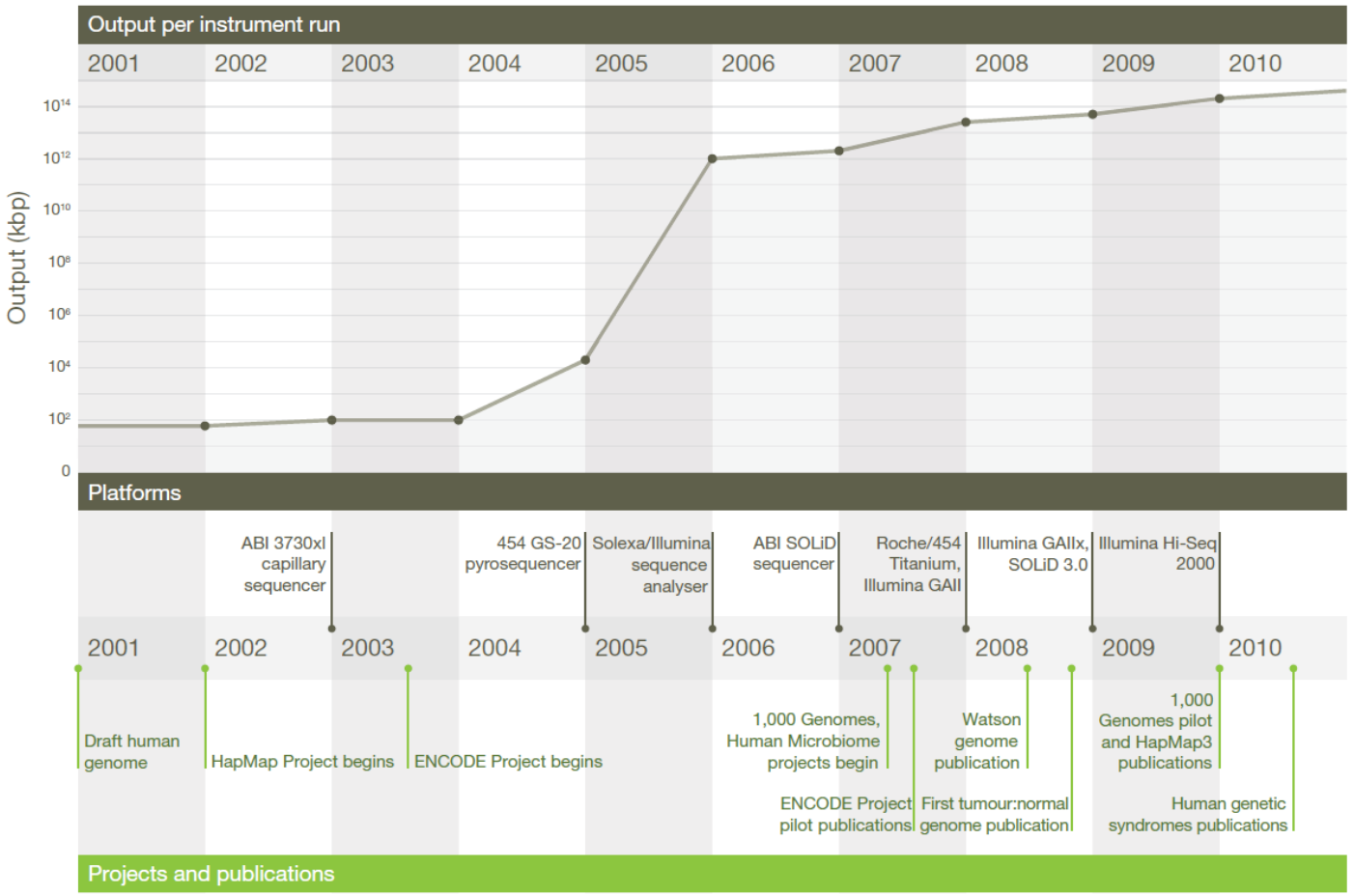
NGS Overview Part I: A Comparison of Next-Generation Sequencing Platforms.

MIN SOO KIM
APRIL 1, 2013
QUANTITATIVE BIOMEDICAL RESEARCH CENTER
DEPARTMENT OF CLINICAL SCIENCES
UT SOUTHWESTERN

OUTLINE

- Background.
- Common Pipeline.
  - Library Prep.
  - Sequencing – Massively Parallel Sequencing.
  - Bioinformatics - Data Analysis .
- Popular Platforms:
  - Roche 454 , AB SOLiD, Illumina(HiSeq,MiSeq).
- Newer Platforms (Third Generation):
  - Ion Torrent, PacBio RS, Oxford Nanopore.

Background



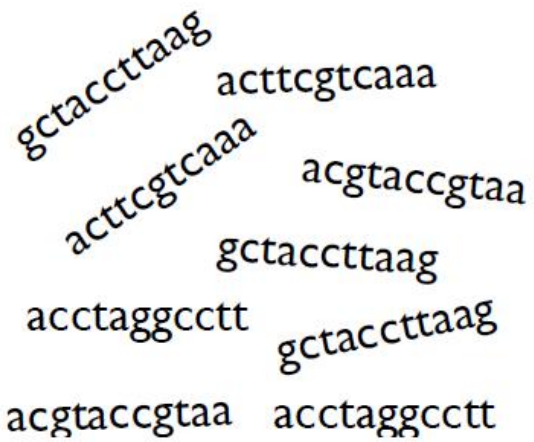
E.R. Mardis, Nature (2011) 470:198 - 203

Next Generation Sequencing Pipeline

DNA Sample

NGS Instrument

Data



Library Preparation

Sequencing

Data Analysis

Library Preparation

- DNA samples are randomly fragmented and platform-specific adaptors are added to the flanking ends to produce a "library".
- Library is then amplified through PCR. (Platform-specific amplification e.g. beads or glass)
- Amplification Introduces Bias:
  - Amplification bias against AT, GC rich regions. (corrected by adding PCR additives.)
  - Alteration of representational abundances(duplicates). Important for quantitative applications like RNA-seq.
    - Overrepresentation of smaller fragments. (corrected by running fewer PCR cycles.)

