



Canadian Centre for
Computational
Genomics

Montreal Genomics Analysis Workshop: RNA-Seq

Day1: Introduction to Next Generation Sequencing
Mathieu Bourgey, PhD

21-22 August 2018

Outline



1. The technology

2. Types of data

3. Conclusions

Technology Revolution



Canadian Centre for
Computational
Genomics

Sequencing genomes in Years



Project cost: Billions \$

Sequencing genomes in HOURS/Minutes !!



©2012 Illumina, Inc. All rights reserved.

Thousands \$

Sequencing: Technological Advances



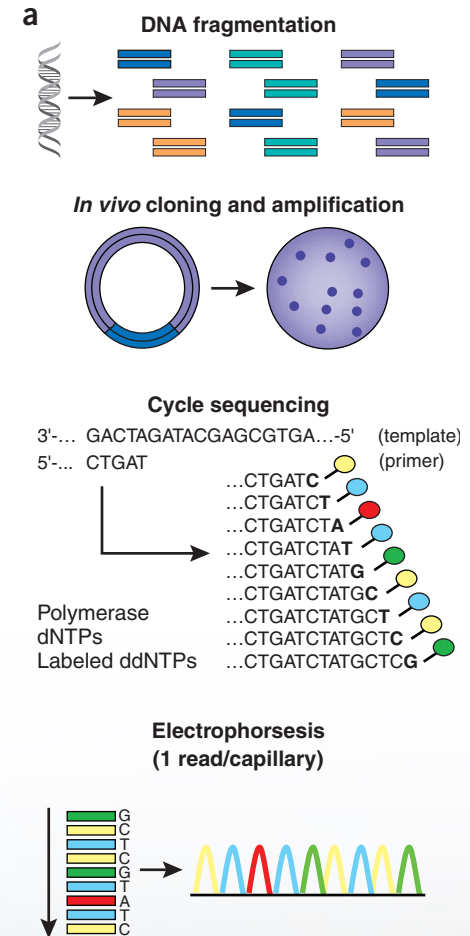
Nb. Sequences/run: 96

Run time: many hours

Limitation: 1 plasmid prep per
tube!

50 cents/sequence

Bacterial genome seq cost : >
\$500k using multiple machines...



From: *Next generation DNA sequencing*, Jay Shendure, Hanlee Ji, 2008

The next wave of DNA sequencing



Canadian Centre for
Computational
Genomics

frequently used terms

- “Massively parallel” sequencing
 - “High-throughput” sequencing
 - “Ultra high-throughput” sequencing
 - “Next generation” sequencing (NGS)
 - “Second generation” sequencing
- **2005: 454 (Roche)**
 - **2006: Solexa (Illumina)**
 - **2007: ABI/SOLiD (Life Technologies)**
 - **2010: Complete Genomics**
 - **2011: Pacific Biosciences**
 - **2010: Ion Torrent (Life Technologies)**
 - **2015: Oxford Nanopore Technologies**

Major Players



Canadian Centre for
Computational
Genomics

Read length

Life technology: SOLiD / ion torrent

Illumina: Novaseq/ Hiseq /
Miseq

Roche: 454

Pacific Bioscience: PacBio

Oxford Nanopore: MinION / GridION

SMALL

MEDIUM

LONG

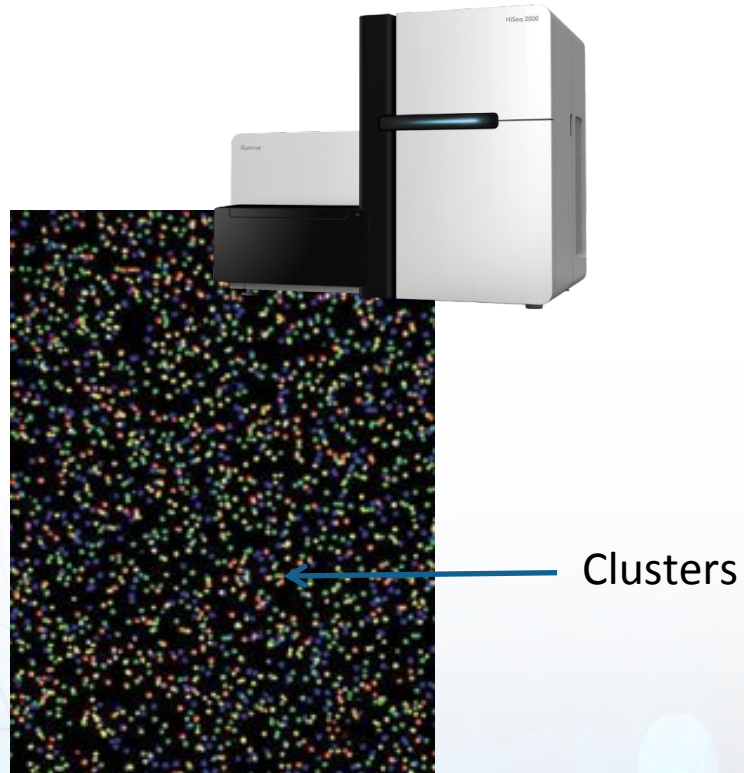


Short Read (Illumina)

