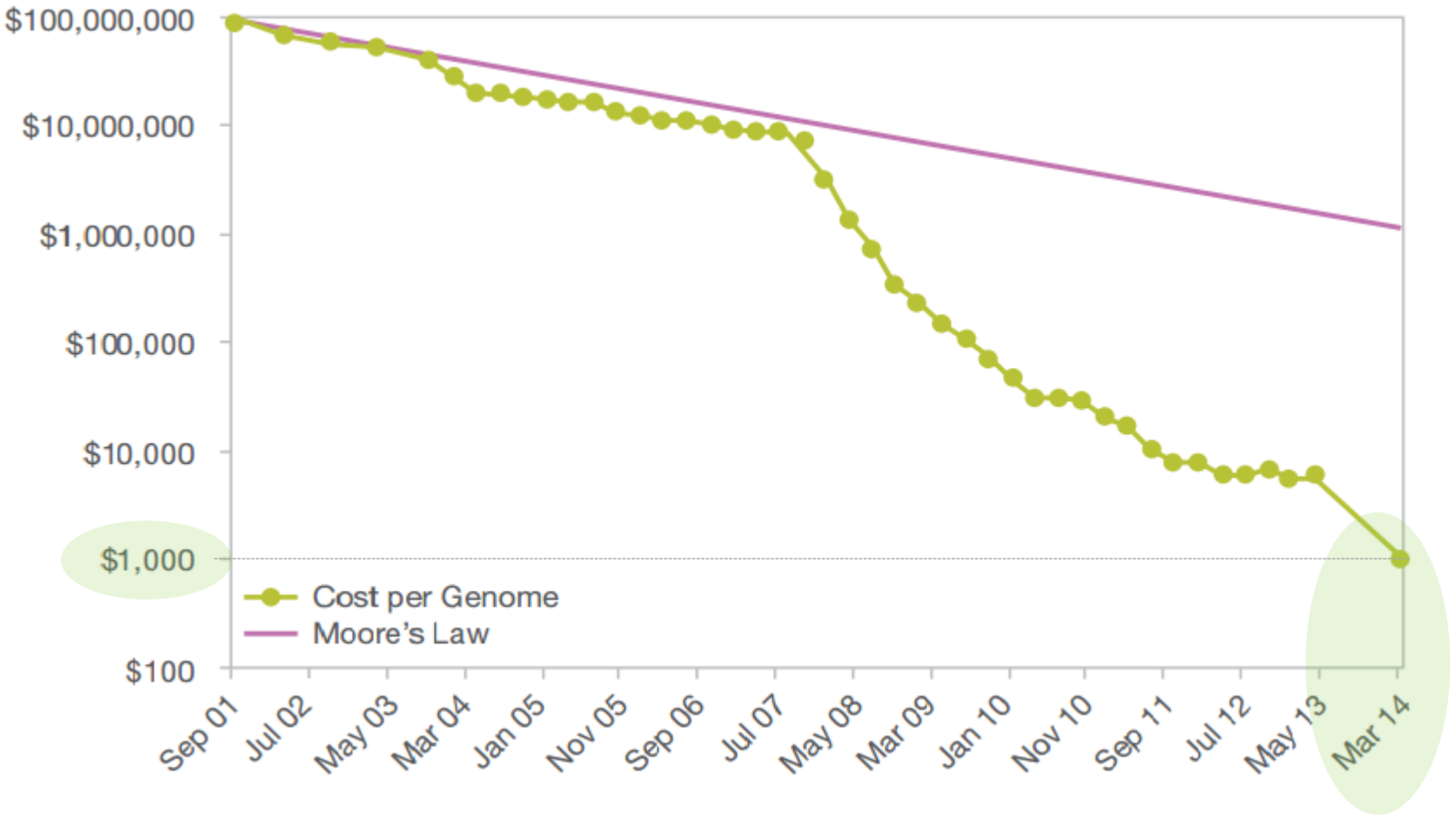
edinburgh genomics.



E.R. Mardis, Nature (2011) 470:198 - 203

Genome Sequencing Cost per Mb (30x)

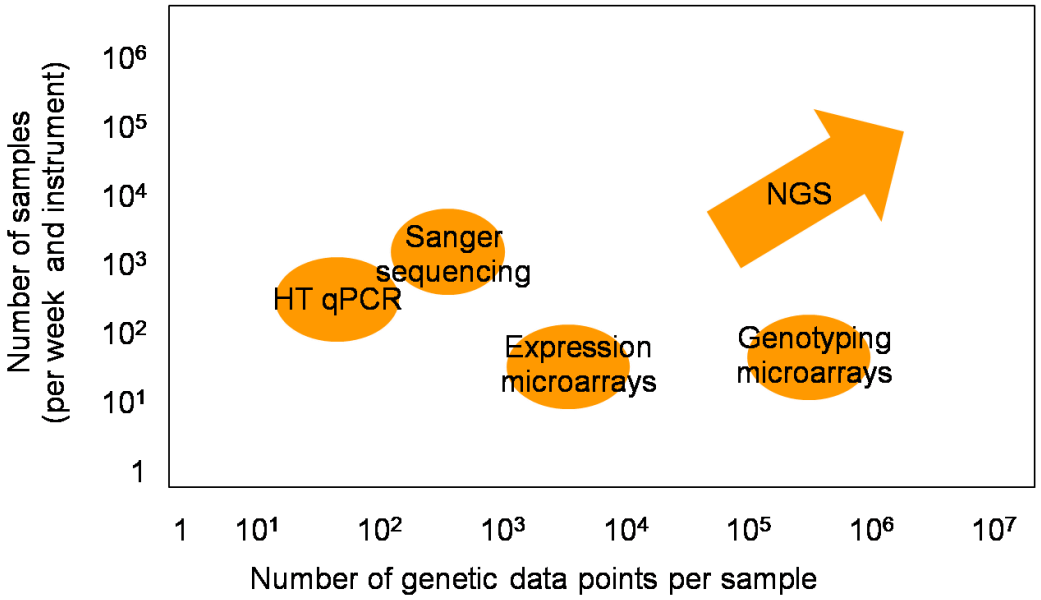
edinburgh genomics.



Relative throughput of HTT

edinburgh genomics.

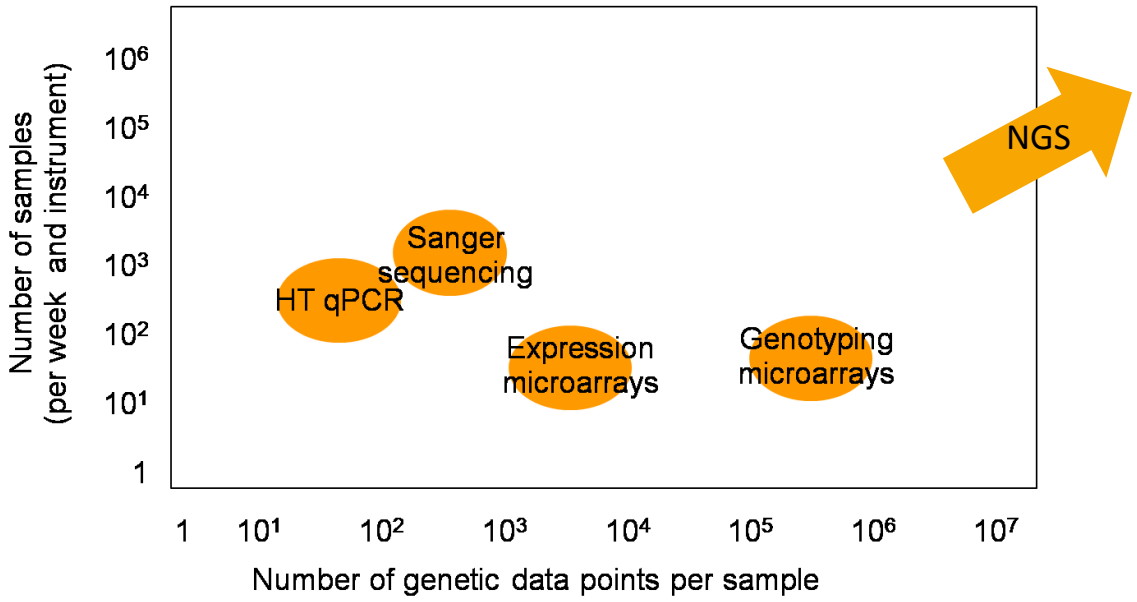
Next Generation Sequencing emerges with a potential of data production that will, eventually wipe out conventional HT technologies in the years coming



Relative throughput of HTT

edinburgh genomics.

Next Generation Sequencing emerges with a potential of data production that will, eventually wipe out conventional HT technologies in the years coming



NGS: Too many sequences to be handled in standard hardware

Many Gbs (Tbs) of Sequences

edinburgh genomics.

- Data management becomes a challenge.
  - Moving data across file systems takes time (several hundred Gbs)

- What quality does the raw data have?
  - Sequencers provide quality values for each bp

- How to do Analysis of the DATA
  - Primary data analysis (QC)
  - Secondary data analysis

Description of NGS technologies

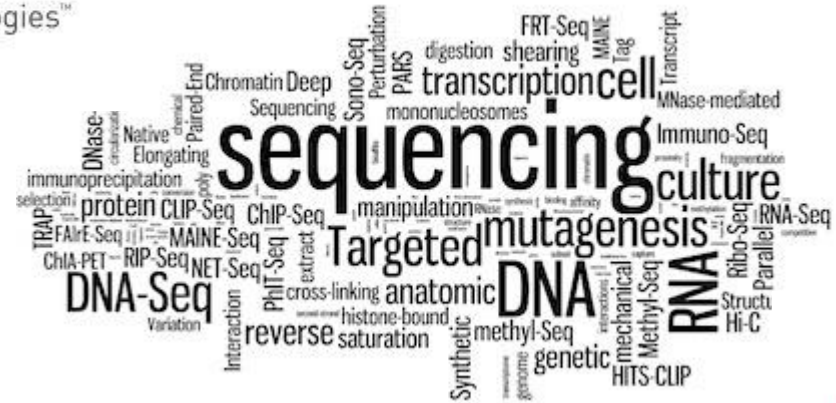
NGS Technologies

edinburgh genomics.

illumina®

applied biosystems™
part of life technologies™

Roche
454
SEQUENCING



ion torrent
by life technologies™

PACIFIC BIOSCIENCES™

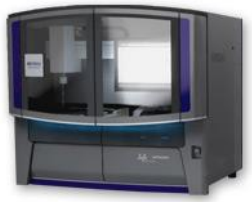
Oxford NANOPORE Technologies

NGS sequencers

edinburgh genomics.

Roche 454 FLX+

Illumina GAIIx



Life Tech SOLID 5500

Life Tech Ion Torrent

Helicos Heliscope

Roche 454 Junior

Illumina MiSeq

NextSeq

Illumina HiSeq

Life Tech Ion Proton

Pacific Biosciences RS

Oxford Nanopore GridIon

Oxford Nanopore MinION

Oxford Nanopore PromethION

Complete Genomics Revolocity

PacBio Sequel

NGS sequencers

edinburgh genomics.

