Weill Cornell Medicine

Core Laboratories Center

"I have samples for sequencing"

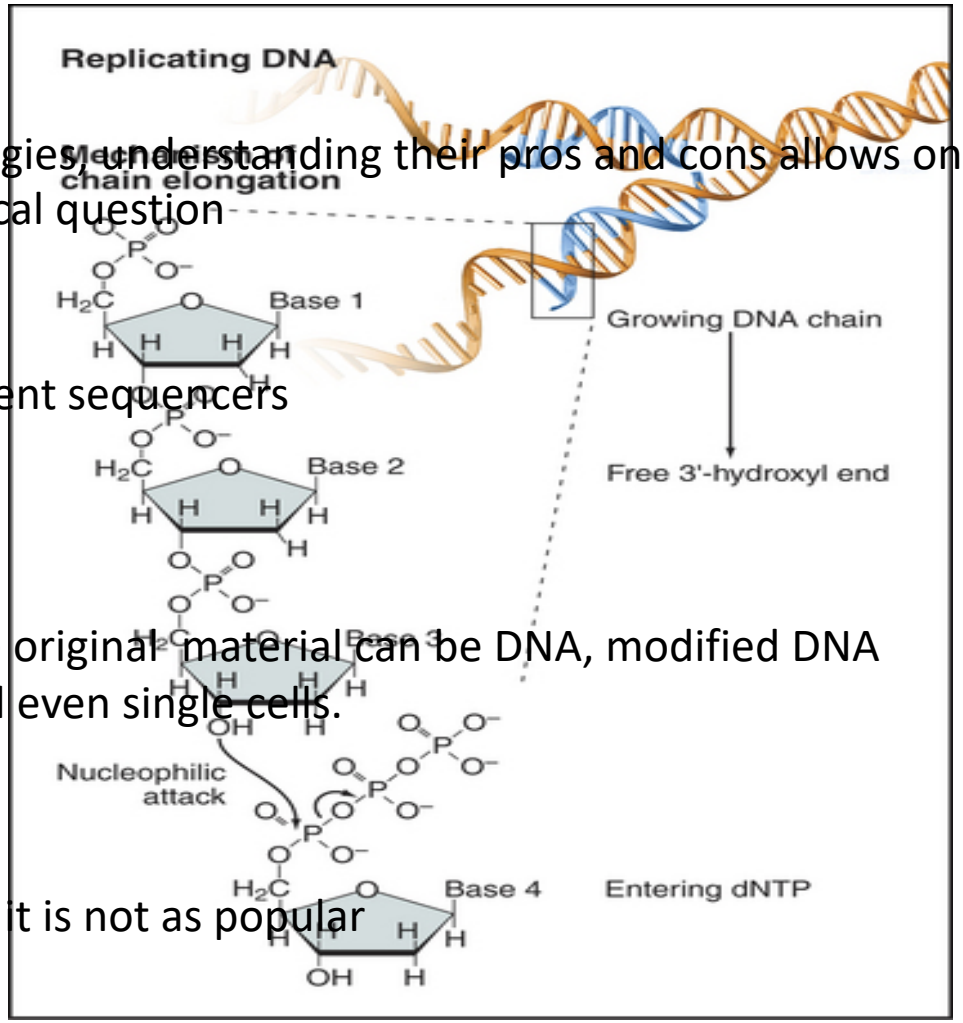
Alicia Alonso, PhD.
Director, Epigenomics Services
Genomics and Epigenomics Core Facility
Assistant Research Professor, Department of Hematology and Oncology

03242020

Notions

- Sequencing is the process of determining the order of the four bases (A,G,C, T) in DNA

- **Most** sequencing is done by using synthetic DNA replication, based on Frederick Sanger's protocol, 1977
  DNA polymerase, primer and tagged nucleotides



- There are different sequencing technologies, understanding their pros and cons allows one to choose the correct one for the biological question

- Within each technology there are different sequencers

- Although sequence is done on DNA, the original material can be DNA, modified DNA (base modification, protein), RNA and even single cells.

- Sequencing can be done from RNA, but it is not as popular

Outline

- Review of the top technologies and sequencers

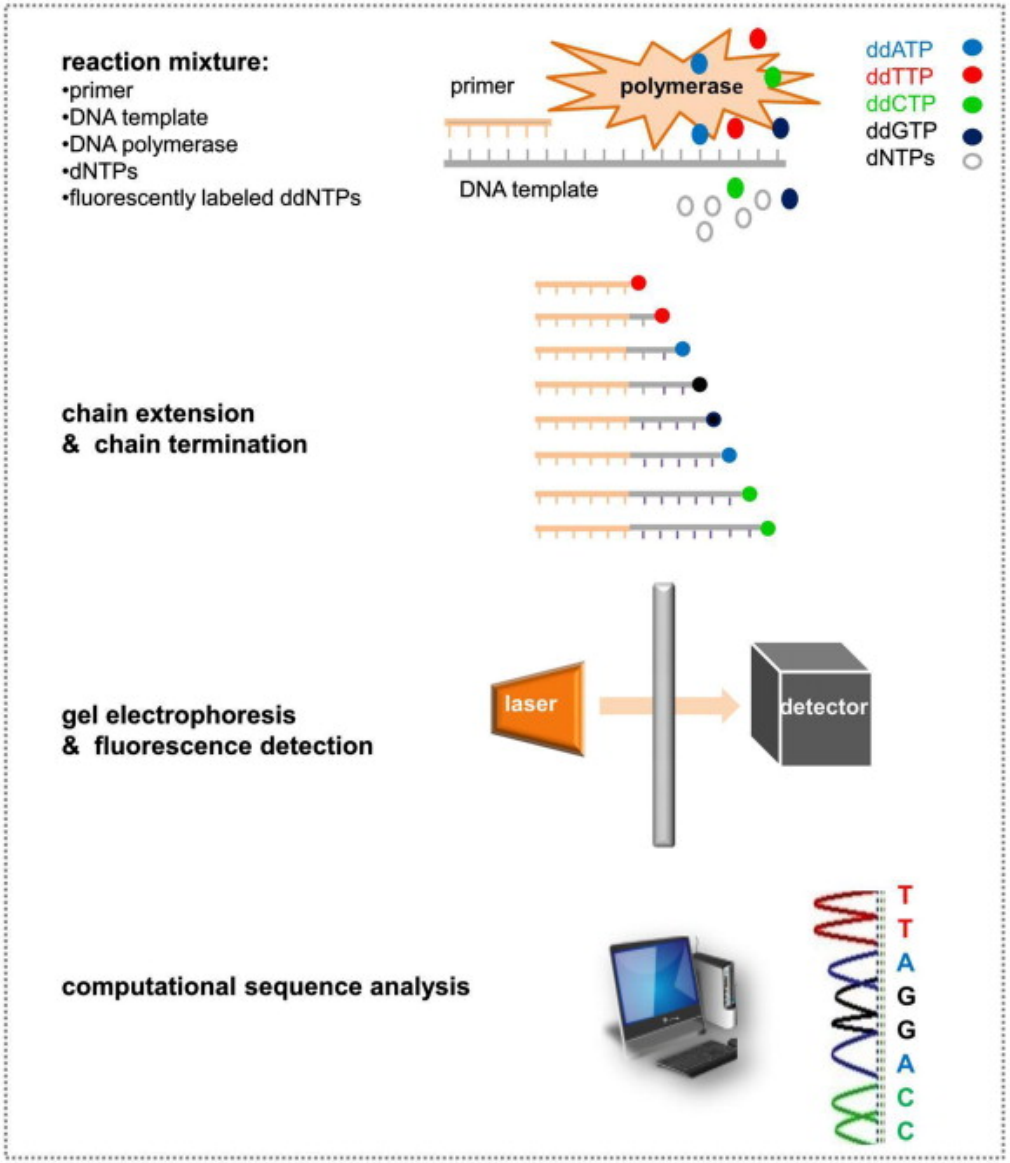
Company	Technology	Sequencer
ThermoFisher	Sanger sequencing	ABI Prism 375
Illumina	SBS –sequencing by synthesis-	HiSeq/NovaSeq
PacBio	SMRT -single molecule real time-	RS/Sequel
Oxford Nanopore	nanopore sequencing	MinION/PromethION
MGI/BGI	DNB -DNA nanoballs-	DNBSEQ

- Instead of the tour of the lab, movies depicting the instruments
  they are corny, but …

- Substrates for sequencing
  - DNA
  - modified DNA
  - RNA
  - single cells

Evolution of Sequencing Technologies

- DNA , primer, dNTPs, dideoxyterminators
- Custom sequencing (targeted)
- Sequencing length: 300-1000bp
- Output: 384 independent samples at a time
- 1 million bases per day
- ABI Prism 375 (Applied Biosystems)

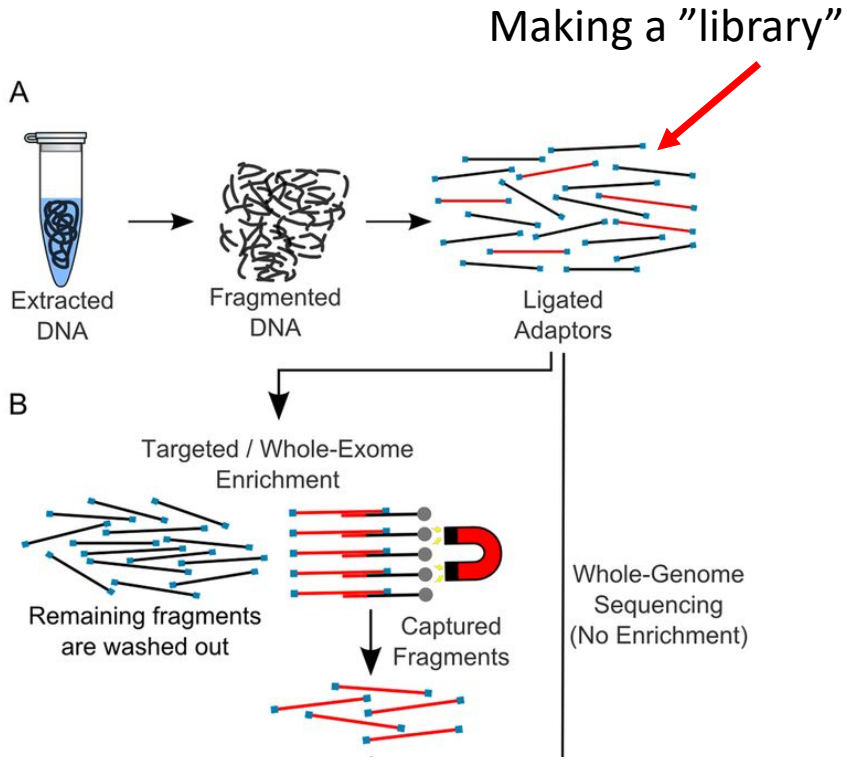


Human Genome Project (1991-2013)

- International team of researchers looking to sequence and map all genes of members of our species, Homo Sapiens, **was done on Sanger sequencing**.

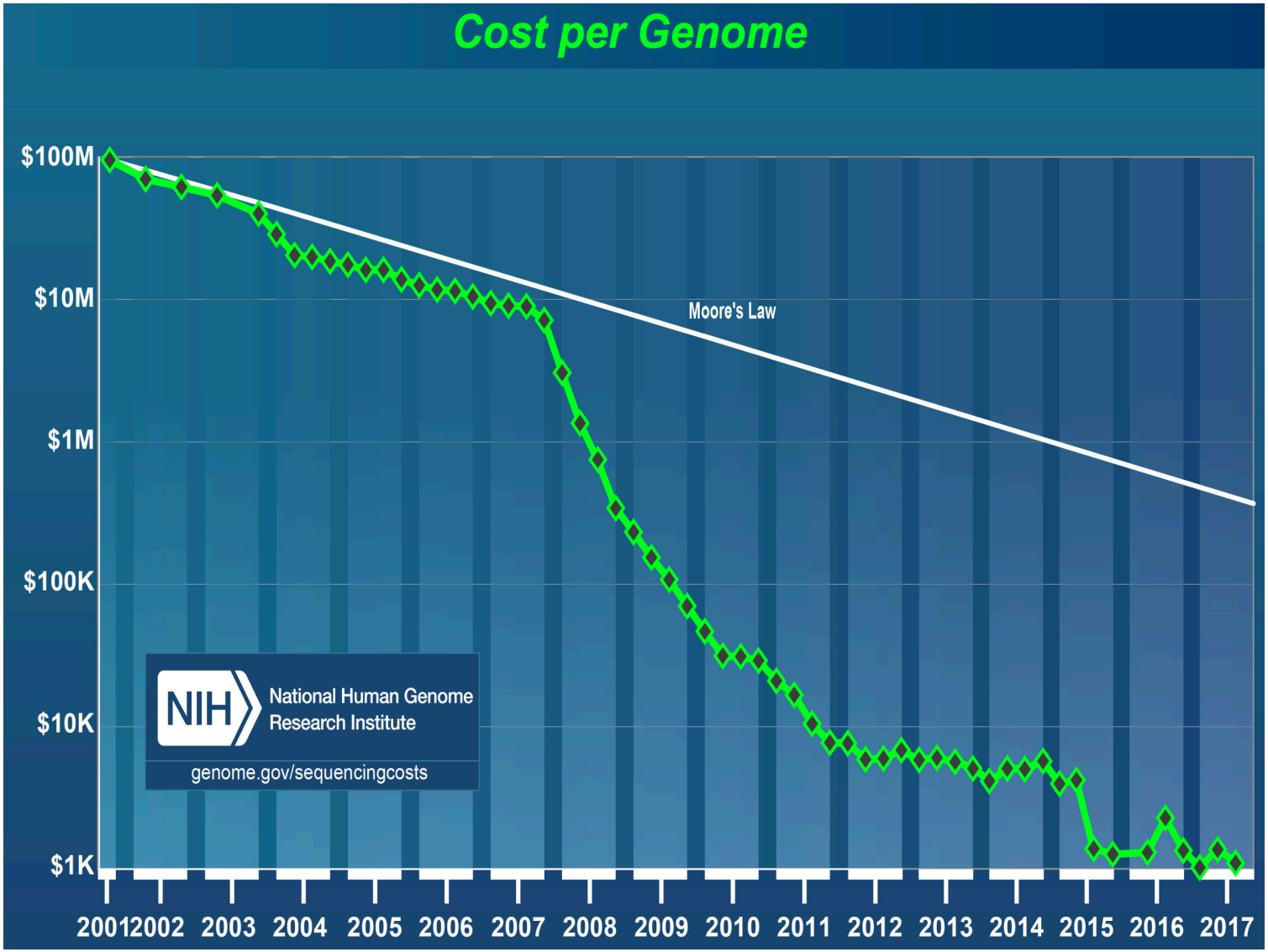
- It was the HGP that spurred the development of faster and cheaper sequencing,
  - massively parallel sequencing

Evolution of Sequencing Technologies



- Generation of many millions of **unique** short reads **to be sequenced** in parallel (150-600bp in length)
  - Speed of sequencing (compared to 1st generation)
  - Cost of sequencing (lower per base)
- Output is detected directly, usually fluorescence (ie no electrophoresis)
- Output in constant increase from 15M – 400M- 1B -20B "short reads" per sequencing run
- Short reads are aligned to a **sequence backbone** – resequencing of known genome to answer biological questions
- Examples of 2nd generation competing technologies
  - Roche/454 (pyrosequencing – extinct- up 1000bp, 1Gb
  - Ion torrent (detects the Hydrogen ion released, change of pH-Life Technologies, 2010) 200-600bp, 10Gb (2-8 hours)
  - ABI/SOLiD 35-75 bp 30Gb/run (each base is read 2x, high accuracy)
  - ***Illumina/Solexa sequencing: Sequencing by Synthesis SBS (50-250bp)***
  - ***BGI/MGI: DNA NanoBalls (upcoming)***

Sequencing costs drop, but it requires high throughput



ABI, 1M reads per run

NovaSeq 20B reads per run

Evolution of Sequencing Technologies

Shortcoming of 2nd generation

- Reads are short and mapped to a backbone
    - Prevent studies of repetitive regions of the genome (centromeres and telomeres)
    - long structural variations
- "Library" preparations require PCR amplification
    - introduction of PCR bias due to DNA polymerase
    - Cannot identify base modifications

Evolution of Sequencing Technologies

Sequencing length: 100kb-2Mb

- Pacific Biosciences (PacBio)

SMRT :single molecule real time sequencing

- Oxford Nanopore:

Nanopores on a lipid membrane

Library prep is PCR-free, only 'adapters' are ligated

Allow

- "de novo" mapping
- Long structural variants
- Sequencing of modifications of the bases (methylation for example, *the 5th base*)

NGS workflow

1. QC of the nucleic material to be processed DNA or RNA or single cell
   quality and quantity
   time frame DNA and RNA: timeless if the quality is bad (2 days to 1 month)
   single cells: immediate

2. "Library" preparation: DNA with adapters that allow attachment to the flow cell
   time frame: can vary from 3 to 6 days including pooling;
   sometimes libraries have to be repeated, if library QC fails

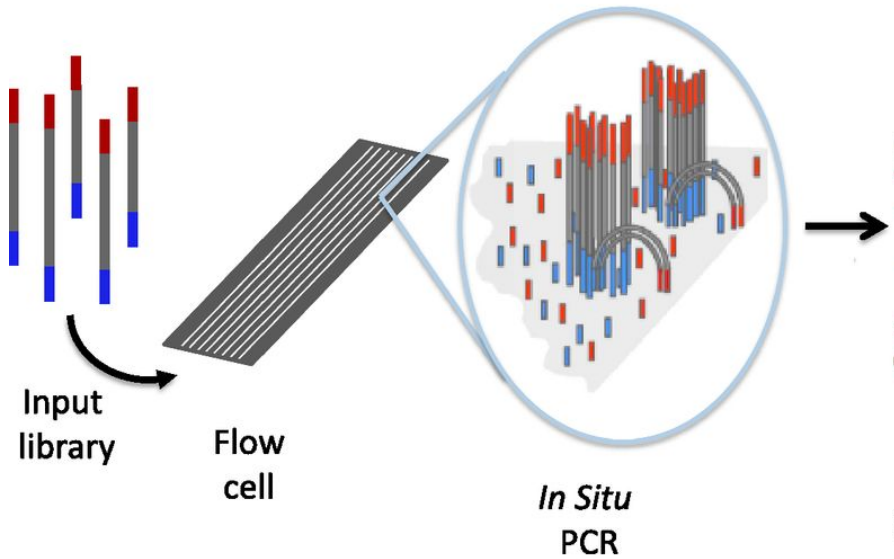
3. Sequencing: "library" is loaded on instrument, minimal hands on, unless run fails (wet lab).
   time frame: variable depending on the Sequencer and the sequencing chemistry
   PE50 on a HiSeq 4000 takes 3 days. If it fails Illumina QC, then more days
   PE50 on the Novaseq takes 1 day

4. Data processing: base calls (bcl files) are made directly from the signal intensity using Illumina's RTA software, using Casava 2.1.7 raw reads (fastq files) and quality scores are generated (dry lab).
   automated takes the same amount of time every time, depending on the run
   time frame: 1-2 days, failures at this step are usually due **Index** issue.