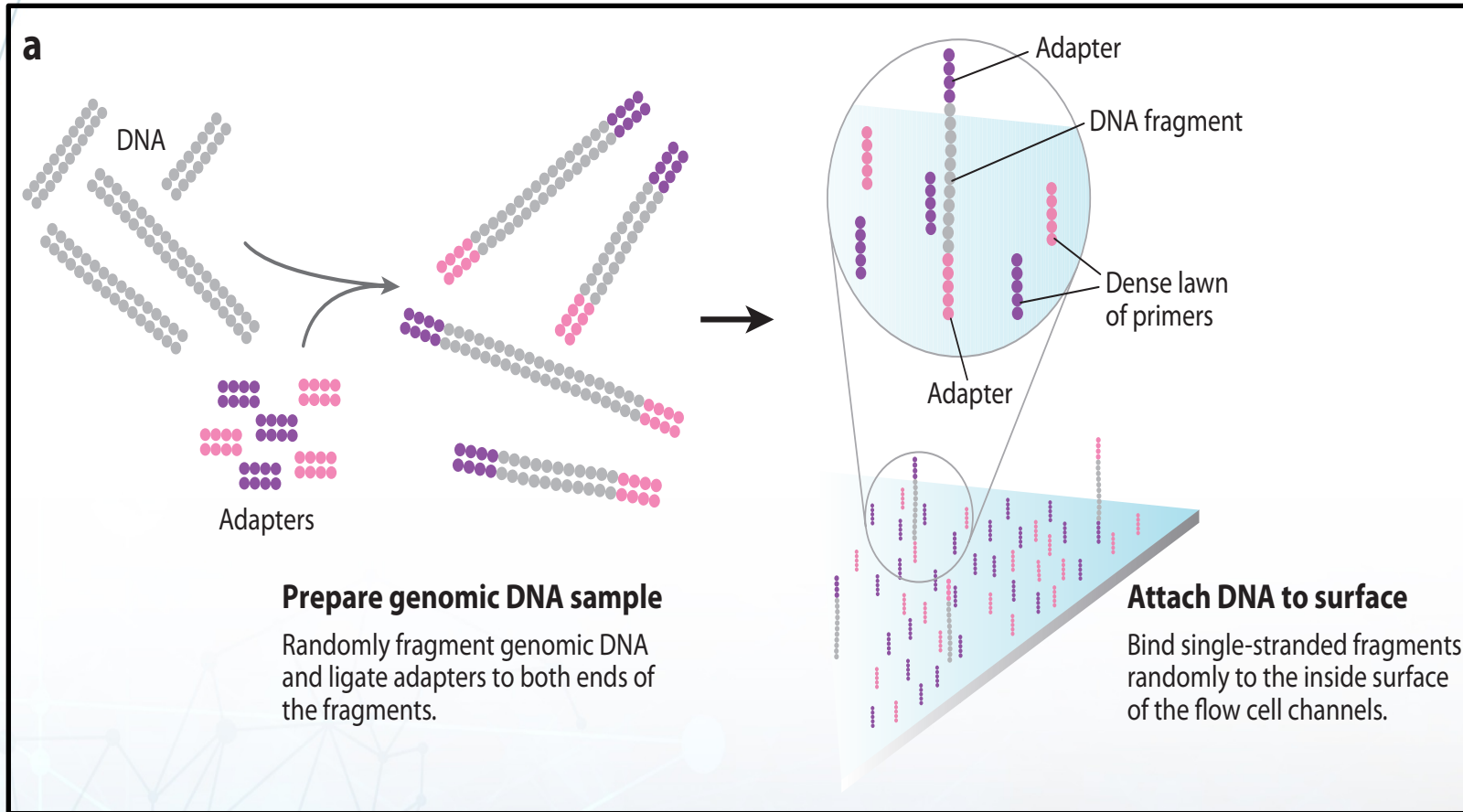
Illumina: How it works



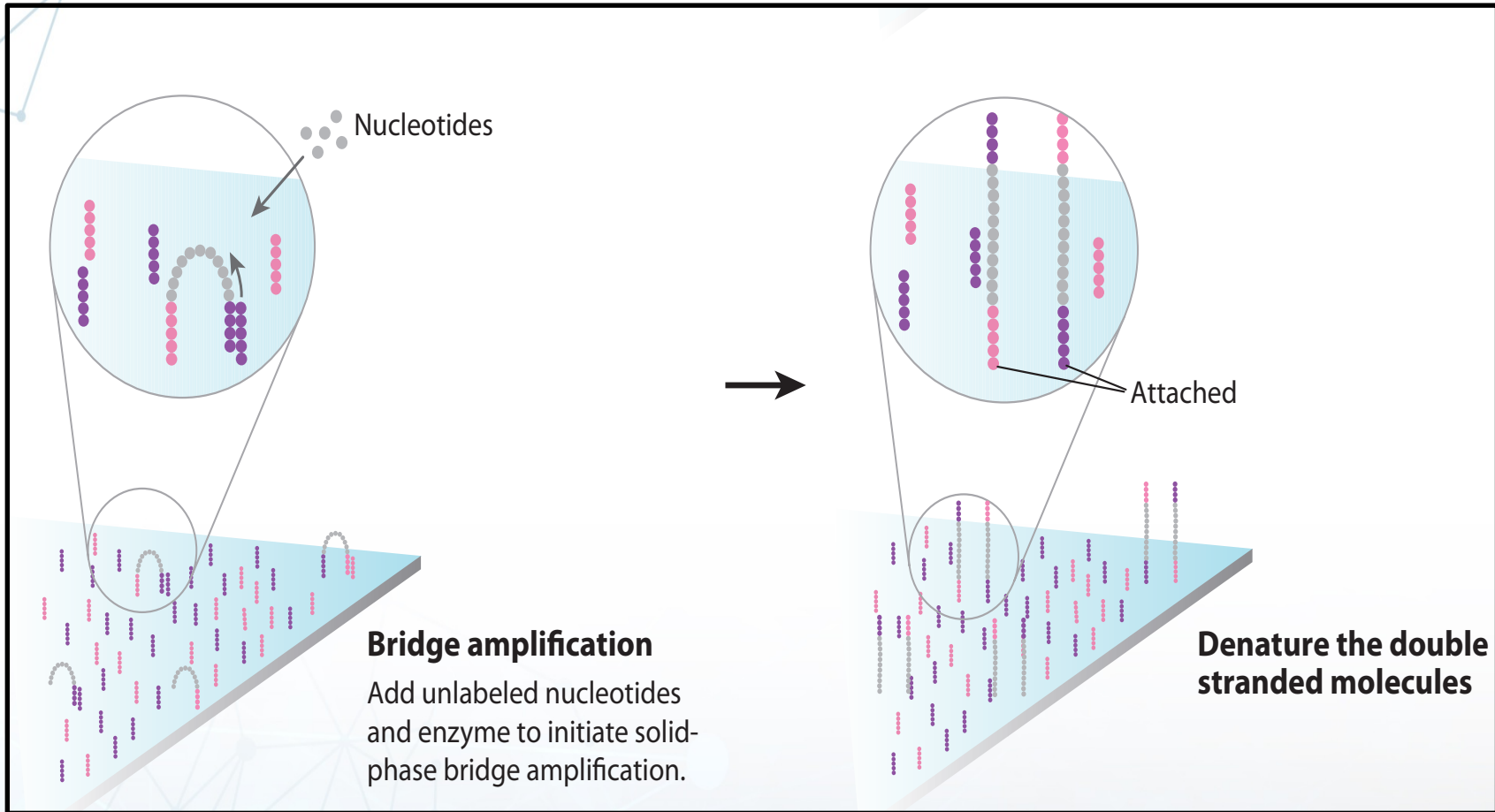
**Illumina sequencing is no longer clone-based : replaced by
Clusters**



Illumina sequencing-by-synthesis

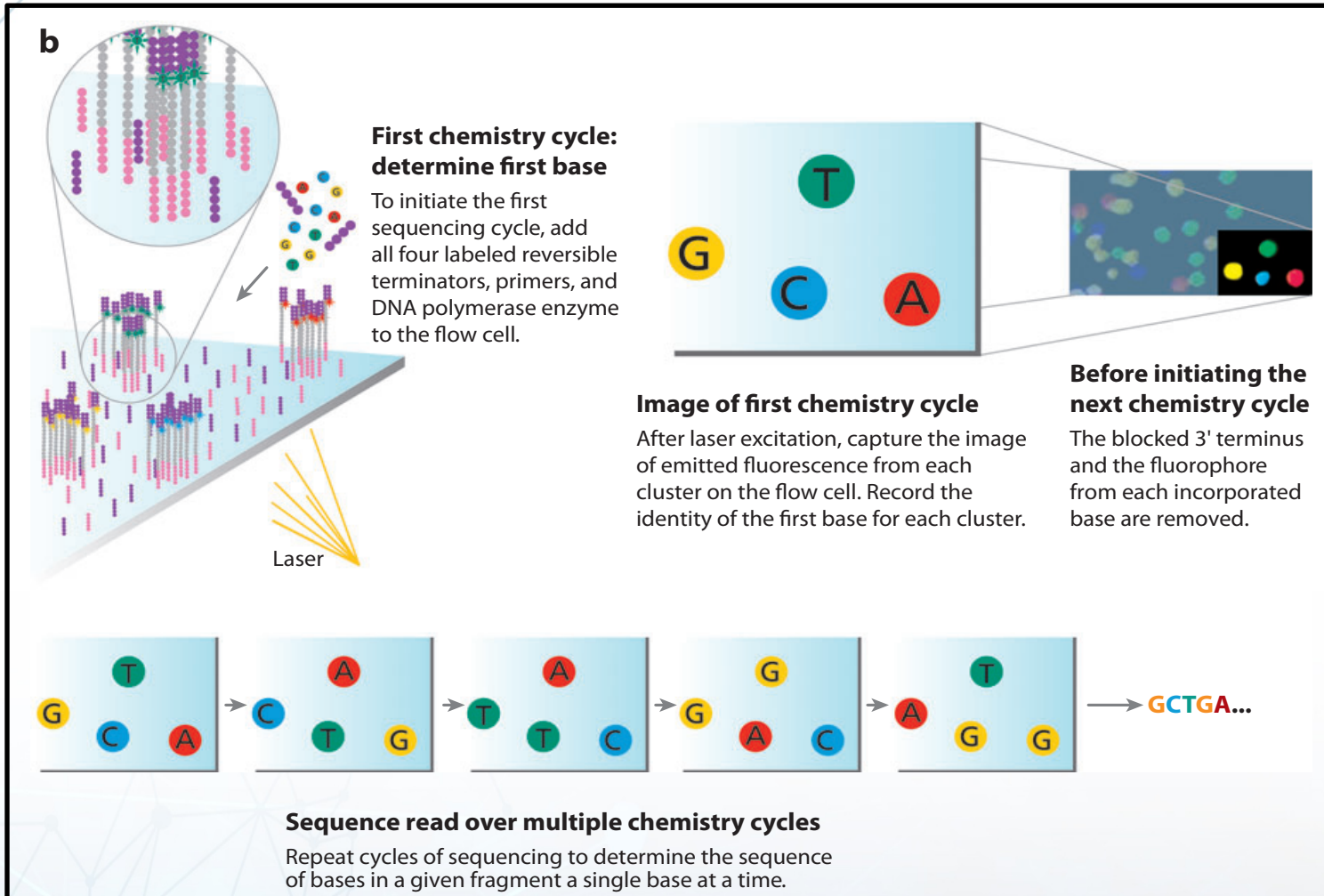


Illumina sequencing-by-synthesis



Next-Generation DNA Sequencing Methods, Elaine Mardis, 2008

Illumina sequencing-by-synthesis

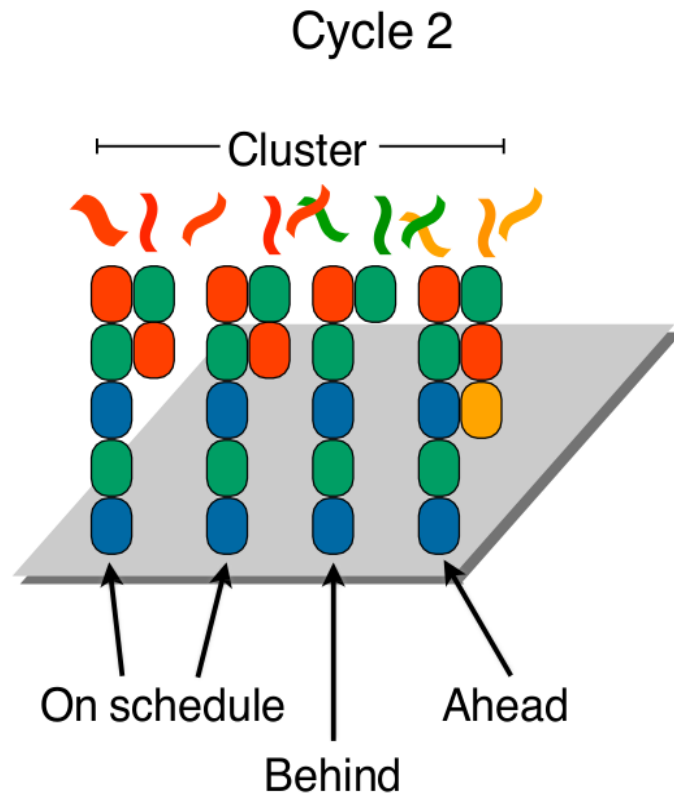


Next-Generation DNA Sequencing Methods, Elaine Mardis, 2008

Sequencing by synthesis: errors



Errors creep in when some templates get “out of sync,” by missing an incorporation or by incorporating 2 or more nucleotides at once



Base caller must deal with this uncertainty. Actual base callers report a *quality score* (confidence level) along with each nucleotide.