Roche 454 FLX+

Illumina GAIIx

Life Tech SOLID 5500

Life Tech Ion Torrent

Helicos Heliscope

Roche 454 Junior

Illumina MiSeq

NextSeq

Illumina HiSeq

Life Tech Ion Proton

Pacific Biosciences RS

Oxford Nanopore GridIon

Oxford Nanopore MinION

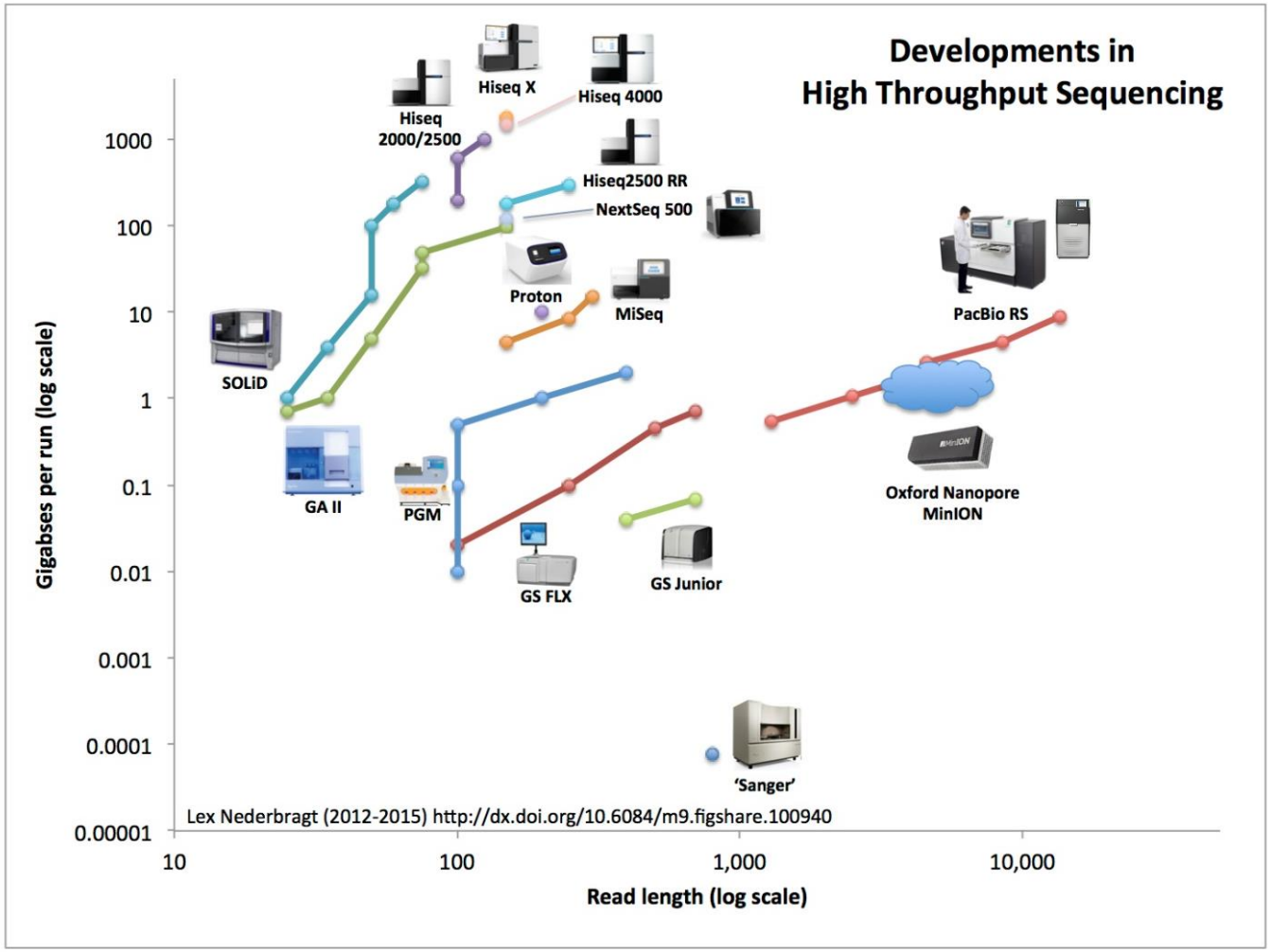
Oxford Nanopore PromethION

Complete Genomics Revolocity

PacBio Sequel

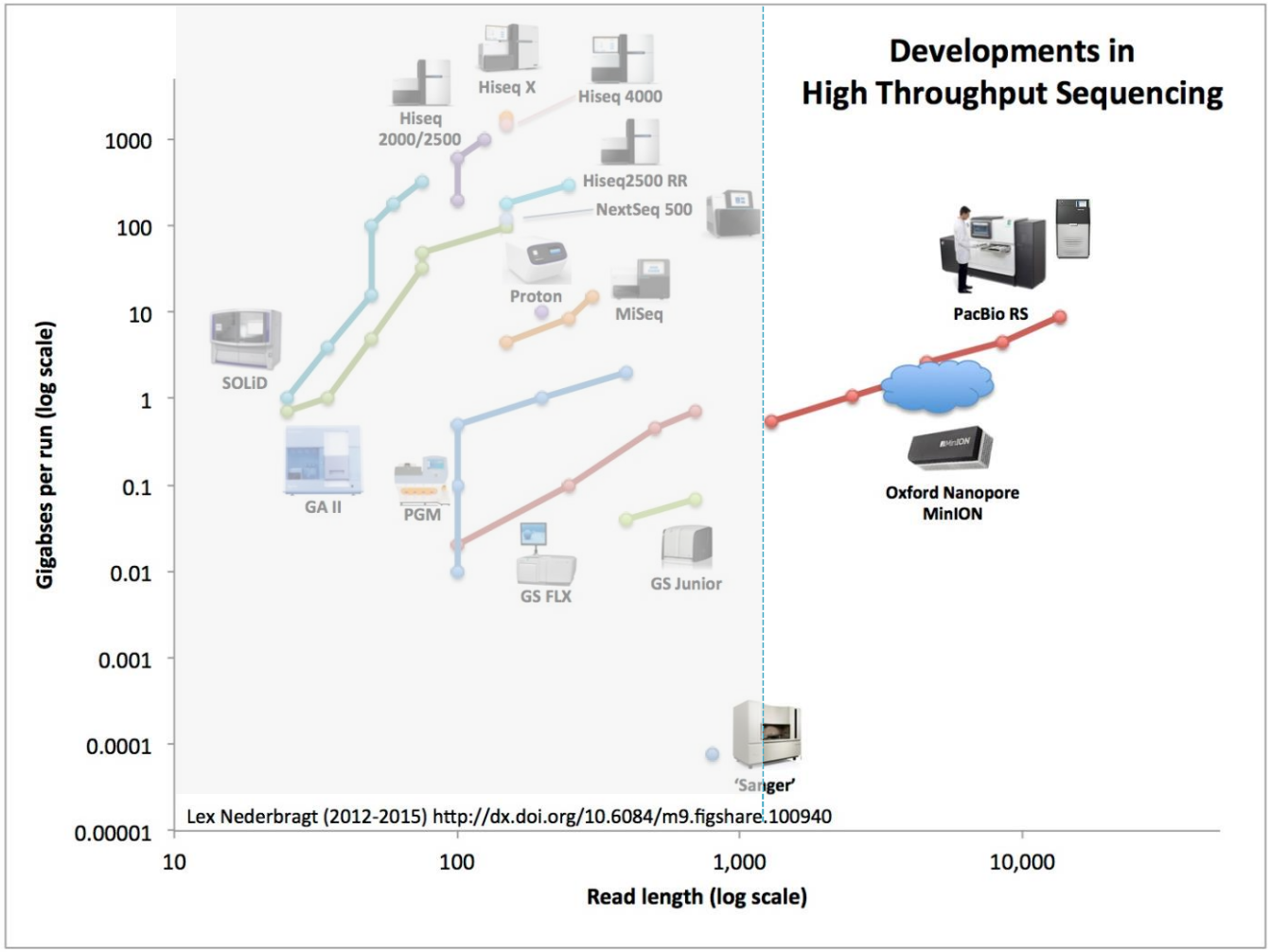
Length & throughput

edinburgh genomics.



Length & throughput

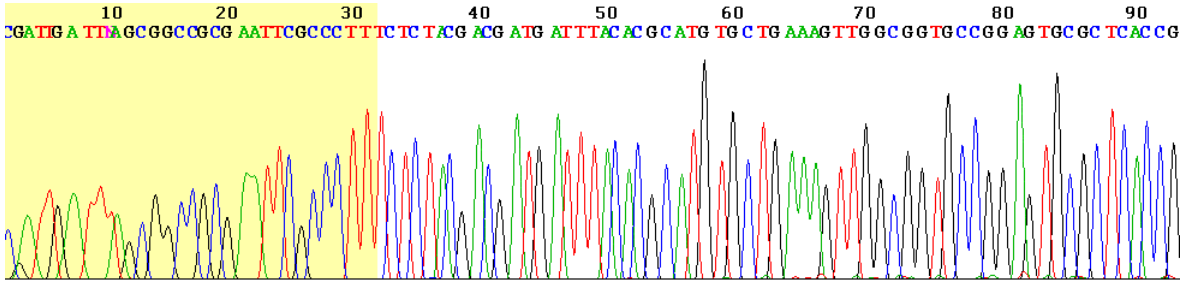
edinburgh genomics.

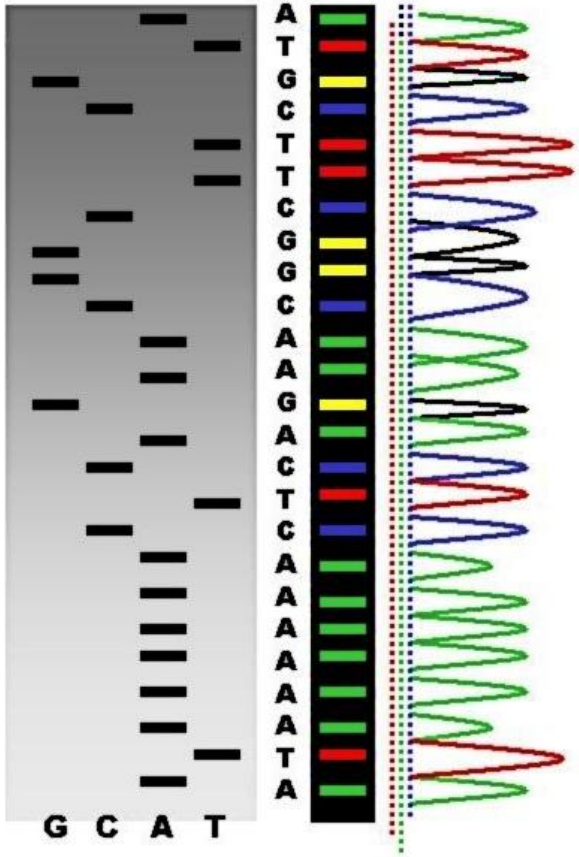


Basics of Sanger Sequencing

edinburgh genomics.

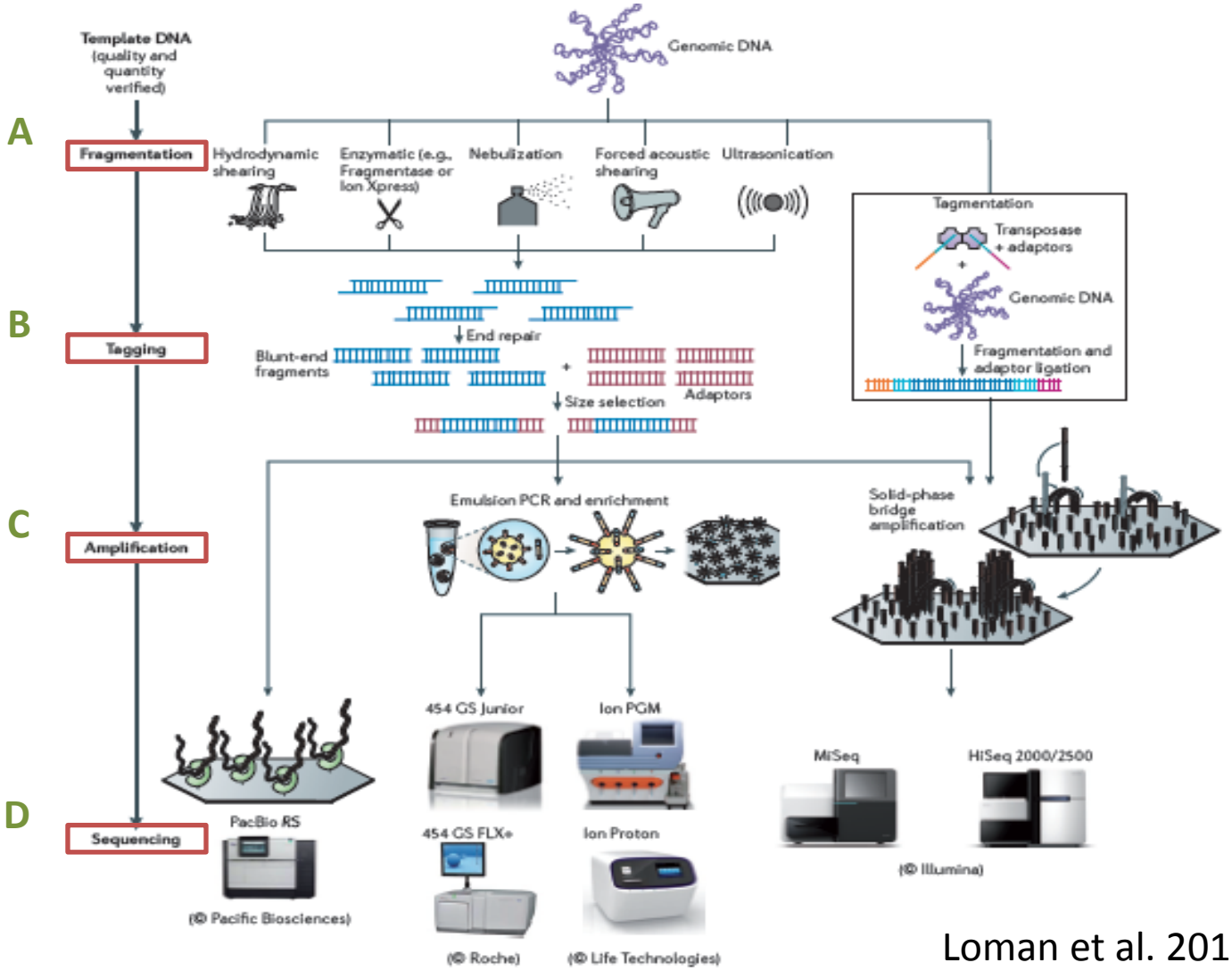
- Clone the DNA.
- Generate a ladder of labeled (colored) molecules that are different by 1 nucleotide.
- Separate mixture on some matrix.
- Detect fluoroscope by laser.
- Interpret peaks as string of DNA.
- Strings are 500 to 1,000 letters long
- 1 machine generates 57,000 nucleotides/run





Basics of NGS Sequencing

edinburgh genomics.