Submitting DNA to Raw Data: 4-8 weeks

- Analysis: timeless if there are no pipelines associated (dry lab or outsource)
  - Bioinformatics quality control of the library – coverage, DNA repertoire, bias
  - Alignment to the reference genome (human 1st 2003, last edit 2013)
  - "Downstream" analysis (Identification of *significative* differences/regional differences)

2nd Gen Illumina SBS technology (reversible terminator chemistry)



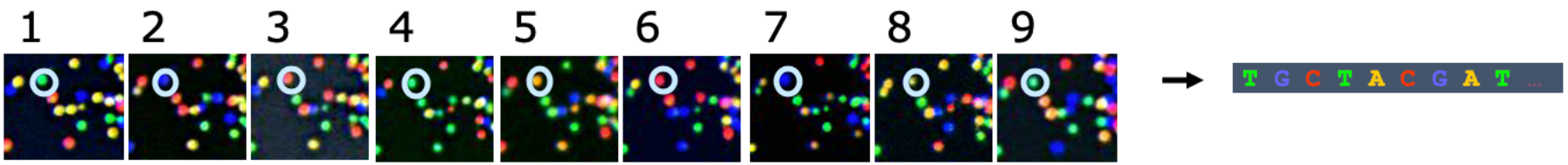
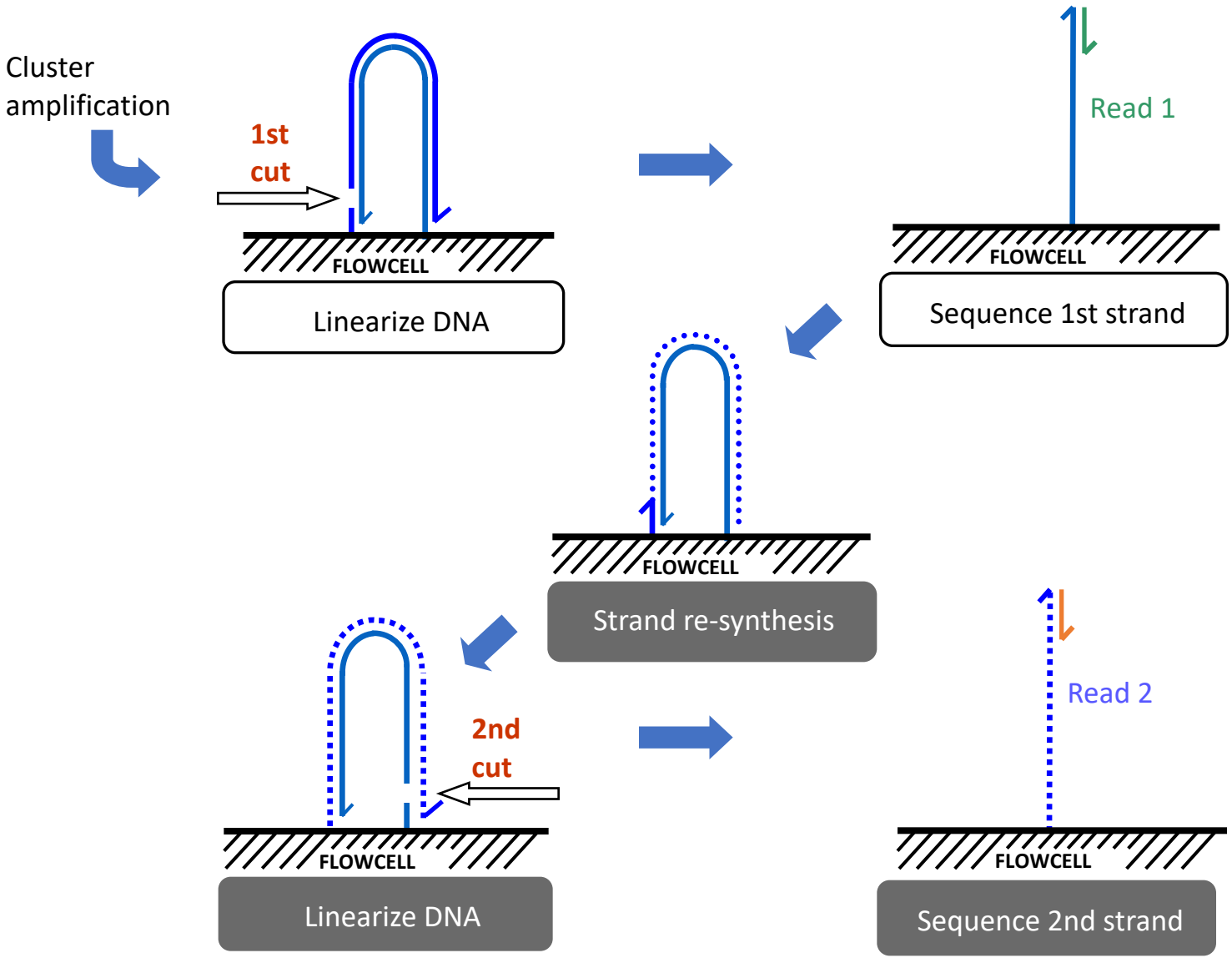


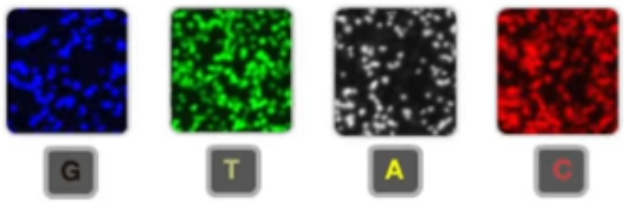
Image acquisition **Base calling**

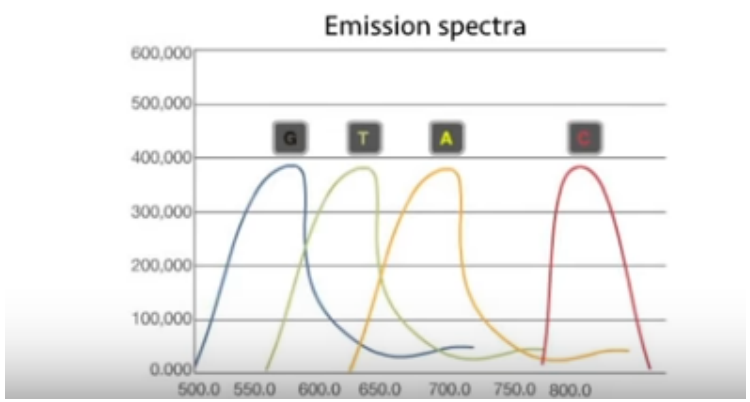
Paired-End Sequencing allows for two looks at the same molecule

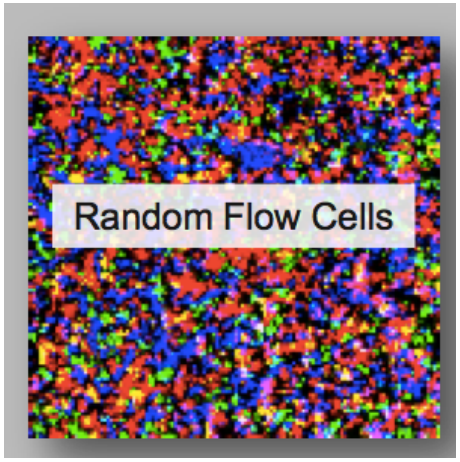


First iteration of SBS: 4-colors and random distribution

One image is taken for each color







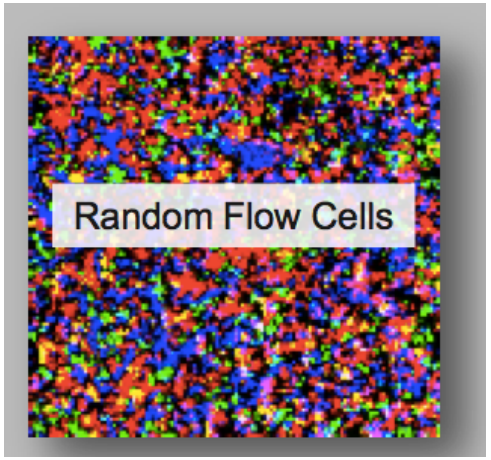
Color overlap

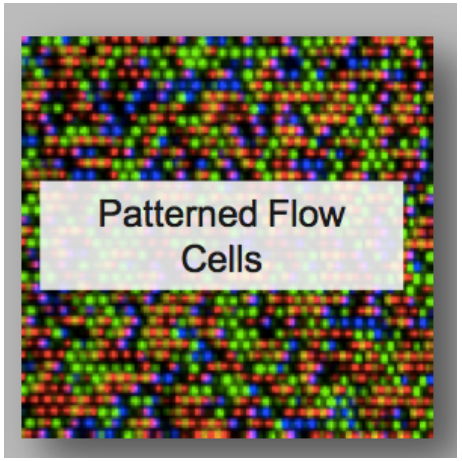
Color compensation

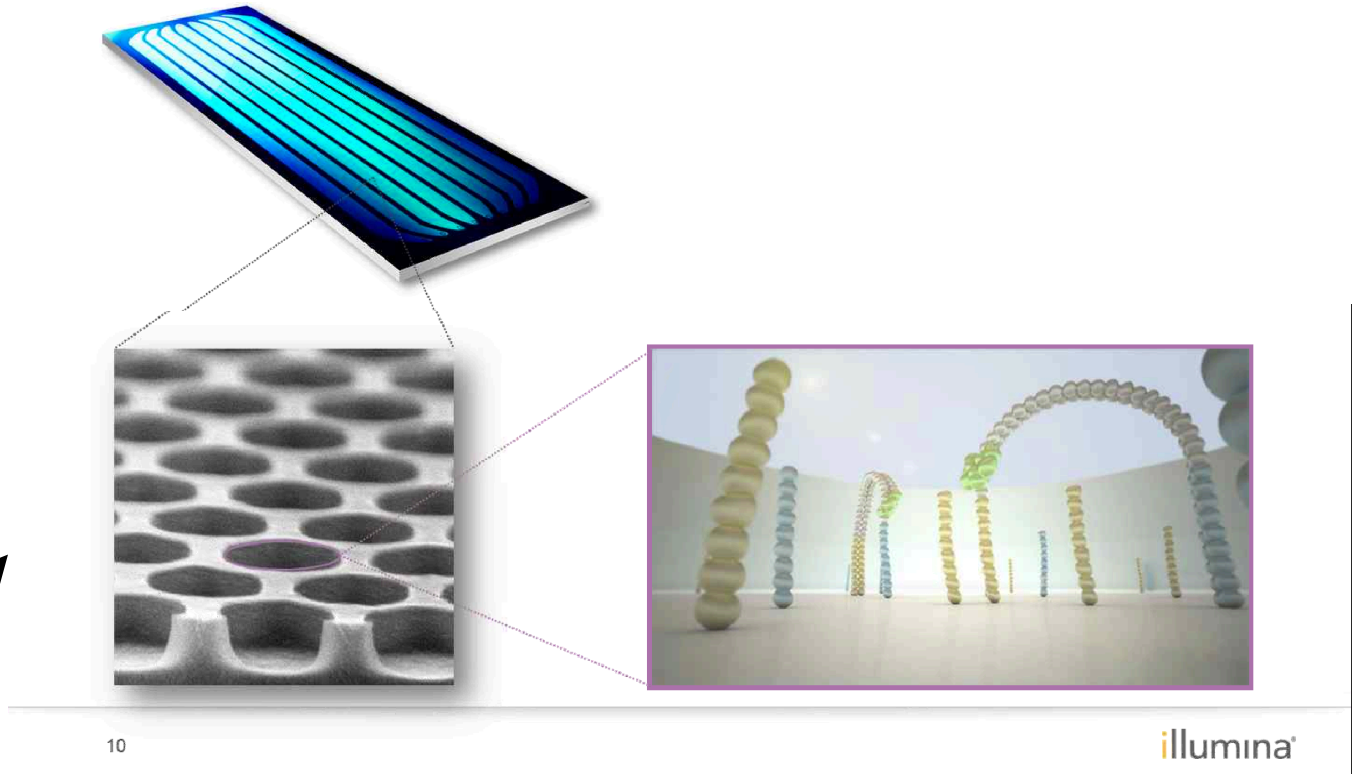
Randomness of clusters: real state is unknown

1st improvement Patterned Flow Cells

Efficient clustering

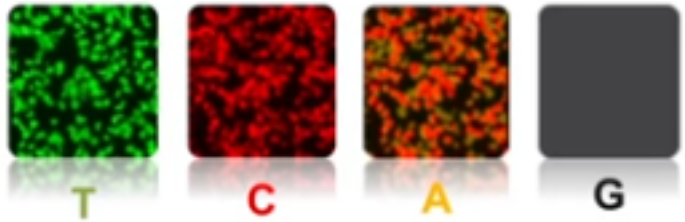


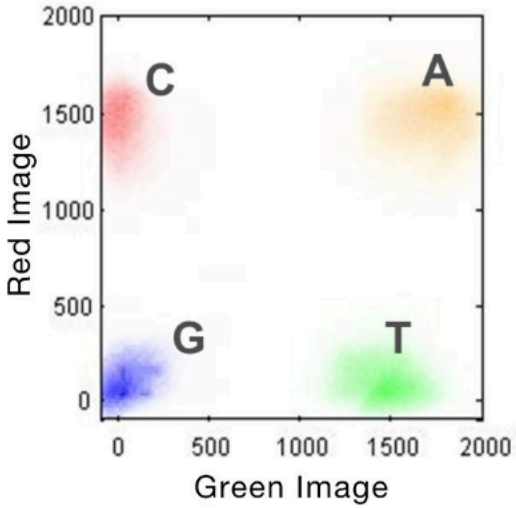




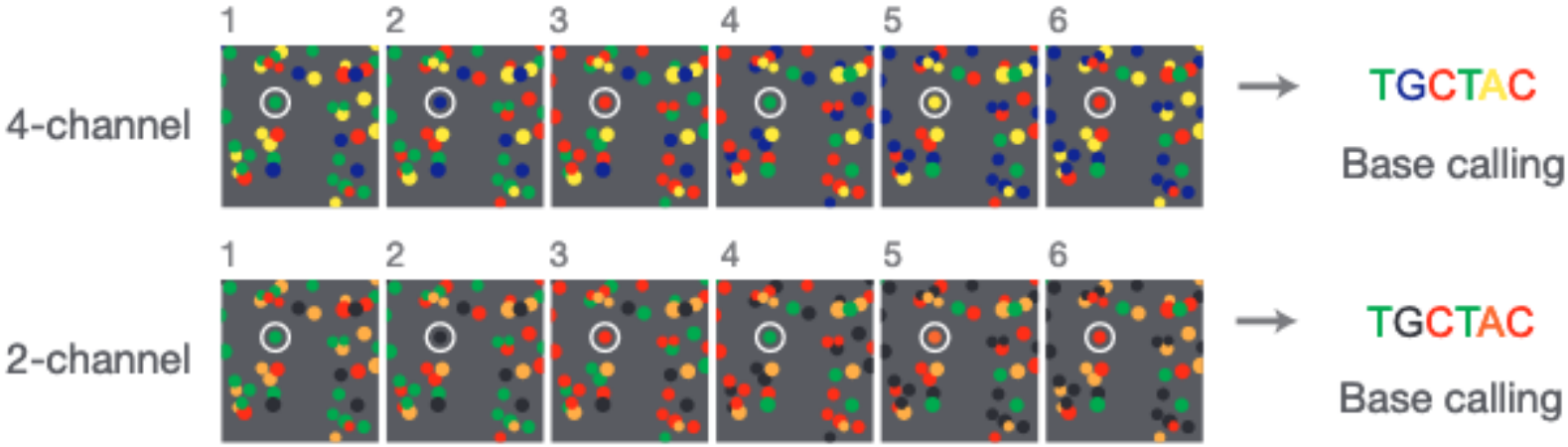
2nd improvement 2-colors, 2 images

Two-color sequencing





Uses a mix of two colors
T are green
C are red
A are red and green (yellow)
G is colorless



2nd Gen Illumina 3rd improvement One color

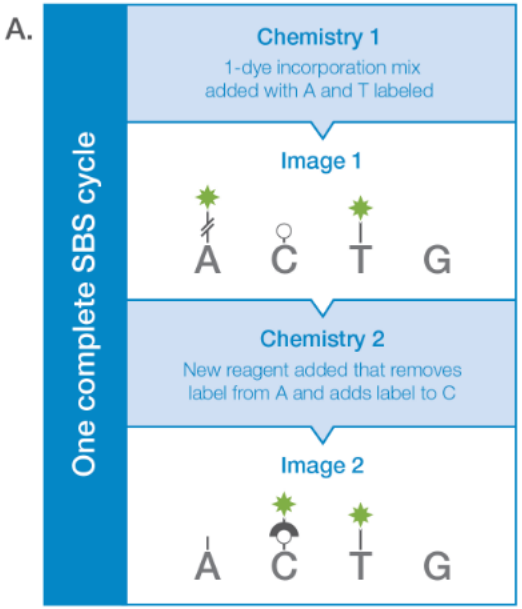
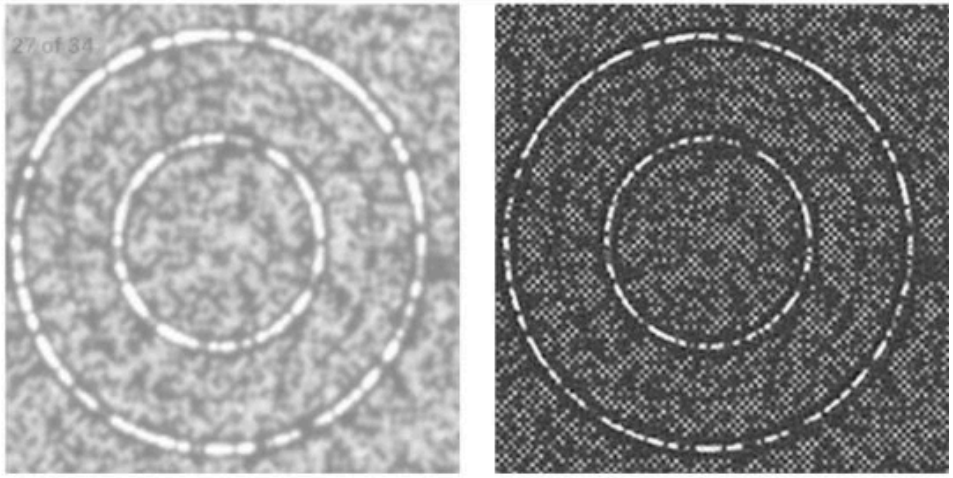


Image 1	Image 2	Result
ON	OFF	A
OFF	ON	C
ON	ON	T
OFF	OFF	G

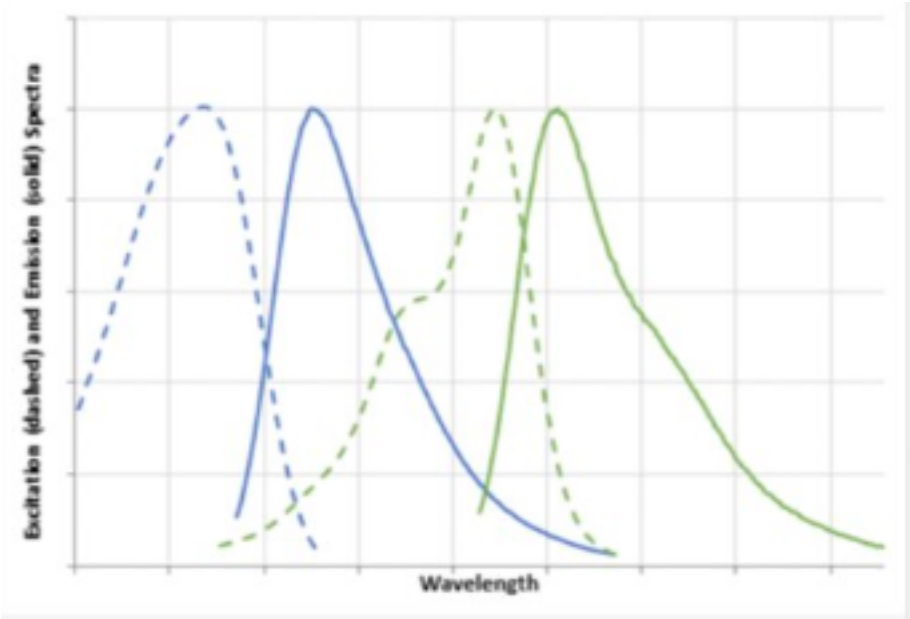
2018

4th improvement: Microscopy and well separated flours

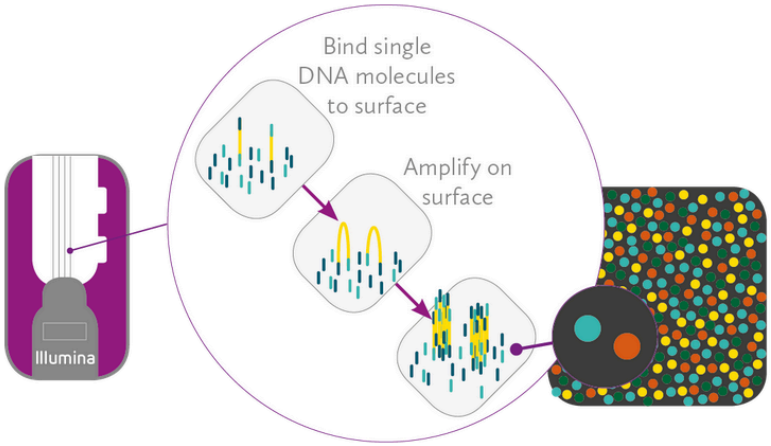
AGBT2020



Diffraction-limited Super Resolution



NGS recap, some concepts



Cluster: clonal group of library fragments

Clusters: varies with the sequencer used and the chemistry used

reads: each clusters produces
ONE single read
ONE pair end read
For example

One lane of MiSeq	~0.03B clusters, ~0.03B reads
One lane of Hiseq4000	~0.4B clusters, ~0.4B reads
One lane of NovaSeq:	~0.4B to 2B clusters, ~0.4B to 2B reads

Sequencing output:
#reads x sequencing length. *For a run of 150 cycles*
0.4Bx150=69Gb
0.4Bx150x2=120Gb

Why bother with single read or pair end?

