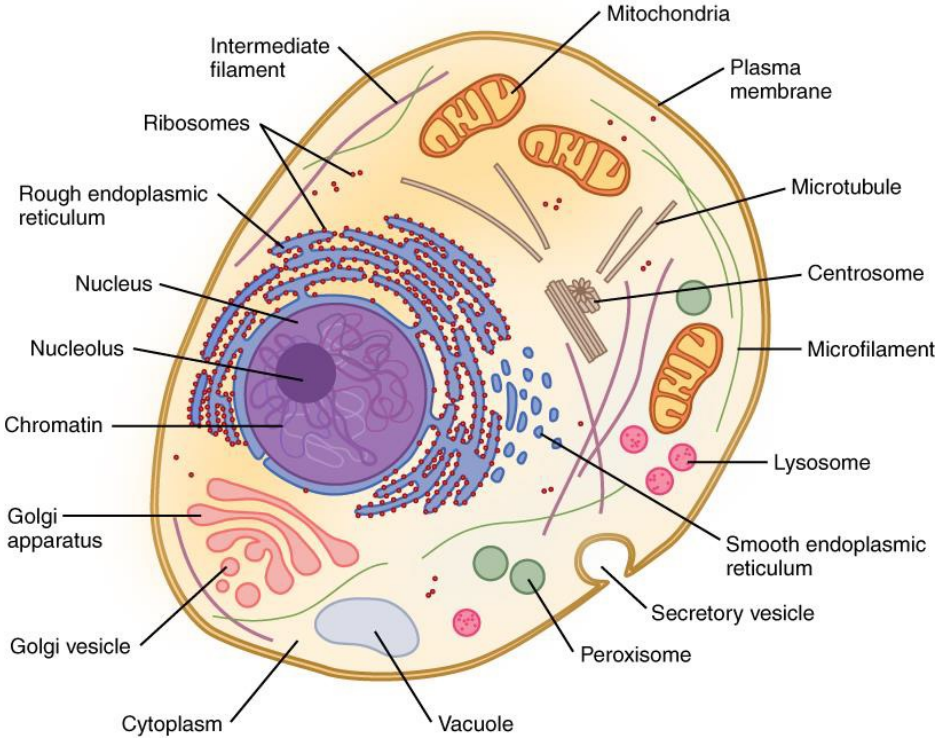
Sequencing for Statisticians

Meng Wang
Senior Bioinformatic Scientist
Stanton Lab, Center for Childhood Cancer Resarch
Nationwide Children's Hospital

NATIONWIDE CHILDREN'S®
When your child needs a hospital, everything matters.

Outline

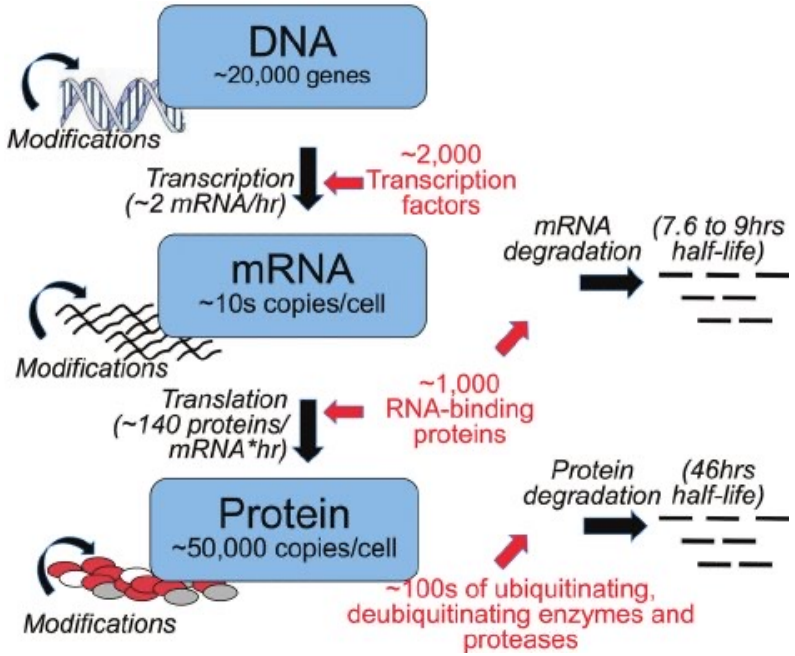
- Background
- Sequencing Platforms
- Different Types of Sequencing



Slides courtesy of Ryan Roberts

Why sequencing?

- Knowledge of our genome
- Transcription differences
- Transcriptional regulations
- Structural Variations
- etc, etc



What can we quantify?

- mRNA (RNA-seq)
- DNA (WGS, ChIP-seq, ATAC-seq, Hi-C)
- Protein (proteinomics)

Slides courtesy of Ryan Roberts

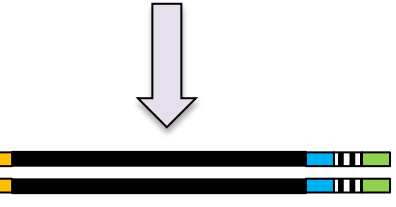
Sequencing Platforms

- Sequencing consists of two steps:
  - Library prep (generating sequencable fragments)
  - Sequencing
- Illumina (dominant platform)
- Pacbio/Nanopore (long read sequencing)