Errors are more common in later sequencing cycles, as proportionally more templates fall out of sync

Illumina sequencing summary



Illumina HiSeq

~3 billion paired 100bp reads
~600Gb, \$10K, 8 days
(or “rapid run” ~90Gb in 1-2 days)

Illumina X Ten

~6 billion paired 150bp reads
1.8Tb, <3 days, ~1000 / genome(\$\$)
(or “rapid run” ~90Gb in 1-2 days)

Illumina NovaSeq

20 billion paired 150bp reads
3Tb < 2days

Advantages:

- Best throughput, accuracy and read length for any 2nd gen. sequencer
- Fast & robust library preparation

Disadvantages:

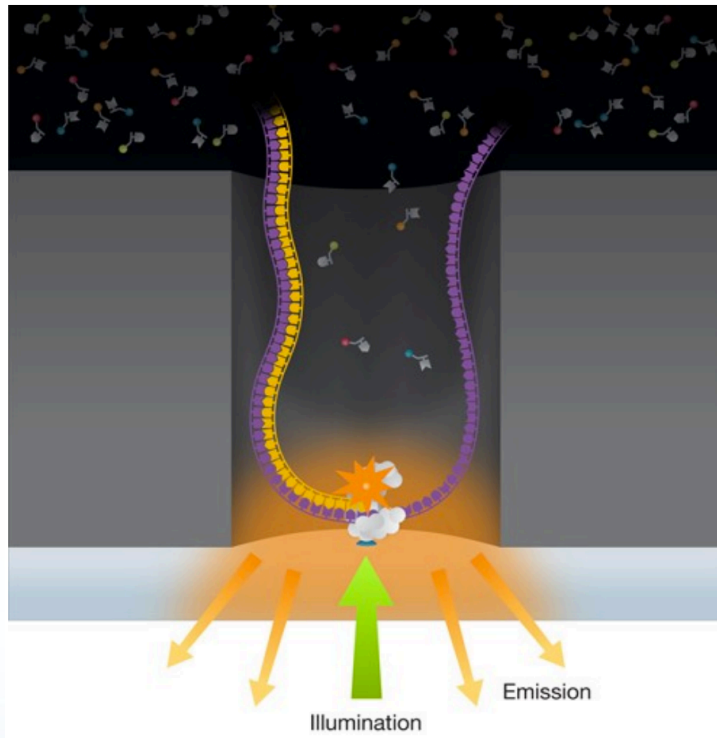
- Inherent limits to read length (practically, 150bp)

Long Reads

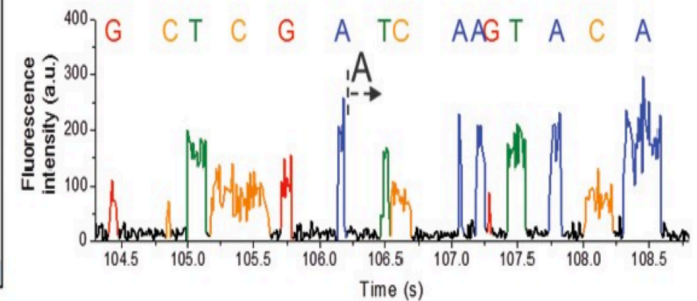
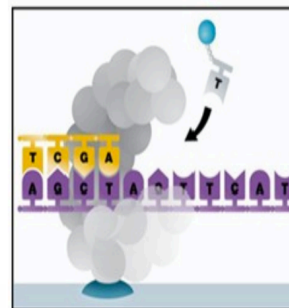


Canadian Centre for
Computational
Genomics

PacBio RS and Sequel systems



4 nucleotides with different fluorescent dye simultaneously present



SMRT Cells containing up to a million ZMWs are processed on PacBio® Systems which simultaneously monitor each of the waveguides in real time.

PacBio

Advantages & limitations

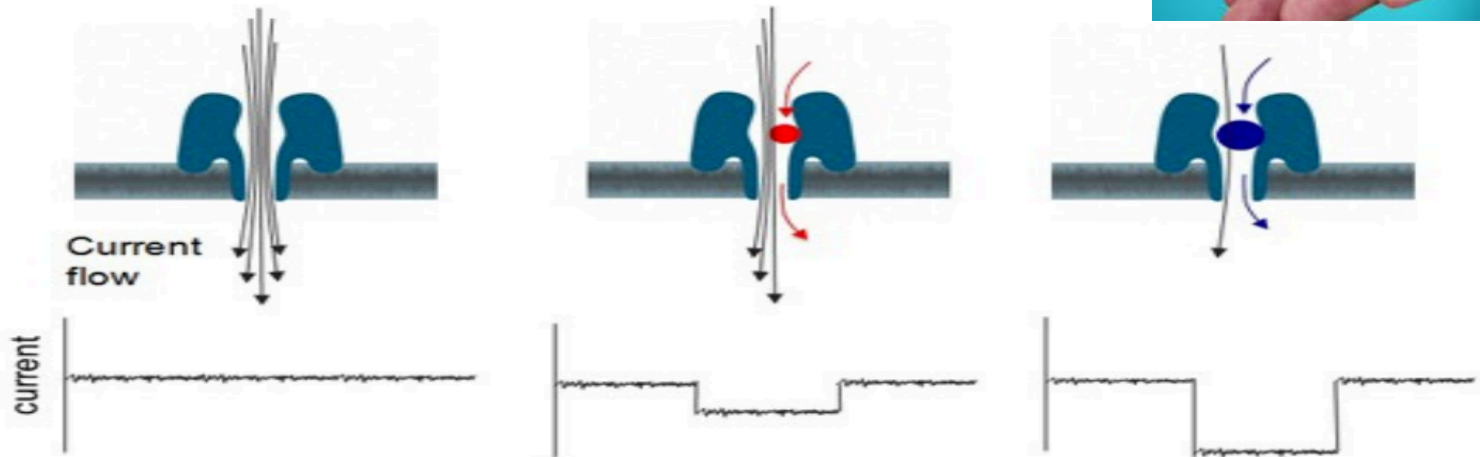


- **Advantages:**
 - Really long reads (up to 70kb)
 - Near random distribution of errors
 - which allows correction in high coverage data
 - No PCR bias
 - Direct detection of modified nucleotides
 - A really high coverage is needed for some modification detection.
 - Circular Consensus Reads (CCS)
 - CCS reads have a low error rate and a length sufficient to solve many long repeats in genomes
- **Limitations:**
 - The amount of input materials
 - The error rate
 - The cost

Nanopore systems



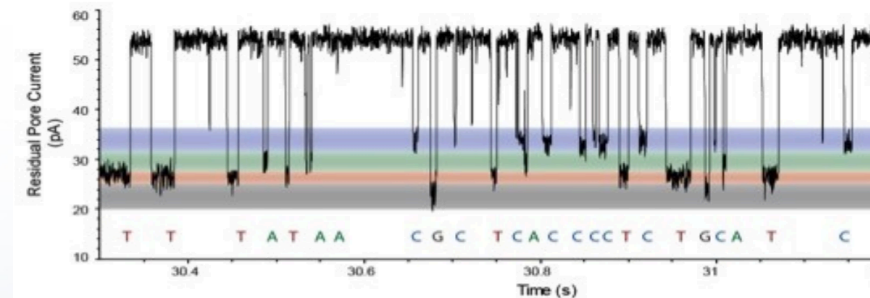
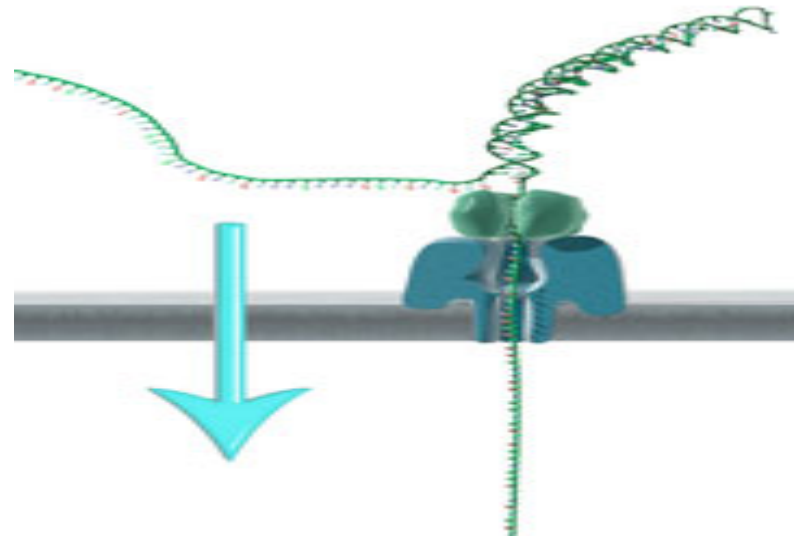
Use nanopore (hemolysin) with inner diameter of 1nm, about 100,000 times smaller than that of a human hair



Nanopore: sequencing



- The DNA sequences are coupled with a zip enzyme which transforms the double helix structure in to a one stranded molecule
- Each different 5-mer going through the pore will a specific modification of the voltage



Nanopore: Advantages & limitations



- **Advantages:**
 - Really long reads (up to 200kb)
 - Low-cost, portable instrument
 - Easy sample prep
 - Can repetitively sequence a given molecule to generate higher quality data
- **Limitations:**
 - The error rate
 - Whole-genome sequencing remains a challenge
 - Performance still being tested and optimized
 - Data processing