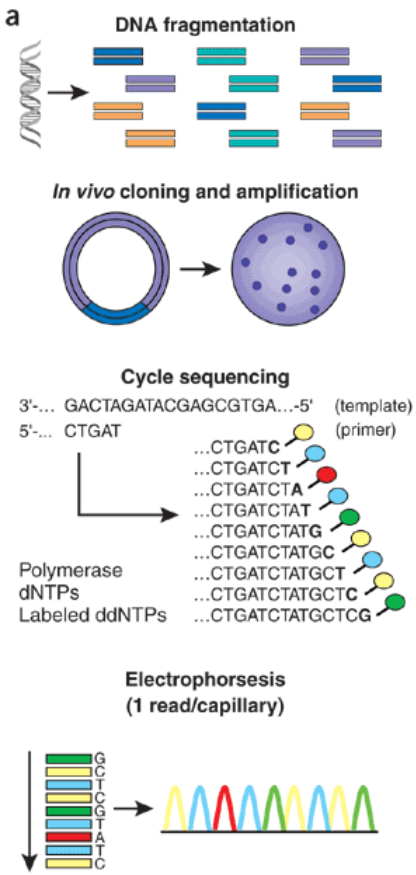
Loman et al. 2012

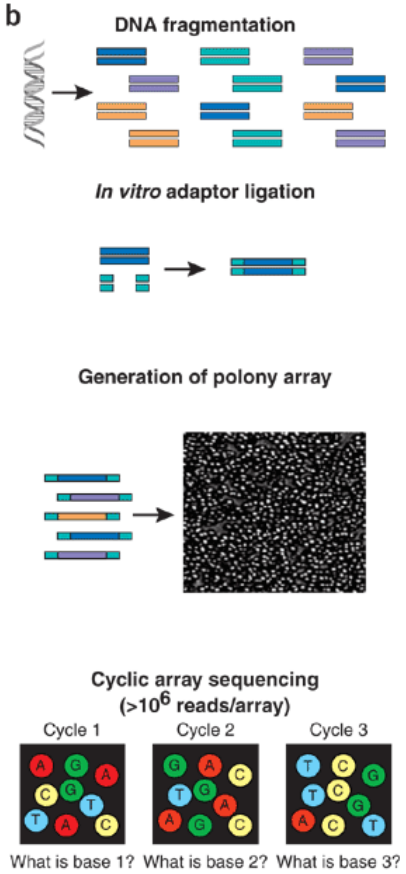
Comparison of Technologies

edinburgh genomics.

Sanger
Max Output
57 Kb run (1h)

NGS
Max Output
1,800 Gb run (3.5 days)





Jay Shendure & Hanlee Ji. Nature Biotechnology 26, 1135 - 1145 (2008)

Illumina HiSeq/MiSeq

Illumina (formerly Solexa)

edinburgh genomics.

Over 90% of all sequencing data is produced on Illumina systems.

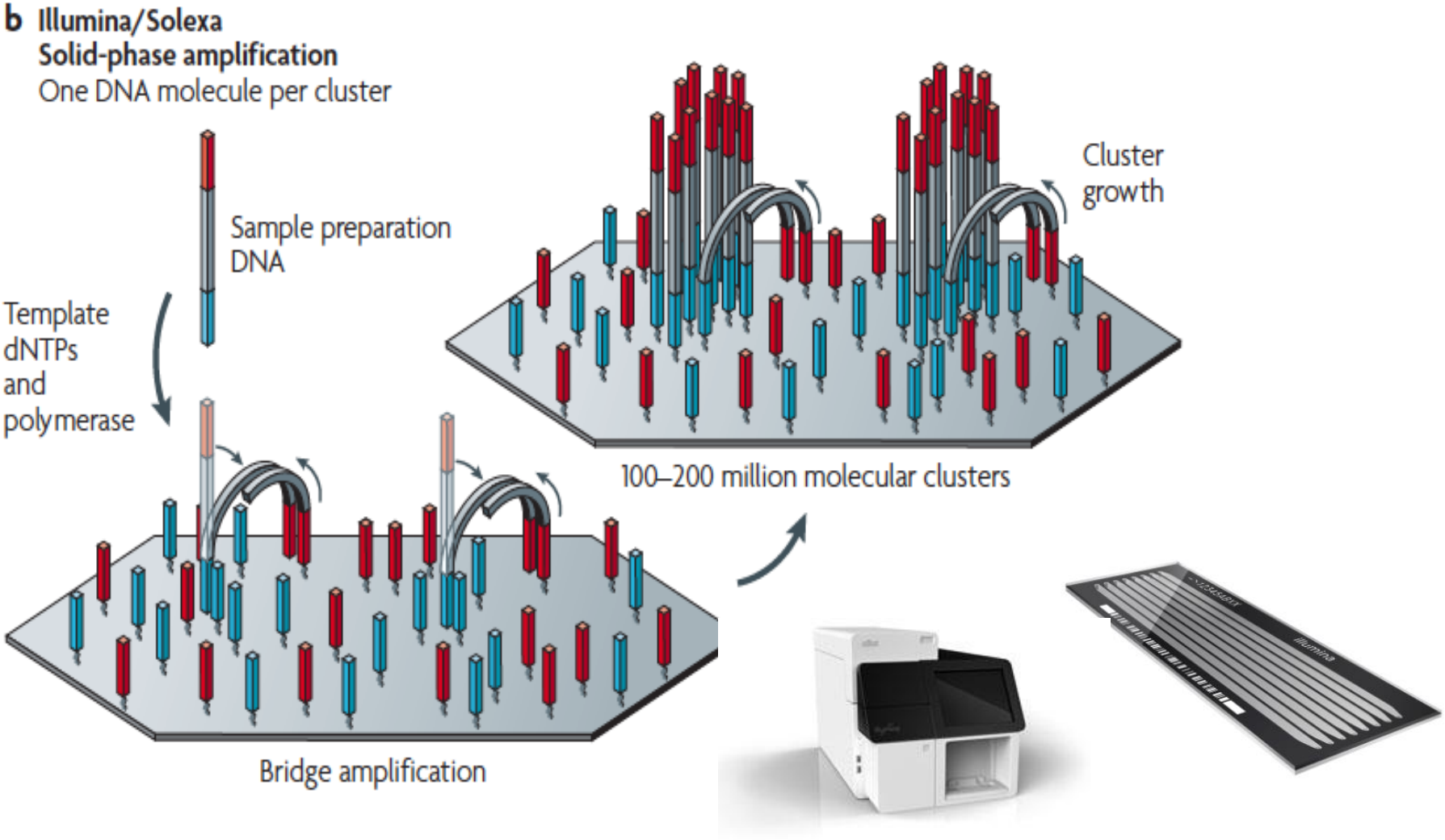
Uses a "sequencing by synthesis" approach:

- **Library**: DNA is broken into small fragments and ligated to adaptors.
- **Amplification**: The fragments are attached to the surface of a flow cell and amplified.
- **Sequencing**: DNA is sequenced by adding polymerase and labeled reversible terminator nucleotides (each base with a different color).
  - The incorporated base is determined by fluorescence.
  - The fluorescent label is removed from the terminator and the 3' OH is unblocked, allowing a new base to be incorporated

- Started with 35 bp, increased now to up to 300 bp
- One run can give up to 10-1,800 Gb, 300-6000 million paired-end reads
- 75-85% of bases at or above Q30
- Substitution errors

PCR bridge amplification

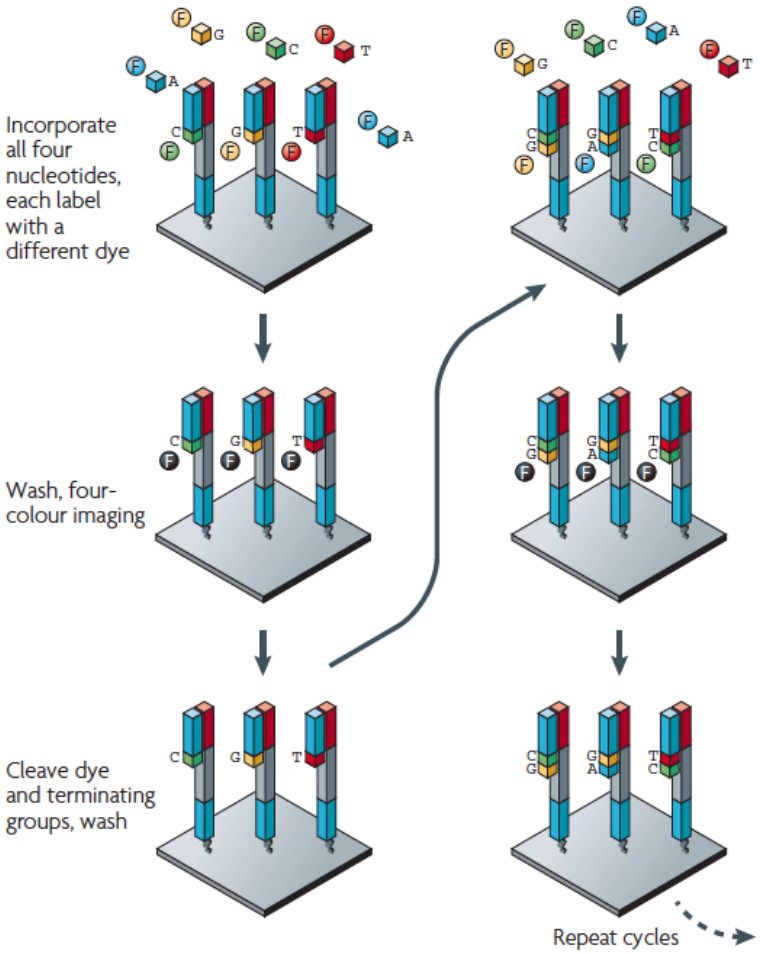
edinburgh genomics.



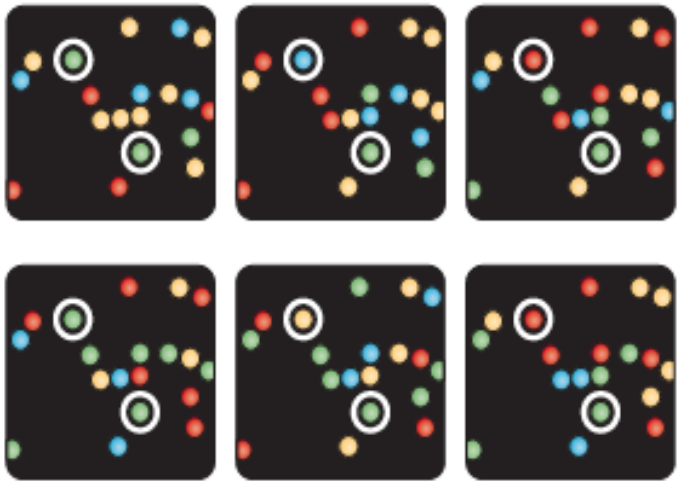
Adapted from Metzker 2010

Sequencing-by-sythesis

edinburgh genomics.



From Michael Metzker, http://view.ncbi.nlm.nih.gov/pubmed/19997069



Top: CATCGT
Bottom: CCCCCC

- Fixed length reads
- Reversible terminators
- The identity of each base of a cluster is read off from **sequential images**
- Illumina sequencing video

Illumina Sequencers

edinburgh genomics.

MiSeq

Max Output	Max Read Number	Max Read Length
15 Gb	25 M	2x300 bp

NextSeq 500/550

Max Output	Max Read Number	Max Read Length
120 Gb	400 M	2x150 bp

www.illumina.com

Illumina Sequencers

edinburgh genomics.

HiSeq 2500*/3000/4000

* Max Output	Max Read Number	Max Read Length
1,000* Gb	4,000* M	2x125* bp

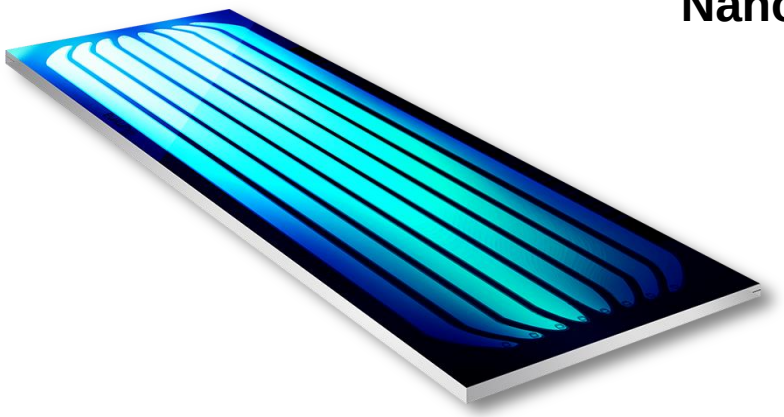
HiSeq X Ten/ X Five

Max Output	Max Read Number	Max Read Length
1,800 Gb	6,000 M	2x150 bp

www.illumina.com

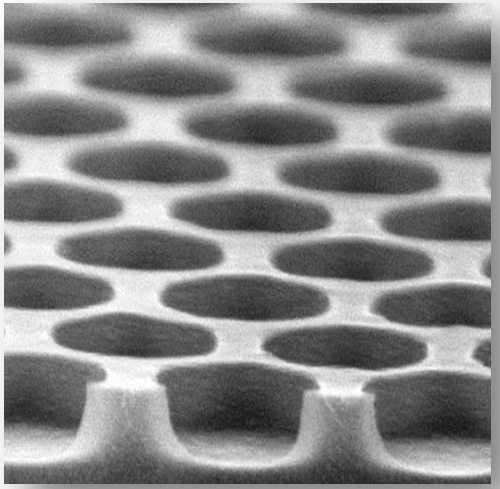
Patterned flow cells

edinburgh genomics.