- Depends of the biological question
  a) Position only and pile up of reads: ChIP, RNAseq differential expression
  b) Mutation analysis: confirmation on both sides of the molecule
- Depends on the $$ available
  a) Single end flow cells are cheaper than pair end flow cells
  b) Less sequencing reagents is necessary

Single end Flow Cell are discontinued
Single end sequencing and short sequencing is also discontinued

Illumina Technology continues improving, the need to MULTIPLEX!

- Coverage (or depth):
  number of reads that are likely to be aligned at a given reference position

$$\textbf{Coverage} = \frac{\text{read length} \times \text{number of reads}}{\text{genome length (haploid)}}$$

- How many reads would one need to sequence ONE human genome at 30x coverage in a PE100 run?

$$\text{number of reads} = \frac{30 \times 3\text{x}10^9}{200}$$

450M reads Pair End

On a HiSeq 2500: two sequencing lanes	~$5,000
On a HiSeq 4000: one sequencing lane	~$2,500
On a Novaseq S4: 1/6 of a sequencing lane	~$1,000

- Multiplexing: being able to physically pool libraries from different origins, sequence and then bioinformatically re-assign each library to its origin

- Illumina's website has coverage depth recommendations

https://emea.illumina.com/science/technology/next-generation-sequencing/plan-experiments/coverage.html?langsel=/gb/

- and a calculator app
- http://support.illumina.com/downloadssequencing_coverage_calculator.html

Illumina Technology continues improving, the need to MULTIPLEX!

Up to 384 independent libraries with unique barcodes

Quantify, calculate molarity

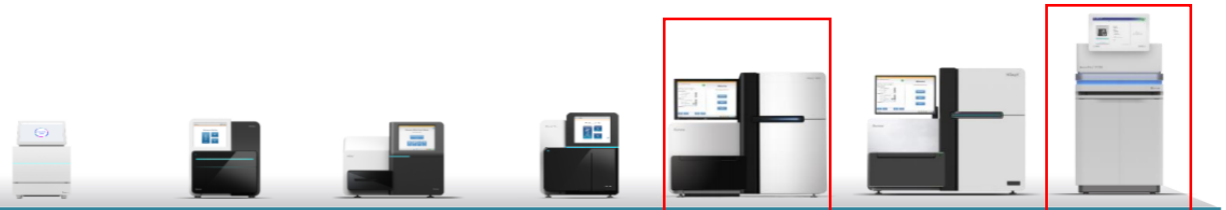
Normalize to same molarity (10nM)

Mix equal volumes

10nM

Quantification, normalization and pooling can be done robotically

Illumina sequencers



Sequencing System	iSeq™	MiniSeq™	MiSeq®	NextSeq®	HiSeq® 4000	HiSeq® X Five/Ten	NovaSeq® 6000
Output per run	1.2 Gb	7.5 Gb	15 Gb	120 Gb	1.5 Tb	1.8 Tb	1 Tb - 6 Tb[1]
Instrument price	$19.9K	$49.5K	$99K	$275K	$900K	$6M[2]/$10M[2]	$985K
Installed base[3]	NA	~600	~6,000	~2,400	~2,300[4]		~285
Clustering	patterned	random	random	random	random & patterned	random	patterned
Chemistry	1-color	4-color	4-color	2-color	4-color	4-color	2-color

2nd Gen

Illumina, HiSeq2500 -Movie 1

2nd Gen

Illumina, NovaSeq –Movie 2

2nd Gen

Illumina, NextSeq2000 –Movie 3

1.30

2nd Gen

MGI/BGI DNBSEQ

	 Sequencers +	 Sequencers +	 Sequencers +	 Sequencers +
Product Model	DNBSEQ-T7	DNBSEQ-G400	DNBSEQ-G50	DNBSEQ-G400 FAST
Features	Ultra-high Throughput	Adaptive	Effective	Fast
Applications	Whole Genome Sequencing,Deep Exome Sequencing,Transcriptome Sequencing,and Targeted Panel Projects.	WGS, WES, Transcriptome sequencing and more	Targeted DNA, RNA, Microbial sequencing	Targeted DNA, RNA, Epigenetics and clinical applications
Flow Cell Type	FC	FCL & FCS	FCS	FCS
Lane/Flow Cell++	1 lane	4 lane & 2 lane	1 lane	2 lane
Operation Mode	Ultra-high Throughput	High Throughput	Medium Throughput	Medium Throughput
Max. Throughput / RUN	6Tb	1440Gb	60Gb	330G
Effective Reads / Flow Cell	5000M	1500-1800M	280-300M	550M
Average run time	PE150 within 24 hours	~38 hours	12-48 hours	12-37 hours
Min. Read Length	PE100	SE50	SE50	SE100
Max. Read Length	PE150	SE400	PE100	PE150

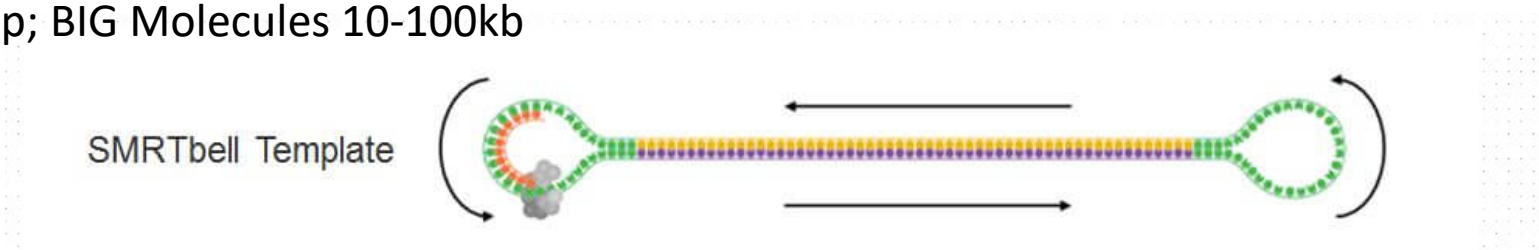
TOP

2nd Gen

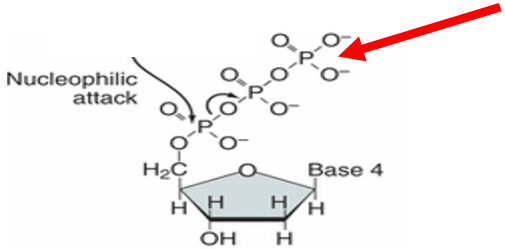
MGI/BGI, DNBSEQ–Movie 4

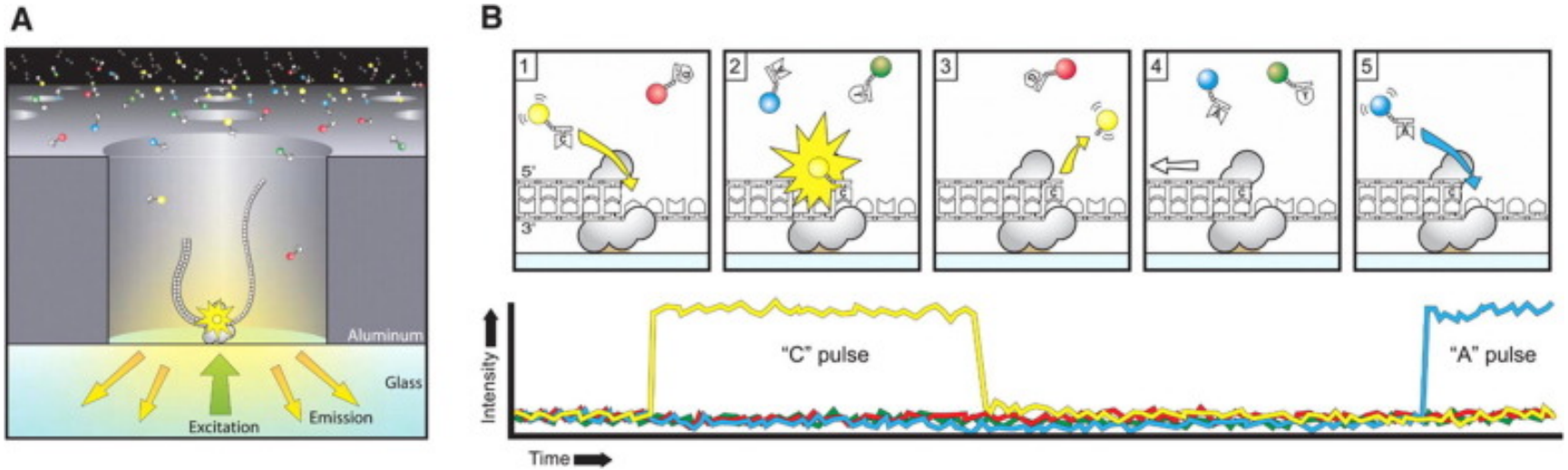
SMRT: single molecule real time sequencing

Library prep; BIG Molecules 10-100kb



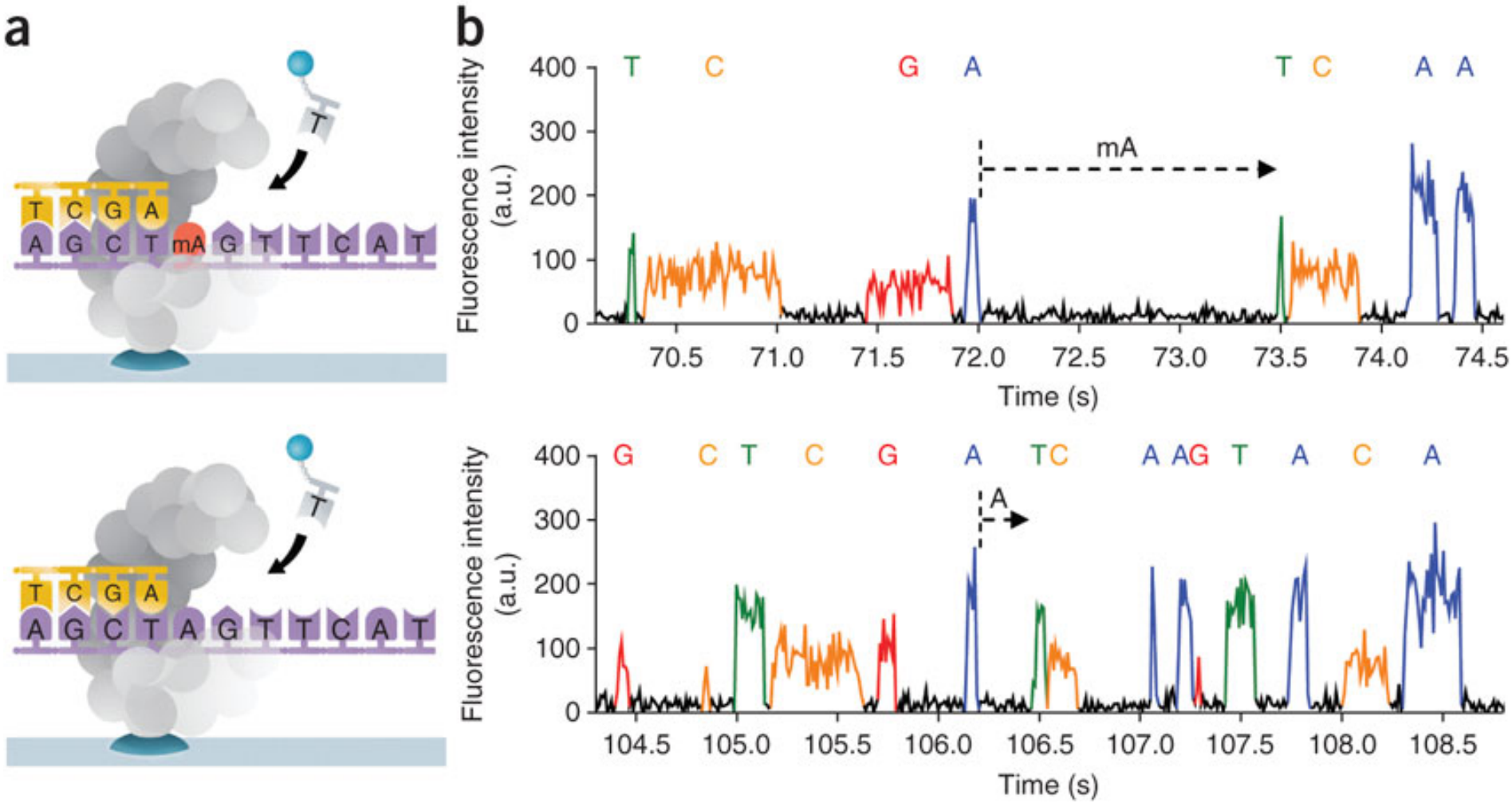
Nucleotides: Fluorescently labeled on the **phosphate,** every base with a different flourophore





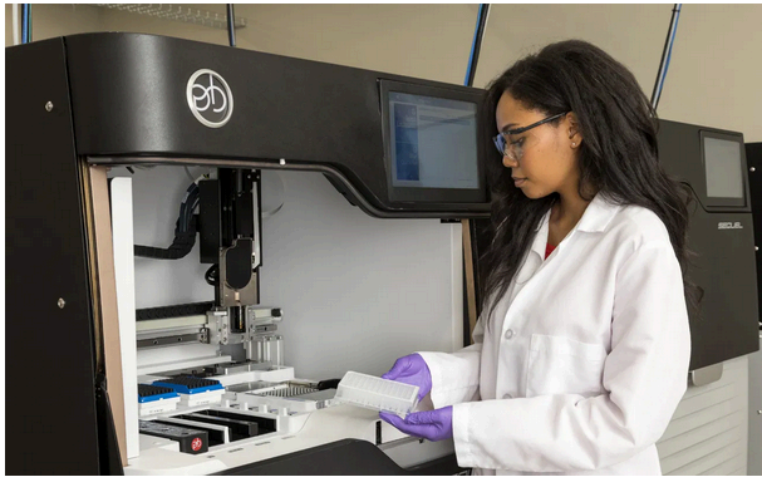
SMRTseq: Direct detection of methylation (length of pulse)

N^6-methyladenosine (^mA)



Flusberg et al., 2010.

PacBio sequencers



	Sequel II System	Sequel (I) System
SMRT Cell	SMRT Cell 8M	SMRT Cell 1M
Average Data Output*	~100Gb	~15Gb
Number of HiFi Reads >99% Accuracy*	Up to 4,000,000	Up to 500,000
Sequencing Run Time per SMRT Cell	Up to 30hrs	Up to 20hrs
Recommended species / genome size	Human (3Gb), Plant, or animal with more than 3Gb of Genome size	Plant, or animal with less than 3Gb of Genome size

*Number of HiFi reads is dependent upon the insert size and sample quality
*Data Output is dependent upon the insert size and sample quality

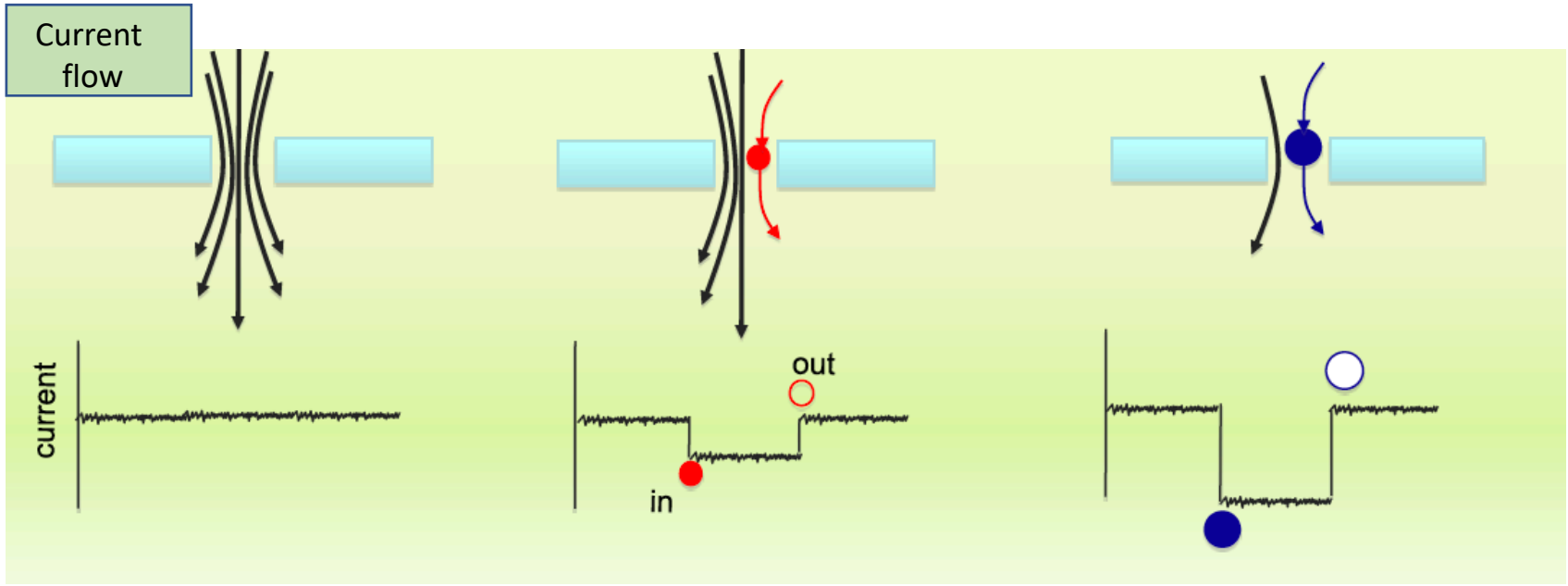
PacBio Sequel Movie 5

Oxford Nanopore WHAT IS A NANOPORE?