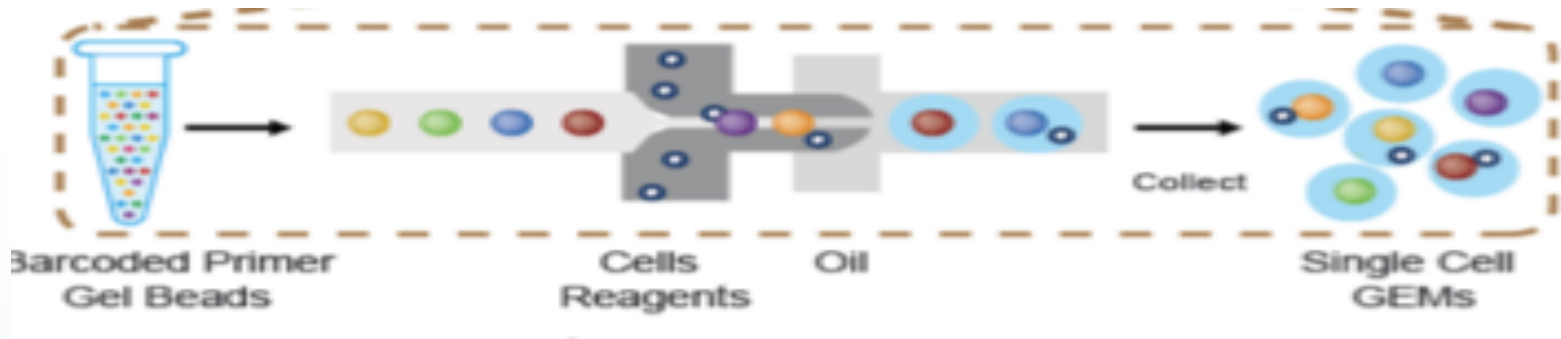
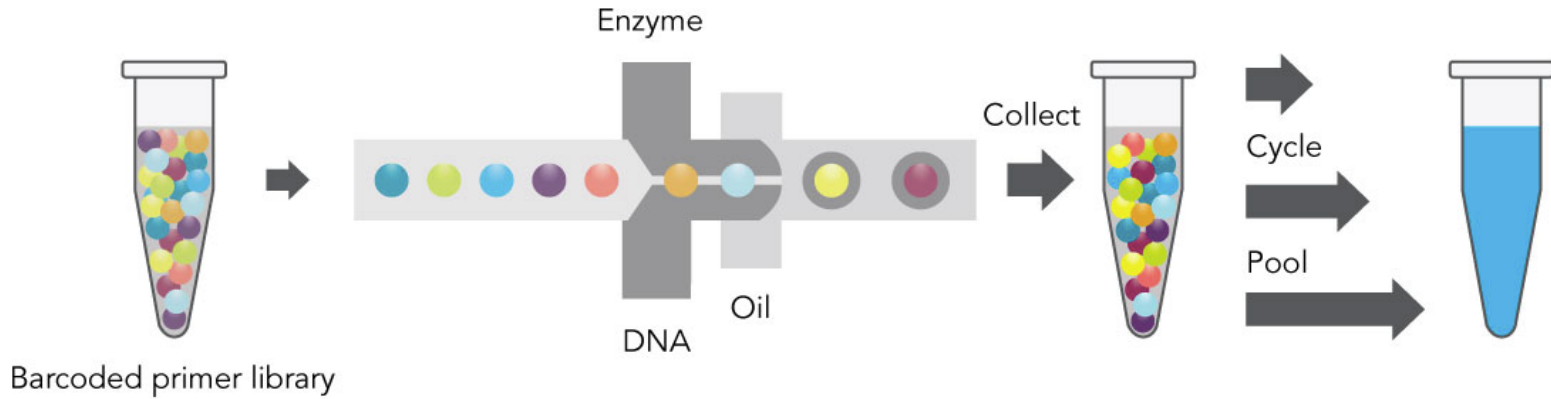


On the side technology

10x Genomics - Technology



Canadian
Comput



10x Genomics: Advantages & limitations











- **Advantages:**

- Compatible with widely used Illumina platform
- Compatible with standard DNA/RNA preps
- Minimal input requirements (1–3 ng)
- DNA: High-quality genome assembly
- scRNA: Large number of cell for a limited cost
- Data processing

- **Limitations:**

- Vulnerable to Illumina biases and limitations
- DNA: Not true long-read and gapped sequence
- scRNA:
 - Depth per cell
 - Only the 3' end of the transcripts is sequenced
- Data processing

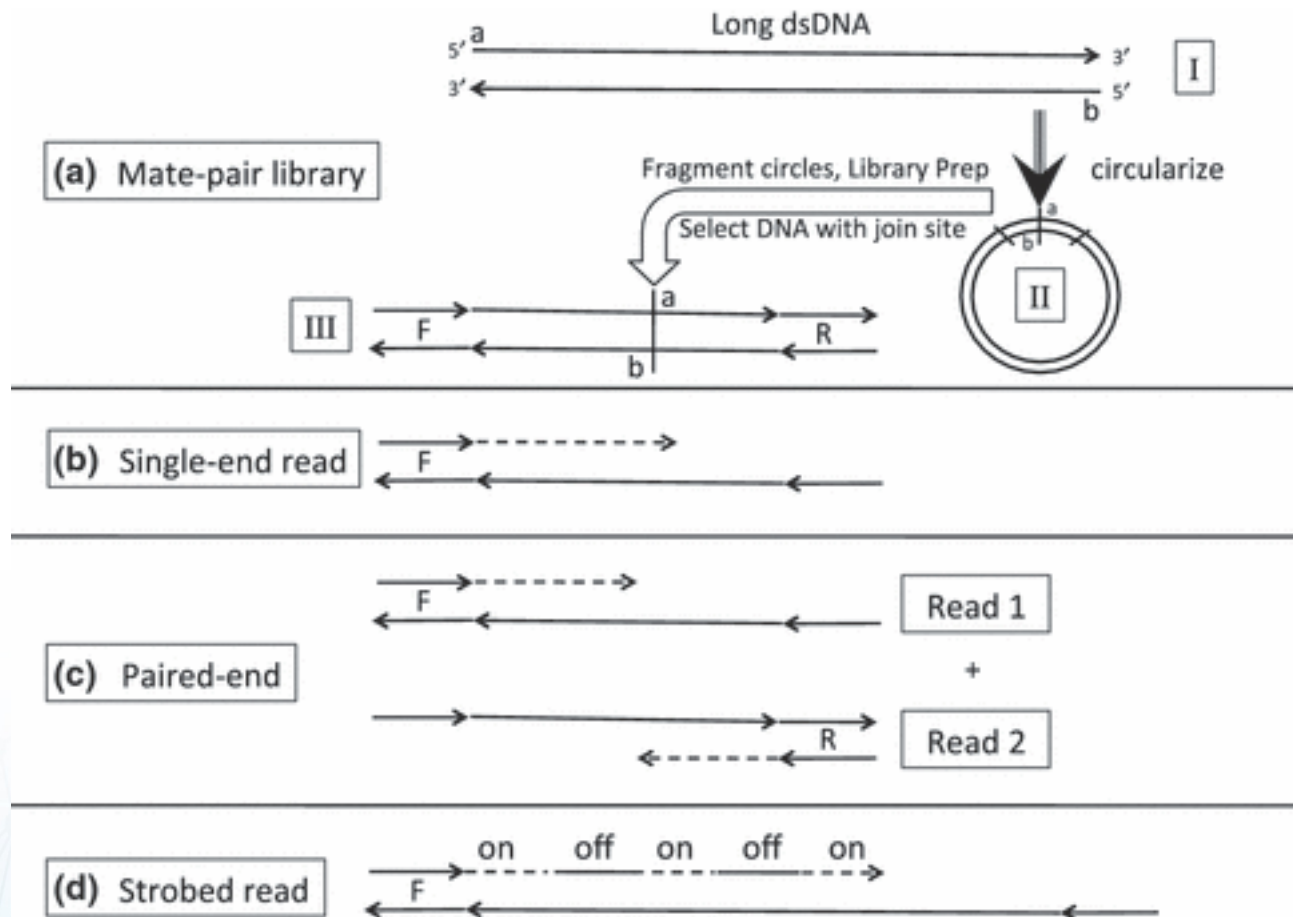
Applications

	Equipment	MUGQIC number	Current Applications
	454	3 (1)	Small <i>de novo</i> genome sequencing Amplicon sequencing Metagenomics Validation
	Ion Torrent	1	
	Illumina MiSeq	2	
	SOLiD	0	Transcriptome sequencing (RNA-Seq), Whole Exome Sequencing, Whole Genome Sequencing, ChIPseq, Whole Genome Bisulfate sequencing, DNase-seq, ...
	Illumina NovaSeq HiSeq 2500/4000/X)	12	
	Pacific Biosciences RS/Sequel	2	Small and medium genomes, Long haplotype sequencing, target sequencing, Epigenomics, Validation
	Nanopore MinION	1	
	10x genomics	1	Whole genome sequencing De novo genome sequencing Single cell sequencing

Some Key Parameters while designing your experiment

- Library type
- Read length
- Error Profile
- Barcoding potential (multiplexing)
- Cost
- Turn around time

Different type of sequencing libraries



From Glenn TC, *Mol Ecol Resour.* 2011 adapted for 2013

What are paired reads?



BICG_2012_Module4.pdf (application/pdf Object) - Mozilla Firefox

File Edit View History Bookmarks Tools Help

ChipSeq expertise (contract) | Bioinforma... BICG_2012_Module4.pdf (application/pdf... TOYOTA CANADA: Configuration/prix

bioinformatics.ca/files/public/BICG_2012_Module4.pdf

18 / 37 147% Collaborate Sign Find

What are Paired Reads?

Paired-end Reads

The diagram illustrates a paired-end read. A horizontal line represents a DNA fragment. The left end is labeled 'ATCAA' and the right end is labeled 'CTAAG'. A red line segment connects the two ends, representing the sequenced region. A bracket below the red line is labeled 'Insert size (IS)'. Above the red line, the text 'Paired-end Reads' and 'DNA fragment' are centered.

ATCAA CTAAG

Insert size (IS)

Slides by M. Brudno

Module bioinformatics.ca

Start 9:28 PM 3/6/2013

Read Length

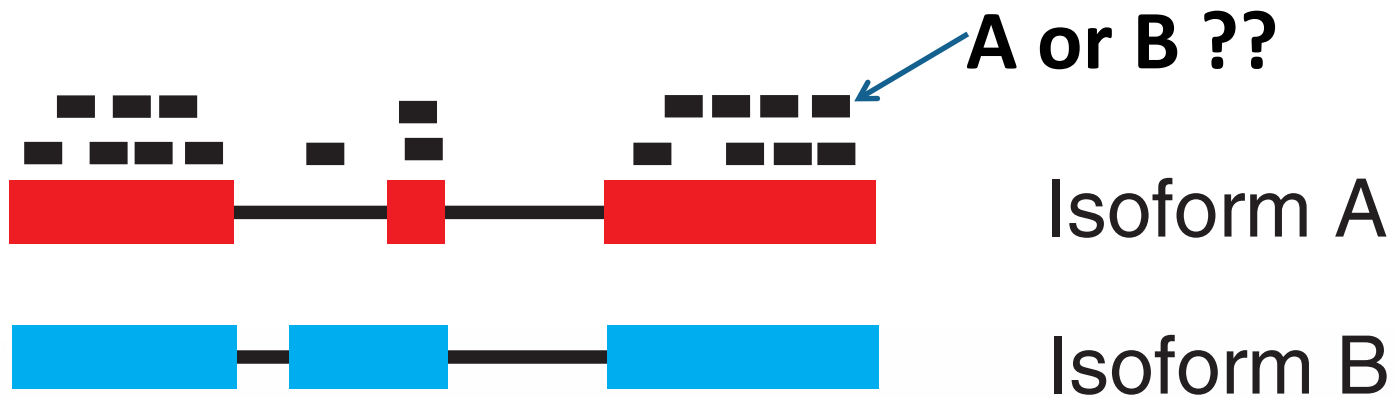


- Illumina HiSeq:
 - up to 250-300 bp for now but the 100-150bp is still the standard
- Pacbio and Minlon:
 - > 50kb but with a very large range of read lengths in the same run.
- Short Reads are sufficient for re-sequencing applications (known genome reference)
- Longer Reads are beneficial for *de novo* genome assemblies