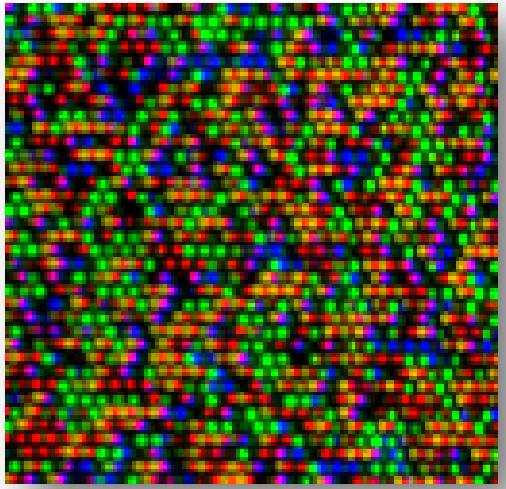


Nanowell substrate | billions of ordered wells

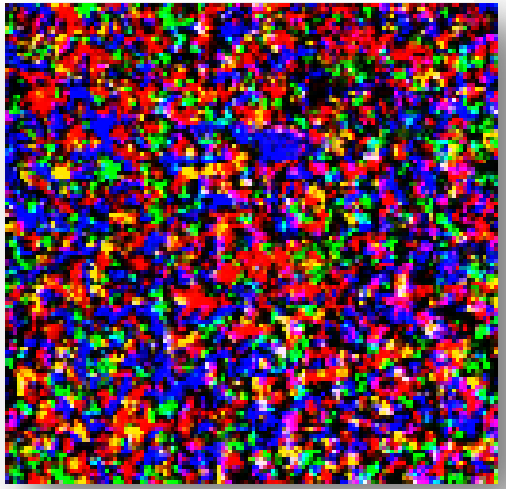
- Defined feature size
- Optimal “fixed” cluster spacing
- Increased cluster density
- Simplified imaging



Ordered spacing



Random spacing



Kinetic Exclusion Amplification

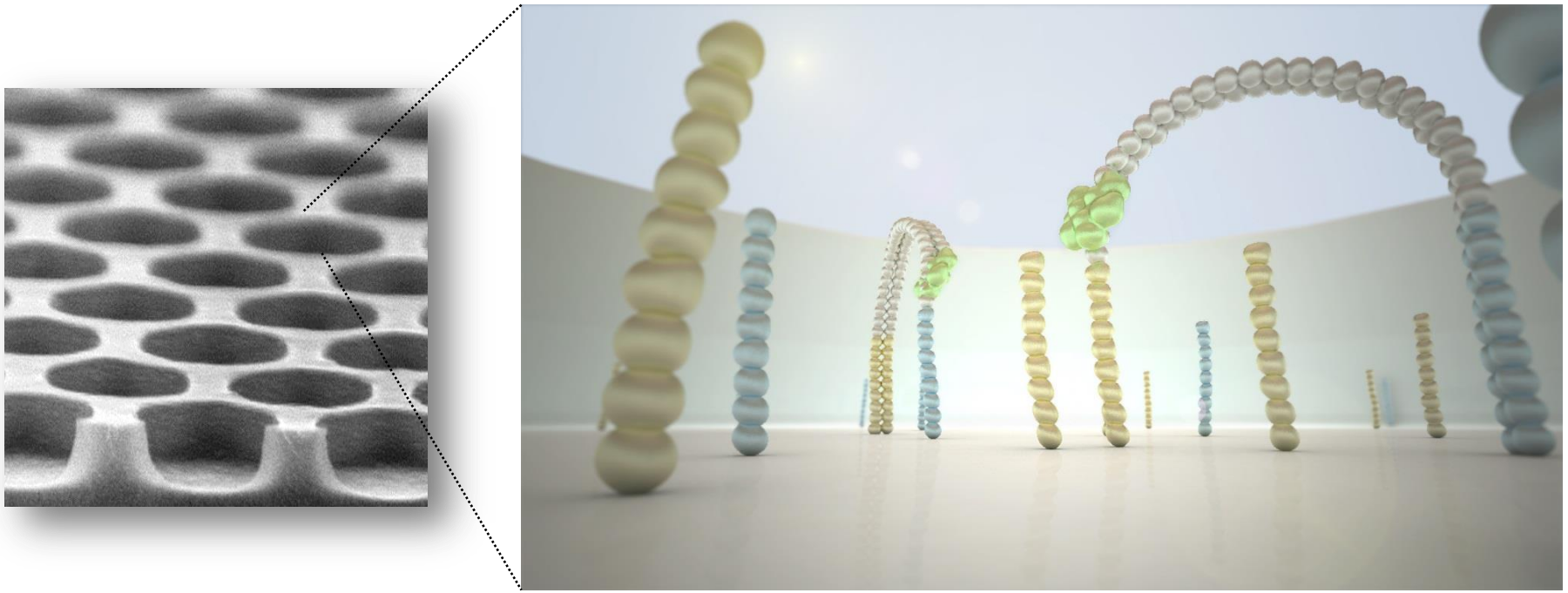
edinburgh genomics.

Kinetic Exclusion Amplification (single template per well)

simultaneous template hybridization and amplification

amplification occurs at 20x the rate of template hybridization

- Patterned cells and KEA video



HiSeq X

edinburgh genomics.

Other species 30x also welcome

HiSeq X Ten/ X Five

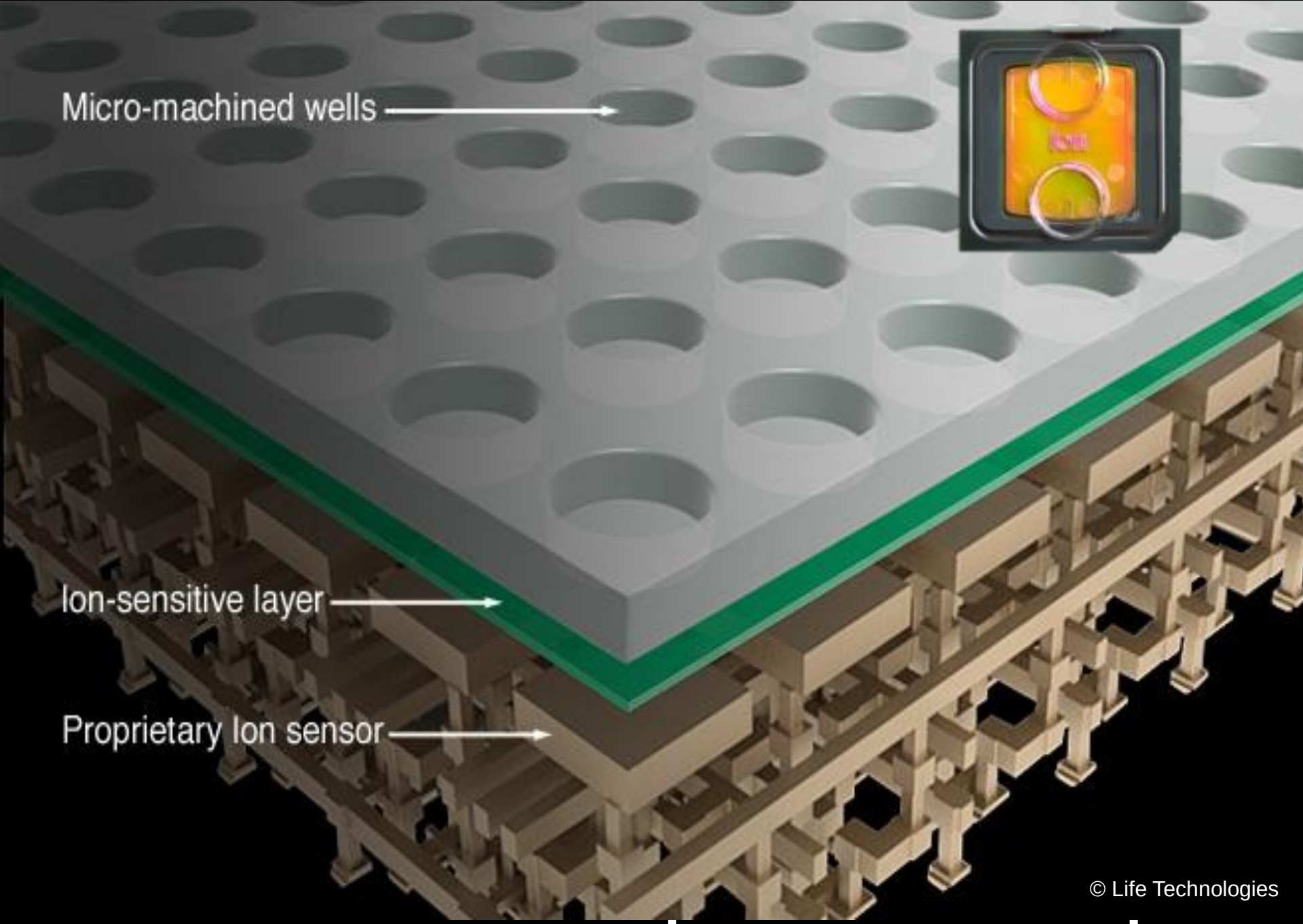
Max Output	Max Read Number	Max Read Length
1,800 Gb	6,000 M	2x150 bp

www.illumina.com

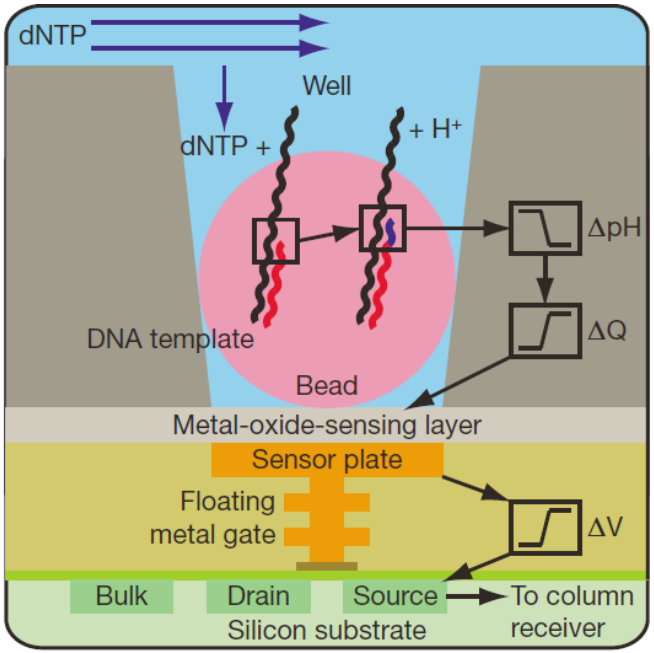
edinburgh genomics.

WuXi AppTec
BAYLOR COLLEGE OF MEDICINE
HUDSONALPHA INSTITUTE FOR BIOTECHNOLOGY
BROAD INSTITUTE
dkfz.
PROVIDENCE Health & Services
MACROGEN
deCODE genetics
HLI
GARVAN INSTITUTE
Novogene 诺禾致源
SciLifeLab
NEW YORK GENOME CENTER
Sidra

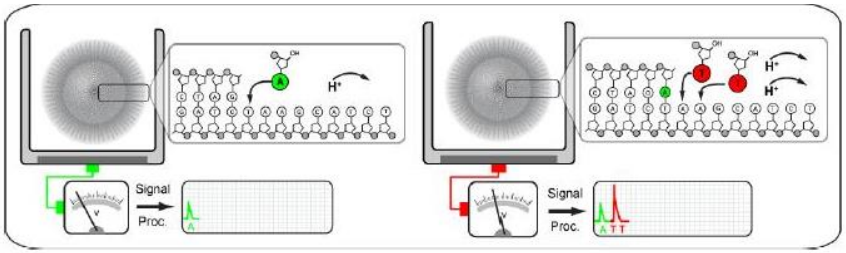
Ion Torrent/Proton



Ion Torrent (Life Technologies)



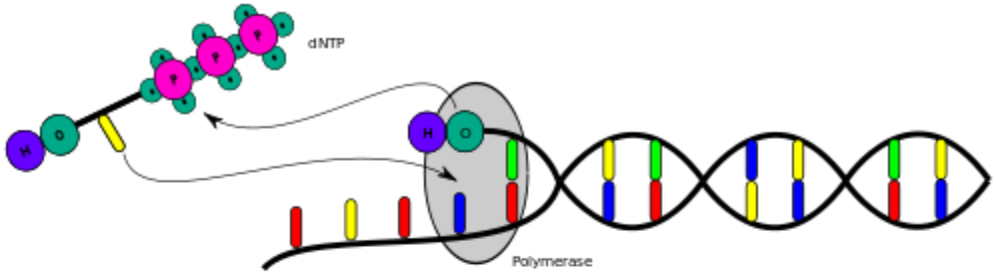
- Similar to pyrosequencing but uses semiconducting chip to detect dNTP incorporation.
- The chip measure differences in pH.
- Different types of chips (throughput/length)
- Shown to have problems with homopolymer reads and coverage bias with GC-rich regions.