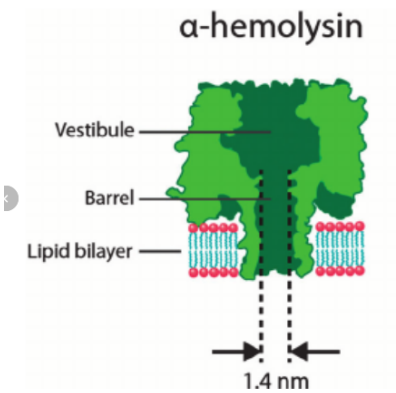
- Nanopore = very small hole
  A transmembrane protein –porin- embedded in lipid membranes creating size dependent porous surfaces 'nanometer holes' (protein channel)

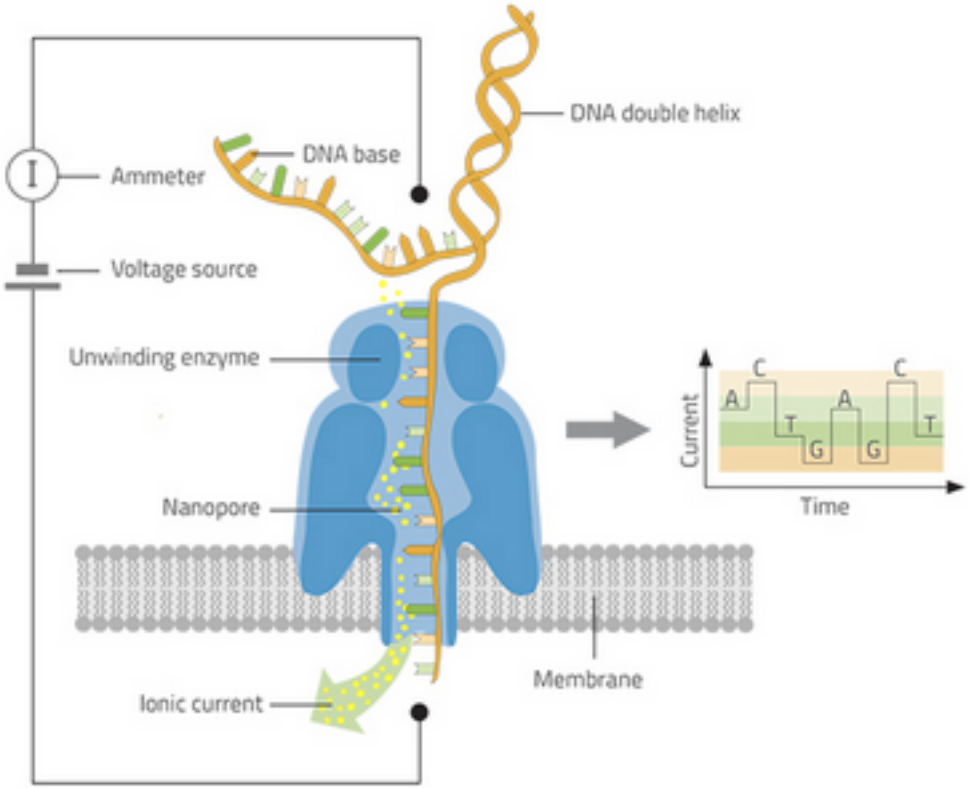
- An electrical current flows through the nanopore

- Introduction of an analyte of interest into the nanopore identifies the "analyte" by the disruption or block to the electrical current



Biological nanopore





This is the case were there is no DNA synthesis.
Modified bases mA, mC have unique pulses
In fact, protein and RNA can be sequenced using nanopores

Oxford Nanopore, Movie 6

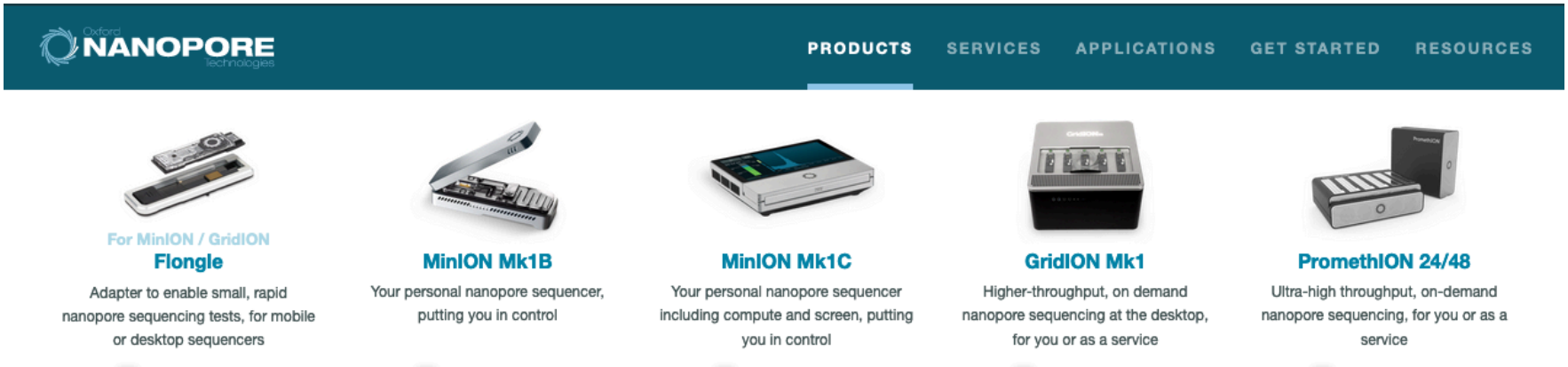
MinION is portable

Antartica

Space

Courtesy, Chris Mason

Oxford Nanopore sequencers



	Flongle	MinION Mk1B	MinION Mk1C	GridION Mk1	PromethION 24/48
Read length	Nanopores read the length of DNA presented to them. Longest read so far: > 2Mb.				
Yield per flow cell, DNA/cDNA	2 Gb	50 Gb	50 Gb	50 Gb	220 Gb
Number of flow cells per device	1	1	1	5	24/48
Yield per device Up to:	2Gb	50 Gb	50 Gb	250 Gb	5.2 Tb/10.5 Tb
Price	From $1,760	From $1,000	From $4,900	From $49,995	From $165,000/$285,000
Suitable applications include	Amplicons	Whole genomes/exomes	Whole genomes/exomes	Larger genomes or projects	Very large genomes or projects
	Panels/targeted sequencing	Metagenomics	Metagenomics	Whole transcriptomes (direct RNA or cDNA)	Population-scale human
	Quality testing	Targeted sequencing	Targeted sequencing	Large numbers of samples	Whole transcriptomes
	Small sequencing tests	Whole transcriptome (cDNA)	Whole transcriptome (cDNA)		Very large numbers of samples
		Smaller transcriptomes (direct RNA)	Smaller transcriptomes (direct RNA)		
		Multiplexing for smaller samples	Multiplexing for smaller samples		
			Particularly suitable for field use		

Evolution of Sequencing Technologies

A

2000
1st Generation
(Sanger sequencing)

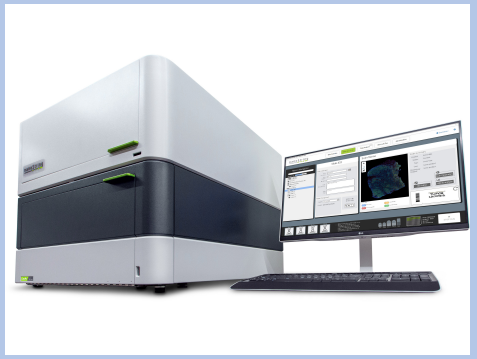
B

2006-2010
2nd Generation
("Next Generation")

C

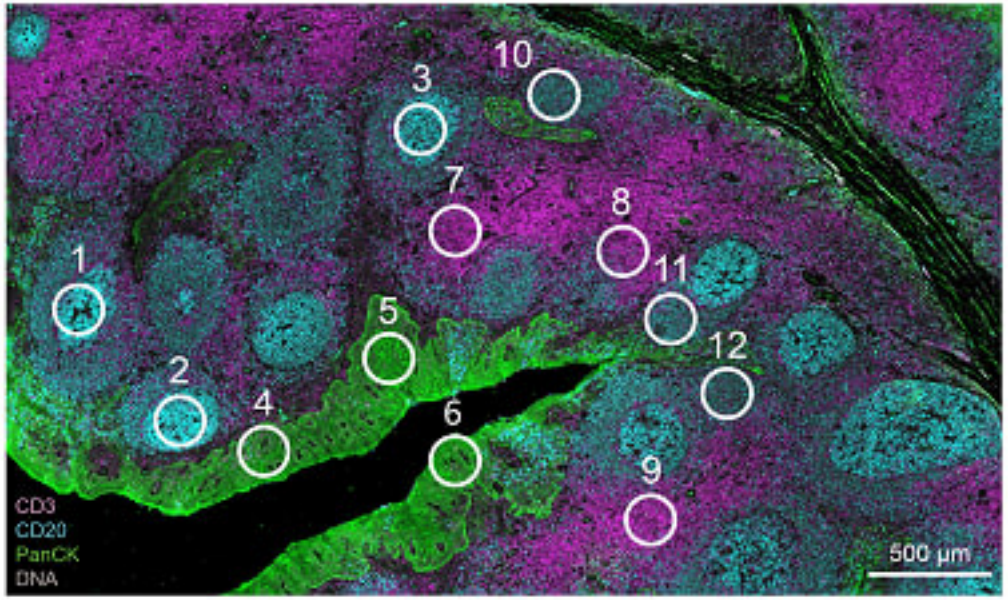
2010-2015
3rd Generation
("Next-Next Generation")

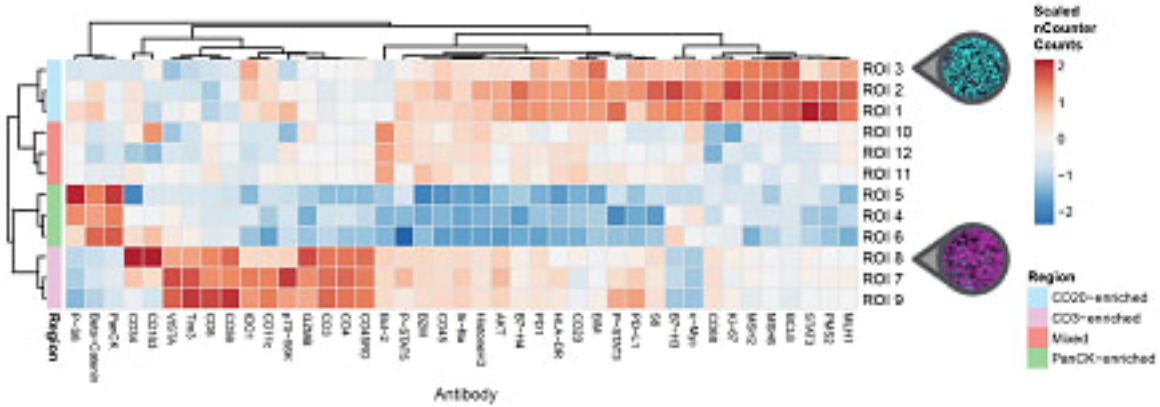
D 2018-2020
4th Generation?
(Spatial Transcriptomics)



Spatial-Transcriptomics (1)

NanoString
GeoMx DSP

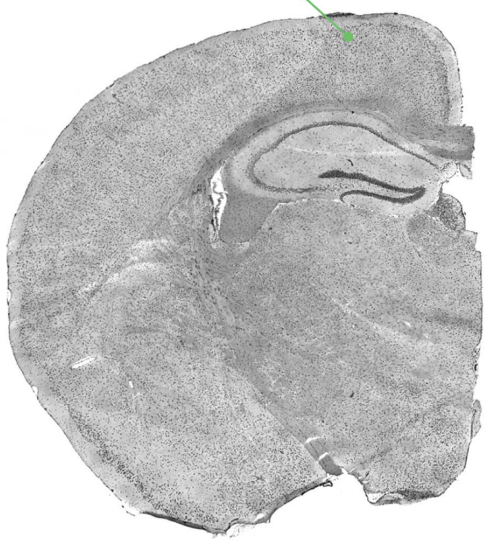




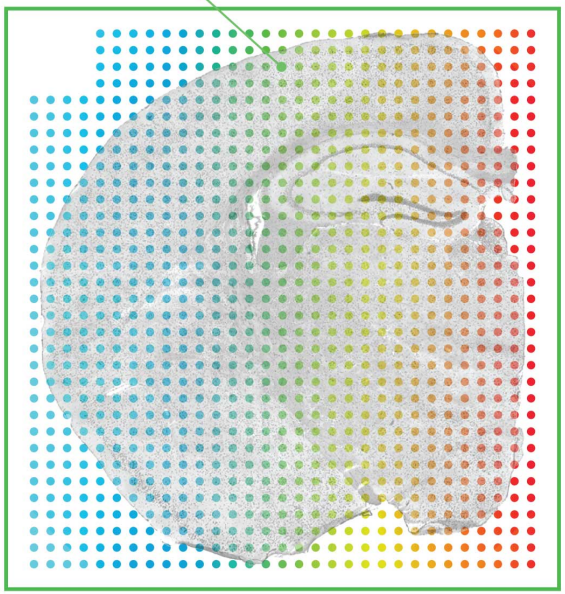
Spatial-Transcriptomics (2)

10x Genomics
Visium

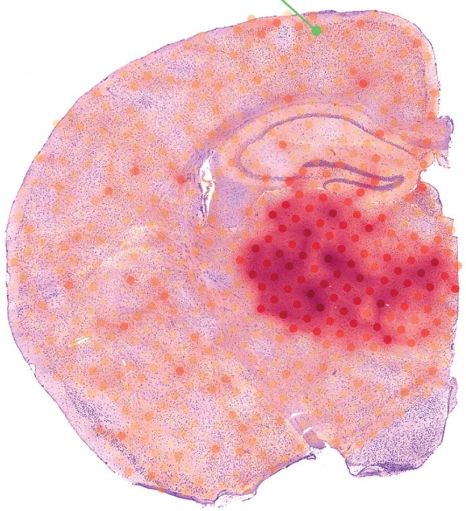
Tissue
Section



Spatial
Transcriptomic
Map

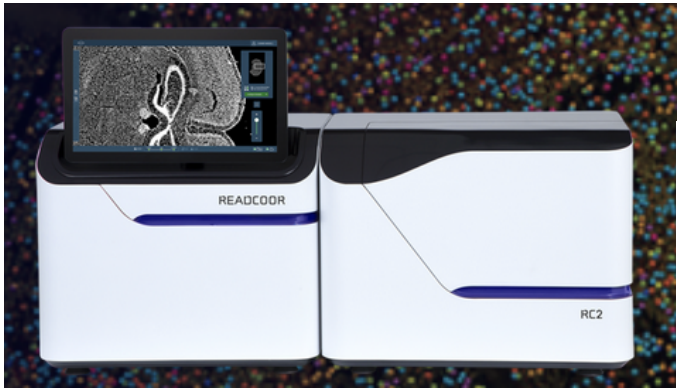


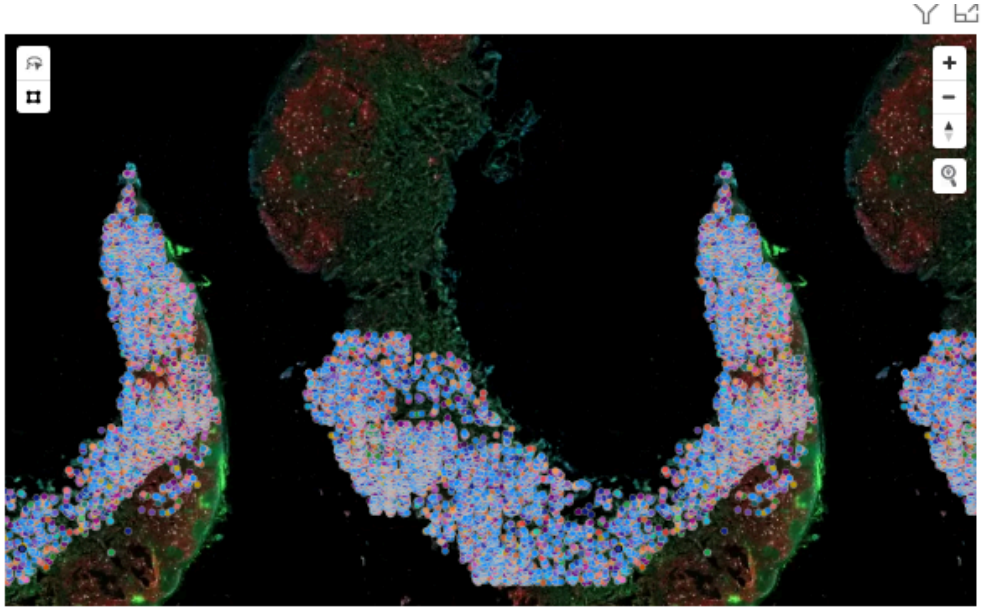
Visualize Expression
of any mRNA

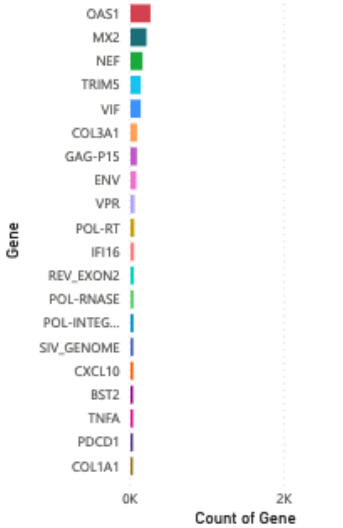


Spatial-Multiomics (3)

ReadCoor
RC2







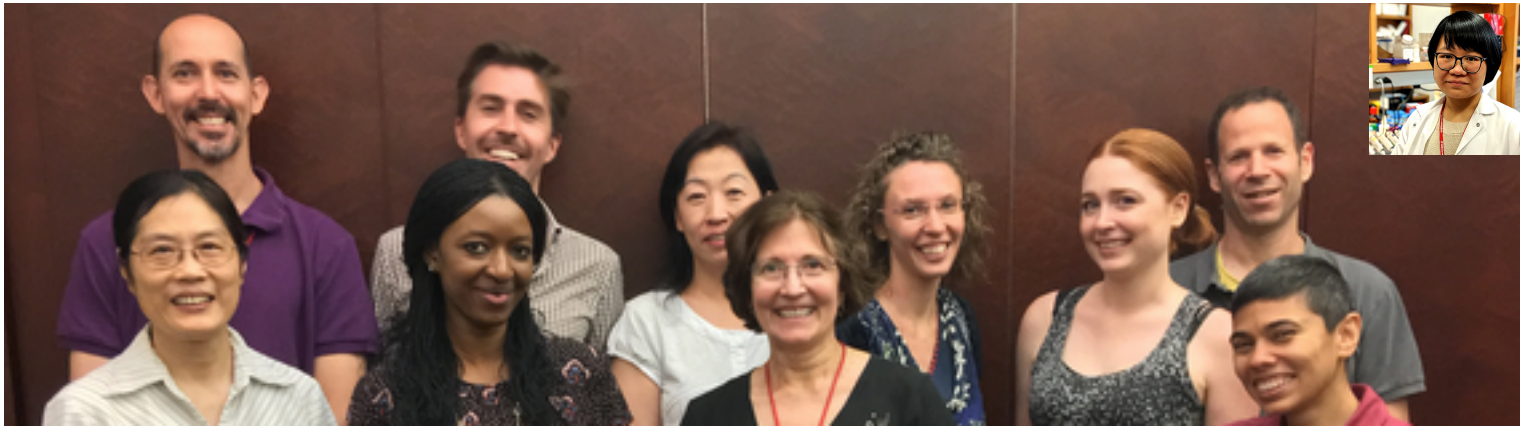
Category	BST2	CCL2	COL1A1	COL1A2	COL3A1	CXCL10	ENV	GAG-P15	GAG-P27	GAG-P8	GAPDH	IDO1	IFI16	IL10	IL1B	ISG
Anti-inflammatory														36		
Anti-viral	40															
Collagen			37	269	92											
Control											282					
Pro-inflammatory		31				43							53		449	
SIV Genome																
Total	**40**	**31**	**37**	**269**	**92**	**43**	**80**	**90**	**33**	**35**	**282**	**36**	**53**	**36**	**449**	**6**

Gene
Search
Select all
BST2
CCL2
COL1A1
COL1A2
COL3A1
CXCL10
ENV
GAG-P15

Category
Search
Select all
Anti-inflammator
Anti-viral
Collagen
Control
Pro-inflammatory
SIV Genome
SIV Transcriptome
T-cell modulation

Meet the Epicore (past and present)

https://epicore.med.cornell.edu/



Director:	Alicia Alonso, PhD
Wet Lab	
Lab Manager	Yushan Li
Research Specialists	Natalie Chow
	Caroline Sheridan
Dry Lab	
Bioinformatics Director:	Doron Betel, PhD
Computational Biologist:	Piali Mukherjee
Software Developers:	Thadeous Kacmarczyk, PhD
	Simon Johnson
Former members	
Research Technician	Yuan Xin
Research Specialist	Marisa Mariani

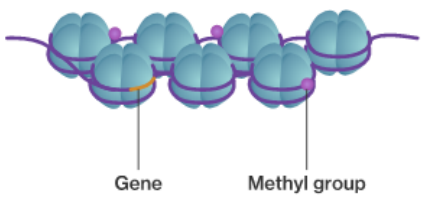
How does one generate an 'epigenomic' library?

Question: is chromatin open or closed?