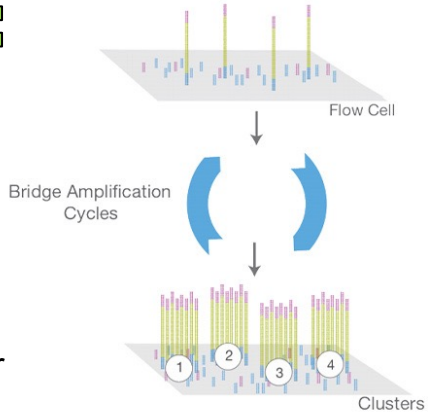
Overview of Illumina Workflow

1 **Sample Prep**

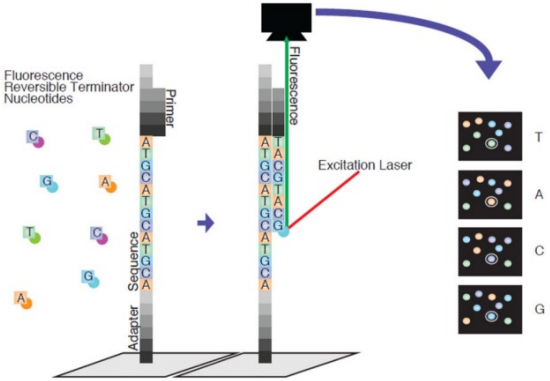
DNA/cDNA



2 **Cluster Generation**



3 **Sequencing**

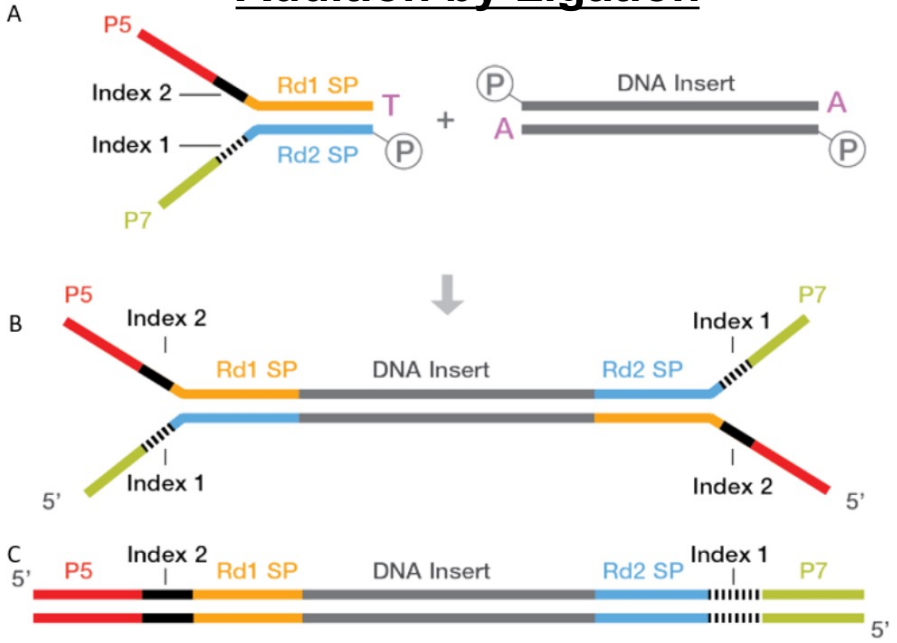


4 **Data Analysis**

Slides courtesy of Anthony Miller

Illumina Adapter – Foundation for Cluster Generation

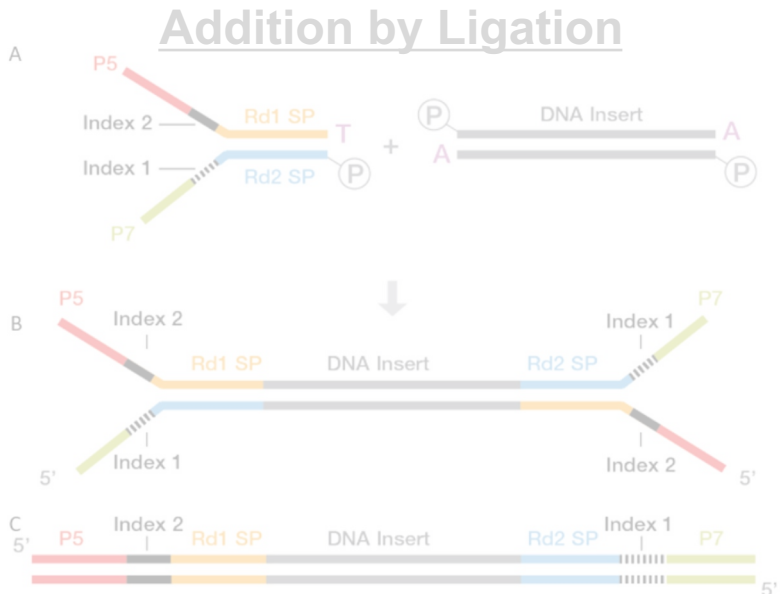
Addition by Ligation

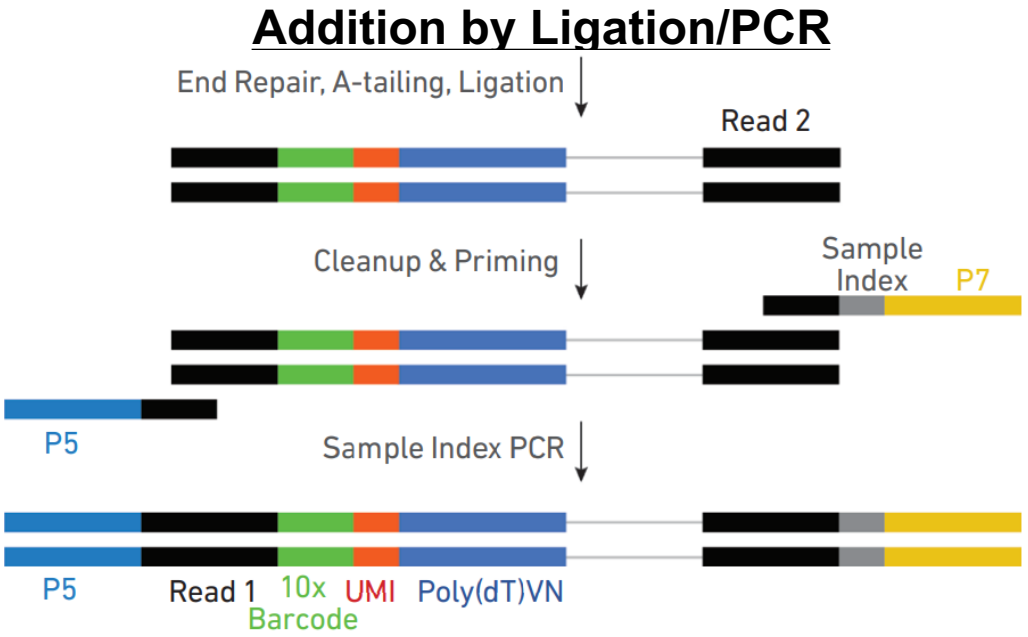


1

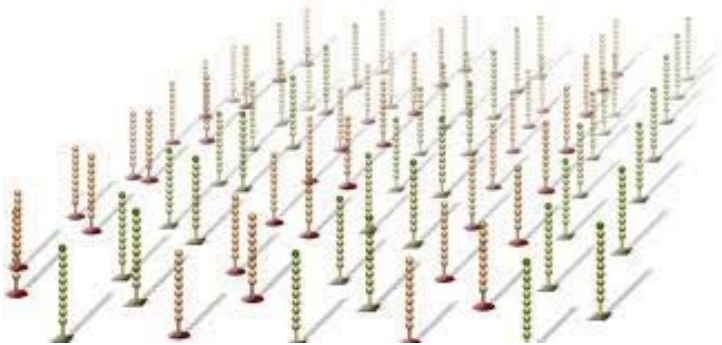
Sample Prep

Illumina Adapter – Foundation for Cluster Generation

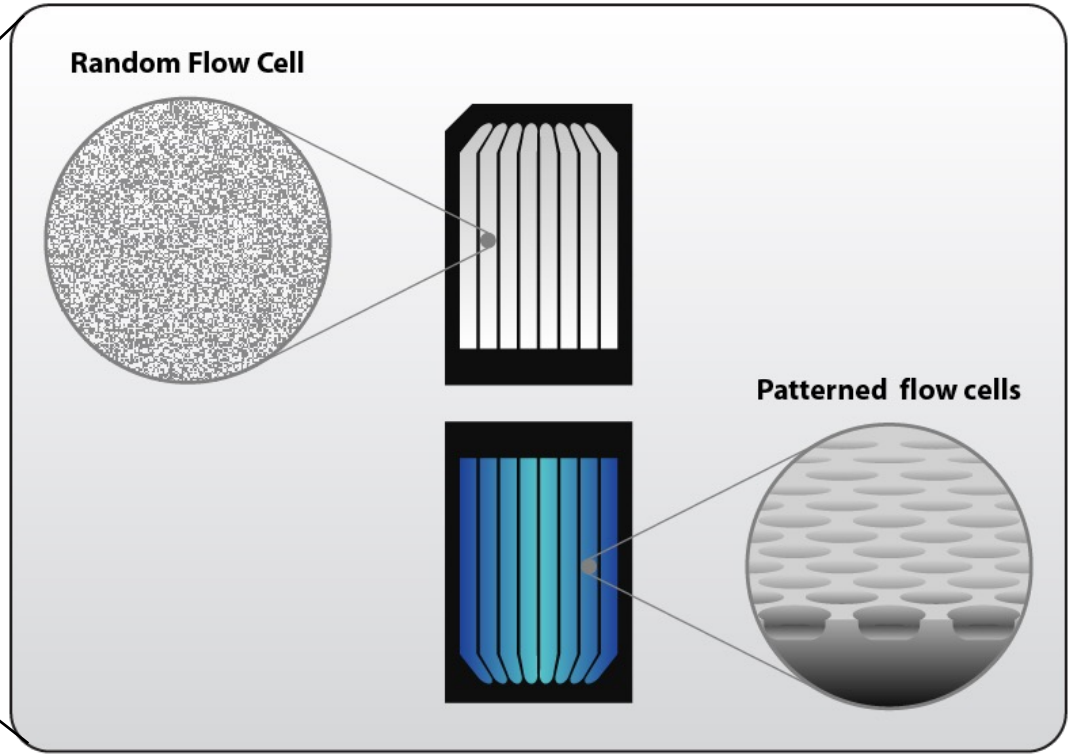


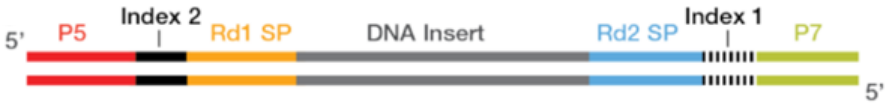


Illumina Flowcell – Cluster Generation



Flow cell contains lawn of oligos complementary to the P5/P7 sequences



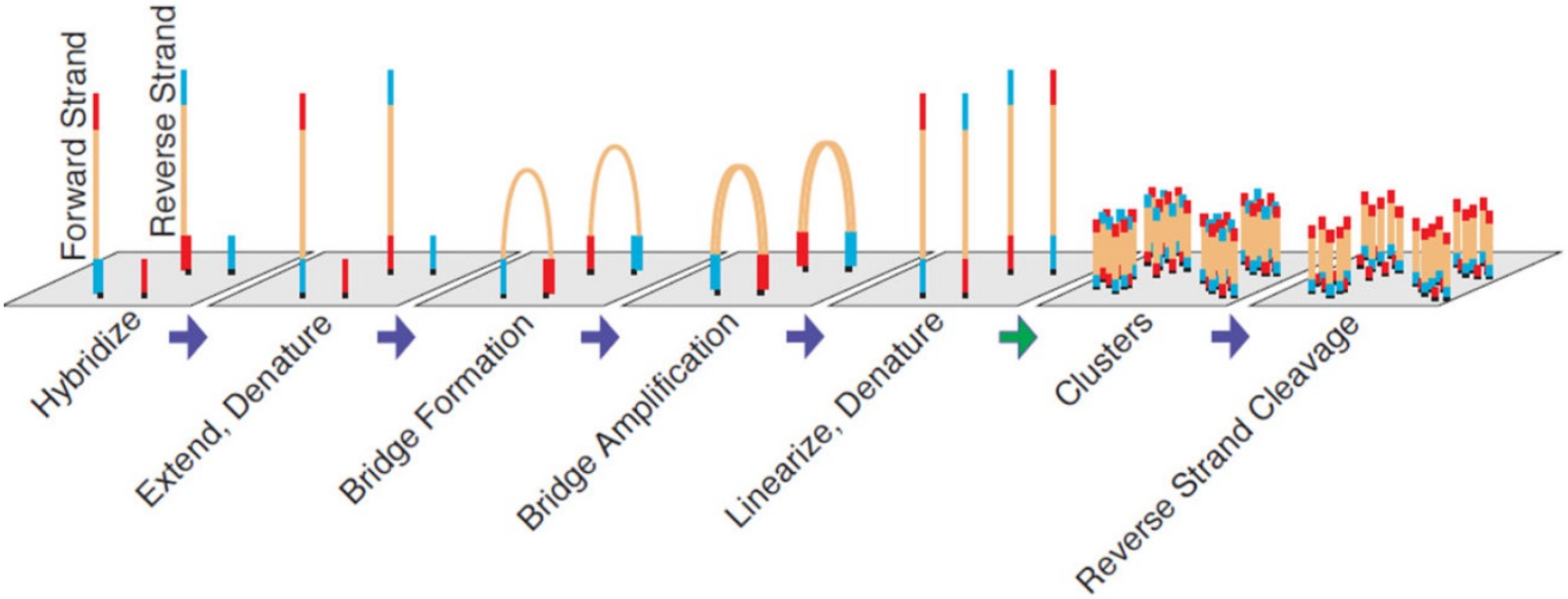


Slides courtesy of Anthony Miller