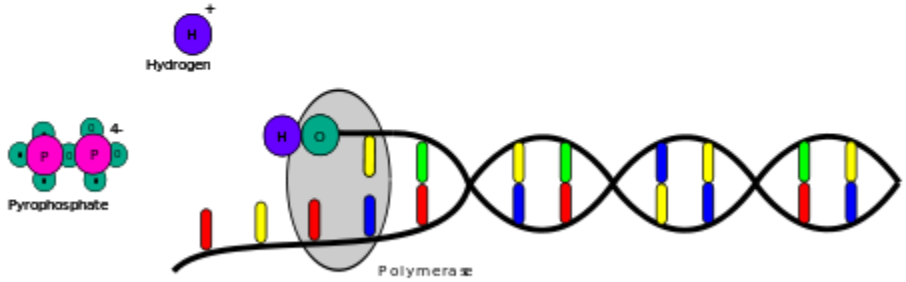
J. M. Rothberg, et al. Nature (2011) 475:348-352

Ion Semiconductor Sequencing

edinburgh genomics.



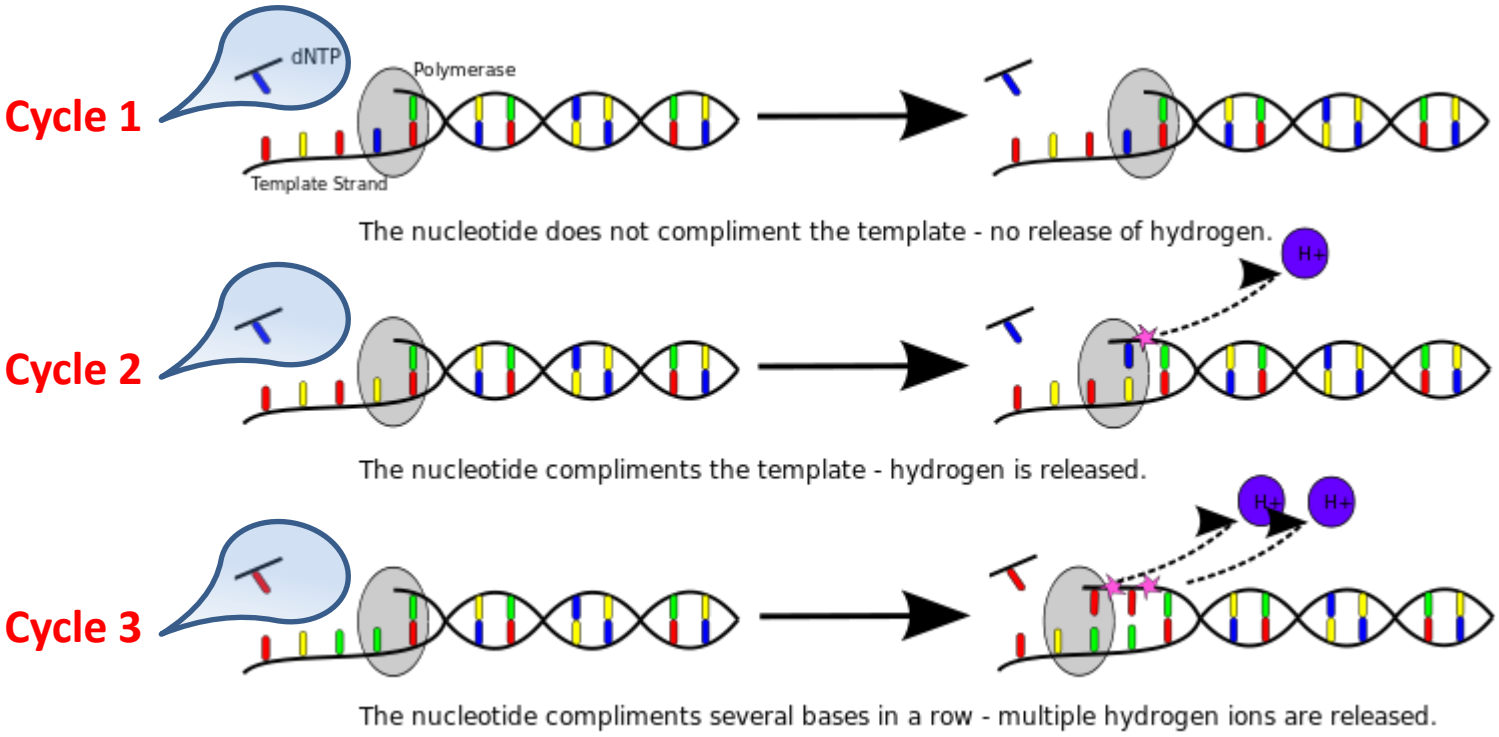
Polymerase integrates a nucleotide.



Hydrogen and pyrophosphate are released.

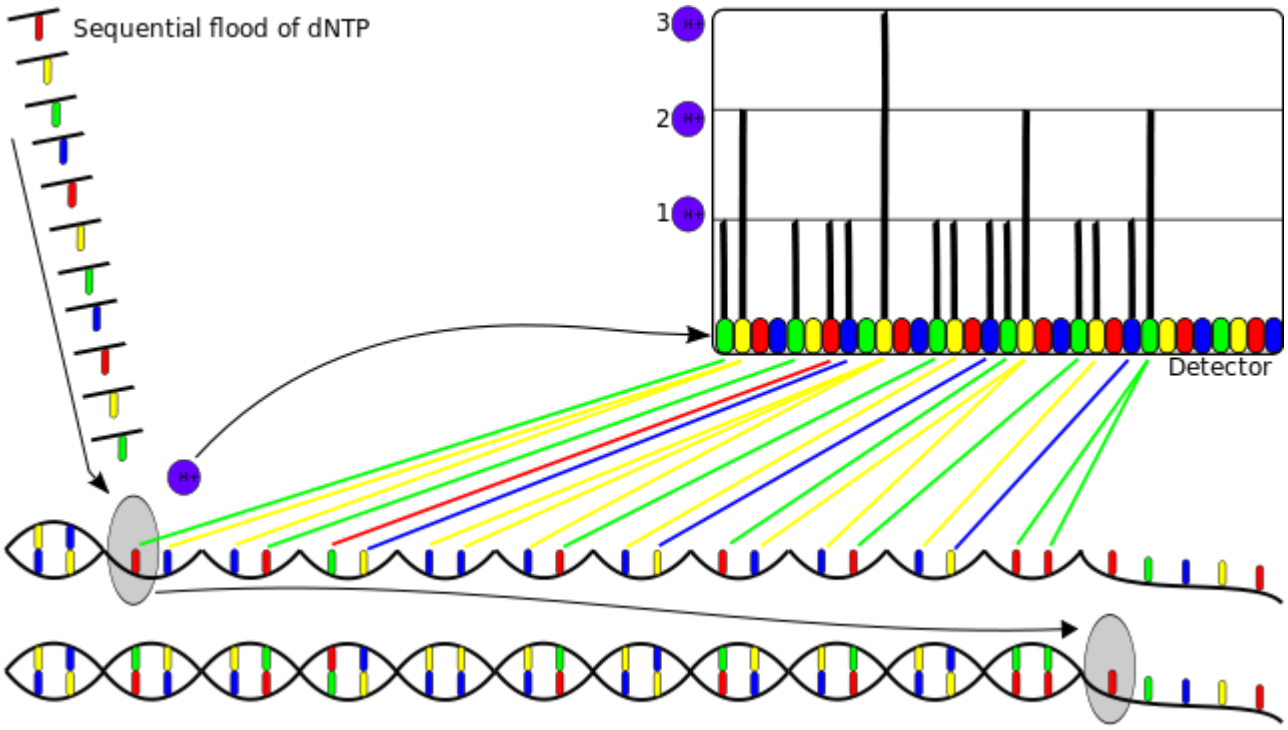
Ion Semiconductor Sequencing

edinburgh genomics.



Ion Semiconductor Sequencing

edinburgh genomics.



Ion PGM Torrent - Proton

edinburgh genomics.

10 Mb to 1 Gb

1 hour/run, > 200 nt lengths

Reads H+ released by DNA polymerase

Chips: 314, 316, 318

Up to 10 Gb

Up to 200 nt, 2-4 h

Reads H+ released by DNA polymerase

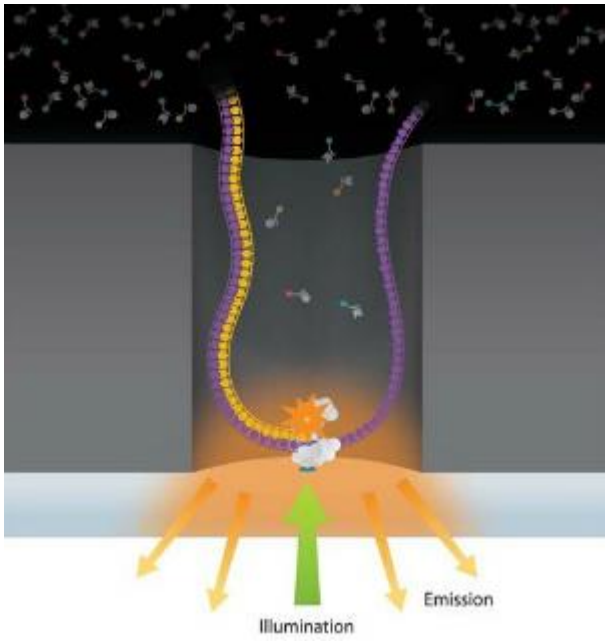
Chips: Proton I, Proton II

Pacific Biosciences

Third Generation: PacBio RS

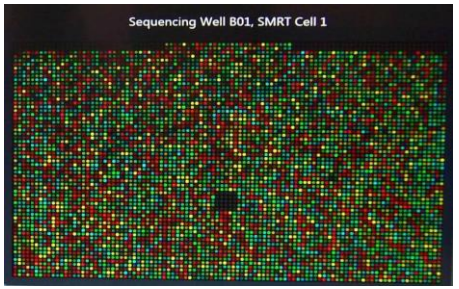
edinburgh genomics.

- **SMRT: Single Molecule Real time DNA synthesis**.
- Single Molecule Sequencing – instead of sequencing clonally amplified templates from beads (Pyro) or clusters (Illumina) DNA synthesis is detected on a single DNA strand.
  - Up to 15,000 nt, 50 bases/second



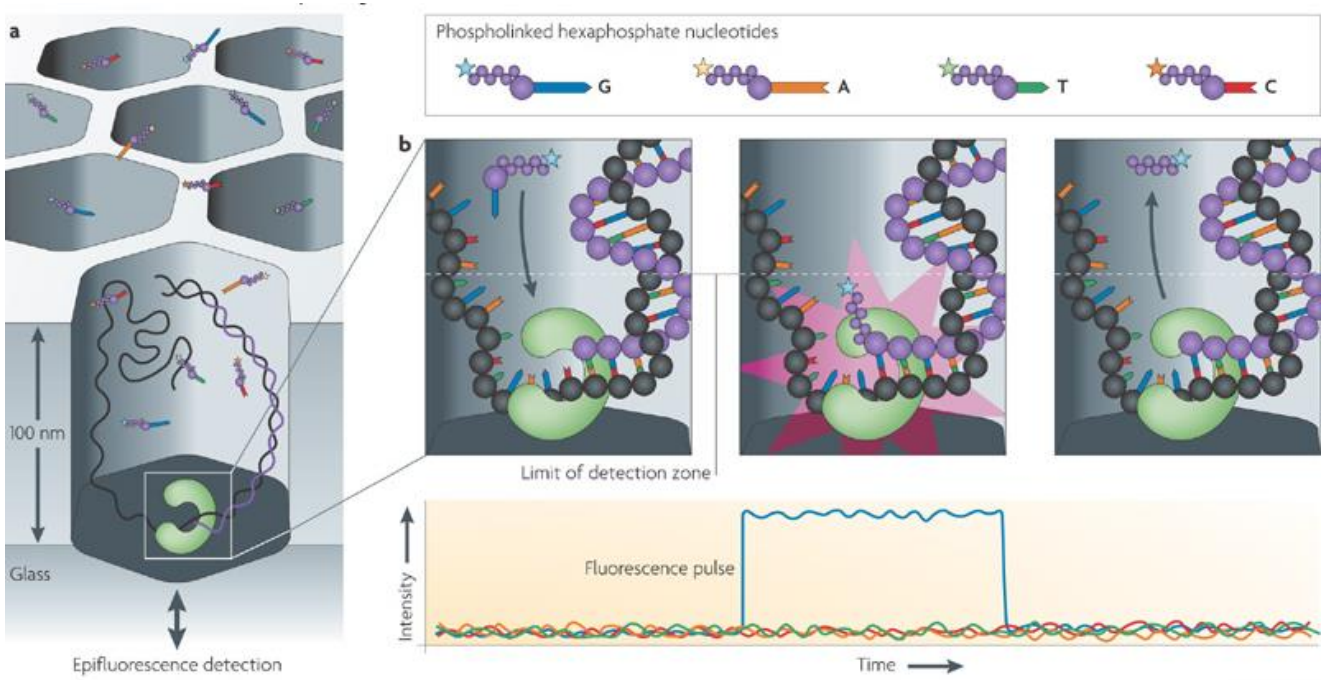
Zero-mode waveguide (ZMW)

- DNA polymerase is affixed to the bottom of a tiny hole (~70nm).
- Only the bottom portion of the hole is illuminated allowing for detection of incorporation of dye-labeled nucleotide.