Massively Parallel Sequencing

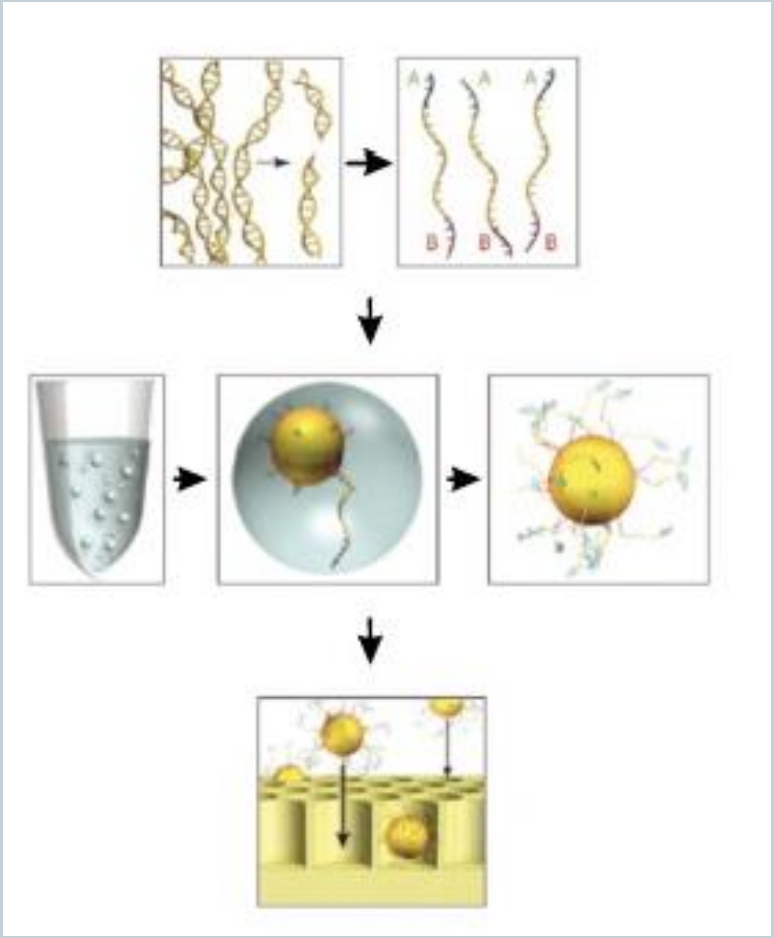
- In Sanger-sequencing the DNA synthesis and detection steps are two separate procedures (slow).
- Next generation sequencing relies on coupling the DNA synthesis and detection (sequencing by synthesis) and multiple sequencing reactions are run simultaneously (Massively Parallel Sequencing).
- For most NGS platforms desynchronization of reads during the sequencing and detection cycle is the main cause of sequencing errors and shorter reads.

Platform	Chemistry	Read Length	Run Time	Gb/Run	Advantage	Disadvantage
454 GS Junior (Roche)	Pyro-sequencing	500	8 hrs.	0.04	Long Read Length	High error rate in homopolymer
454 GS FLX+ (Roche)	Pyro-sequencing	700	23 hrs.	0.7	Long Read Length	High error rate in homopolymer
HiSeq (Illumina)	Reversible Terminator	2*100	2 days (rapid mode)	120 (rapid mode)	High-throughput / cost	Short reads Long run time (normal mode)
SOLiD (Life)	Ligation	85	8 days	150	Low Error Rate	Short reads Long run time
Ion Proton (Life)	Proton Detection	200	2 hrs.	100	Short Run times	New*
PacBio RS	Real-time Sequencing	3000 (up to 15,000)	20 min	3	No PCR Longest Read Length	High Error Rate

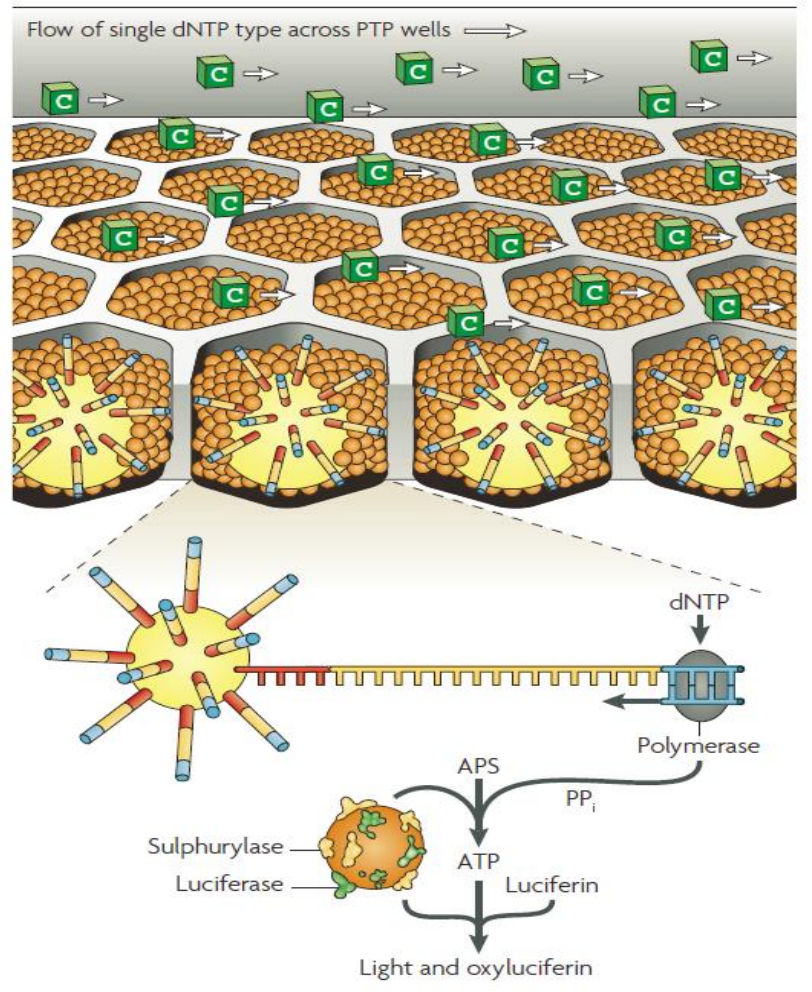
Roche 454 – Library Prep.