- Closed chromatin = repression of transcription
- Open chromatin = accessibility to transcription factors or actively transcribed

OFF

ACCESS TO DNA BLOCKED, GENE INACTIVE

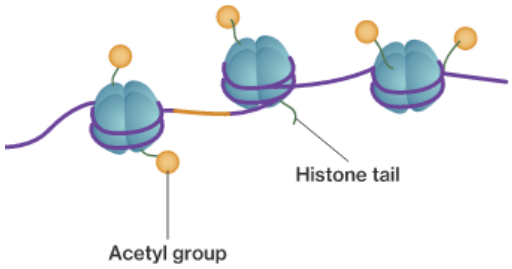


>95% of the DNA is compacted and transcriptionally repressed

nucleosome remodeling

ON

ACCESS TO DNA POSSIBLE, GENE ACTIVE



2-3% of the DNA is "open" and transcriptionally permissive

- DNA methylation
- DNA conformation and protein occupancy
- Nucleosome free regions
- Protein-mediated DNA interactions (3D)

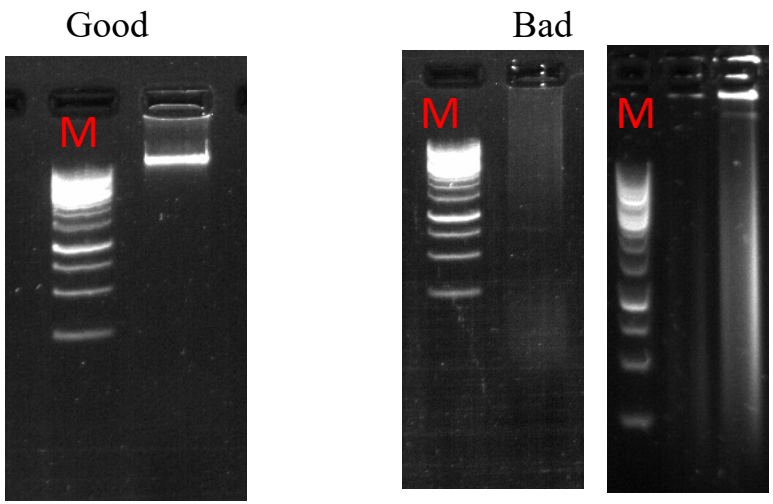
Most common assays to study epigenetic changes (as per Epicore offering)

- Methylation sequencing (Cytosine, CpG ie ^{m}CpG)
- Chromatin immunoprecipitation (ChIPseq)
- Assay for transposase associated chromatin (ATACseq)

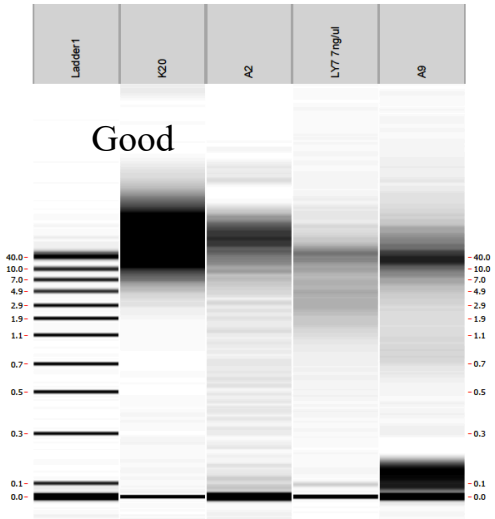
Genomic DNA sample QC: ~40kb, no RNA contamination

whole genome sequencing, targeted sequencing, methylation sequencing)

Using an agarose gel



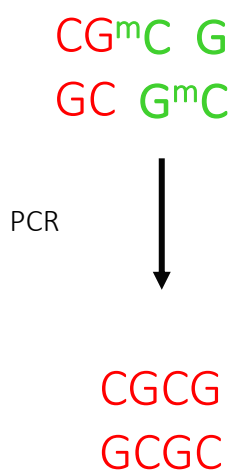
Using the Perkin Elmer Labchip GX



A methylated C cannot be identified by sequencing by synthesis methods

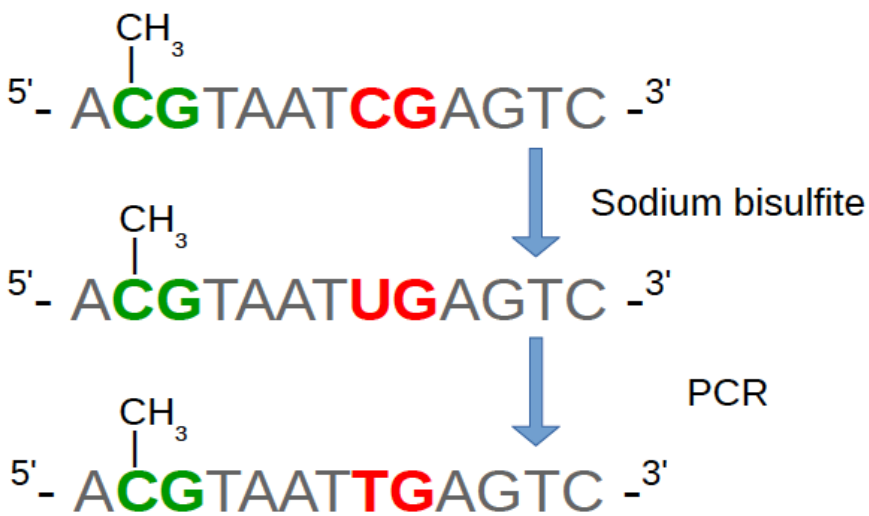
NGS workflow

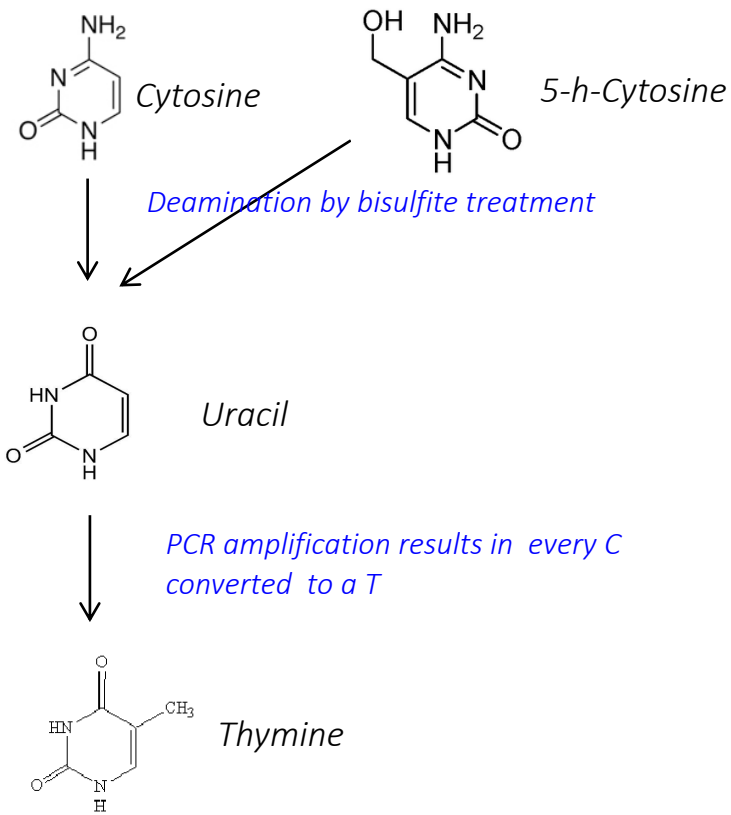
- Creating a "DNA library" that is suitable for sequencing:
  - DNA must be between 200-1000bp
  - Must have adapters (will tether the DNA to the Illumina flowcell)
  - Usually requires a **PCR amplification** step to obtain enough material for sequencing



PacBio

Methylation sequencing refers to a chemical step prior to sequencing (Bisulfite conversion - BS)





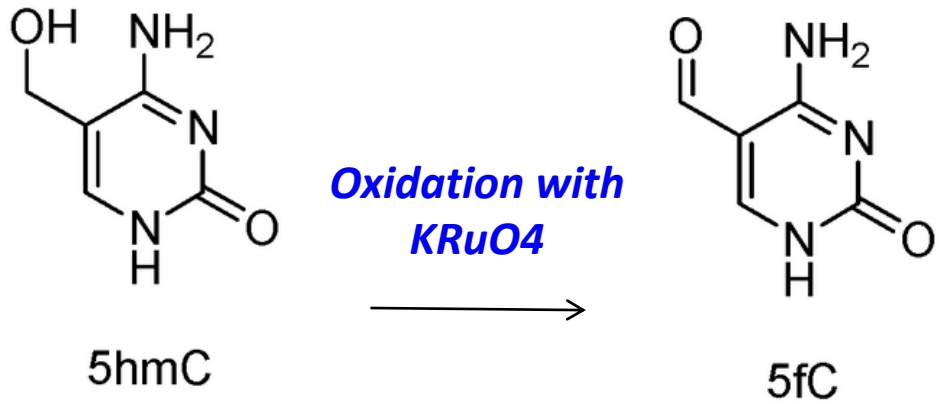
This chemical step gives the combined measure of 5mC and 5hmC, which are about 1% of the total Cs (my guess: 95% of the data published is combined 5mC and 5hmC)

BS conversion reduces the complexity of the 4-coded genome (mostly) to a 3-letter code

How does one differentiate between a CpG that is methylated or hydroxymethylated?

Hydroxymethylation sequencing requires an oxidation step prior to sequencing (oxBS)

Step 1: oxidation of 5 hmC to 5 fC



Step 2: bisulfite conversion of 5fC to Uracil

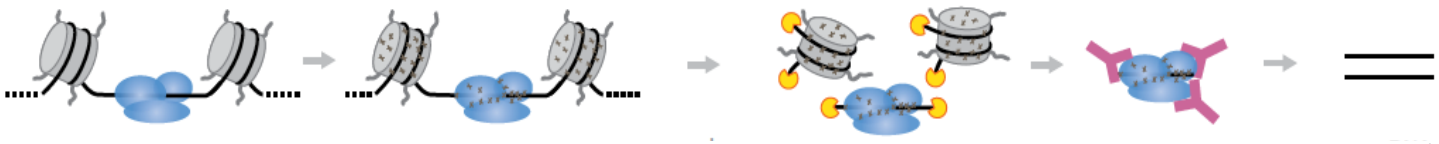
Step 3: PCR amplification of Uracil to Thymine

- *oxBS libraries give the true 5mC content*
- *the bioinformatic differential between a BS and oxBS library gives the 5hmC content*

Most common assays to study epigenetic changes (as per Epicore offering)

- Methylation sequencing
- Chromatin immunoprecipitation (ChIPseq)
- Assay for transposase associated chromatin (ATACseq)

Chromatin immunoprecipitation (ChIP)



DNA protein complex → Crosslink protein and DNA → Sonicate the crosslinked material to ~500bp → Immunoprecipitate → Amplification and sequencing

Analysis: ENRICHMENT
- Alignment to reference genome
- Peak calling
  - gene annotations
  - motif analysis
  - pathway enrichment

ChIP is fickle

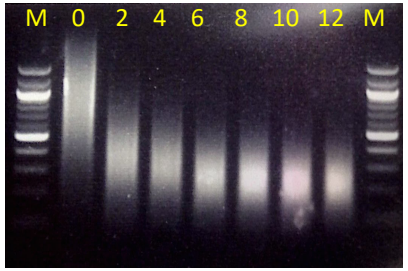
- Depends on cross-linking the **protein to the DNA**
  - (formaldehyde treatment)
  - Conditions vary per protein studied
  - Requires **triplicate** experiments for accuracy
- Depends on **DNA shearing (post successful cross-linking)**
  - Most common method is **sonication,** which needs to be optimized **per cell**, **per protein studied**
- Depends on **antibodies**
  - Affinity/binding needs optimization
    - Commercially vs in-house
    - Requires pre-immune Ab
- Signal depends on how abundant and how well the proteinof interest binds to DNA

Quality Control of ChIP DNA

QC2: qualitative measurement

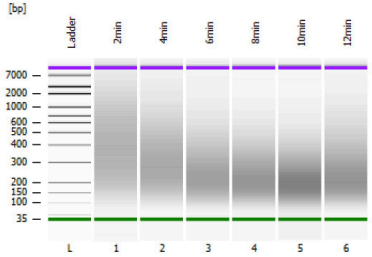
10% of total material must be in the size range of the DNA fraction required for library preparation (130-230bp)

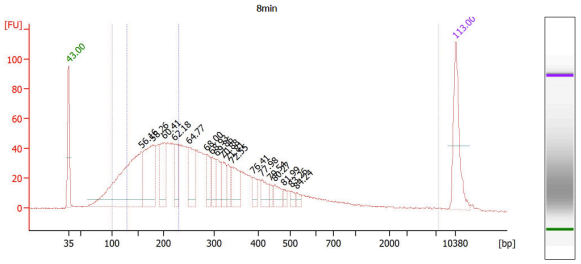
1.5% agarose gel



500bp

Agilent BioAnalyzer High Sensitivity

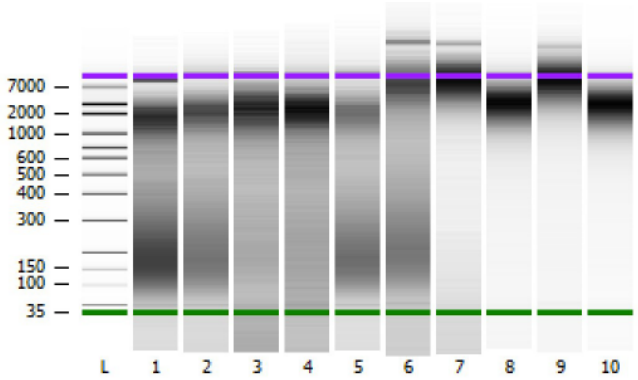


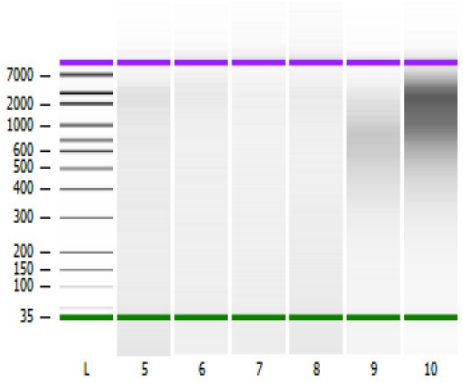


Region table for sample 4 : 8min

From [bp]	To [bp]	Corr. Area	% of Total	Average Size [bp]	Size distribution in CV [%]	Conc. [pg/µl]	Molarity [pmol/l]	Co lor
100	7,500	1,795.4	95	390	100.0	1,824.23	12,742.0	
130	230	675.3	36	182	15.5	754.80	6,491.2	

Real samples look like this



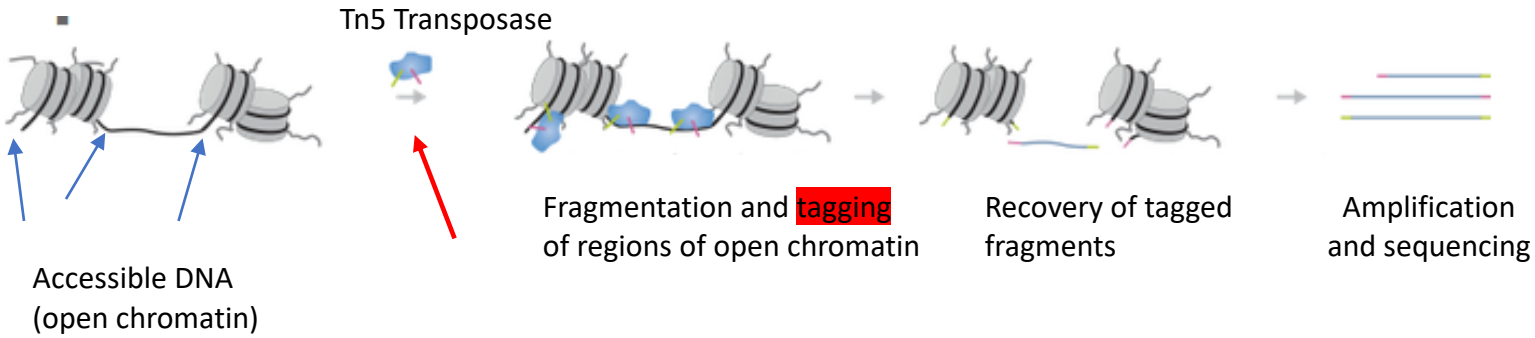


Epigenomics Core

Most common assays to study epigenetic changes (as per Epicore offering)

- Methylation sequencing
- Chromatin immunoprecipitation (ChIPseq)
- Assay for transposase associated chromatin (ATACseq)

Assay for transposase associated chromatin (ATAC-seq)



Analysis: ENRICHMENT

- Alignment to reference genome
- Peak calling
  - gene annotations
  - motif analysis
  - pathway enrichment

ATAC-seq is also fickle

- Starts from **nuclei**
  - Releasing intact nuclei from a cell is dependent on **cell membrane composition**, which is different per cell type
  - Requires **triplicate** experiments for accuracy

Advantages

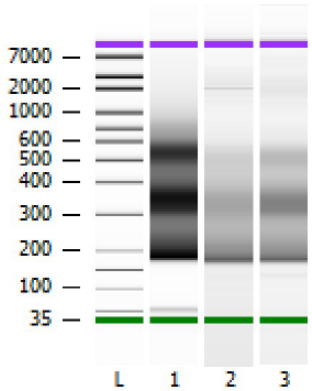
- **Easiest** library preparation ever, if your nuclear prep is good!

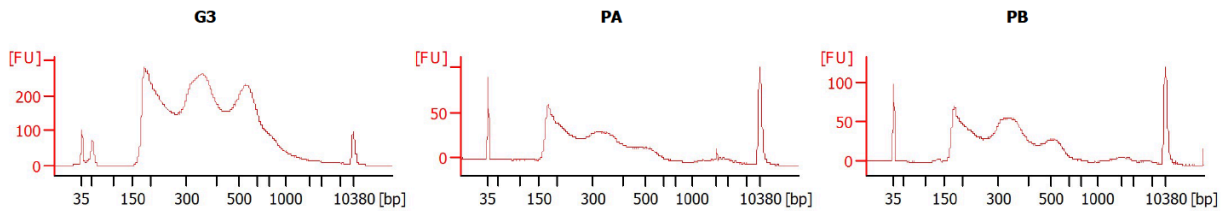
- Choice technique for "active" chromatin, but not for "closed" chromatin
- Requires ~50,000 cells
- Independent on **antibodies**
  - Easier to optimize

Quality control of ATAC-seq libraries is at the sequencing stage!

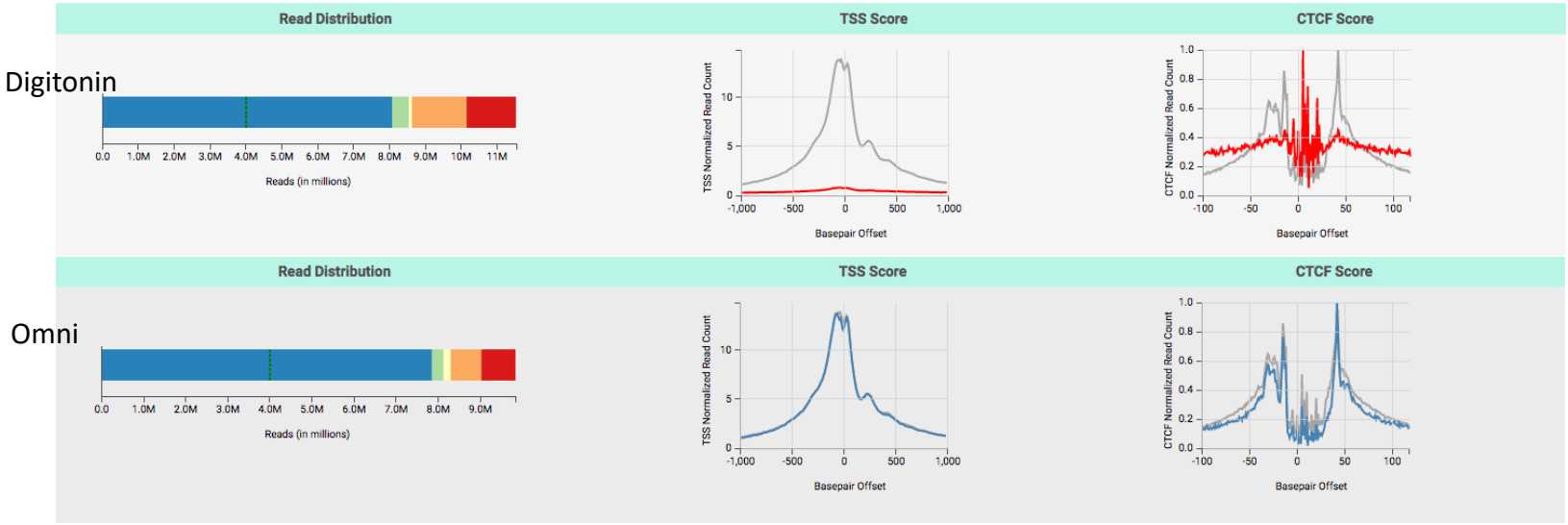
Bio

Bioanalyzer QC

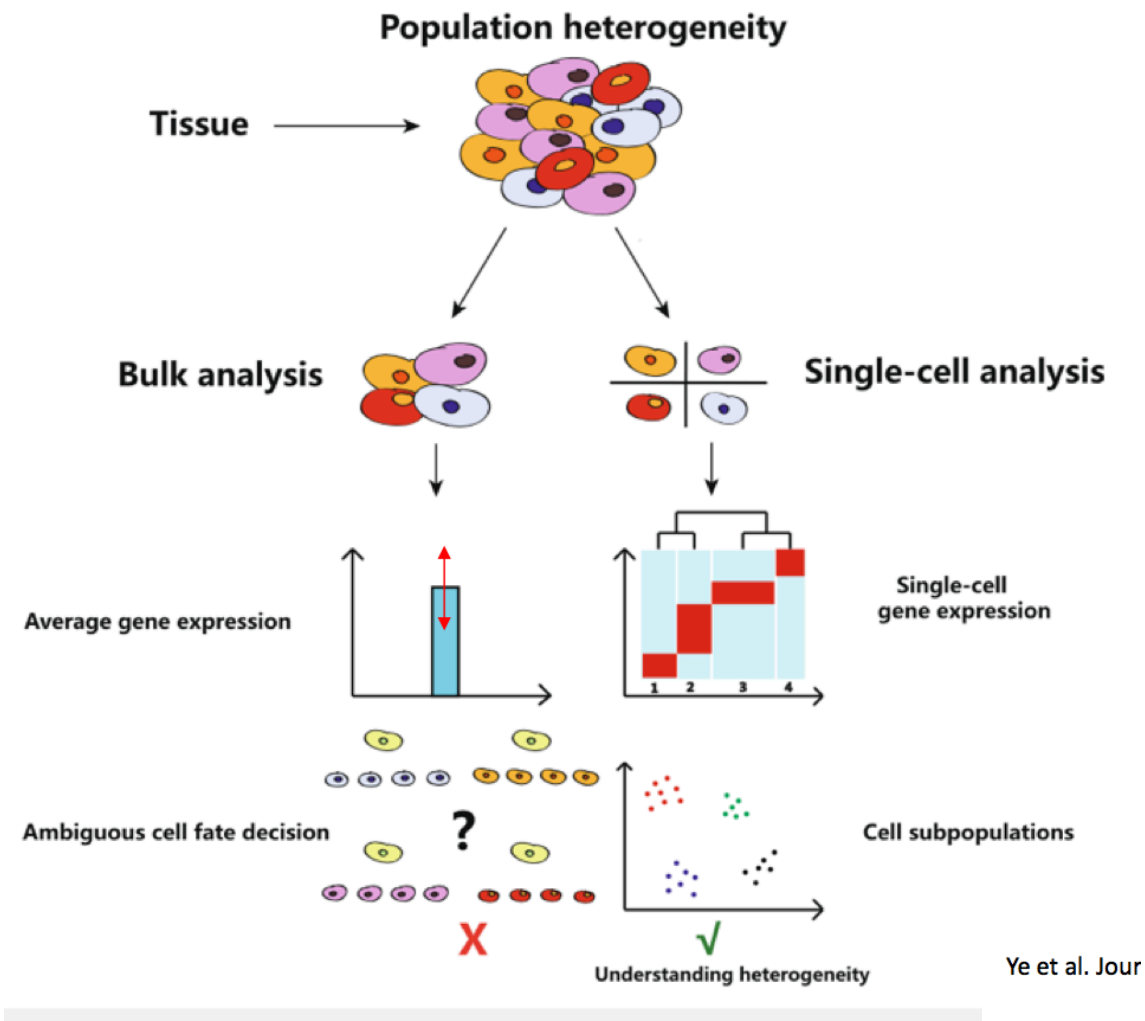




QC using Epinomics website (*www.epinomics.co/*)Greenleaf initiative)