2 Cluster Generation

3 Sequencing

Illumina Flowcell – Cluster Generation

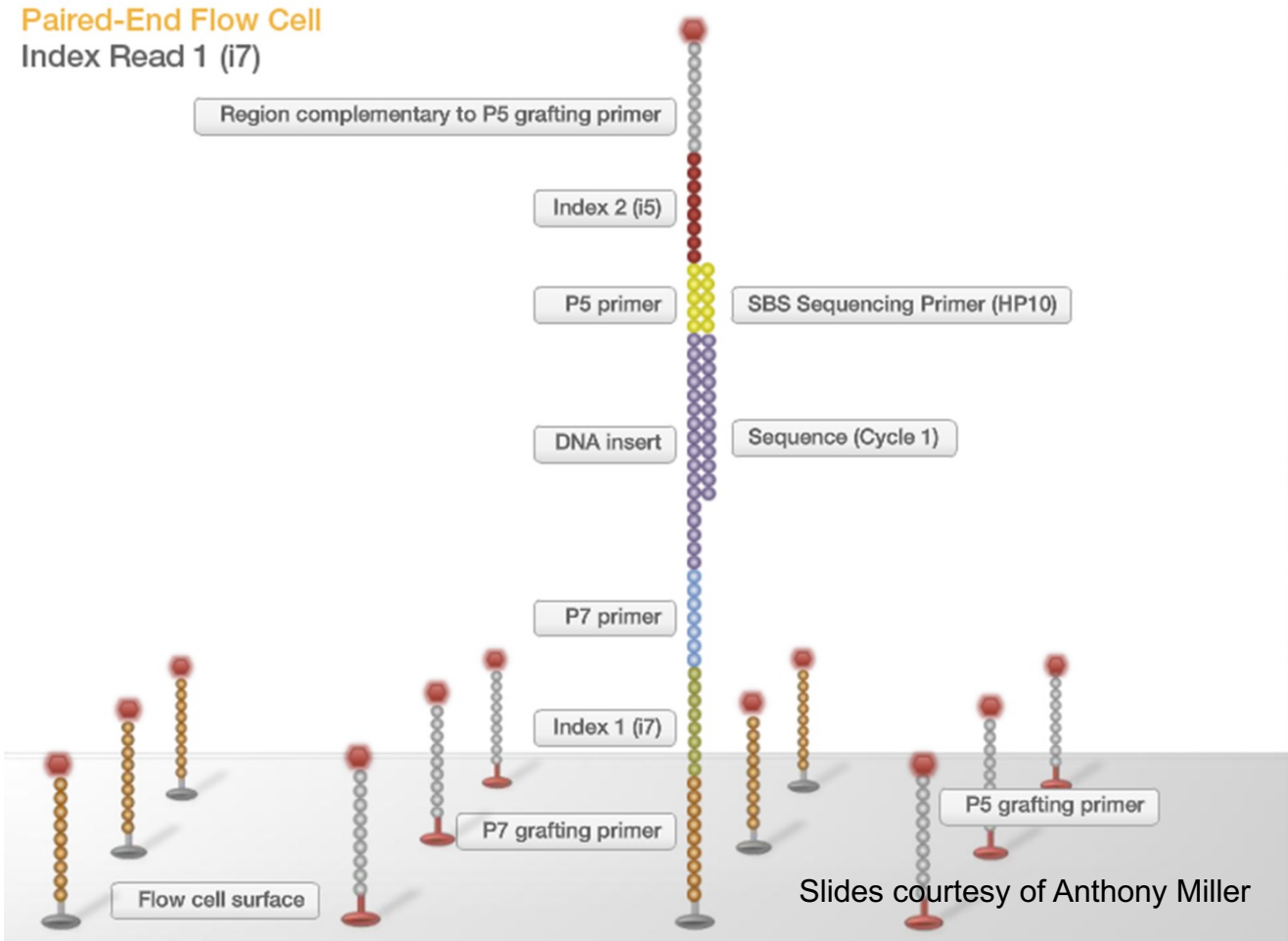


P7 is blue in above diagram

Slides courtesy of Anthony Miller

Illumina Sequencing By Synthesis (SBS) – Read 1

- Initiated by HP10 primer (Rd1 SP)
- Fluorescently labeled and reversibly terminated nucleotides
- Clusters are excited by light, fluorescent signal emitted
- Terminator remove for next round of nucleotide addition

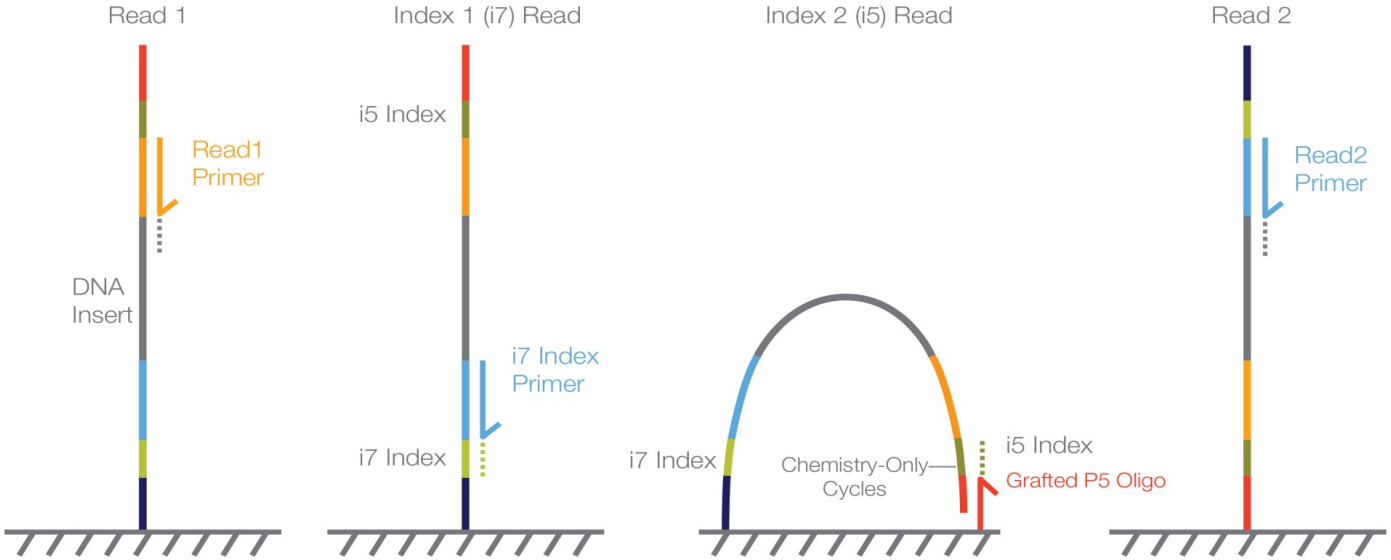


Slides courtesy of Anthony Miller

Illumina Sequencing By Synthesis (SBS) – Index(s) and Read 2

NovaSeq 6000 w/ v1 chem, MiniSeq w/ rapid chem, MiSeq, HiSeq 2000/2500

Figure 2 Dual-Indexed Sequencing on a Paired-End Flow Cell (Workflow A)



After Rd1 SBS, product washed way

After i7 Index SBS, product washed way and 3' end deprotected