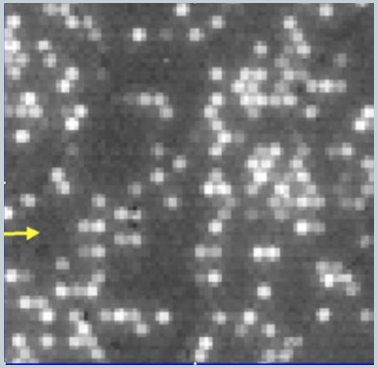
Emulsion PCR (emPCR)

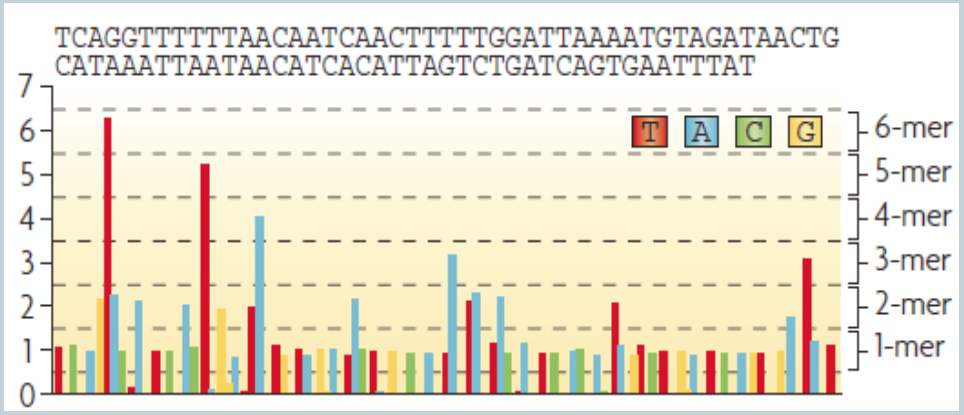
1. DNA fragmentations and adaptor ligation.
2. DNA fragments are added to an oil mixture containing millions of beads.
3. Emulsion PCR results in multiple copies of the fragment.
4. Beads are deposited on plate wells ready for sequencing.