Read Length



Longer reads are also good in transcriptomics:



Error Profile



NGS reads have errors; diff. technologies, different rates

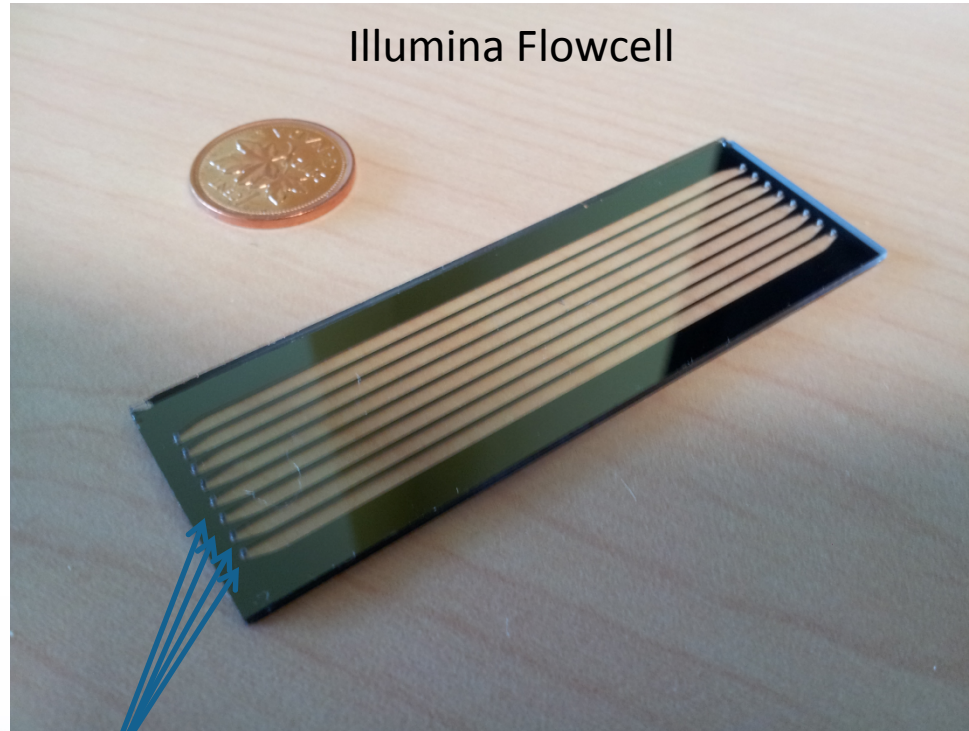
instrument	Nanopore	Pacbio	Ion Torrent	454	Illumina	SOLiD
single-Pass Error rate %	~12 (1-3)	~13 (~1)	~1	~0.1	~0.1	~0.1

Source: 2014 NGS Field Guide, Glenn TC.

How to deal with errors:

1. Remove it: it works for technologies with semi-random error distribution and with higher throughput
2. Correct it : it works for non-random errors but needs high depth of sequencing or hybrid sequencing design

Multiplexing (Barcoding)

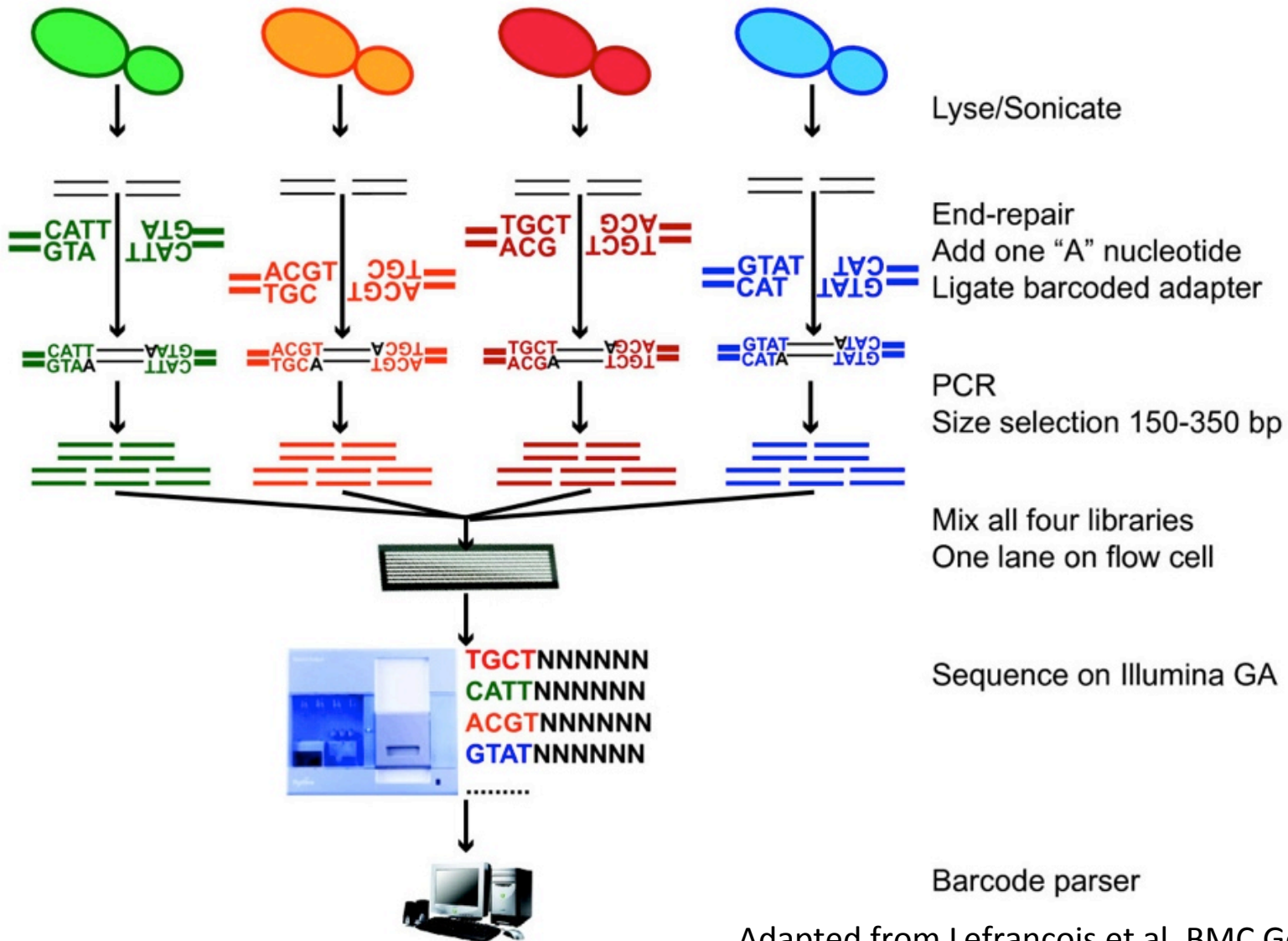


Illumina Flowcell

8 lanes
150M 2x100 bp reads
each

What if only 50M reads per
samples are sufficient?

Multiplexing (Barcoding)



Outline



1. The technology

2. Types of data

3. Conclusions

What is the NGS short read problem all about ?



- Strings of 100 to \approx 50kb letters
- Puzzle of 3,000,000,000 letters
- Usually have 120,000,000,000 letters you need to fit
- Many pieces don't fit :
 - sequencing error/SNP/Structural variant
- Many pieces fit in many places:
 - Low complexity region/microsatellite/repeat



DNaseq

Why DNAseq?



- Whole genome sequencing:
 - Whole genome SNV detection
 - Structural variant
 - Capture the regulatory region information
 - Cancer analysis
 - De novo genome assembly
- Whole exome sequencing:
 - Cheaper
 - Captures only the coding region information
 - Rare diseases analysis

DNAseq – SNP Discovery



GTTACTGTCGTTGTAATACTCCACGATGTC
GTTACTGTCGTTGTAATACTCCACGATGTC
GTTACTGTCGTTGTAATACTCCACGATGTC
GTTACTGTCGTTGTAATACTCCACAAATGTC
GTTACTGTCGTTGTAATgCTCCACGATGTC
GTTACTGTCGTTGTAATACTCCACAAATGTC
GTTACTGTCGTTGTAATACTCCACGATGTC
GTTACTGTCGTTGTAATACTCCACaATGTC
GTTACTGTCGTTGTAATACTCCACaATGTC
GTTAaTGTCGTTGTAATACTCCACGATGTC
GTTACTGTCGTTGTAcTACTCCACGATGTC
GTTACTGTCGTTGTAATACTCCACaATGTC



sequencing errors

SNP



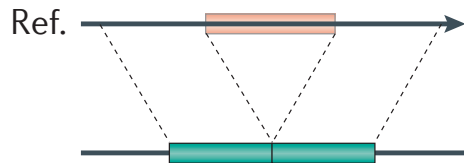
An accurate SNP discovery is closely linked with a good base quality and a sufficient depth of coverage