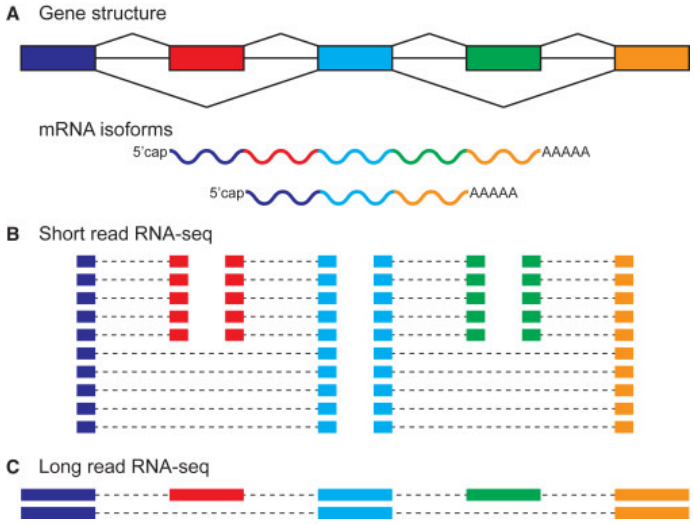
i5 Index SBS, polymerase extends forming double strand, linearized, 3' end block, original strand cleaved

Read2 primer initiates Rd 2 SBS

Slides courtesy of Anthony Miller

Alternative Splicing

- Genes are not continuous coding.
  - Exons: coding regions
  - Introns: silent.
- Different combinations of exons yields **isoforms**.
- This phenomenon is called alternative splicing.

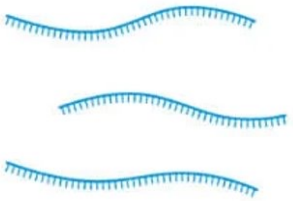


Types of Sequencing

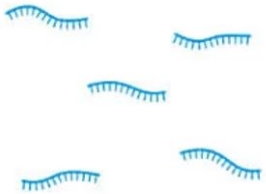
- RNA: quantifying transcription
- WGS: sequencing the genome
- Epigenetics (sequencing DNA):
  - ChIP-seq: transcription factors and histone modification
  - ATAC-seq: chromatin accessibility
  - Hi-C: 3D chromatin structure.
  - bisulfite sequencing (DNA methylation).