Roche 454 – Pyrosequencing

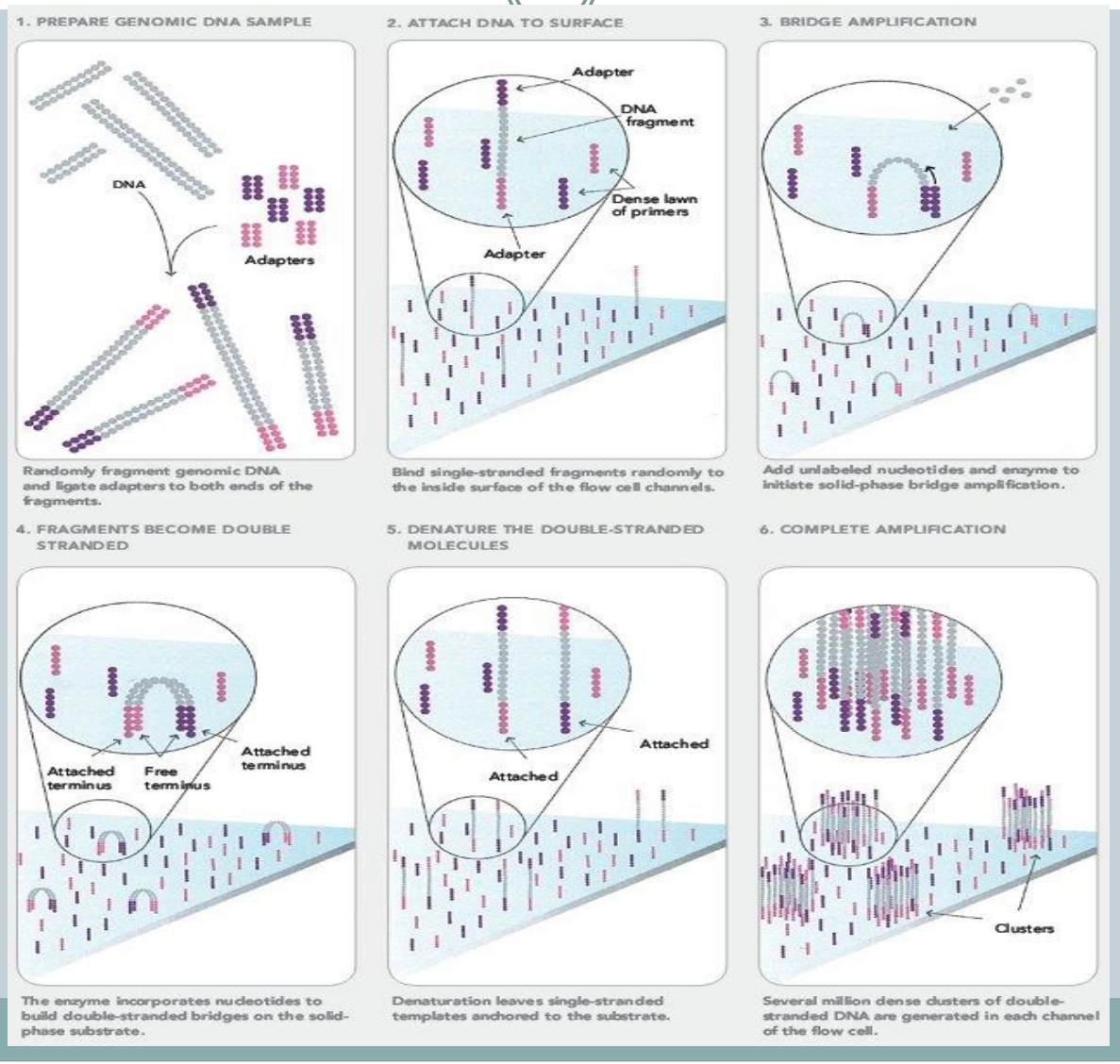




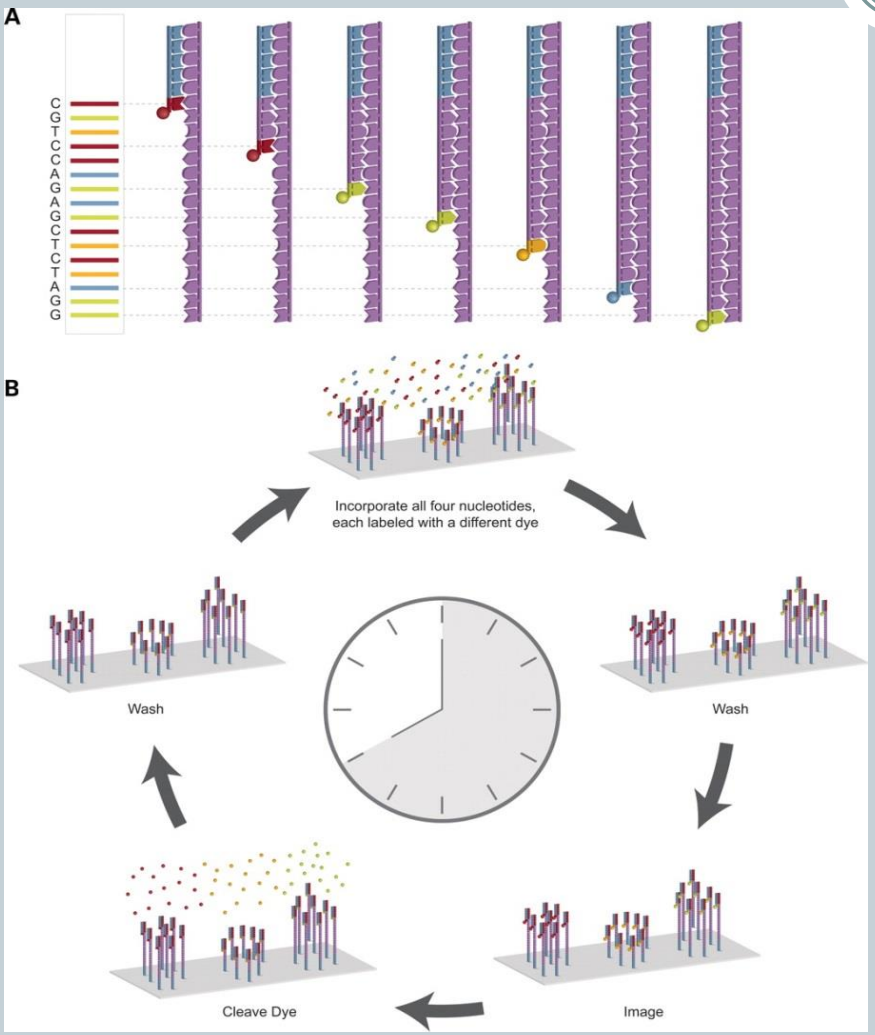


M.L. Metzker, Nature Review Genetics(2010)

Illumina – Library Prep.