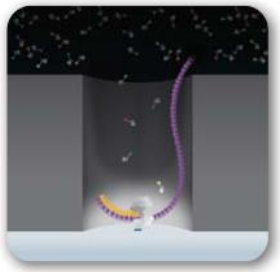


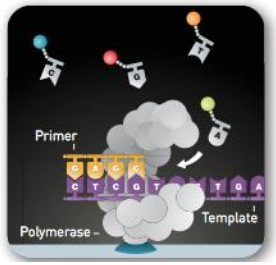
Third Generation: PacBio RS

edinburgh genomics.



[Link to PacBio movie]



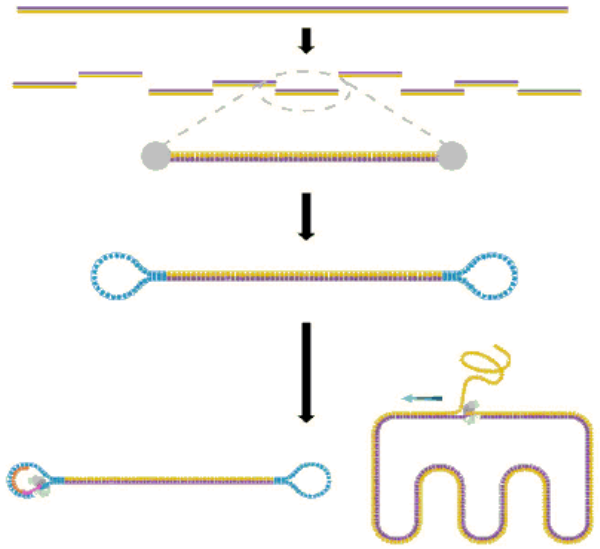


From Michael Metzker, http://view.ncbi.nlm.nih.gov/pubmed/19997069

Third Generation: PacBio RS

edinburgh genomics.

- Real-time Sequencing
  - Unlike reversible termination methods (Illumina) the DNA synthesis process is never halted. Detection occurs in real-time.



Library Prep.

- DNA template is circularized by the use of “bell” shaped adapters.
- As long as the polymerase is stable this allows for continuous sequencing of both strands.

Third Generation: PacBio RS

edinburgh genomics.

Advantages

- No amplification required.
- Extremely long read lengths.
- Average 2500 nt. Longest 15,000 nt.

Disadvantages

- High error rates.
- Error rate of ~15% for Indels. 1% Substitutions.

Third Generation: Sequel

edinburgh genomics.

Advantages

- No amplification required.
- Extremely long read lengths.
- Average 2500 nt. Longest 15,000 nt.
- **7 times more throughput than an RS II**
- **Half price of an RS II.**

Disadvantages

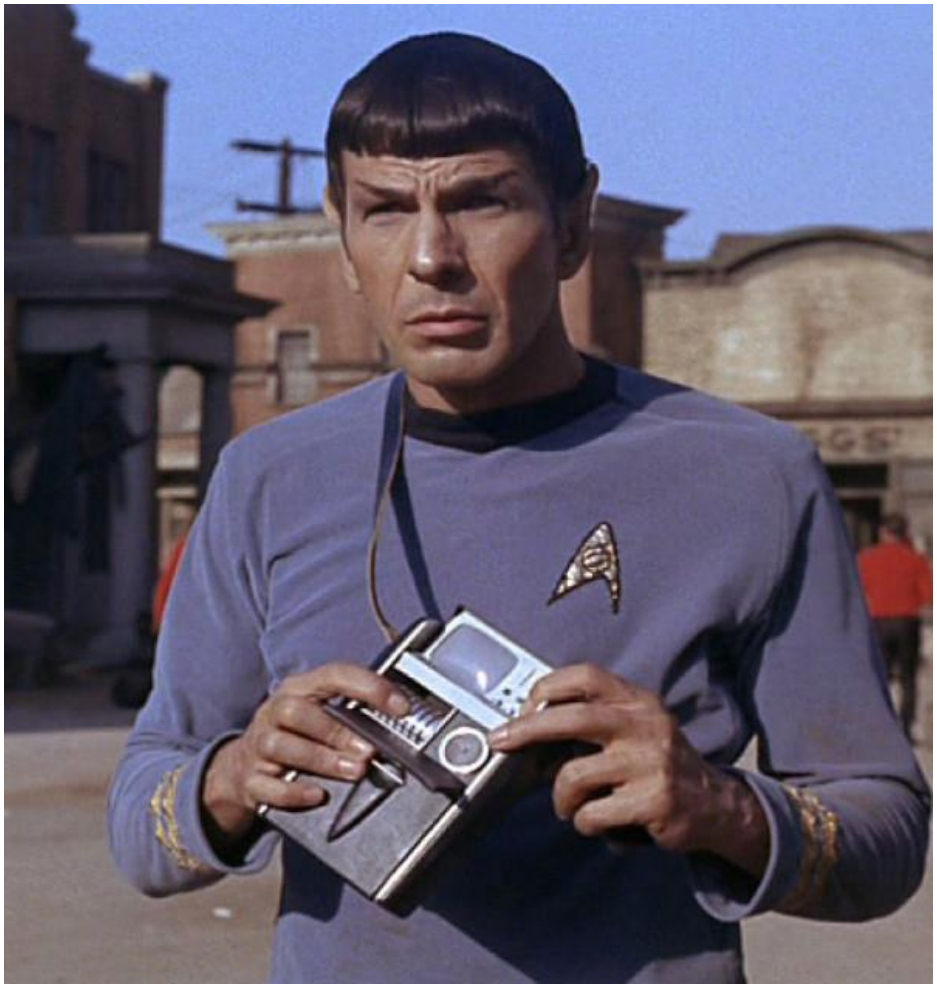
- High error rates.
- Error rate of ~15% for Indels. 1% Substitutions.

Oxford Nanopore MinION

Oxford Nanopore: MinION

The MinION is a memory key–sized disposable unit that can be plugged into a laptop for under $1,000, according to the company.

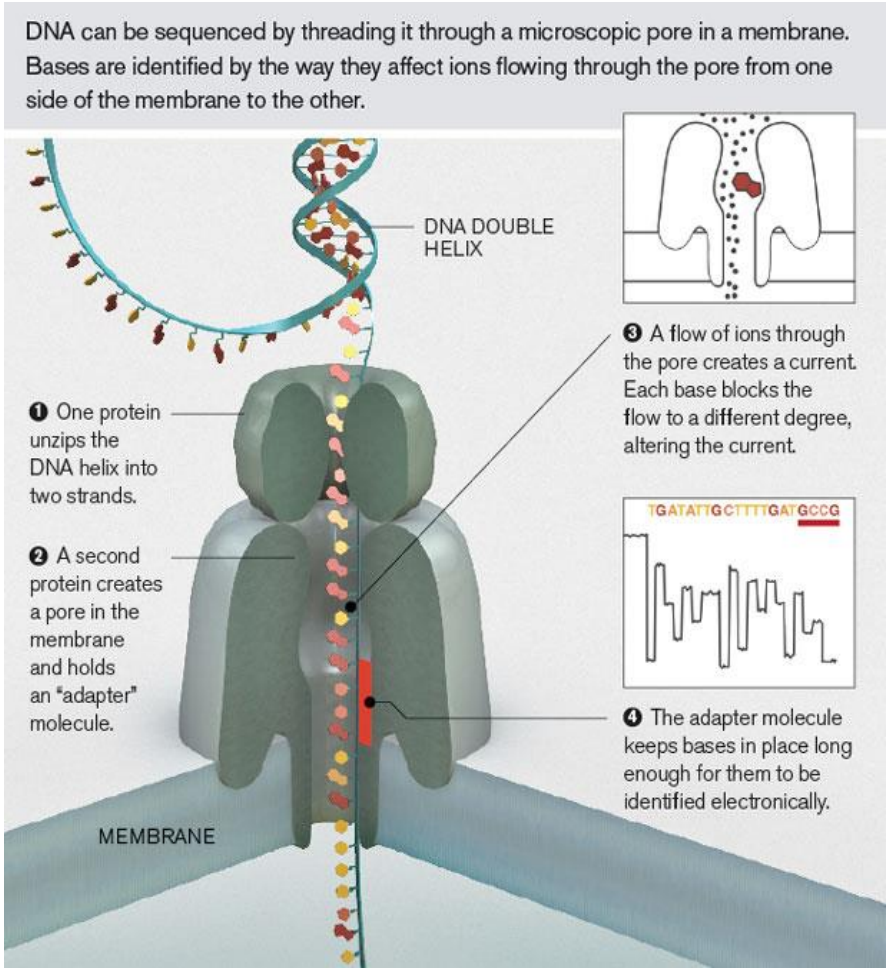
- Announced Feb. 2012 at ABGT conference.
- Portable device
- Disposable device with a sensor chip and nanopores
- Plugging directly into a USB port
- Real-time sequencing data
- DNA sequencing or protein sensing
- MiniON Access Programme (Nov 2013-Jan 2014)



Tricorder

It is a multifunction hand-held device used for sensor scanning, data analysis, and recording data [Wikipedia]

Third Generation: Oxford Nanopore



http://www2.technologyreview.com

- Measure changes in ion flow through nanopore.
- Potential for long read lengths and short sequencing times.

Link to MinION movie

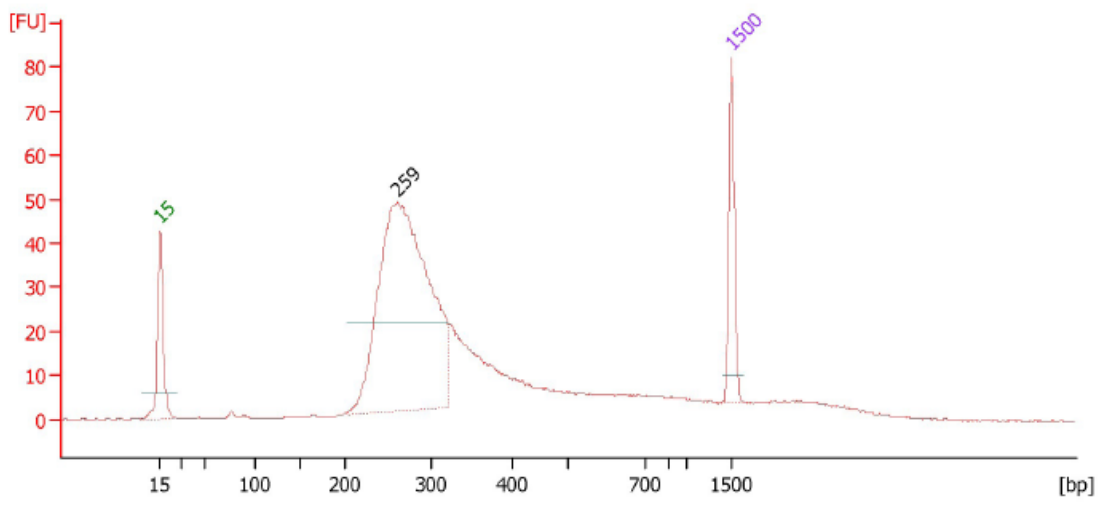
PromethION system

edinburgh genomics.