Bulk vs Single Cell Transcriptomics

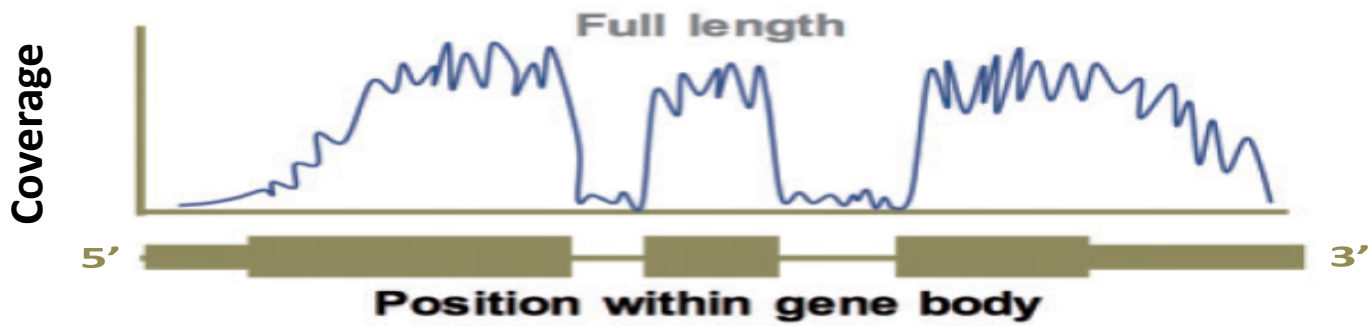


Ye et al., J. Hematology Onc. 2017

Transcriptome analysis

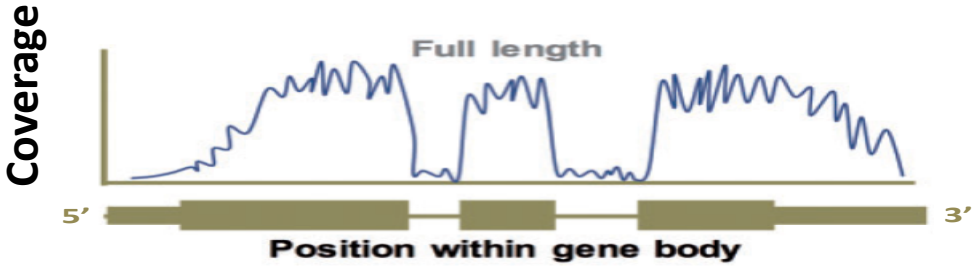
RNA-seq
Bulk

Stranded mRNA-seq	*(150ng)*	*($250)*
Total RNA-seq	*(150ng)*	*($480)*
Low Input RNA-seq	*(10ng)*	*($480)*
Total Low Input RNA-seq	*(10ng)*	*($550)*



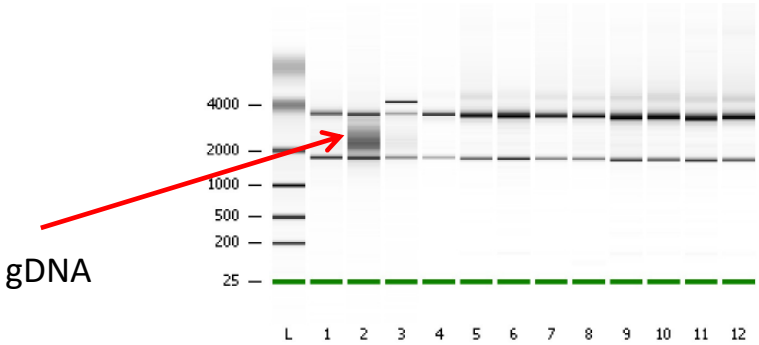
Quality Control for RNA

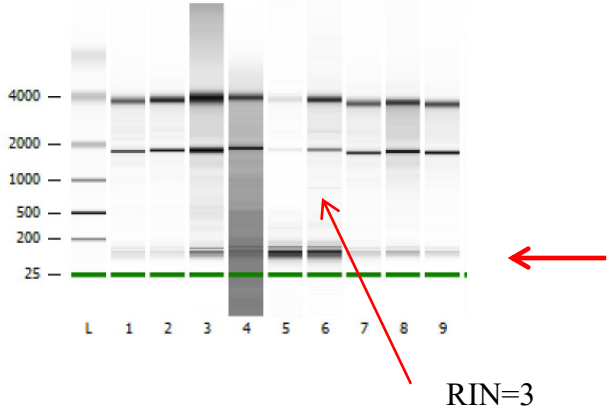
Transcriptome analysis



RIN number (RNA integrity number)

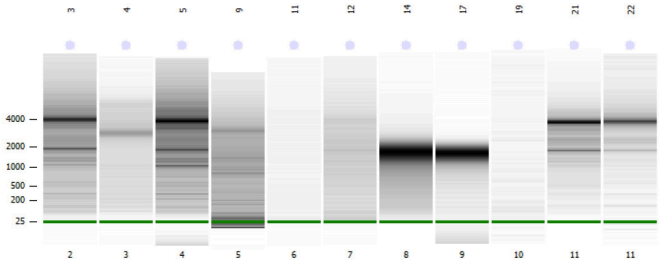
250ng RNA, RIN >- 8





Total stranded RNA prep

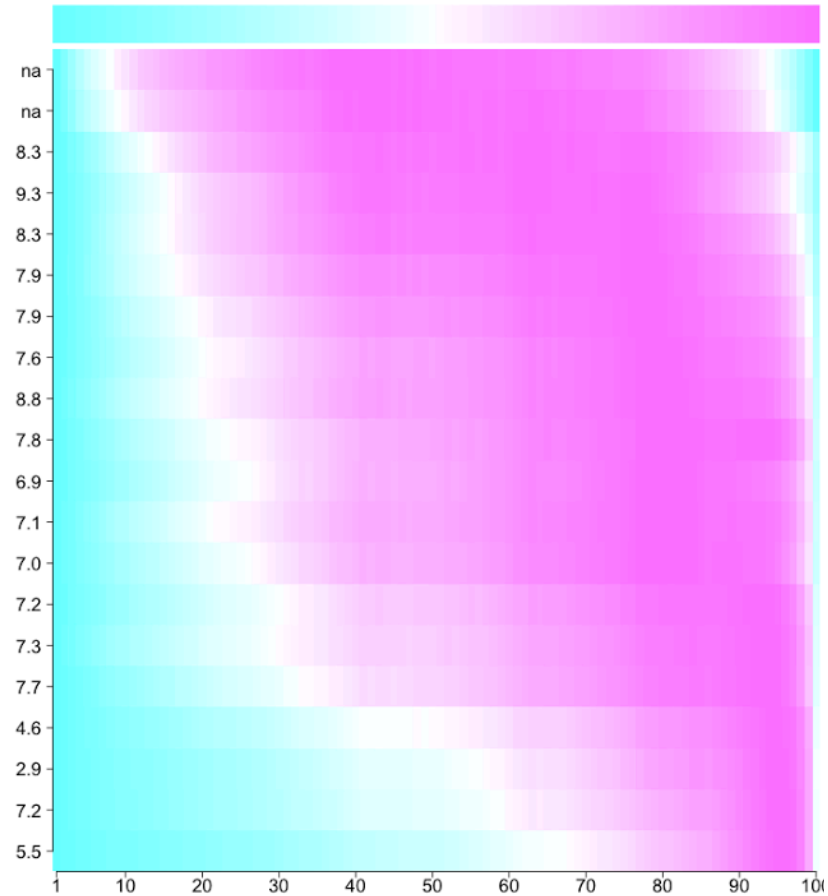
Requires 500ng RNA, RIN >- 2



Effect of RIN on gene body counts

Gene body coverage heat map

RIN



Gene body coverage (5’ to 3’)

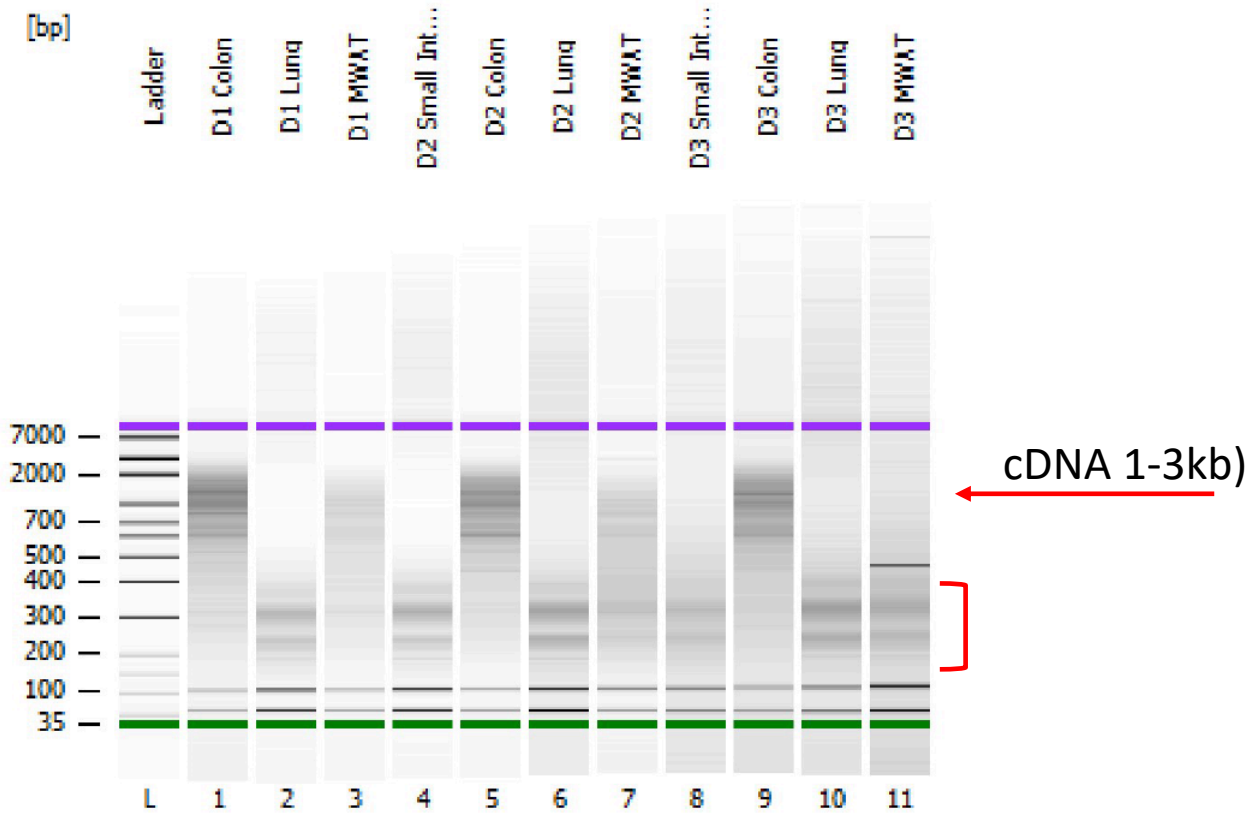
Piali Mukherjee

Quality Control for RNA (low input)

UltraLowInput RNAseq (SMART technology)

Requires 1-10ng RNA, RIN >- 8; or 10-500 cells

Quality control is done after cDNA preparation



Single Cell Experimental Considerations

- Sample preparation, best practices
  - Prepare single cell suspension
  - Cell viability > 70%? (Dead Cell Removal Protocol)
  - Use a cell strainer to remove clumps or debris from washed cells
  - Store cells suspension on ice until you are ready to load

- How many cells do U want to target and number of samples
  - What is the question?
  - How rare is the cell type of interest? Does it have highly expressed markers?
  - Are cell or sample number limiting?

- Sequencing depth
  - Requires foreknowledge of both total mRNA content in individual cells and the diversity of mRNA
  - What are you looking for and why?
  - Are you searching for a rare population
  - Are you studying heterogeneity of single genes within a population

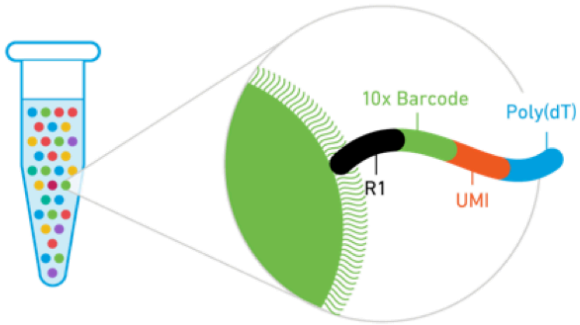
10 x Chromium Platform:
Single Cell 3’ Digital Expression 500-10,000 cells

3 Components:

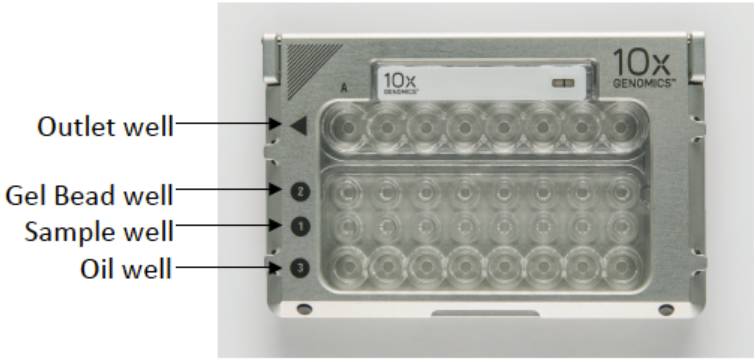
1. Gel beads with 750,000 unique barcodes

2. Cell suspension with RT reagents

3. Partitioning oil



Single-use microfluidics chip



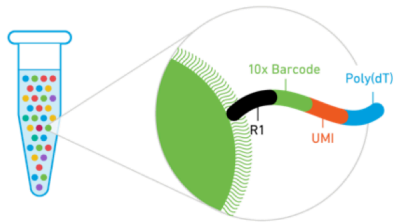
10 X Chromium Platform:
Single Cell 3' Digital Expression 500-10,000 cells

3 Components:

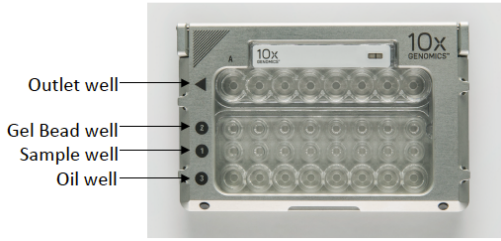
1. Gel beads with 750,000 unique barcodes