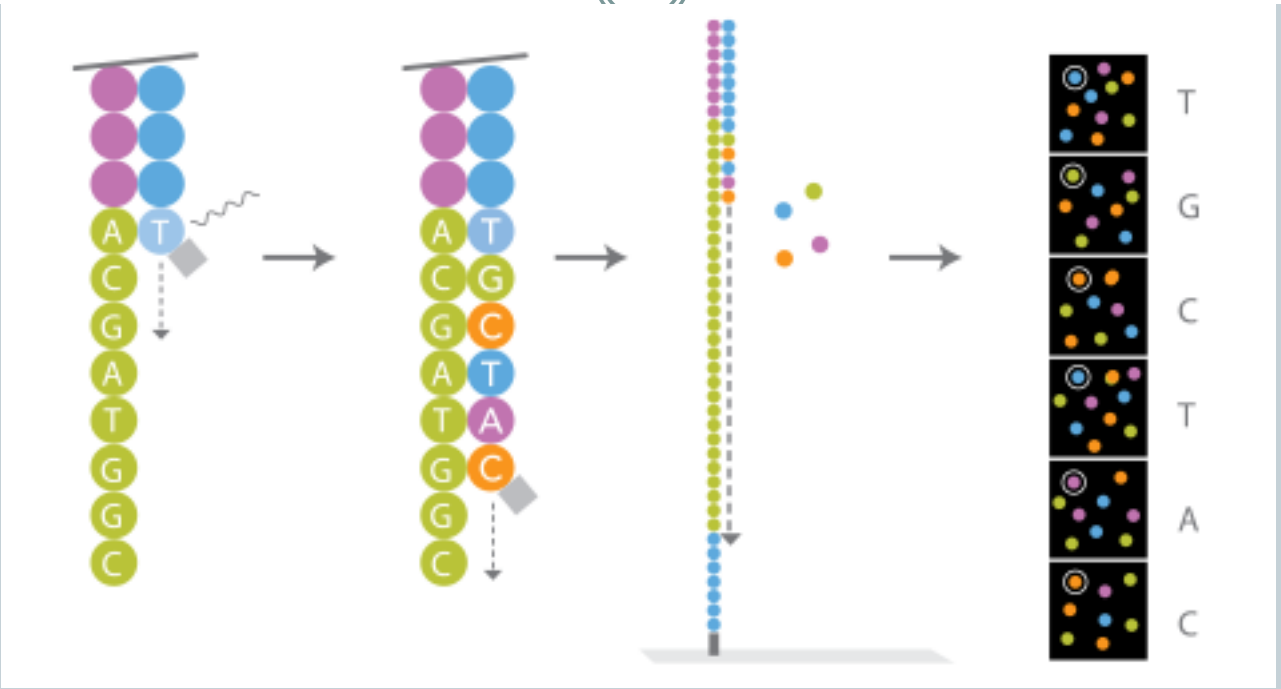
Illumina – Sequence by Synthesis



1. Add dye-labeled nucleotides.
2. Scan and detect nucleotide specific fluorescence.
3. Remove 3' – blocking group (Reversible termination).
4. Cleave fluorescent group.
5. Rinse and Repeat.

http://hmg.oxfordjournals.org

Illumina – Sequencing by Synthesis



Platform	Run Time	Yield (GB)	Error Type	Error Rate
GAIIx	14 days	96	Sub	0.1
HiSeq 2000/2500	10/2 days	600/120	Sub	0.03*
MiSeq	1 day	2	Sub	0.03*

Illumina

Advantages

- High throughput / cost.
- Suitable for a wide rage of applications most notably whole genome sequencing.

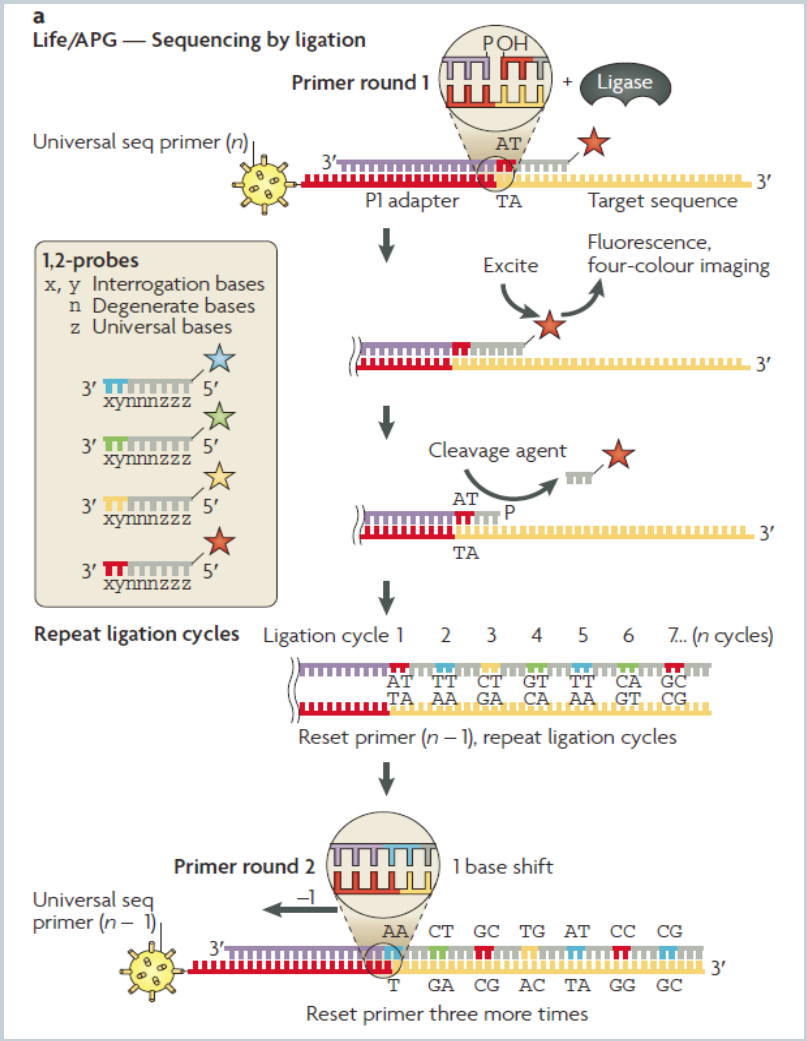
Disadvantages

- Substitution error rates (recently improved).
- Lagging strand dephasing causes sequence quality deterioration towards the end of read.