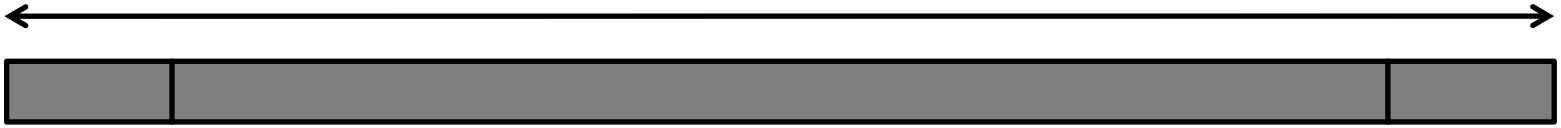
- Scale up system
- Like 96 minION in parallel
- You can use multichannel to load samples into flowcells
- Same chemistry as MinION
- Same software to analyse data
- PromethION early access program PEAP
  - $ 75K deposit

Anatomy of an NGS Library

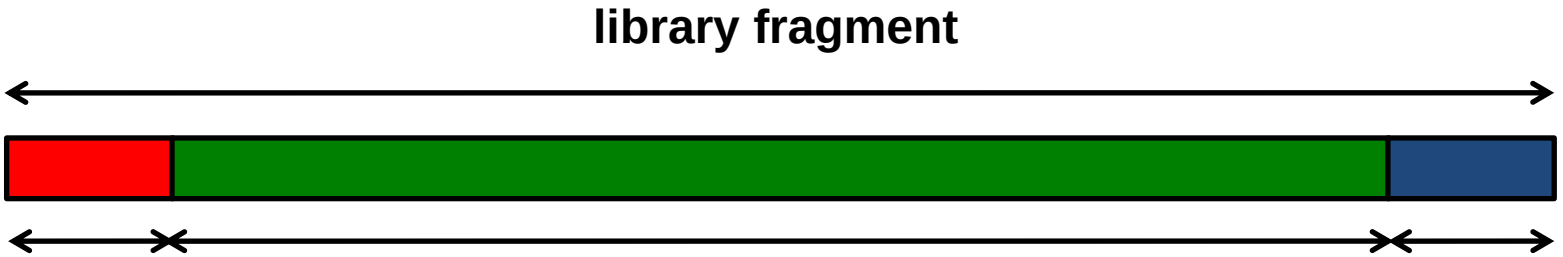
Illumina NGS Library



library fragment



Illumina NGS Library

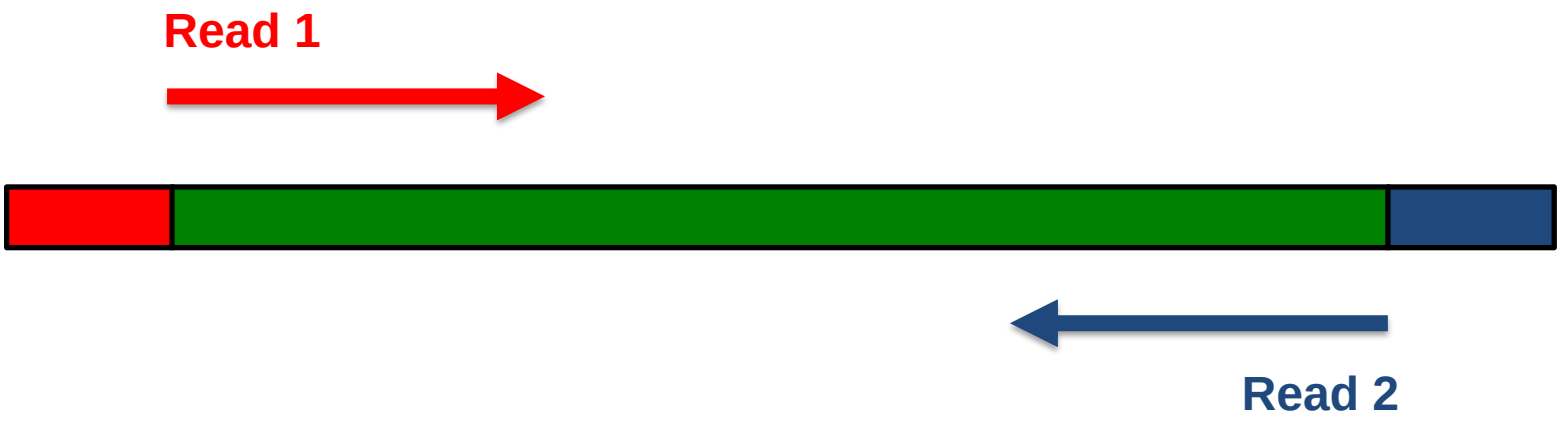


flowcell/bead binding sequences
amplification primers
sequencing primers
indexing primers

flowcell/bead binding sequences
amplification primers
sequencing primers
indexing primers

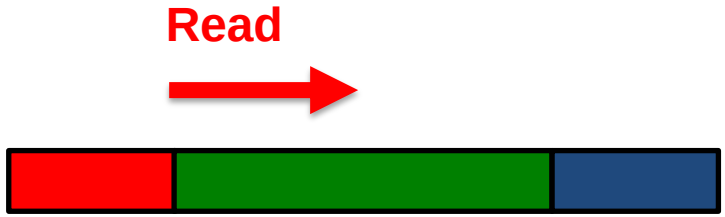
Illumina NGS Library

Illumina NGS Library

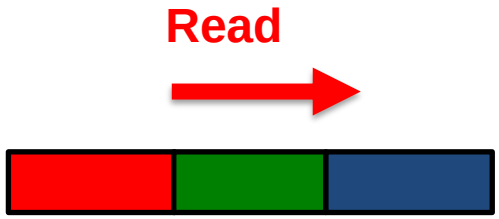


Illumina NGS Library

edinburgh genomics.



- single-end read
- partial sequence from library fragment

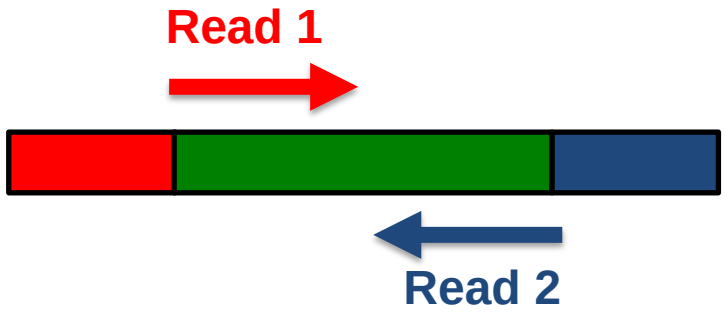


- single-end read
- complete sequence from library fragment
- partial (or complete) adapter sequence

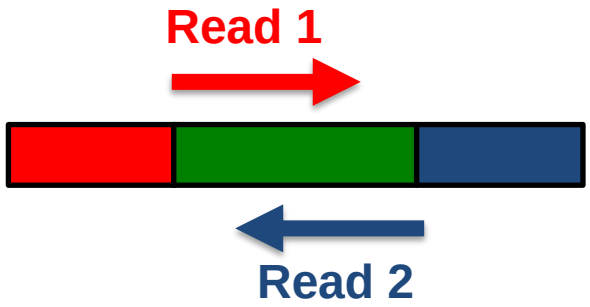
- single-end read
- no library fragment
- partial (or) complete adapter sequence

Illumina NGS Library

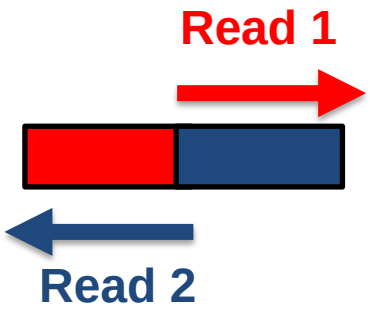
edinburgh genomics.



- paired-end read
- non-overlapping reads
- partial sequence from library fragment



- paired-end read
- complete sequence from library fragment
- overlapping reads



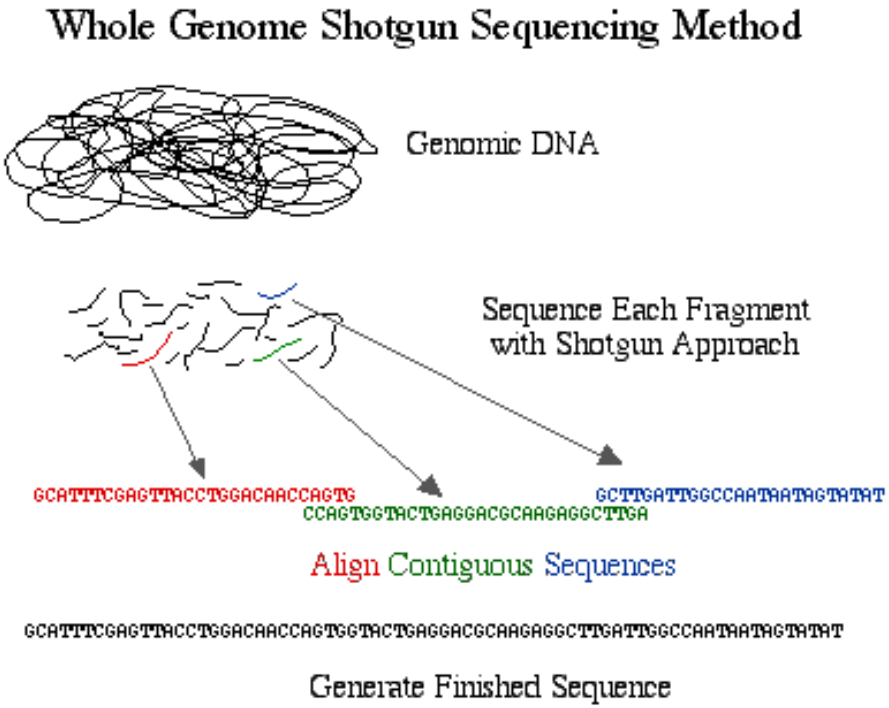
- paired-end read
- no library fragment
- non-overlapping reads
- partial (or) complete adapter sequence

NGS for variant calling

DNA Sequencing - 1

edinburgh genomics.

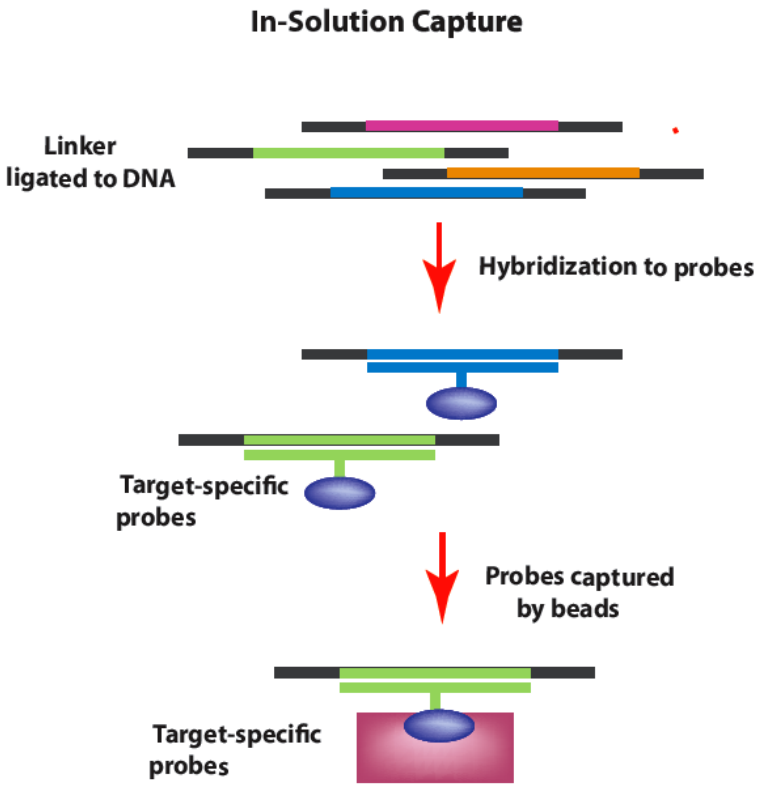
- **Whole GENOME Resequencing**
  - Need reference genome
  - Variation discovery



DNA Sequencing - 2

edinburgh genomics.

- **Targeted Resequencing**
  - Specific regions in the genome
  - Need reference genome
  - Need custom probes complementary to the genomic regions
    - Nimblegen
    - Agilent
- **Custom genes panel sequencing**
  - Allows to cover high number of genes related to a disease
  - Low cost and quicker than capillary sequencing
  - *E.g. Disease gene panel*
- **Whole EXOME Resequencing**
  - Available for Human and Mouse
  - Variation discovery on ORFs
    - 2% of human genome (lower cost)
    - 85% disease mutation are in the exome

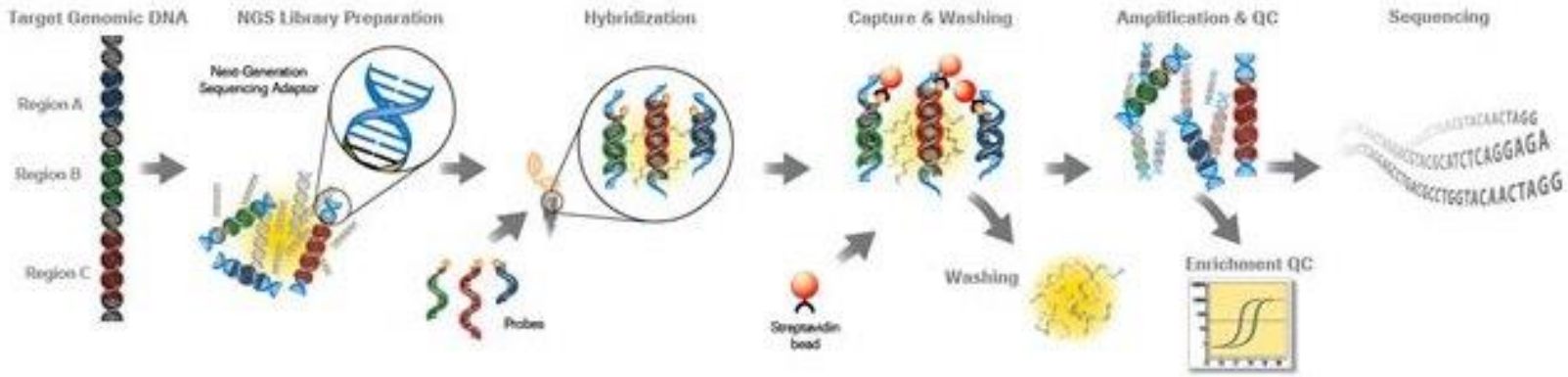


DNA Sequencing - 3

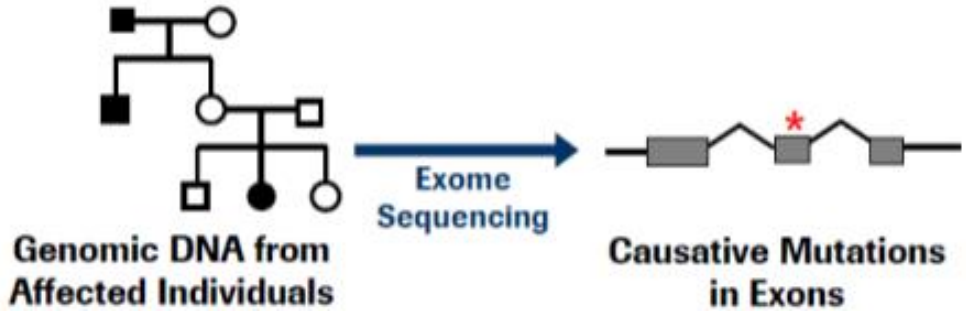
edinburgh genomics.