RNA Sequencing

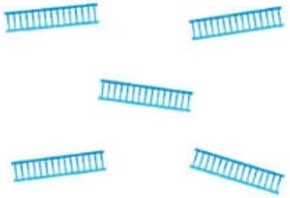
1. Isolate RNA from samples



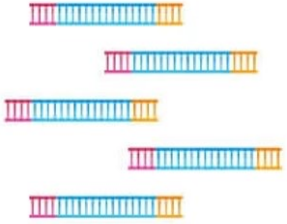
2. Fragment RNA into short segments



3. Convert RNA fragments into cDNA

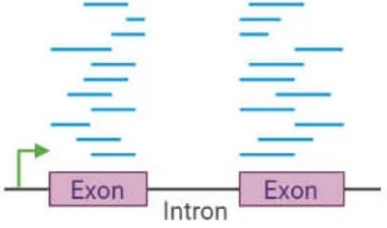


4. Ligate sequencing adapters and amplify



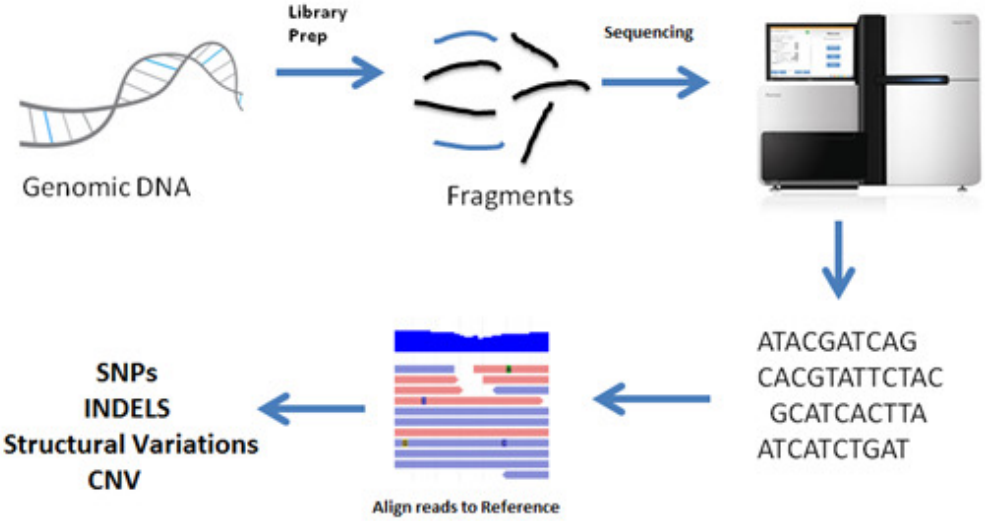
5. Perform NGS sequencing

6. Map sequencing reads to the transcriptome/genome



Slides courtesy of Ryan Roberts

Whole Genome Sequencing

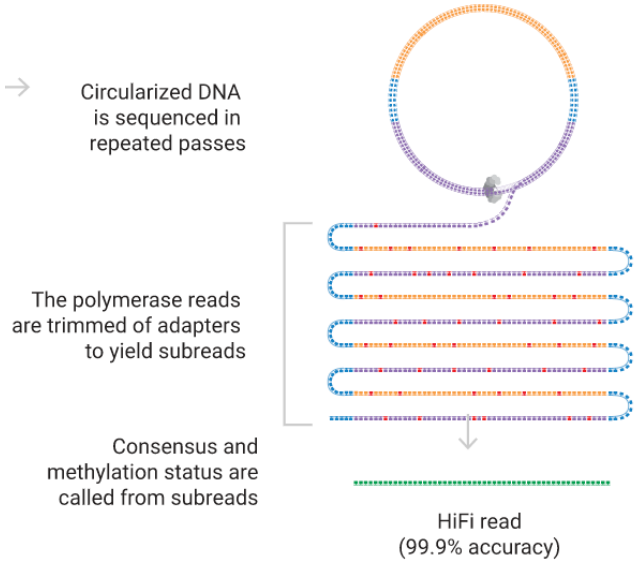


PacBio HiFi Sequencing

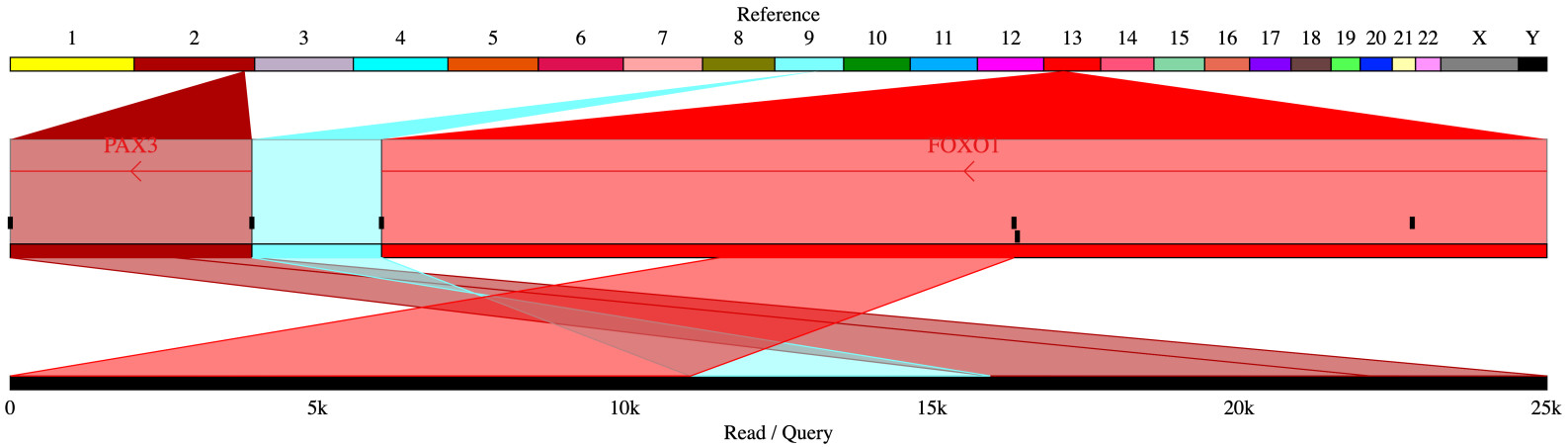
- Illumina platforms are restricted to 300bp reads, and 500bp fragments.
- Great overall, but not that good for isoform or structure variation detection.

PacBio Long Read

- 10k-30kb read length.
- Amazing for:
  - genome assembly,
  - CNV and SV calls,
  - isoform analysis (RNA),
  - Microsatellite repeats,
  - Haplotyping!!!
- Single cell version is being developed.
- Terrible error rate one pass (10%)
- Consensus (10 times or more) reads are very accurate.