SOLiD (Life/AB)

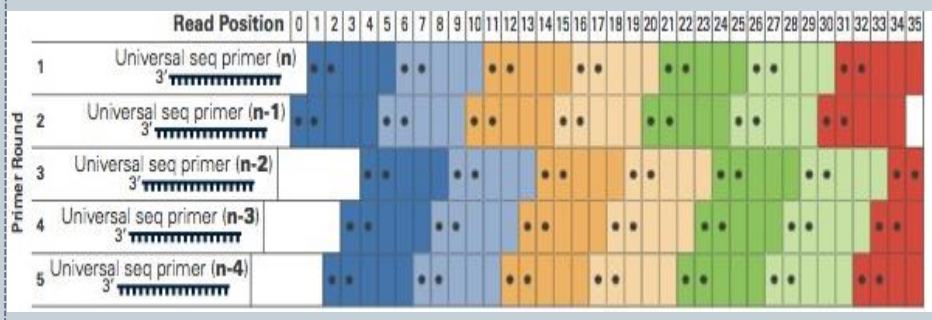
- SOLiD: Sequencing by Oligonucleotide Ligation and Detection.
- Library Prep: emPCR
- "2-base encoding" – Instead of the typical single dNTP addition, two base matching probes are used. (possible 16 probes).
- Color Space – four color sequencing encoding further increases accuracy.

SOLiD (Life/AB)

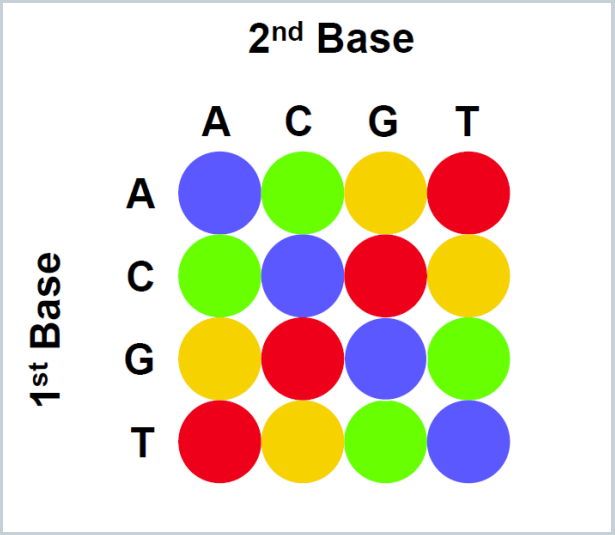


1. Anneal primer and hybridize probe(8-mer).
2. Ligation and Detection.
3. Cleave fluorescent tail(3-mer).
4. Repeat ligation cycle.

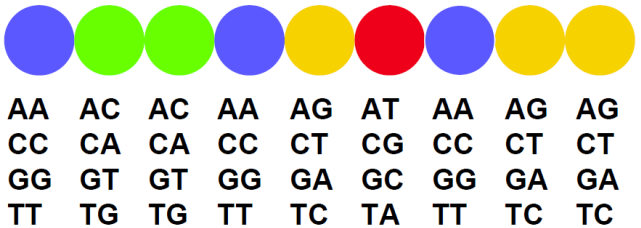
Repeat steps 1-4 with (n-1) primer.



SOLiD – Color Space



- 16 possible base combination are represented by 4 colors.