DNaseq – structural variants

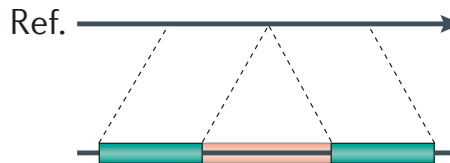


(Re-)sequence genomes to compare to a reference

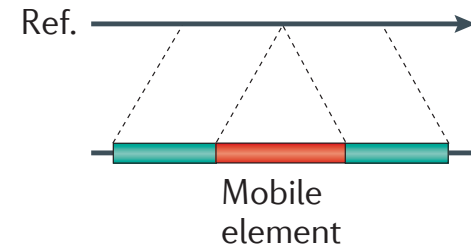
Deletion



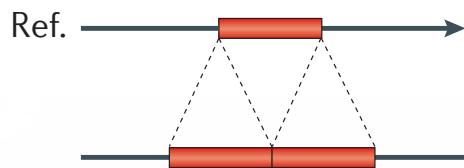
Novel sequence insertion



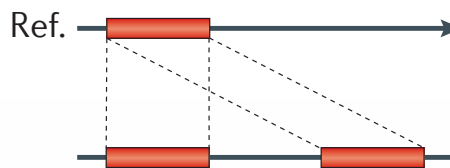
Mobile-element insertion



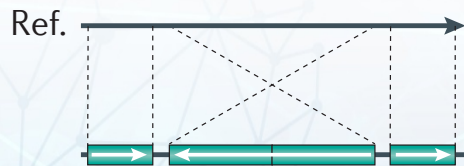
Tandem duplication



Interspersed duplication



Inversion



Translocation

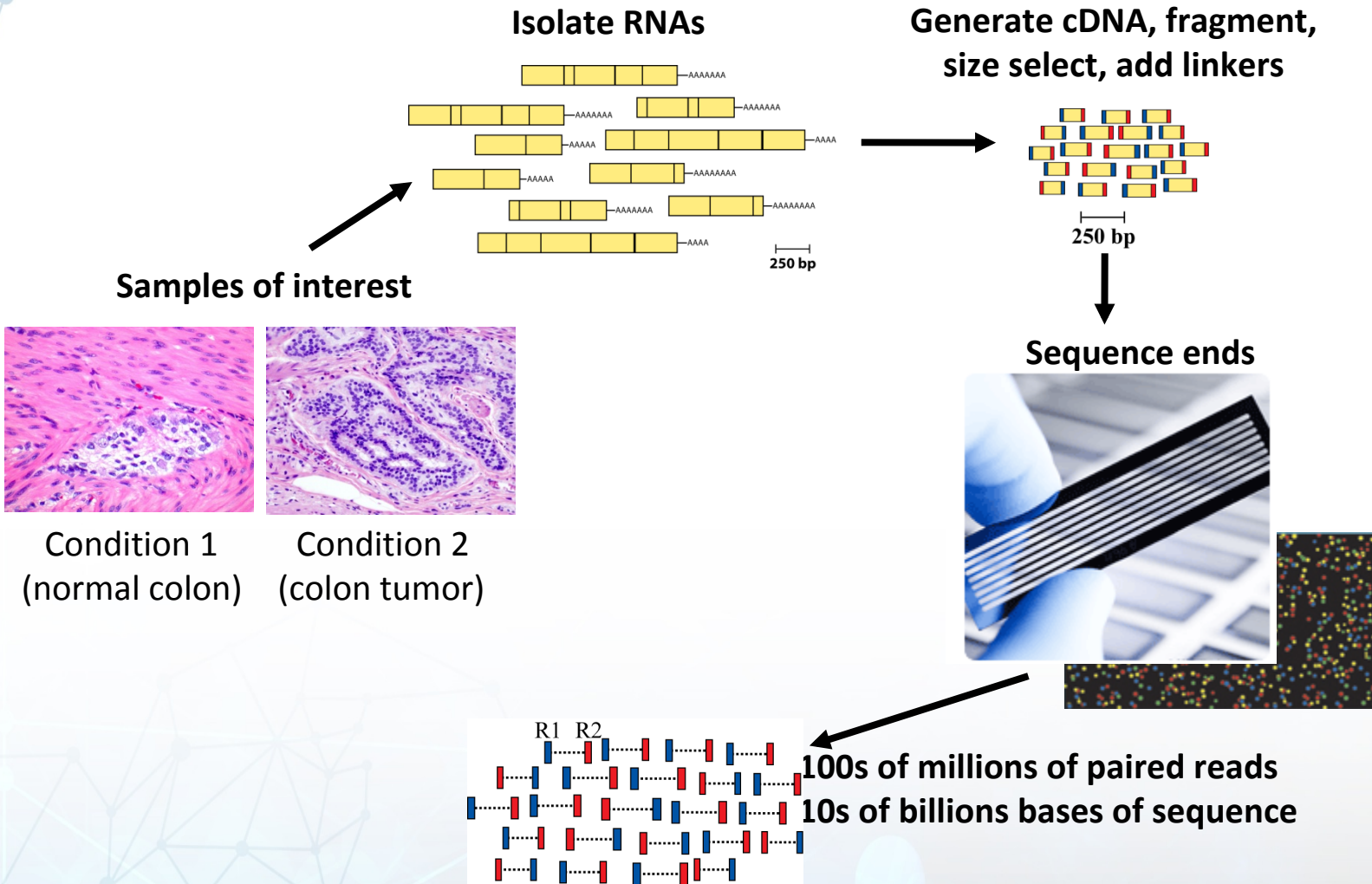


Genome structural variation discovery
and genotyping



RNAseq

RNA sequencing



RNAseq Challenges



- RNAs consist of small exons that may be separated by large introns
 - Mapping reads to the genome is challenging
 - Ribosomal and mitochondrial genes are misleading
- RNAs come in a wide range of sizes
 - Small RNAs must be captured separately
- RNA is fragile and easily degraded
 - Low quality material can bias the data

Why sequence RNA?

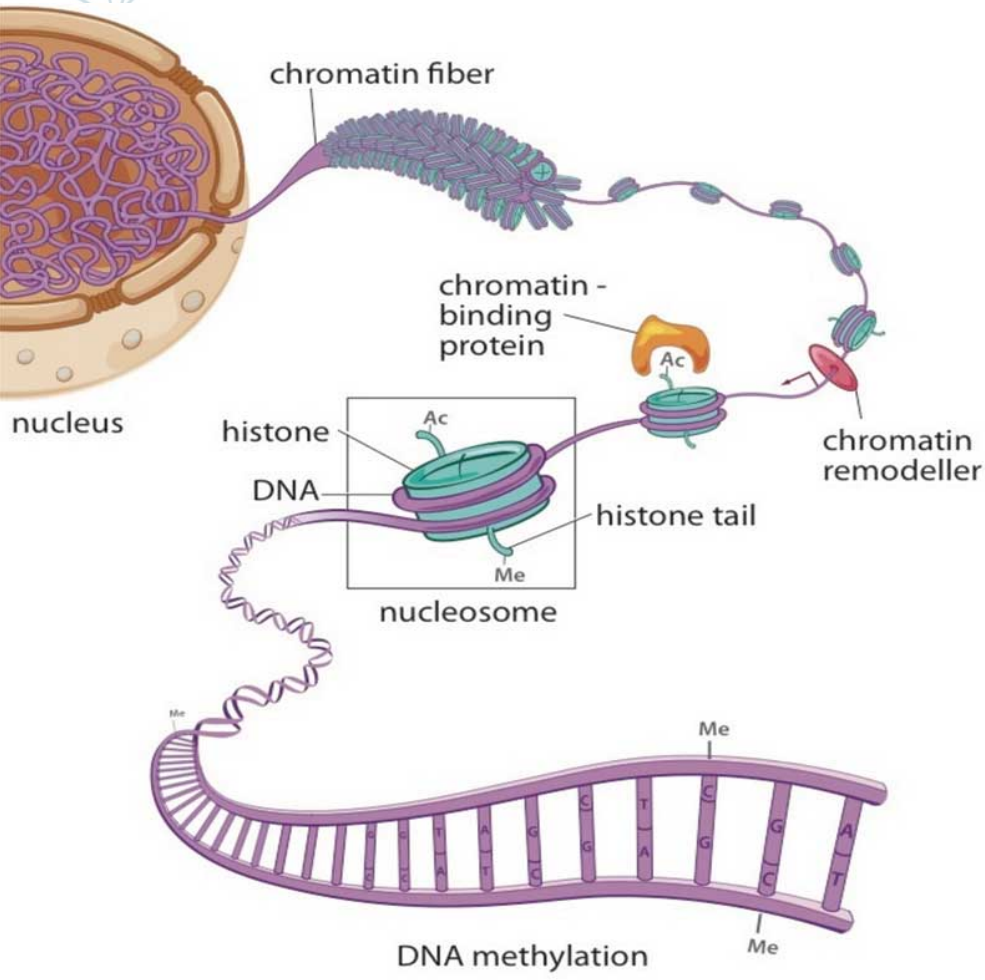


- Functional studies
 - Genome may be constant but experimental conditions have pronounced effects on gene expression
- Some molecular features can only be observed at the RNA level
 - Alternative isoforms, fusion transcripts, RNA editing
- Interpreting mutations that do not have an obvious effect on protein sequence
 - ‘Regulatory’ mutations
- Prioritizing protein coding somatic mutations (often heterozygous)



Epigenomics

Epigenetics



Studies changes in gene expression which are not encoded by the underlying DNA sequence