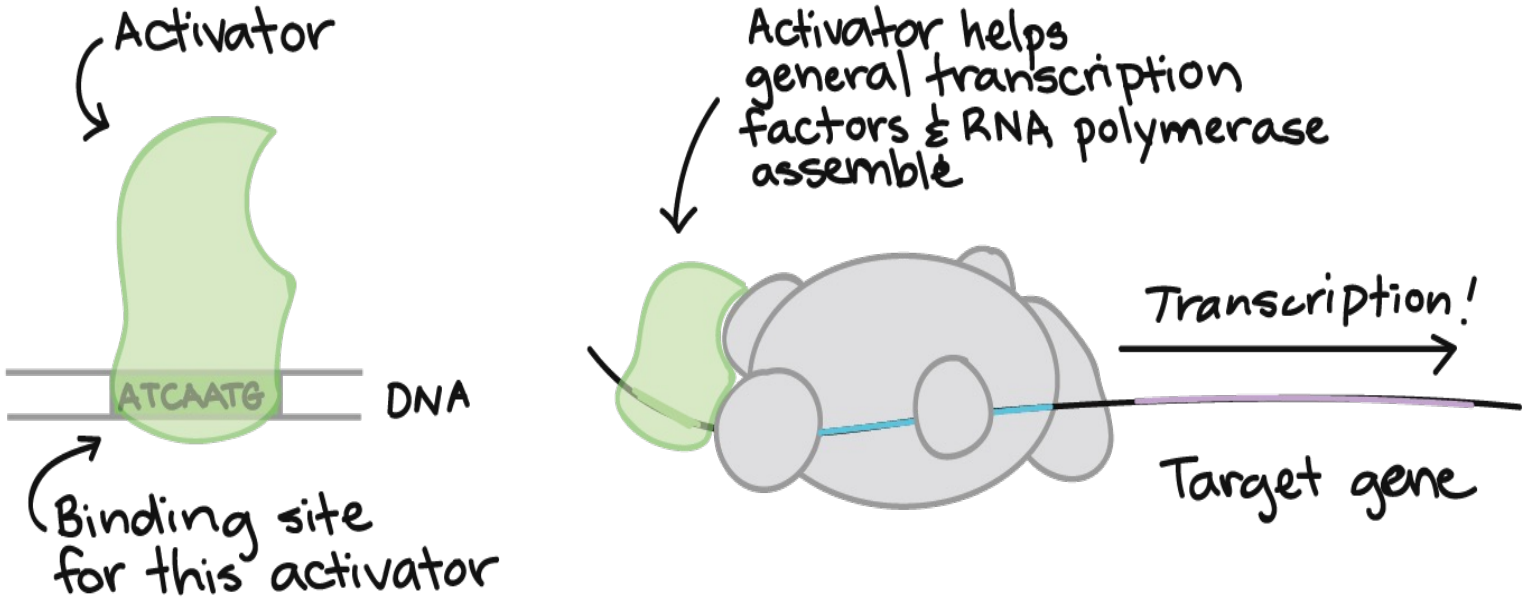
PacBio Mapping of a cancer fusion gene



unpublished

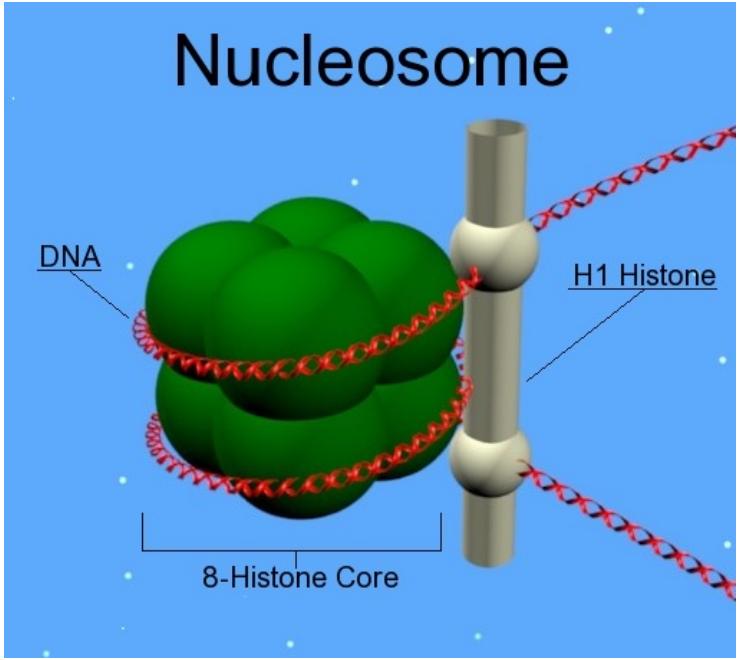
Transcription Factor

Proteins that modulate gene transcription



Nucleosomes

- Nucleosome: basis of chromatin: 147bp of DNA wrapped around a nucleosome (8 histone).
- Heterochromatin = tightly packed nucleosomes + DNA wrapped around it, usually repressed.
- Euchromatin = ”free” chromatin, usually transcriptionally active.



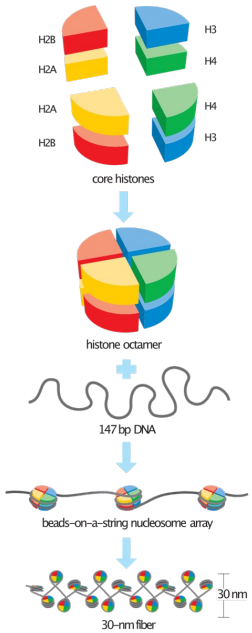
wikipedia

Nucleosomes

- Histones can be modified chemically. Those modifications activate or repress expression.

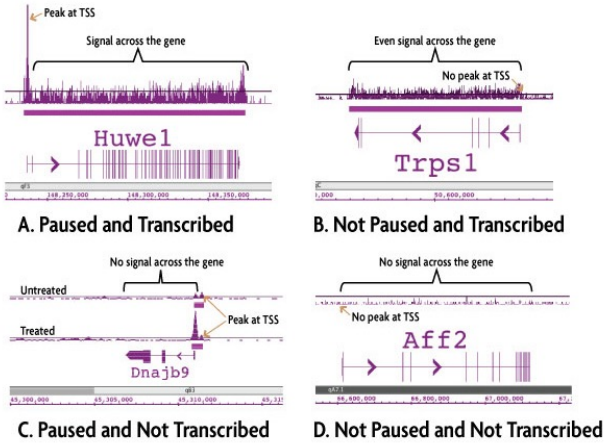
- For example:
  - H3K9me3 = repressive (3rd histone, 9th lysine, methylated 3 times)
  - H3K27ac = active.