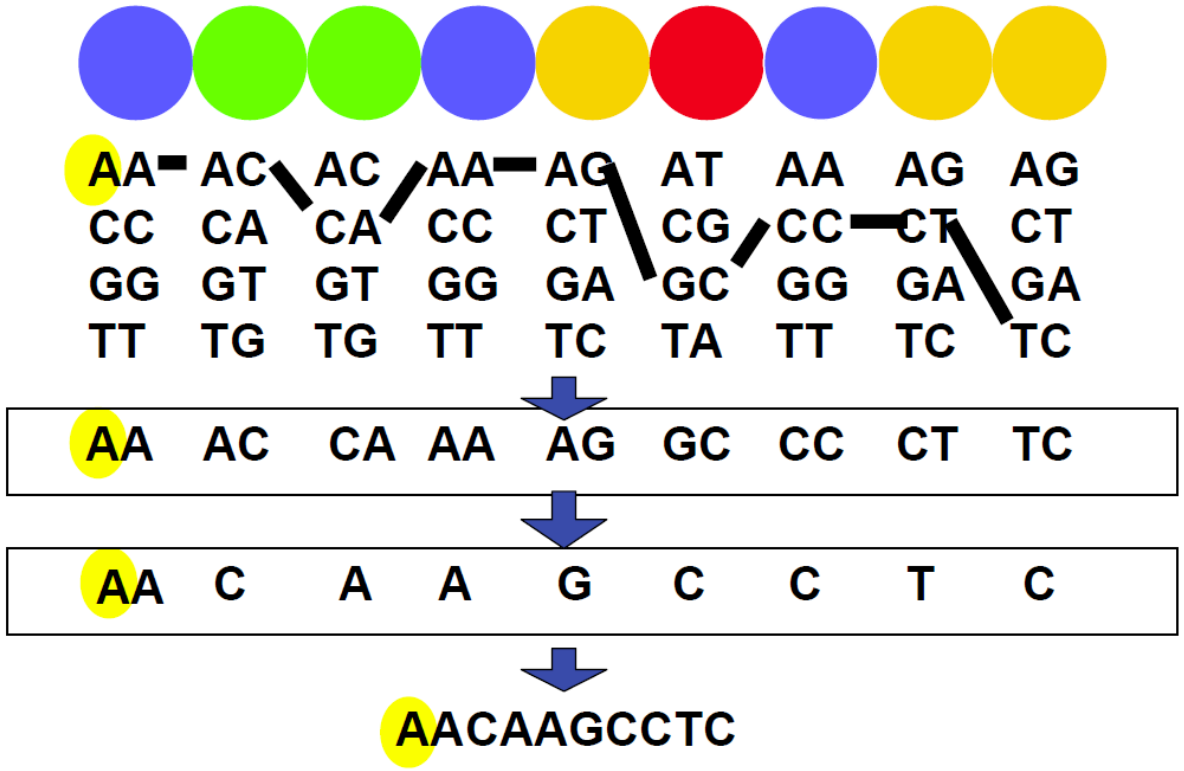


AA	AC	AC	AA	AG	AT	AA	AG	AG
CC	CA	CA	CC	CT	CG	CC	CT	CT
GG	GT	GT	GG	GA	GC	GG	GA	GA
TT	TG	TG	TT	TC	TA	TT	TC	TC

Cannot determine any of the bases

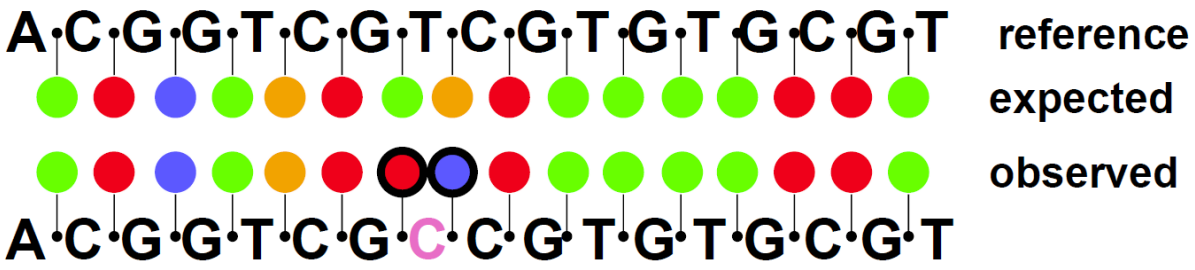
- All possible sequencing combination need to be decoded.

SOLiD – Color Space Decoding



SOLiD – SNP Detection

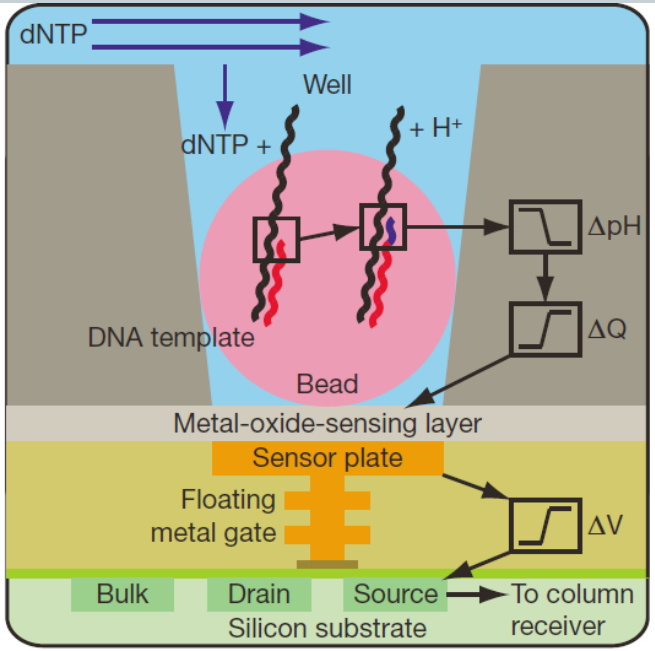
Advantages of 2 base pair encoding
Real SNP

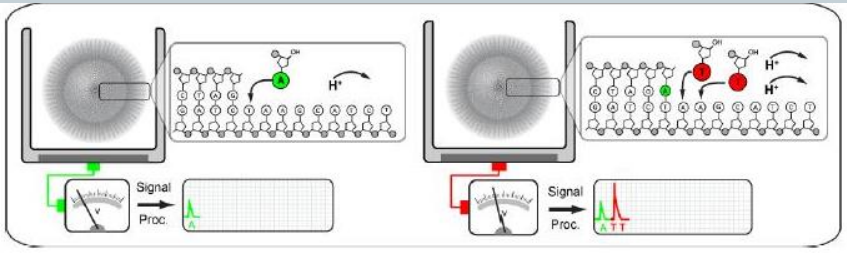


Two color changes represent only a single mismatch to reference sequence (SNP)

Summary: SOLiD has one of the lowest error-rates (~0.01) due to 2-base encoding. It is however still limited by short read lengths (35 bp / 85 bp for PE).

Ion Torrent (Life Technologies)



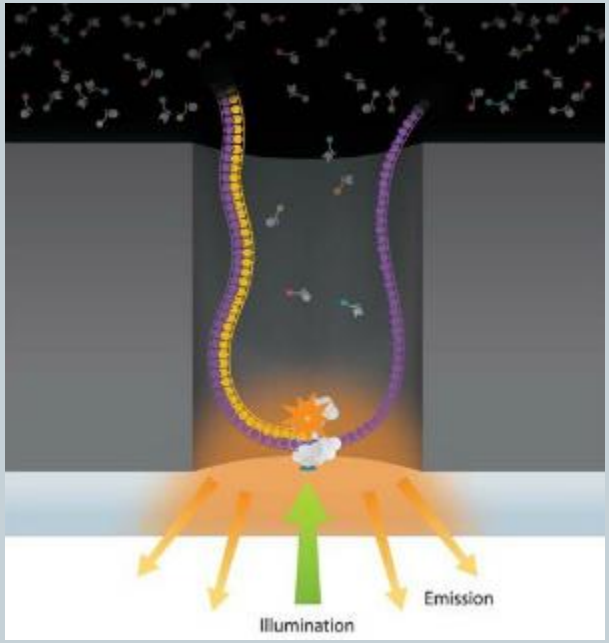


- Similar to pyrosequencing but uses semiconducting chip to detect dNTP incorporation.
- The chip measure differences in pH.
- Shown to have problems with homopolymer reads and coverage bias with GC-rich regions.
- Ion Proton™ promises higher output and longer reads.