Don't sequence all, just what you need

- Focus on the Most Relevant Portion of the Genome



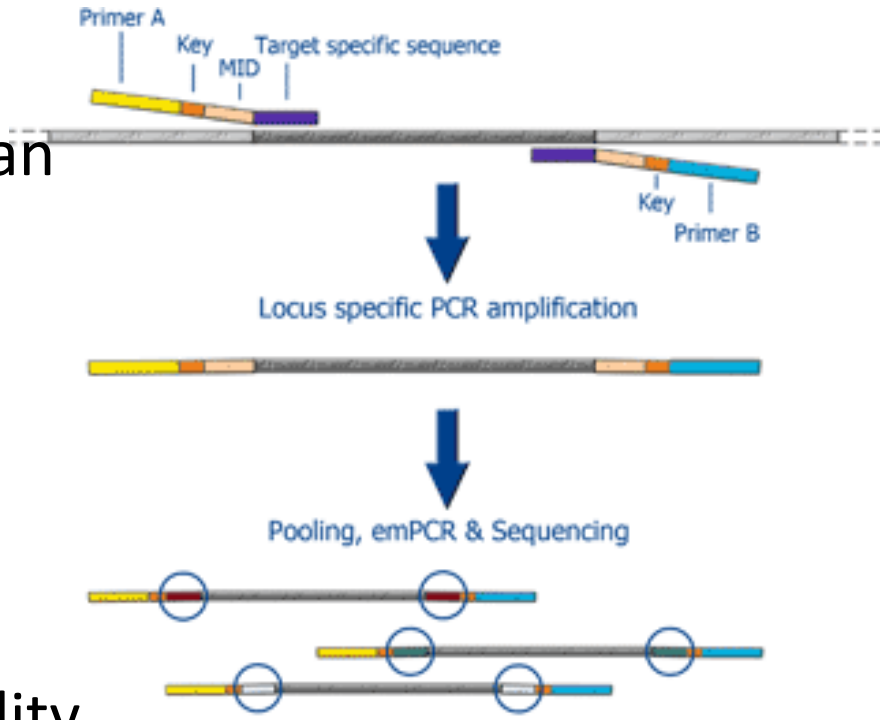
- Capture all exons in the genome: EXOME
  - the most functionally relevant ~2% of the genome.
  - where the majority of known inherited disease-causing mutations reside.



DNA Sequencing - 4

edinburgh genomics.

- **Amplicon sequencing**
  - Sequencing of regions amplified by PCR.
  - Shorter regions to cover than targeted capture
  - No need of custom probes
  - Primer design is needed
  - High fidelity polymerase
  - Multiplexing is needed
  - Low complexity. Lower quality



NGS Data Analysis

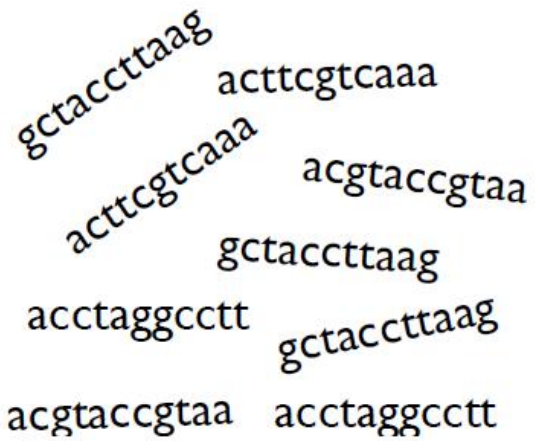
edinburgh genomics.

- Similar pipeline -

DNA Sample

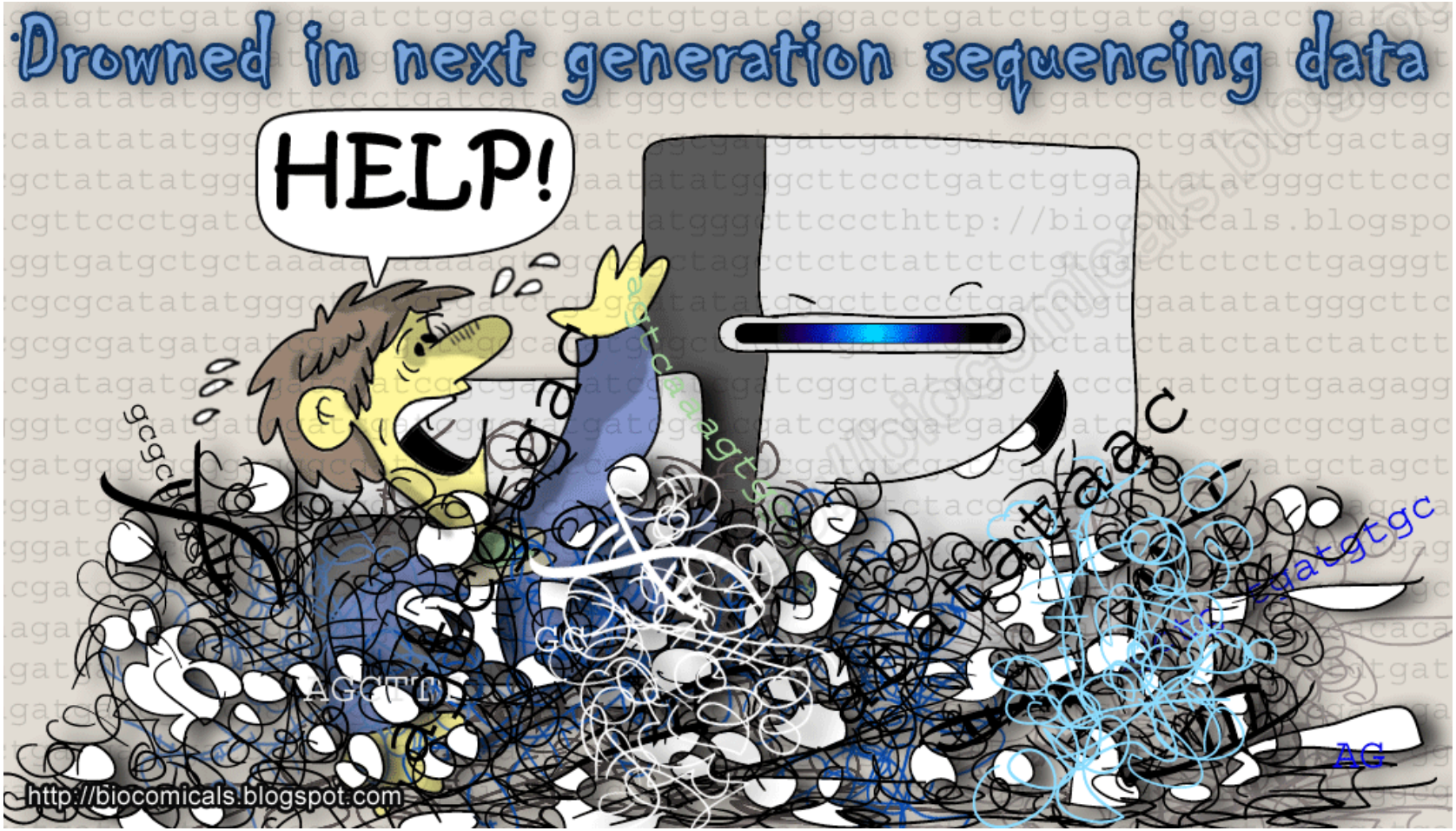
NGS Instrument

Data



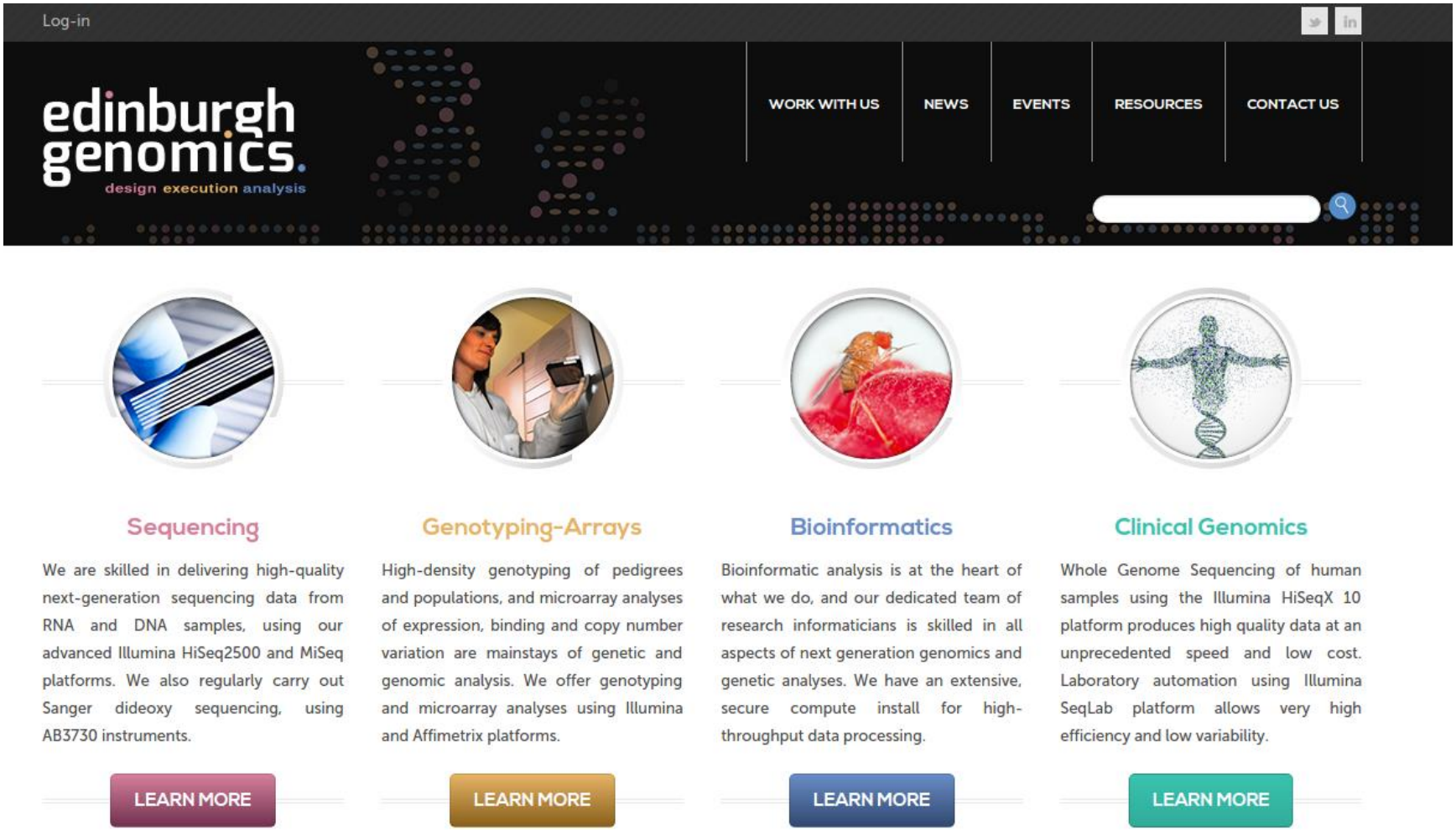
Library Preparation → Sequencing → Data Analysis

NGS is relatively cheap but you have to think how you are going to analyze HUGE AMOUNTS of DATA


Drowned in next generation sequencing data
HELP!
http://biocomicals.blogspot.com

http://genomics.ed.ac.uk

edinburgh genomics.




THANK
YOU
GRACIAS
ARIGATO
SHUKURIA
BOLZÏN
MERCI
BÏYAN
SHUKRIA
DANKSCHEEN
JUSPAXAR
TASHAKKUR ATU
YAQHANYELAY
SUKSAMA
EKHMET
MEHRBANI
GRAZIE
KOMAPSUMNIDA
GOZAIMASHITA
EFCHARISTO
TINGKI
SPASSIBO
SNACHALHUYA
NUHUN
CHALTU
WABEEJA
MAITEKA
YUSPAGARATAM
HUI
GUI
HÄTUR
EKOJU
SIKOMO
MAKETAI
MINMONCHAR
SPASIBO
DENKAUJA
NENACHALHIYA
UNALCHÉESH
MERSI
ATTO
ANIHA
DHANYAVAAD
MAAKE
LAH
PALDIES
MERASTAWHY
GAEJTHO
AGUYJE
FAKAAUE
SANCO
TAVTAPUCH
MEDAWAGSE
BAIKA