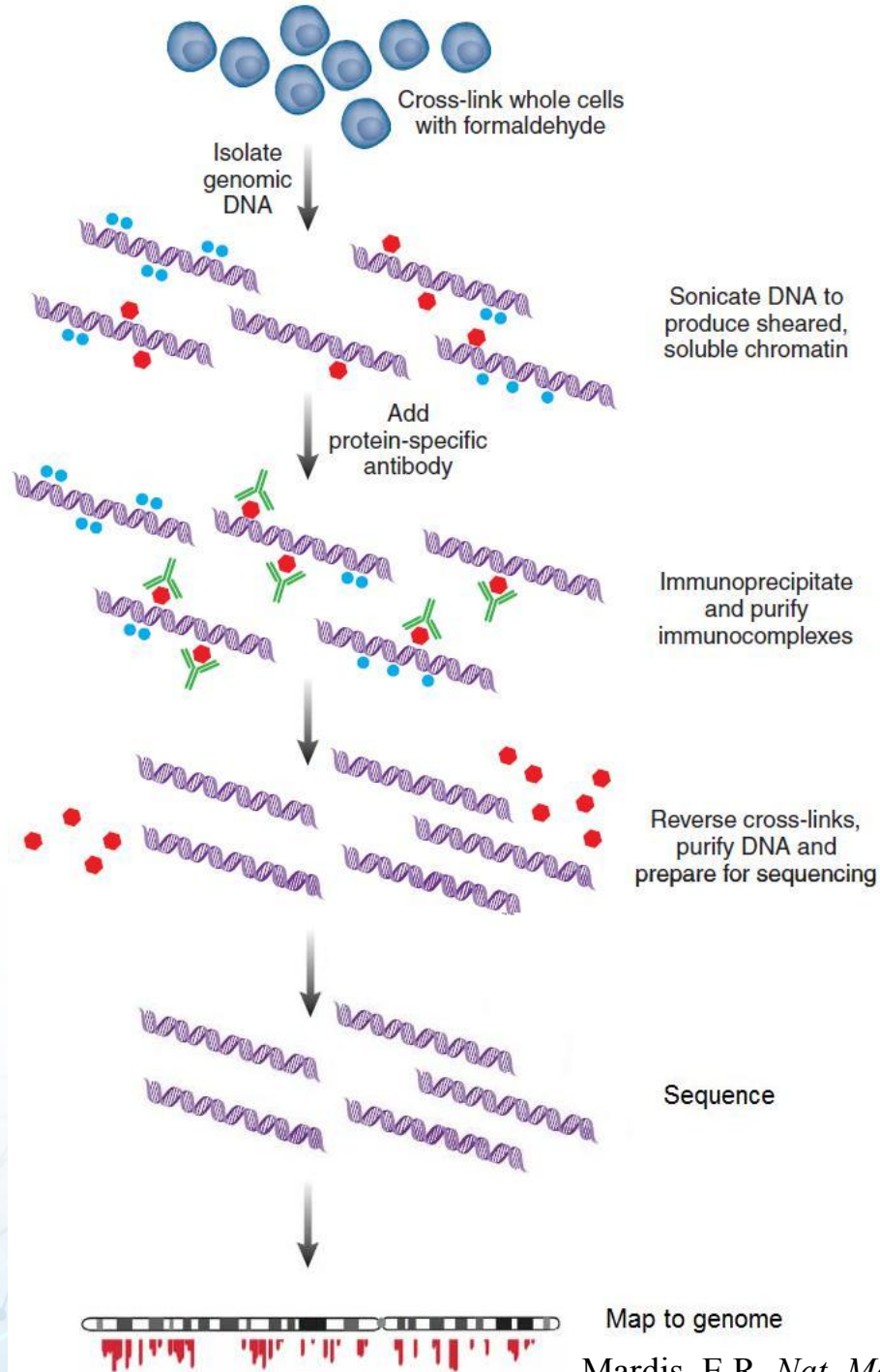
1) histone modification (accessibility/compaction)

2) DNA methylation

What is ChIP-Sequencing?



- Combination of chromatin immunoprecipitation (ChIP) with ultra high-throughput massively parallel sequencing
- Allows mapping of protein–DNA interactions *in vivo* on a genome scale
- Why run a ChIP-seq experiment:
 - Transcription factors and other chromatin-associated proteins influence phenotype
 - Can be evaluated for the entire genome in a single experiment





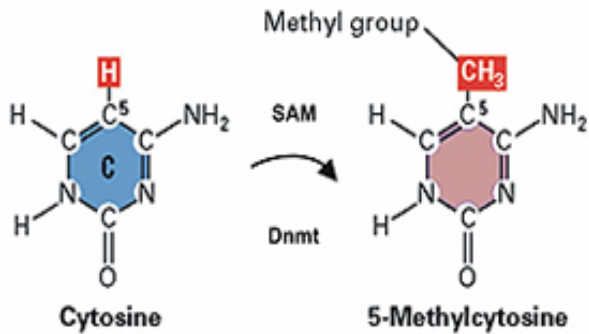
Methylseq

Why Methyseq ?

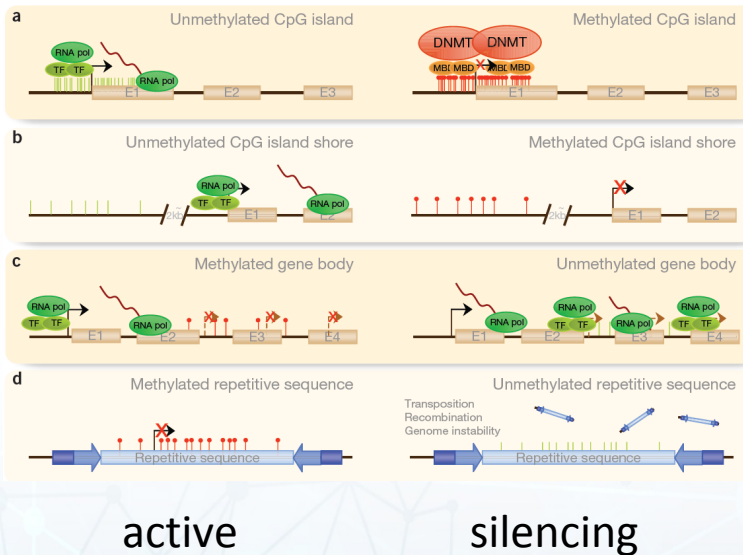


- Cytosine methylation can significantly modify temporal and spatial gene expression and chromatin remodeling.
- Whole-genome bisulfite sequencing (WGBS) provides a comprehensive view of methylation patterns at single-base resolution across the genome.

DNA Methylation: Background

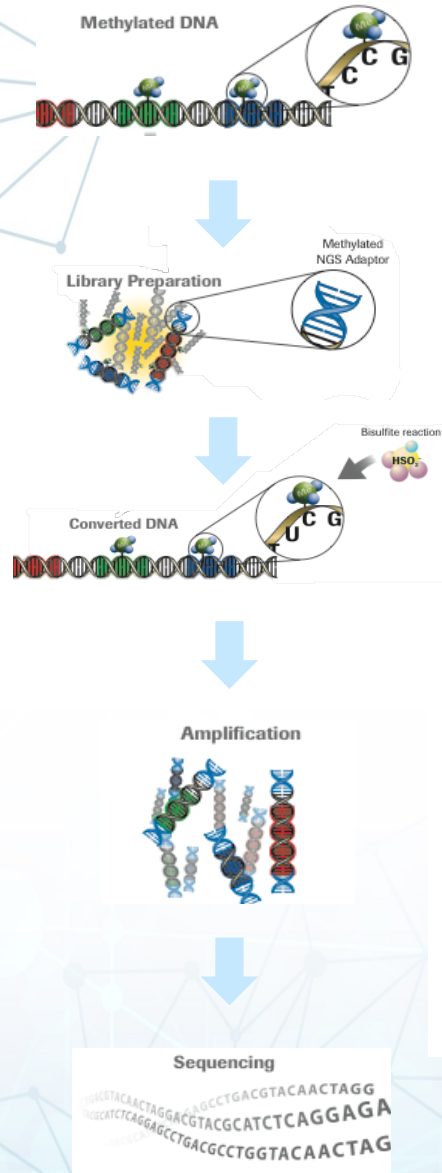


- DNA methylation is one of the most commonly occurring epigenetic events in the mammalian genome
- DNA methylation plays a role in **silencing of genes**, and in X-chromosome inactivation
- DNA methylation plays a role in the establishment and maintenance of **imprinted genes**



Portela et al. 2010, Nat Biotech 28 (10), 1057

Bisulfite Sequencing



Watson >>**AC^mGTT**CGCTT**GAG**>>

Crick <<**TGC^mAAG**CGAACTC****<<

C^m methylated
C Un-methylated

1) Denaturation



Watson >>**AC^mGTT**CGCTT**GAG**>>

Crick <<**TGC^mAAG**CGAACTC****<<

2) Bisulfite Treatment



BSW >>**AC^mGTT**UGU**TTGAG**>>

BSC <<**TGC^mAAG**UGAAUTU****<<

3) PCR Amplification



BSW >>**AC^mGTT**TGTT**TTGAG**>>

BSC <<**TGC^mAAG**TGAATT****<<

BSWR <<**TG CAAACAAACTC**<<

BSCR >>**ACG TTC**ACTTAA****>>

Whole-genome bisulfite sequencing (WGBS): detect DNA methylation at single base resolution genome-widely.