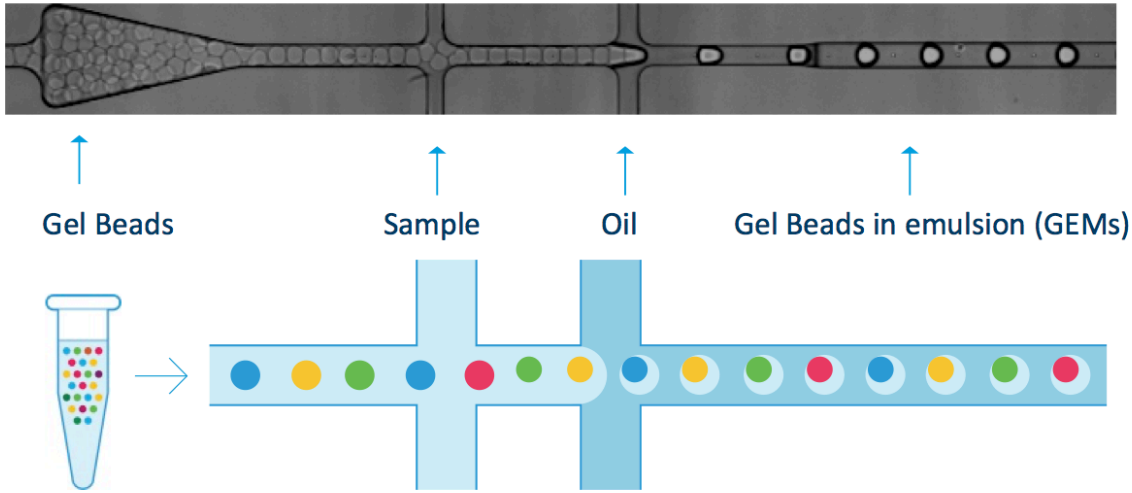
2. Cell suspension with RT reagents

3. Partitioning oil

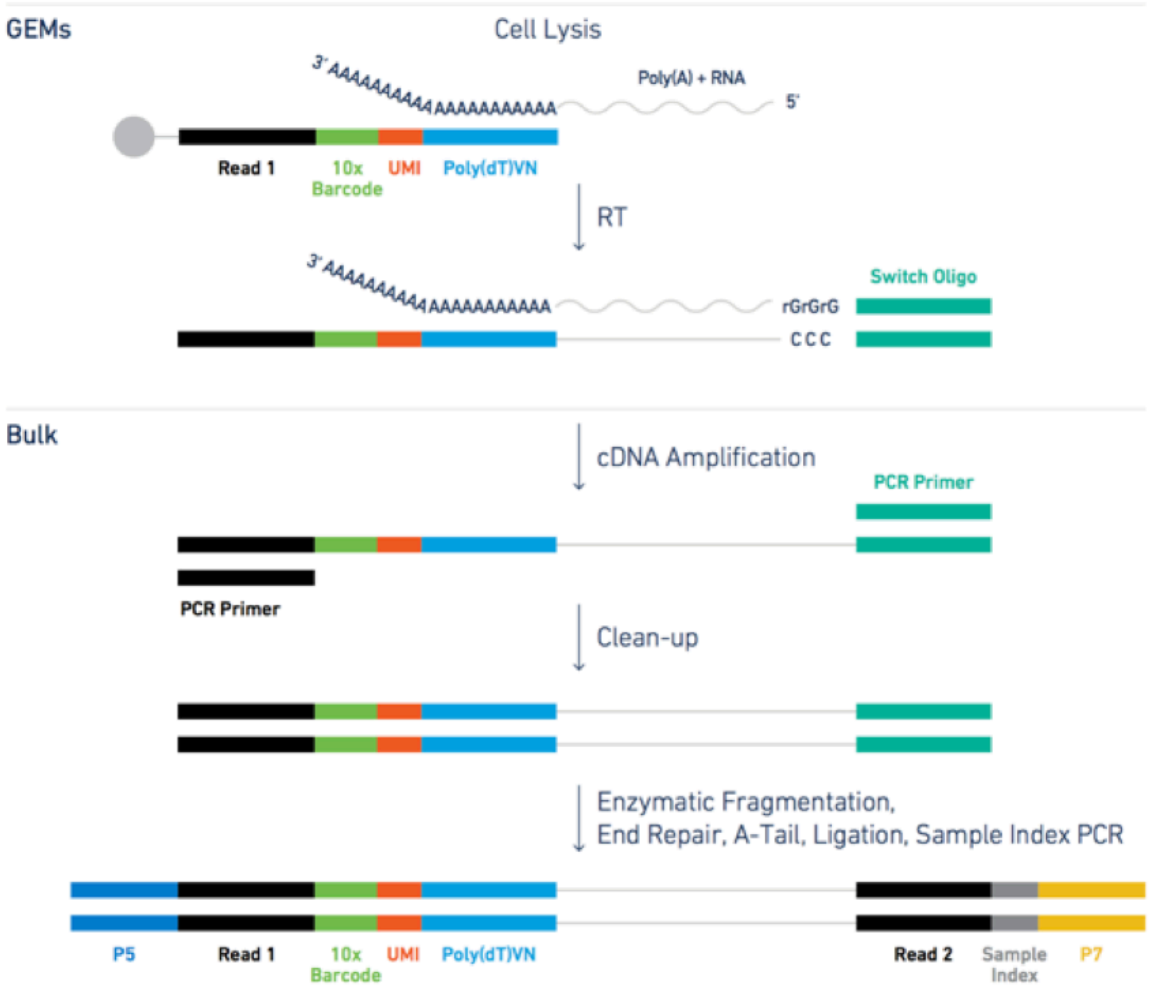


Single-use microfluidics chip



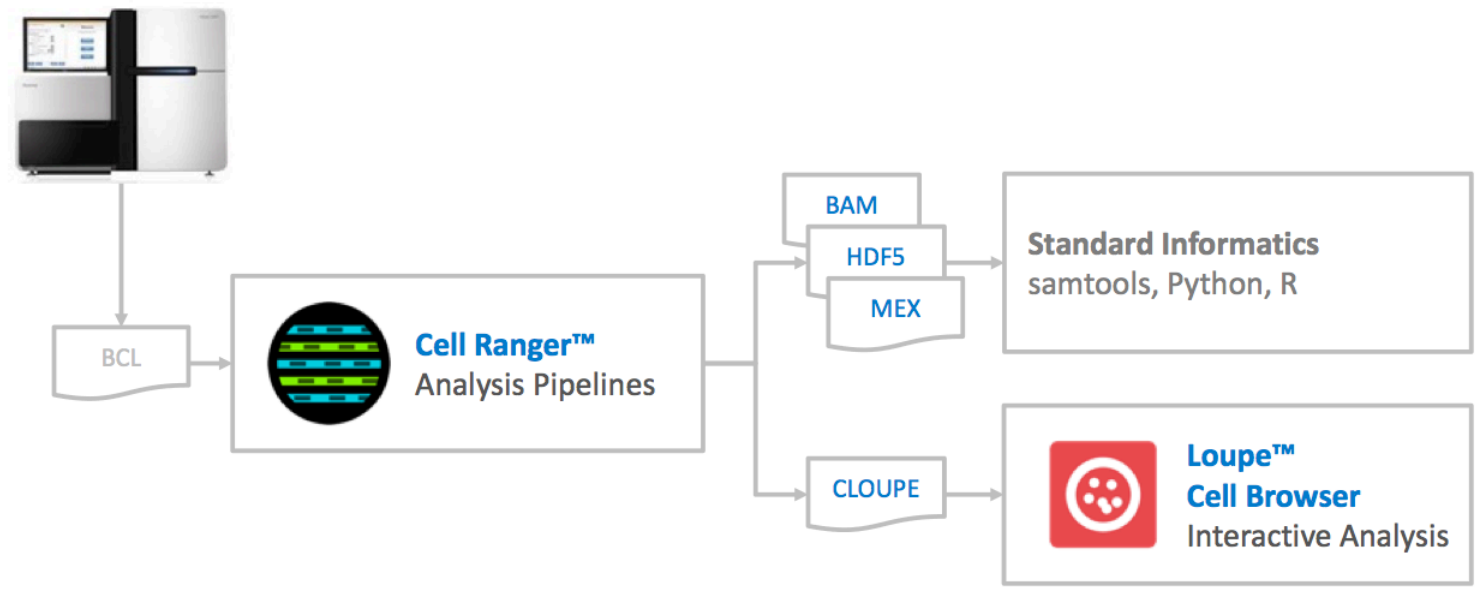


Chromium platform, library preparation



Chromium platform, Bioinformatics Support

- Sequence Chromium libraries
- Cell Ranger™ pipeline converts sequence data to single cell gene expression profiles
- Loupe™ Cell Browser enables interactive analysis



10 X Chromium Platform: Single Cell 3' Digital Expression

Cell Ranger analysis pipeline

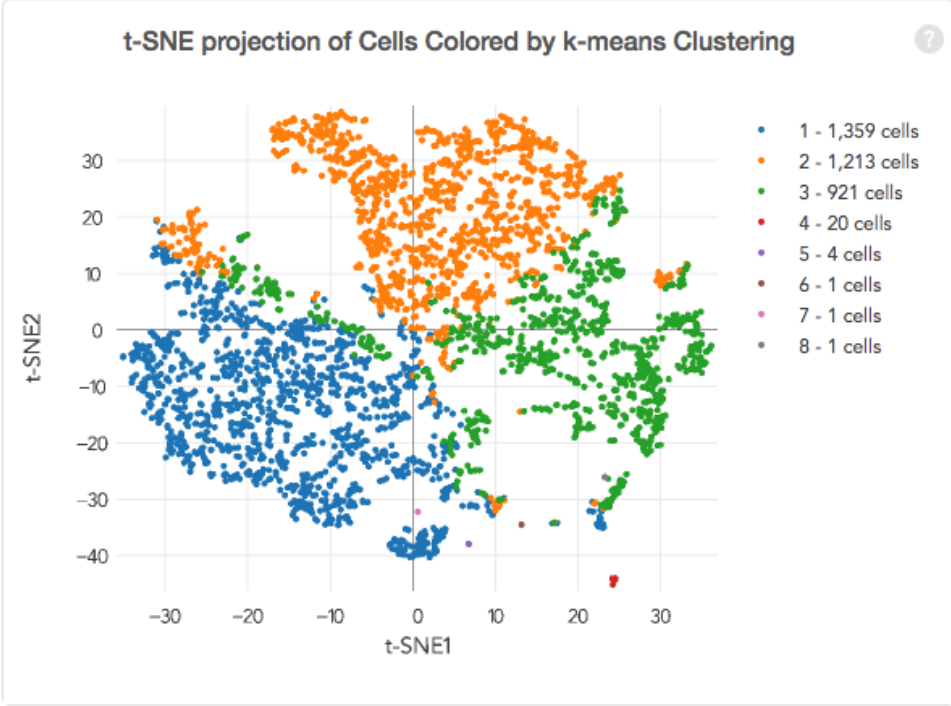
Estimated Number of Cells
3,520

Mean Reads per Cell **41,267**

Median Genes per Cell **2,151**

Sequencing

Number of Reads	145,260,501
Valid Barcodes	98.3%
Reads Mapped Confidently to Transcriptome	55.5%
Reads Mapped Confidently to Exonic Regions	58.6%
Reads Mapped Confidently to Intronic Regions	15.3%
Reads Mapped Confidently to Intergenic Regions	3.4%
Sequencing Saturation	61.4%
Q30 Bases in Barcode	95.9%
Q30 Bases in RNA Read	87.5%
Q30 Bases in Sample Index	95.6%
Q30 Bases in UMI	96.4%



10 x Chromium Platform: Single Cell 3’ Digital Expression

Sometimes, it fails …

Estimated Number of Cells
3,574

Mean Reads per Cell	Median Genes per Cell
48,598	3,255

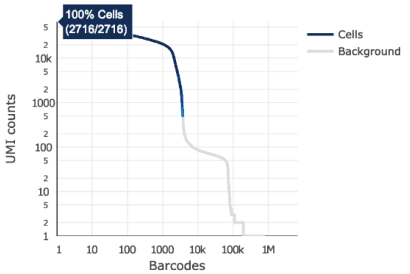
Sequencing

Number of Reads	173,689,296
Valid Barcodes	98.5%
Sequencing Saturation	46.1%
Q30 Bases in Barcode	98.9%
Q30 Bases in RNA Read	61.3%
Q30 Bases in Sample Index	95.5%
Q30 Bases in UMI	98.7%

Mapping

Reads Mapped to Genome	90.7%
Reads Mapped Confidently to Genome	88.0%
Reads Mapped Confidently to Intergenic Regions	4.8%
Reads Mapped Confidently to Intronic Regions	19.2%
Reads Mapped Confidently to Exonic Regions	64.1%
Reads Mapped Confidently to Transcriptome	61.4%
Reads Mapped Antisense to Gene	0.6%

Cells



Estimated Number of Cells	3,574
Fraction Reads in Cells	90.9%
Mean Reads per Cell	48,598
Median Genes per Cell	3,255
Total Genes Detected	21,018
Median UMI Counts per Cell	14,527

Sample

Name	1412A
Description	
Transcriptome	hg19
Chemistry	Single Cell 3' v3
Cell Ranger Version	3.0.0

Estimated Number of Cells
184

Mean Reads per Cell	Median Genes per Cell
959,423	1,856

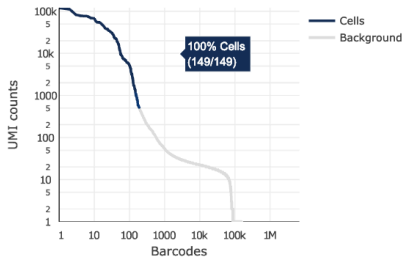
Sequencing

Number of Reads	176,533,869
Valid Barcodes	98.6%
Sequencing Saturation	95.8%
Q30 Bases in Barcode	98.5%
Q30 Bases in RNA Read	61.7%
Q30 Bases in Sample Index	95.4%
Q30 Bases in UMI	98.1%

Mapping

Reads Mapped to Genome	92.1%
Reads Mapped Confidently to Genome	88.4%
Reads Mapped Confidently to Intergenic Regions	5.3%
Reads Mapped Confidently to Intronic Regions	19.3%
Reads Mapped Confidently to Exonic Regions	63.7%
Reads Mapped Confidently to Transcriptome	61.1%
Reads Mapped Antisense to Gene	0.7%

Cells



Estimated Number of Cells	184
Fraction Reads in Cells	83.7%
Mean Reads per Cell	959,423
Median Genes per Cell	1,856
Total Genes Detected	16,925
Median UMI Counts per Cell	5,787

Sample

Name	1429A
Description	
Transcriptome	hg19
Chemistry	Single Cell 3' v3
Cell Ranger Version	3.0.0

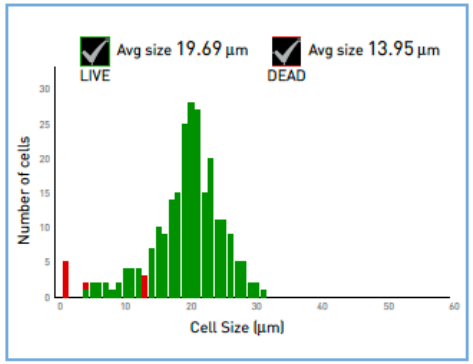
QC of single cell InvitroGen Cell Countess

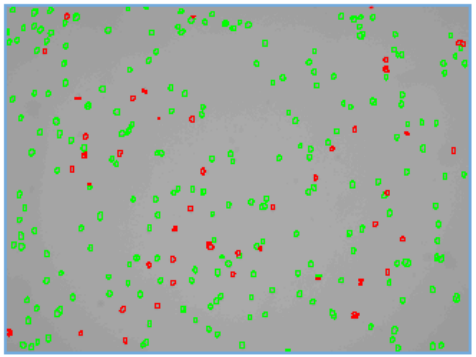
Countess II Live/Dead Report

File name: AH-6187 1_R.pdf
Date: 01.09.2020 00:12:10 AM

Results:

		Concentration
Total		1.72×10^6/mL
Live	81%	1.40×10^6/mL
Dead	19%	3.23×10^5/mL





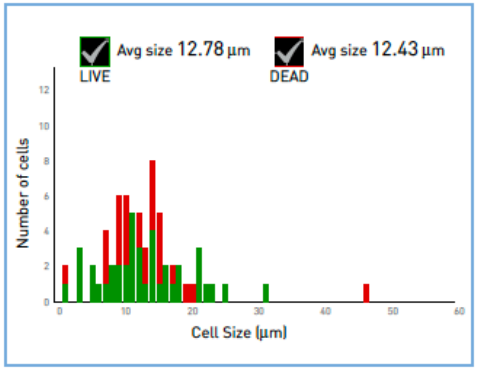
81% live

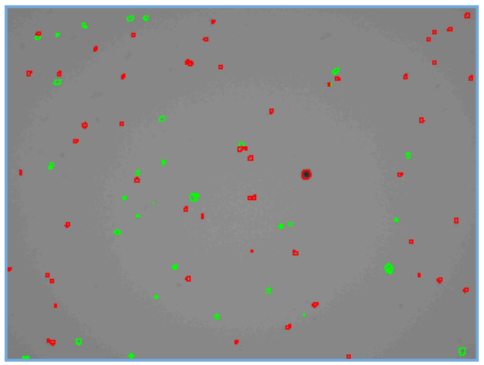
Countess II Live/Dead Report

File name: 11.21.19 Tumor PT_R.pdf
Date: 11.22.2019 06:40:55 AM

Results:

		Concentration
Total		5.63×10^5/mL
Live	42%	2.35×10^5/mL
Dead	58%	3.28×10^5/mL





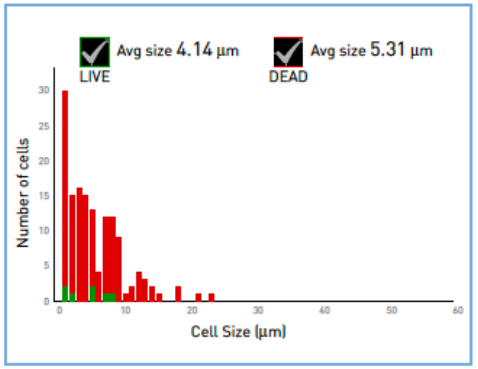
42% live

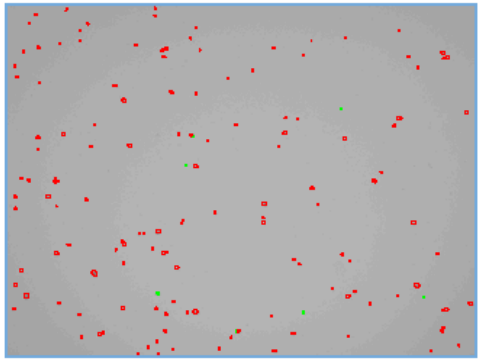
Countess II Live/Dead Report

File name: AH-6261_R.pdf
Date: 03.03.2020 01:43:42 AM

Results:

		Concentration
Total		8.80×10^5/mL
Live	5%	4.11×10^4/mL
Dead	95%	8.39×10^5/mL

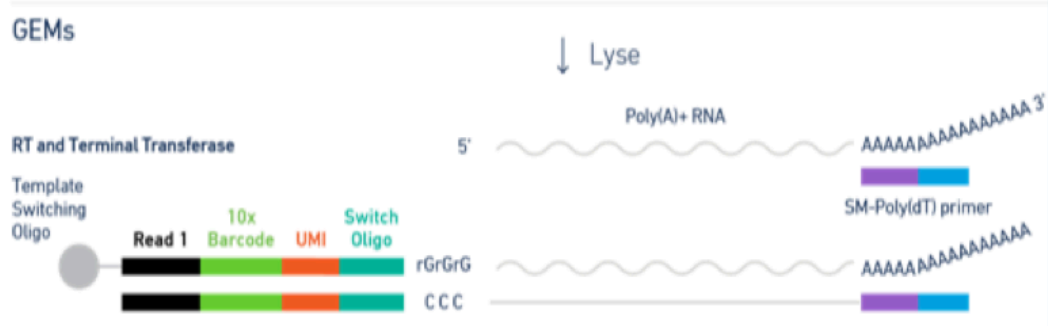




5% live

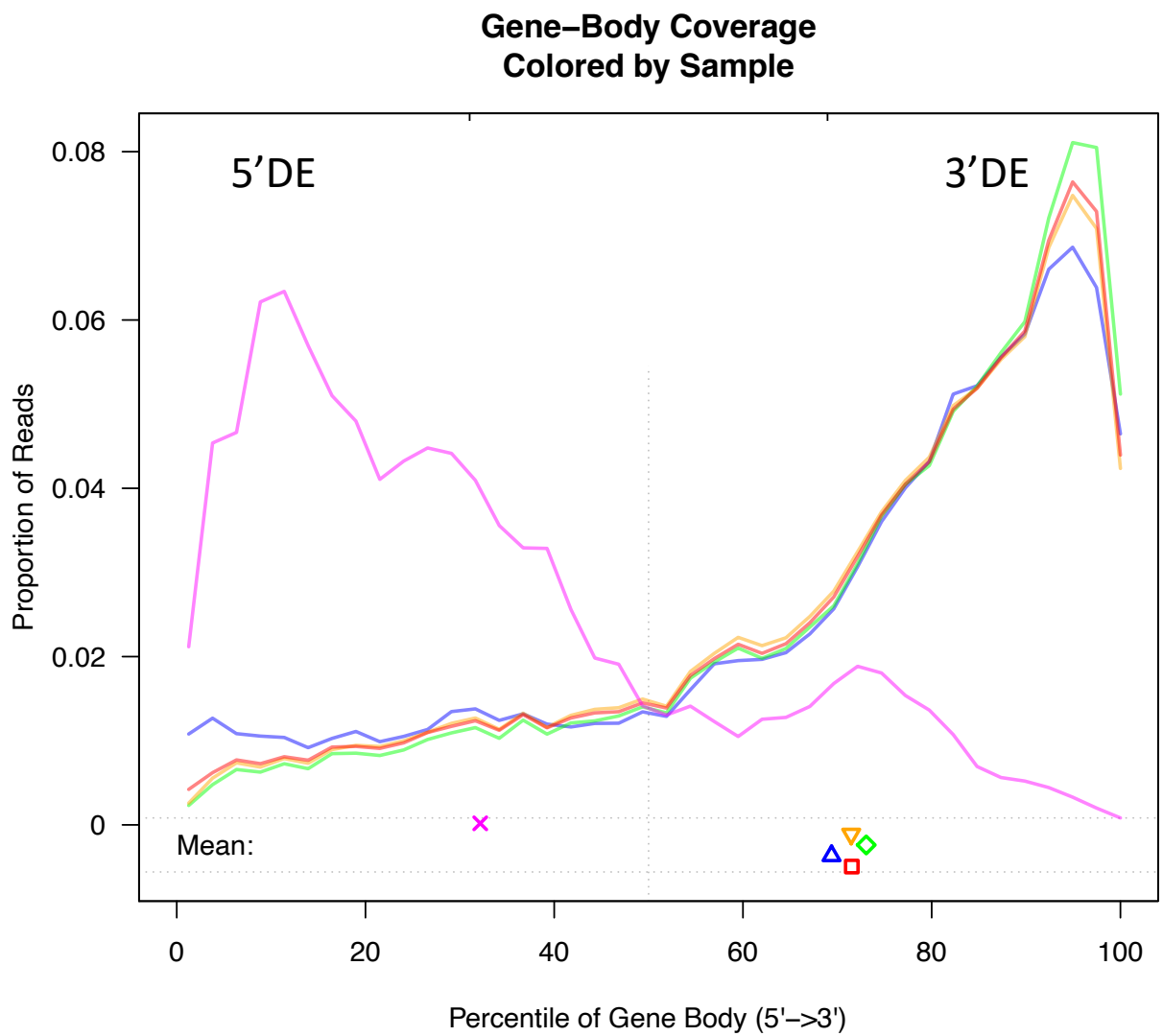
10 X Chromium Platform:
Single Cell 5’ Digital Expression and VDJ repertoire