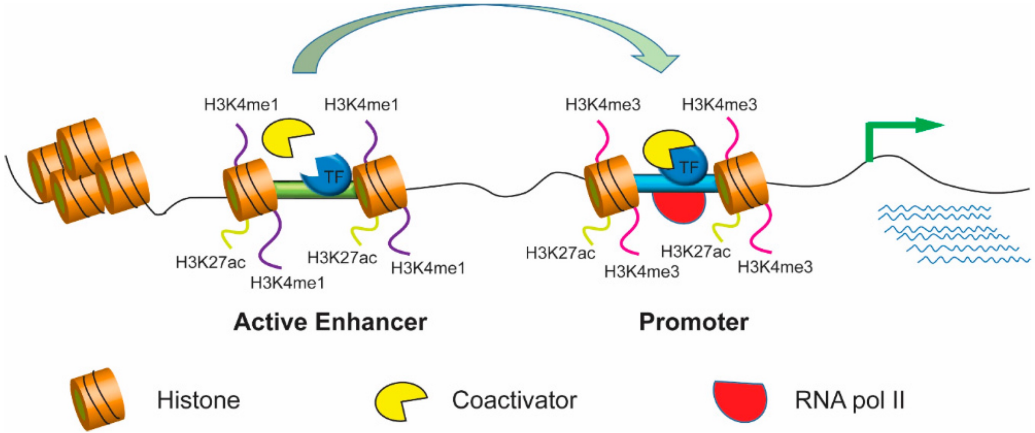
- Histone Modification are protein and therefore can be assessed with ChIP-Seq.



wikipedia

Transcription Regulation

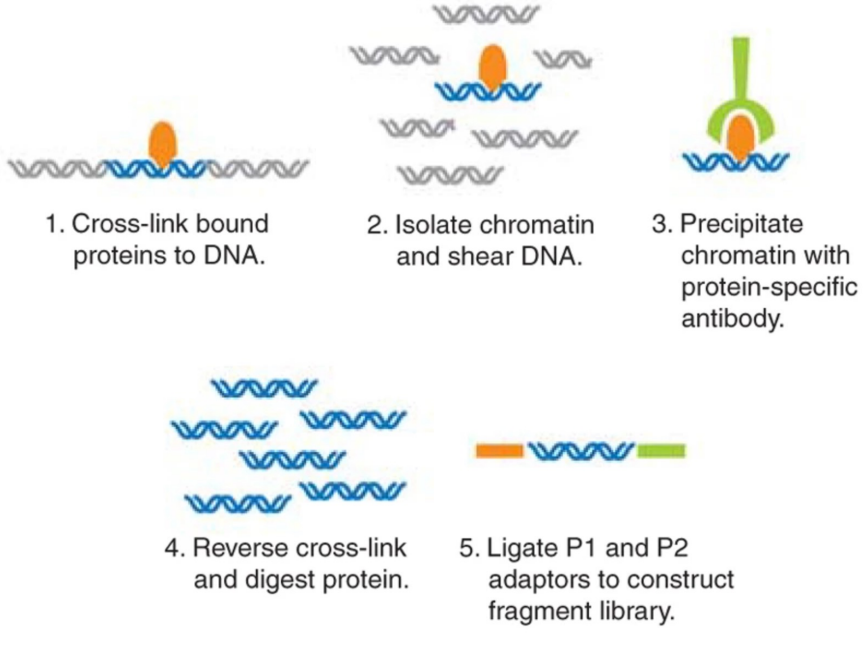




RNA Pol II ChIP-Seq

ChIP-Seq

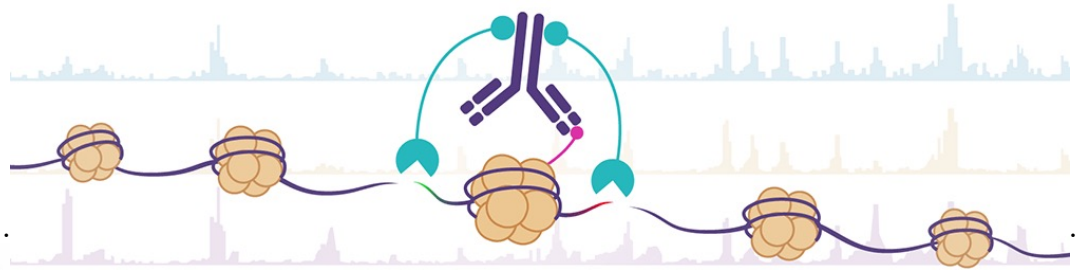
- Chromatin ImmunoPrecipitation:
  1. cross link
  2. sonication
  3. ChIP
  4. Remove protein



Shah, 2009

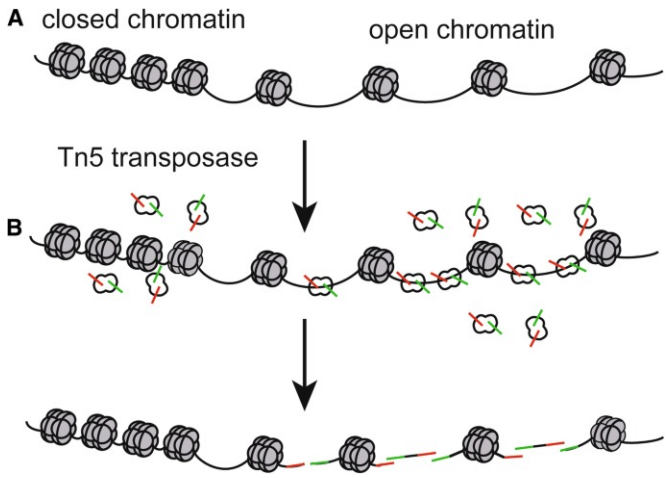
Cut-and-Tag

- Alternative to ChIP-seq
- Uses Tn5 enzyme instead of sonication
  - Cut where there isn't a nucleosome.