Outline



1. The technology

2. Types of data

3. Conclusions

Sequencing technology summary



- **Illumina:**
 - 100-200bp reads
 - Up to 600Gbp per run*
 - Very low error rate (<1% bases miscalled)

- **Pacbio/Oxford Nanopore:**
 - Single molecule sequencing (no amplification)
 - >50kb bp reads
 - 5-10 Gbp per run*
 - Higher error rate (5-15%)
 - Can detect modified bases

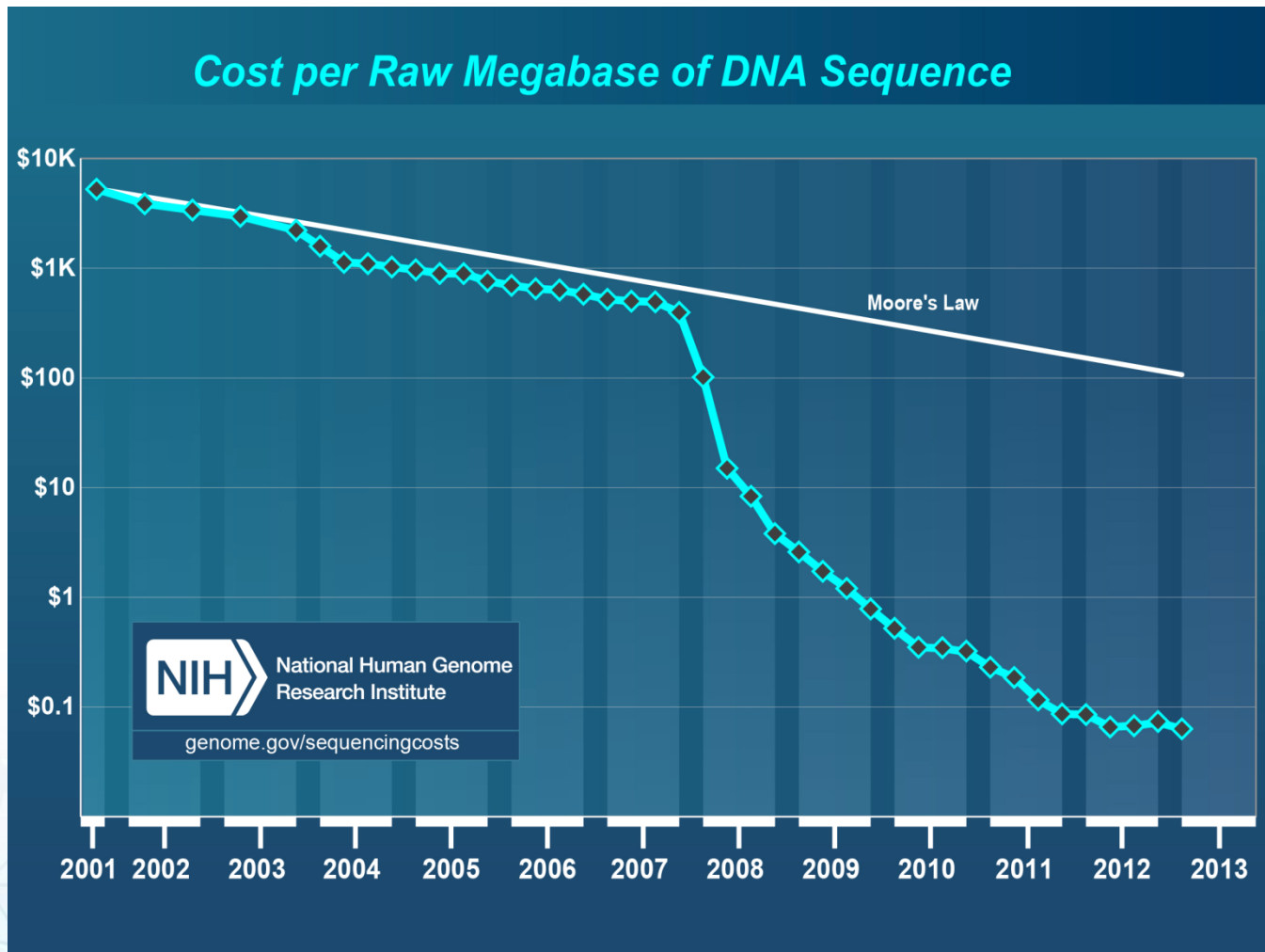
Notes



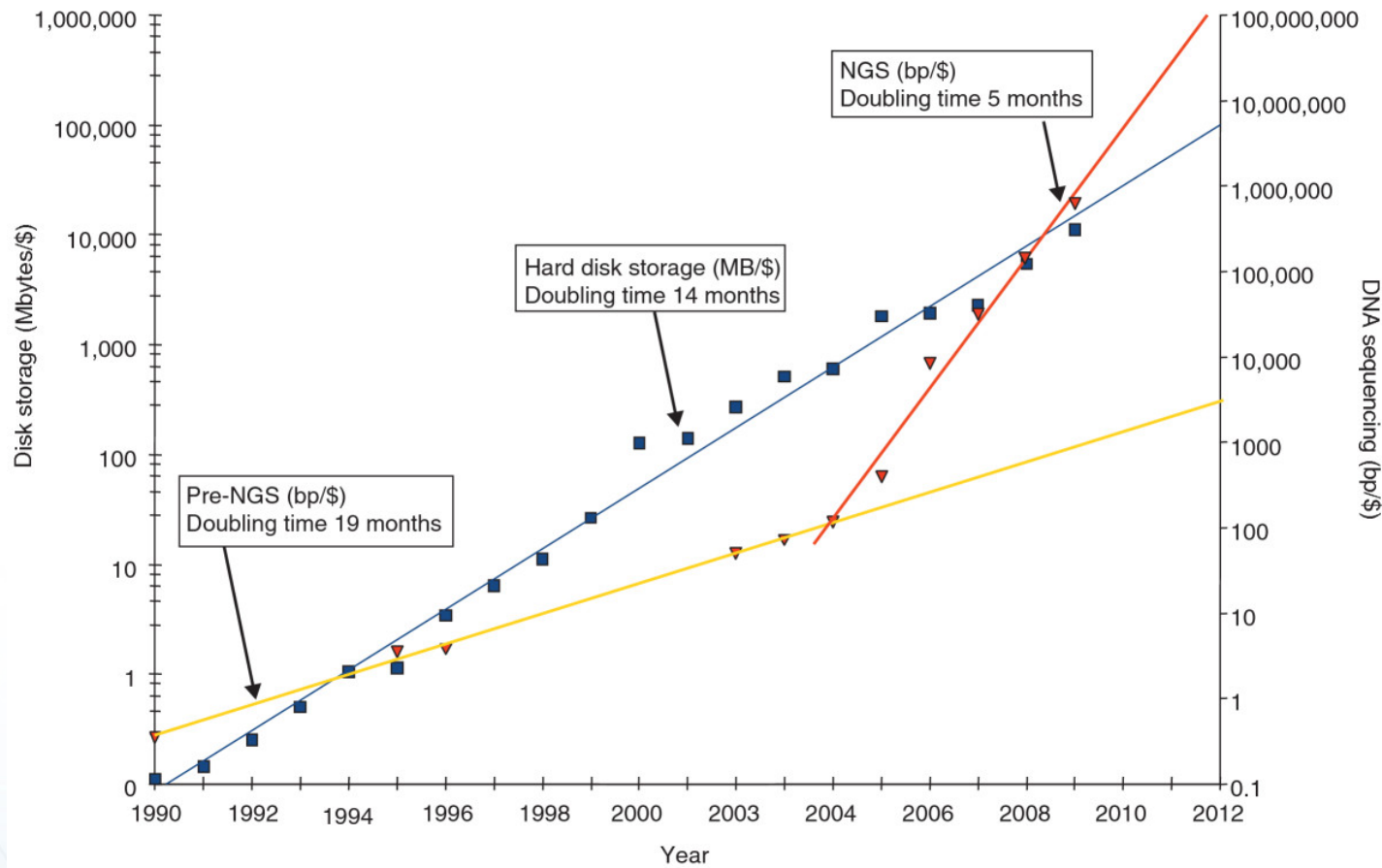
- NGS offers a variety of technologies and methods
- A good knowledge of errors and technicality allows a better choice of analysis and a better understanding of results
- NGS analyses requires both mathematics and informatics skills
- The major challenge is actually link to the analysis, the compute and storage capacities

Cost of sequencing

Good news: Cost of sequencing rapidly decreasing



Next-generation sequencing (NGS)



Will computers crash genomics?

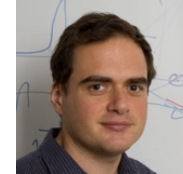


Canadian Centre for
Computational
Genomics



Pennisi, Science, 2011

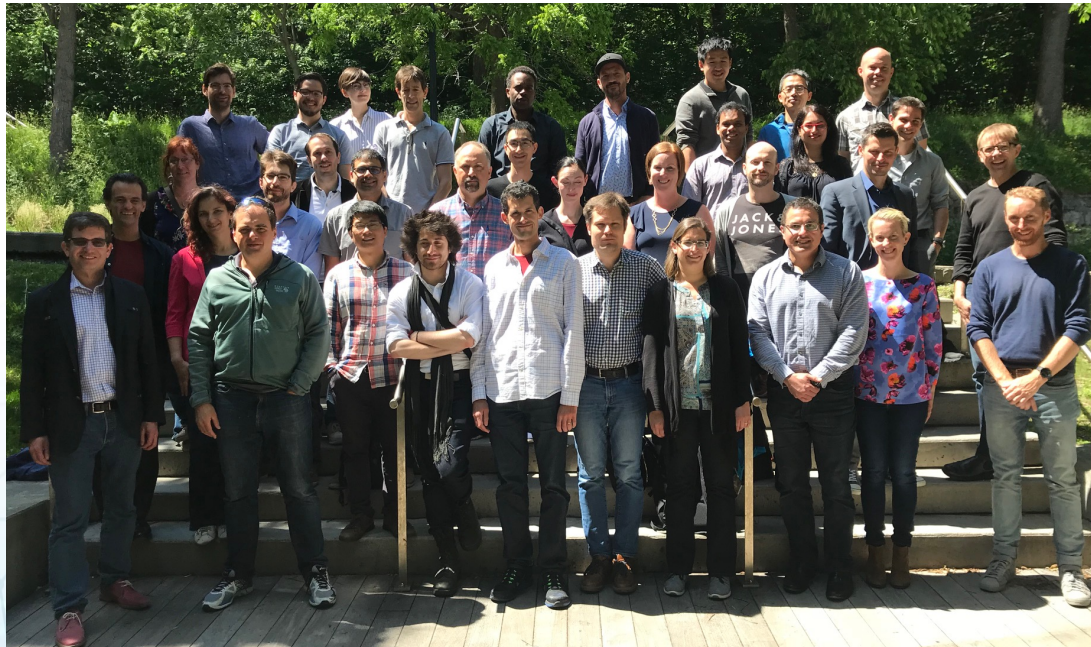
About us



GenomeCanada

SickKids®

*C3G provides bioinformatics **analysis**, **HPC** services and solutions for the life science research community.*



" *The \$1,000 genome, the \$100,000 analysis?*" Elaine R. Mardis



Genome Québec



Genome Canada



Ontario Genomics



Canadian Centre for
Computational
Genomics



Thank you!



compute | calcul
canada | canada