SC5’



- UMI and Cell Barcode at 5’ position
- Primers contain sequences for universal PCR amplification for DGE
- Can use V(D)J primers to enrich for immune repertoire

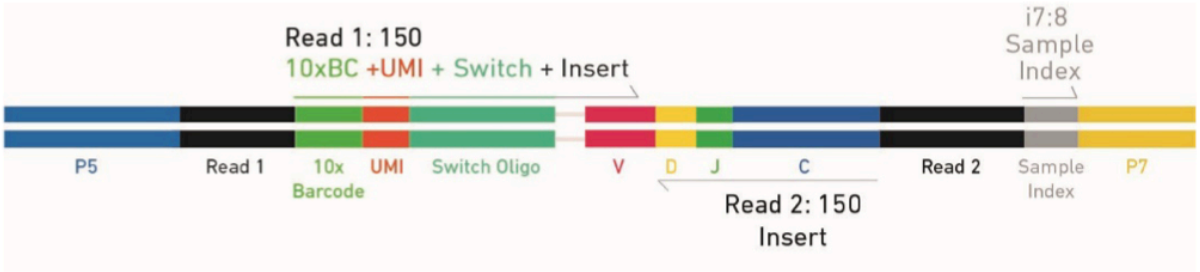
10 X Chromium Platform:
Single Cell 5’ Digital Expression vs 3’ Digital Expression



10 X Chromium Platform:
Single Cell 5' Digital Expression and VDJ repertoire

Two libraries, VDJ & DGE linked together by 10X Barcode

V(D)J Enriched Library Structure:



5,000 reads per cell

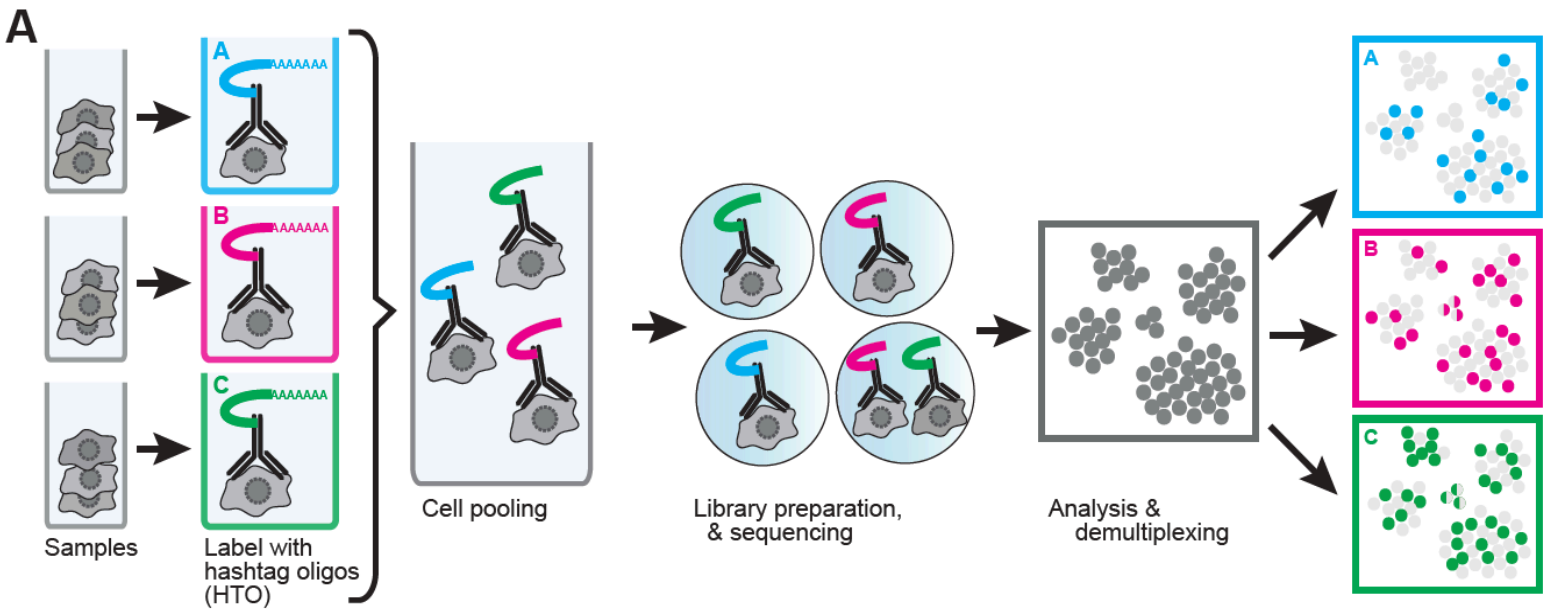
5' Gene Expression Library Structure:

50,000 reads per cell

For a 6000 cell experiment
$0.9 per cell, $5432

10 X Chromium Platform: Cell "hashing" 2020

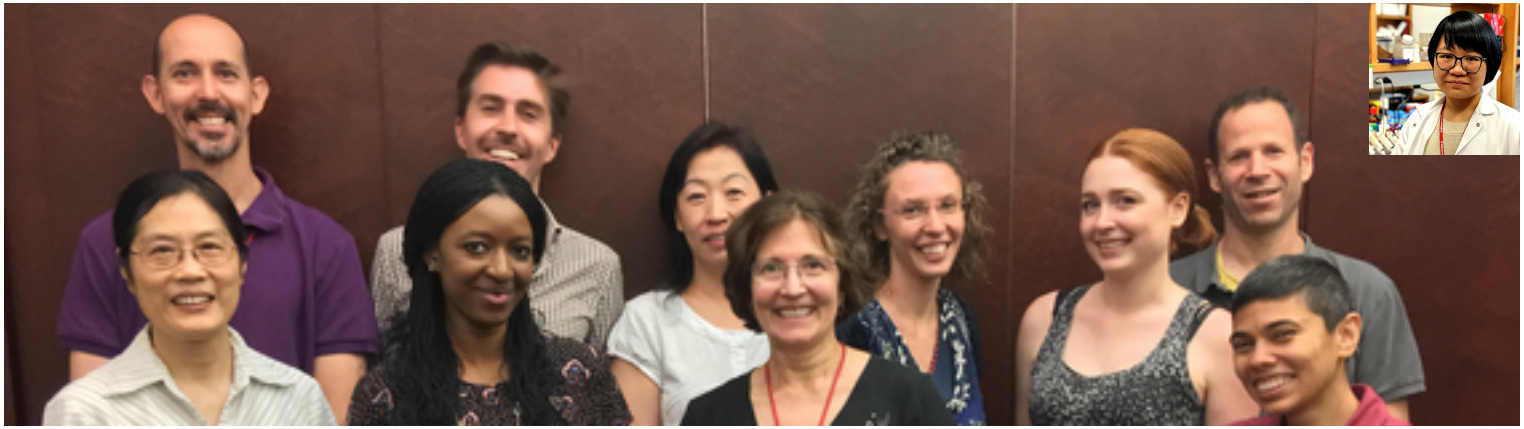
Cell "hashing" with barcoded antibodies enables multiplexing and doublet Detection for single cell genomics. NYGC



Potentially will lower library costs

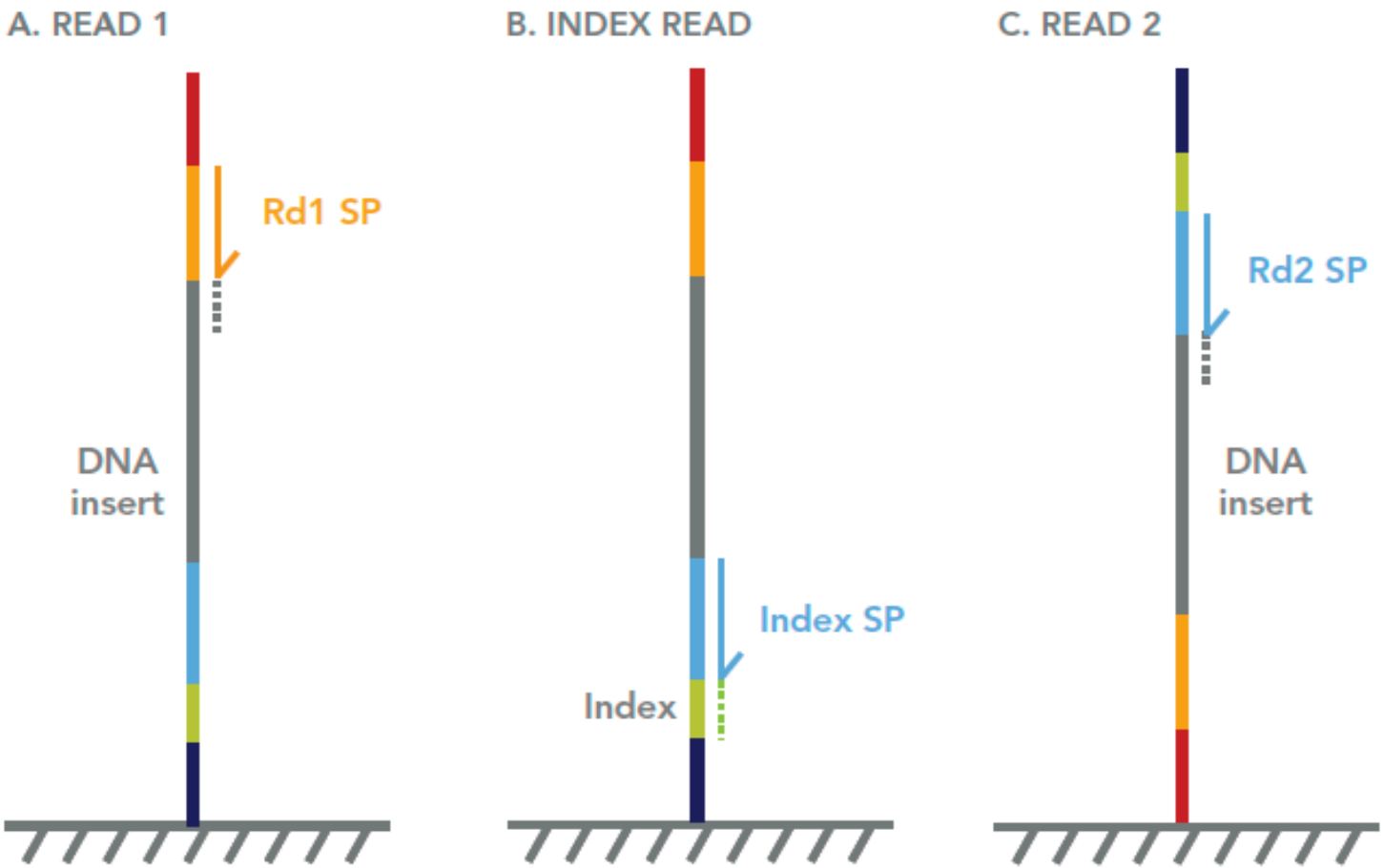
Meet the Epicore (past and present)

https://epicore.med.cornell.edu/



Director:	Alicia Alonso, PhD
Wet Lab	
Lab Manager	Yushan Li
Research Specialists	Natalie Chow
	Caroline Sheridan
Dry Lab	
Bioinformatics Director:	Doron Betel, PhD
Computational Biologist:	Piali Mukherjee
Software Developers:	Thadeous Kacmarczyk, PhD
	Simon Johnson
Former members	
Research Technician	Yuan Xin
Research Specialist	Marisa Mariani

Indexed sequencing method is now standard for single and paired reads



Data processing (demux report)

Flowcell Summary

Clusters (Raw)	Clusters(PF)	Yield (MBases)
1,448,964,720	1,047,850,753	132,029

Lane Summary

Lane	Project	Sample	Barcode sequence	PF Clusters	% of the lane	% Perfect barcode	% One mismatch barcode	Yield (Mbases)	% PF Clusters	% >= Q30 bases	Mean Quality Score
4	Project_EC-MDL-5785	Sample_1_Pt_10_LN_C2-5GEX_S17	TTCAGGTG	15,287,467	4.60	98.29	1.71	1,926	100.00	87.19	36.58
4	Project_EC-MDL-5785	Sample_1_Pt_10_LN_C2-5GEX_S18	ACGGACAT	20,320,393	6.12	98.31	1.69	2,560	100.00	88.21	36.88
4	Project_EC-MDL-5785	Sample_1_Pt_10_LN_C2-5GEX_S19	GATCTTGA	15,451,081	4.65	97.88	2.12	1,947	100.00	87.65	36.72
4	Project_EC-MDL-5785	Sample_1_Pt_10_LN_C2-5GEX_S20	CGATCACC	20,675,988	6.23	98.02	1.98	2,605	100.00	87.63	36.70
4	Project_EC-MDL-5785	Sample_2_Pt_13_BMA-5GEX_S13	CCTCATTC	19,353,590	5.83	97.44	2.56	2,439	100.00	87.35	36.62
4	Project_EC-MDL-5785	Sample_2_Pt_13_BMA-5GEX_S14	AGCATCCG	113,509	0.03	6.39	93.61	14	100.00	83.14	35.52
4	Project_EC-MDL-5785	Sample_2_Pt_13_BMA-5GEX_S15	GTGGCAAT	11,446,782	3.45	98.13	1.87	1,442	100.00	87.70	36.74
4	Project_EC-MDL-5785	Sample_2_Pt_13_BMA-5GEX_S16	TAATGGGA	16,362,901	4.93	97.67	2.33	2,062	100.00	87.57	36.70
4	Project_EC-MDL-5785	Sample_3_Pt_13_PB_C1a-5GEX_S10	AGGCCCGA	23,173,971	6.98	97.90	2.10	2,920	100.00	87.62	36.71
4	Project_EC-MDL-5785	Sample_3_Pt_13_PB_C1a-5GEX_S11	TACGTGAC	17,192,787	5.18	98.33	1.67	2,166	100.00	83.94	35.73
4	Project_EC-MDL-5785	Sample_3_Pt_13_PB_C1a-5GEX_S12	GTTTATCT	13,890,552	4.18	98.13	1.87	1,750	100.00	87.54	36.70
4	Project_EC-MDL-5785	Sample_3_Pt_13_PB_C1a-5GEX_S9	CCAAGATG	16,317,332	4.91	98.21	1.79	2,056	100.00	87.85	36.76
4	Project_EC-MDL-5785	Sample_4_Pt_13_PB_C1b-5GEX_S5	TATGATTC	13,876,961	4.18	97.62	2.38	1,748	100.00	88.52	36.95
4	Project_EC-MDL-5785	Sample_4_Pt_13_PB_C1b-5GEX_S6	CCCACAGT	20,094,411	6.05	98.00	2.00	2,532	100.00	88.36	36.90
4	Project_EC-MDL-5785	Sample_4_Pt_13_PB_C1b-5GEX_S7	ATGCTGAA	15,173,413	4.57	98.18	1.82	1,912	100.00	88.42	36.94
4	Project_EC-MDL-5785	Sample_4_Pt_13_PB_C1b-5GEX_S8	GGATGCCG	18,655,252	5.62	97.45	2.55	2,351	100.00	86.81	36.48
4	Project_EC-MDL-5785	Sample_5_Pt_16_PB_C3-5GEX_S1	ACTTCATA	15,771,239	4.75	98.02	1.98	1,987	100.00	88.23	36.88
4	Project_EC-MDL-5785	Sample_5_Pt_16_PB_C3-5GEX_S2	GAGATGAC	17,069,548	5.14	95.24	4.76	2,151	100.00	84.50	35.85
4	Project_EC-MDL-5785	Sample_5_Pt_16_PB_C3-5GEX_S3	TGCCGTGG	9,952,240	3.00	97.42	2.58	1,254	100.00	87.67	36.71
4	Project_EC-MDL-5785	Sample_5_Pt_16_PB_C3-5GEX_S4	CTAGACCT	23,092,433	6.96	98.08	1.92	2,910	100.00	88.96	37.08
4	default	Undetermined	*unknown*	8,731,527	2.63	100.00	NaN	1,100	5.47	82.92	35.47

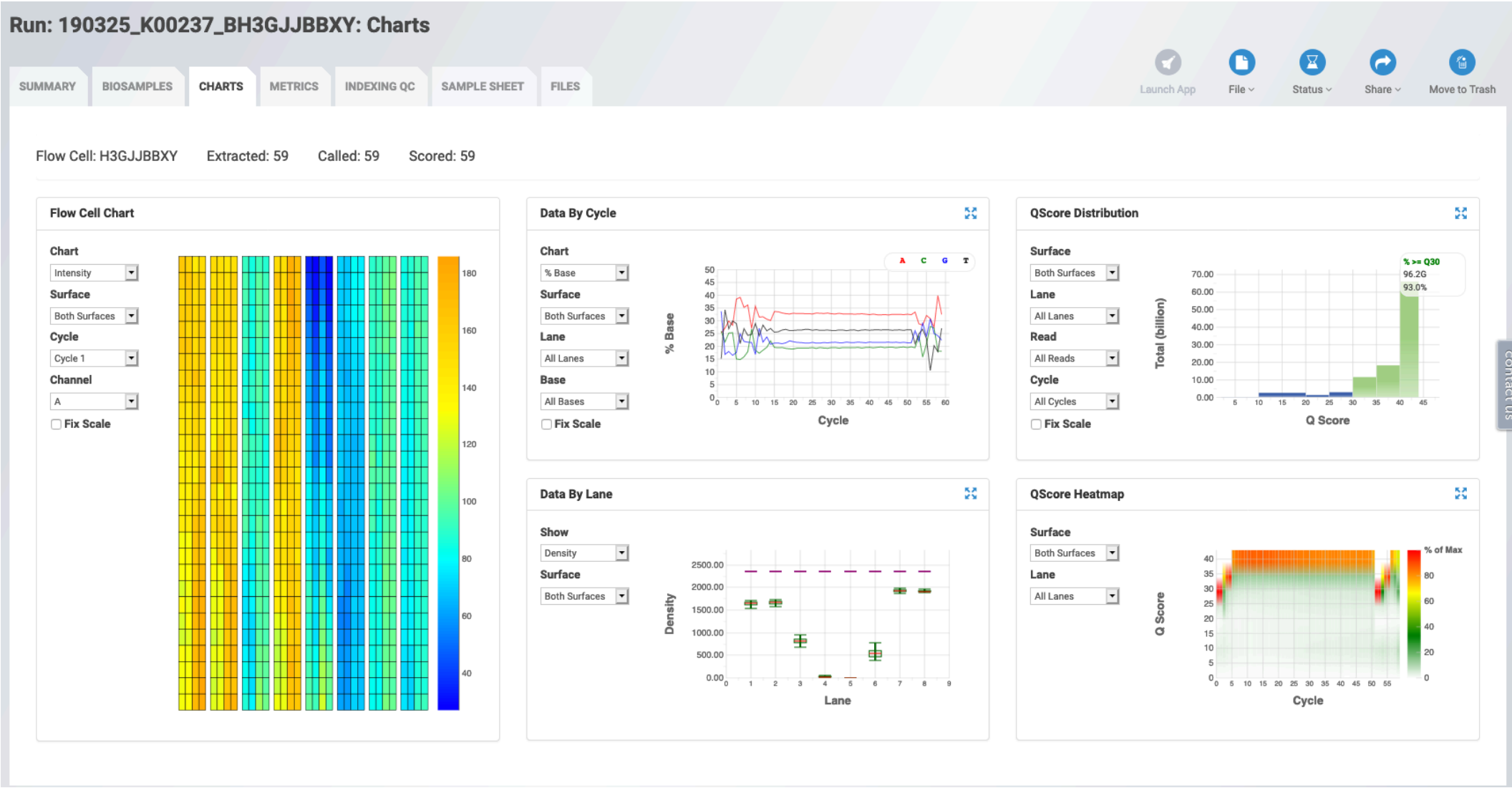
Sequence data (base call files) bcl- are demultiplexed and converted to **fastq** files using Illumina bcl2fastq software

Raw data: FASTQ file

Sequencing run quality control

Some lanes on this run need to be repeated

4 colors, patterned



Sequencing run quality control

Patterned , 2 colors (ie NovaSeq)

Flow Cell: HMFYYDSXX Extracted: 318 Called: 318 Scored: 318