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

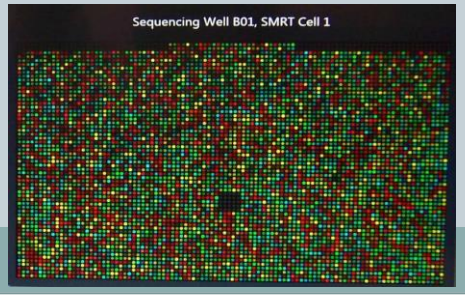
PacBio RS

- Single Molecule Sequencing – instead of sequencing clonally amplified templates from beads (Pyro) or clusters (Illumina) DNA synthesis is detected on a single DNA strand.



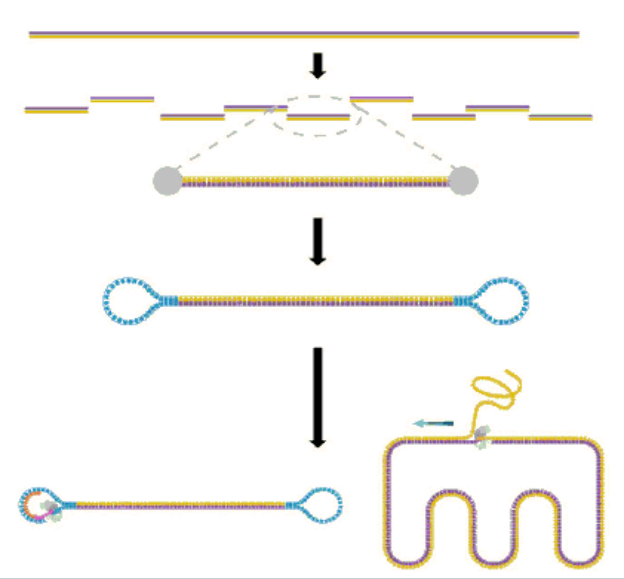
Zero-mode waveguide (ZMW)

- 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.



PacBio RS

- 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.

PacBio RS

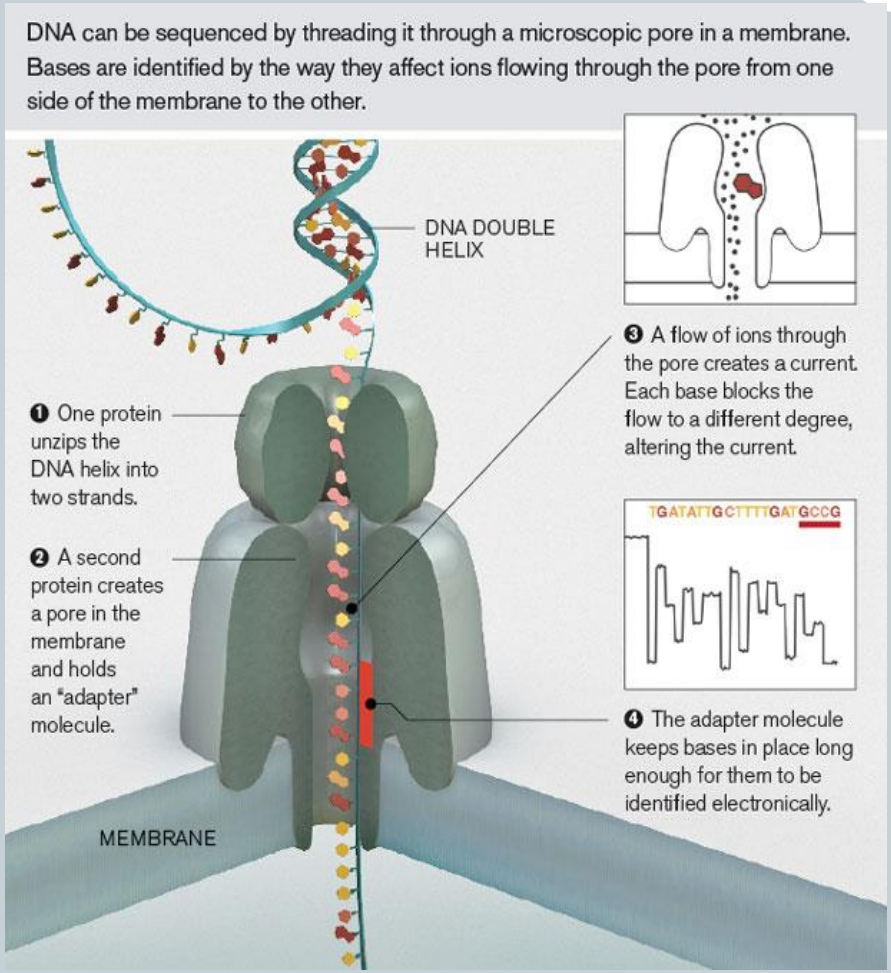
Advantages

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

Disadvantages

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

Oxford Nanopore



- Announced Feb. 2012 at ABGT conference.
- Measure changes in ion flow through nanopore.
- Potential for long read lengths and short sequencing times.

http://www2.technologyreview.com

NGS Overview Part II

- Applications
  - Whole Genome Sequencing.
  - Exome-sequencing
  - Target Resequencing
  - RNA-seq
  - Chip-seq
  - SNP/Indel/Structural Variation Discovery.
- Experimental Design.

Thank you.

Acknowledgements

Dr. Tae Hyun Hwang

References

Mardis, E. R. *A decade's perspective on DNA sequencing technology.* Nature (2011) 470:198 - 203

Metzker, M.L. *Sequencing technologies – the next generation*. Nature Review Genetics(2010) 11:31 – 46

N. J. Loman, C. Constantinidou, Jacqueline Z. M. Chan, M. Halachev, M. Sergeant, C. W. Penn, E. R. Robinson & M. J. Pallen. *High-throughput bacterial genome sequencing: an embarrassment of choice, a world of opportunity*. Nature Review Microbiology 10, 599-606.