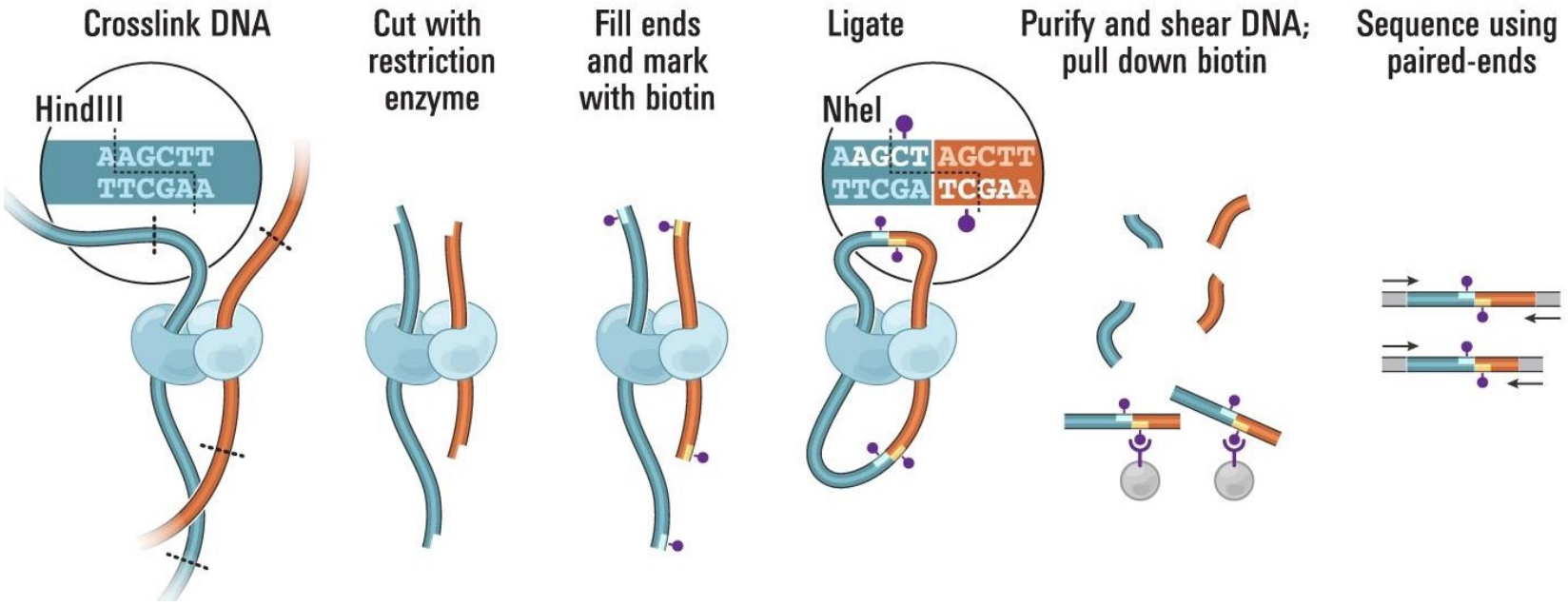


ATAC-seq

- Assay for Transposase Accessible Chromatin
- Sub-nucleosome (<150bp), mono-nucleosome (~150bp), di-nucleosome(300bp) reads.
- Nucleosome positioning, eviction etc is very important in epigenetics.
- cut-and-run: MNase instead of Tn5.
  - MMase: cut into a single nucleotide and start digesting them, until stopped by a nucleosome.



Hi-C



NATIONWIDE CHILDREN'S
When your child needs a hospital, everything matters.

Lieberman-Aiden et al. 2009

Variation

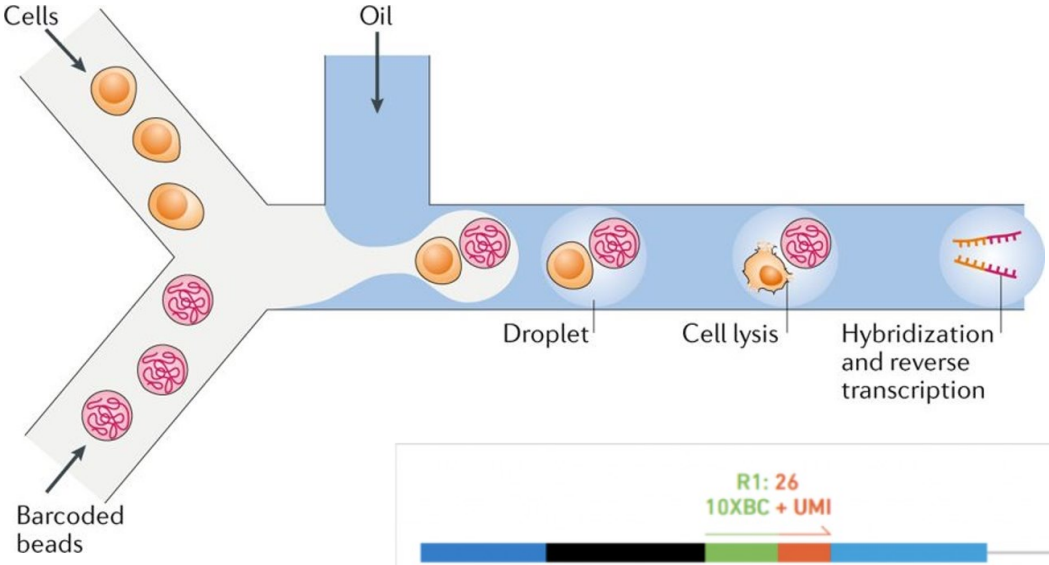
- Combine Hi-C + ChIP = Hi-ChIP
  - long-range interactions/loops
- Cut&Tag, Cut&Run
  - Alternative to ChIP and ATAC-seq

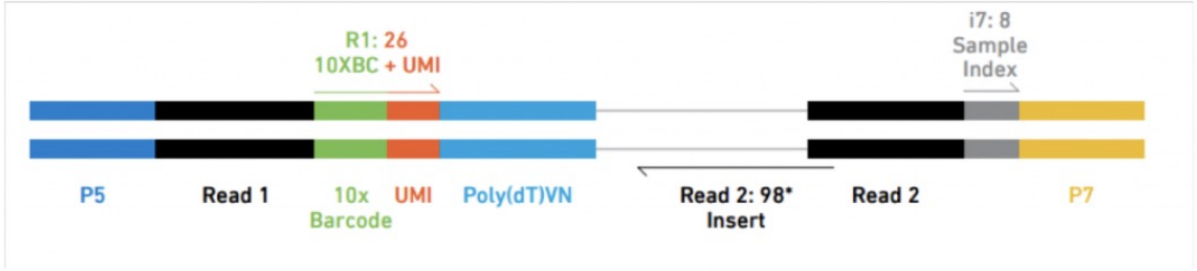
And many more…

Single Cell

- sc-RNA-seq
- sc-ChIP (not really)
- sc-ATAC-seq
- sc-Hi-C
- spatial transcriptomics

sc-RNA-seq (Droplet based)



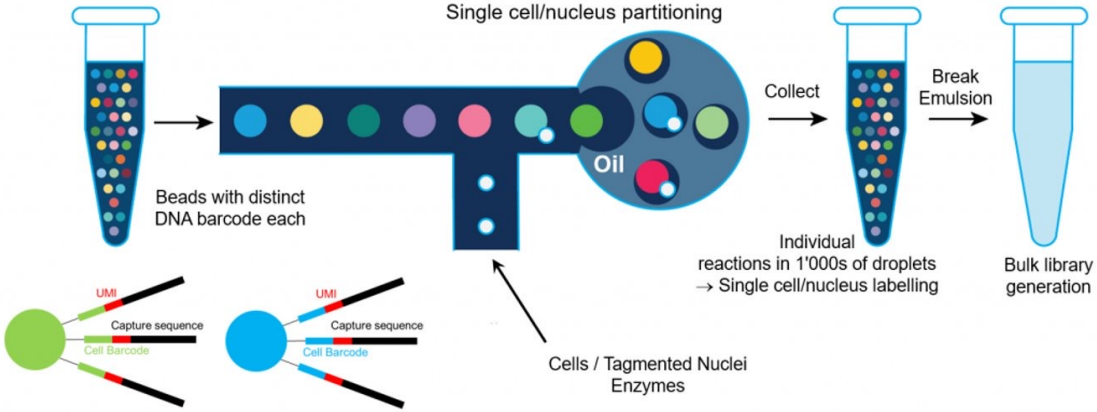


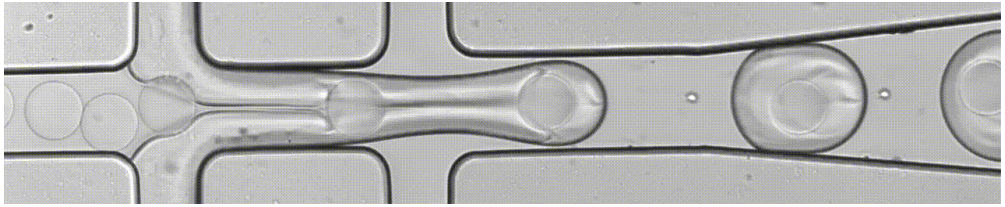
10X GENOMICS

Slides courtesy of Katie Miller

NATIONWIDE CHILDREN'S
When your child needs a hospital, everything matters.

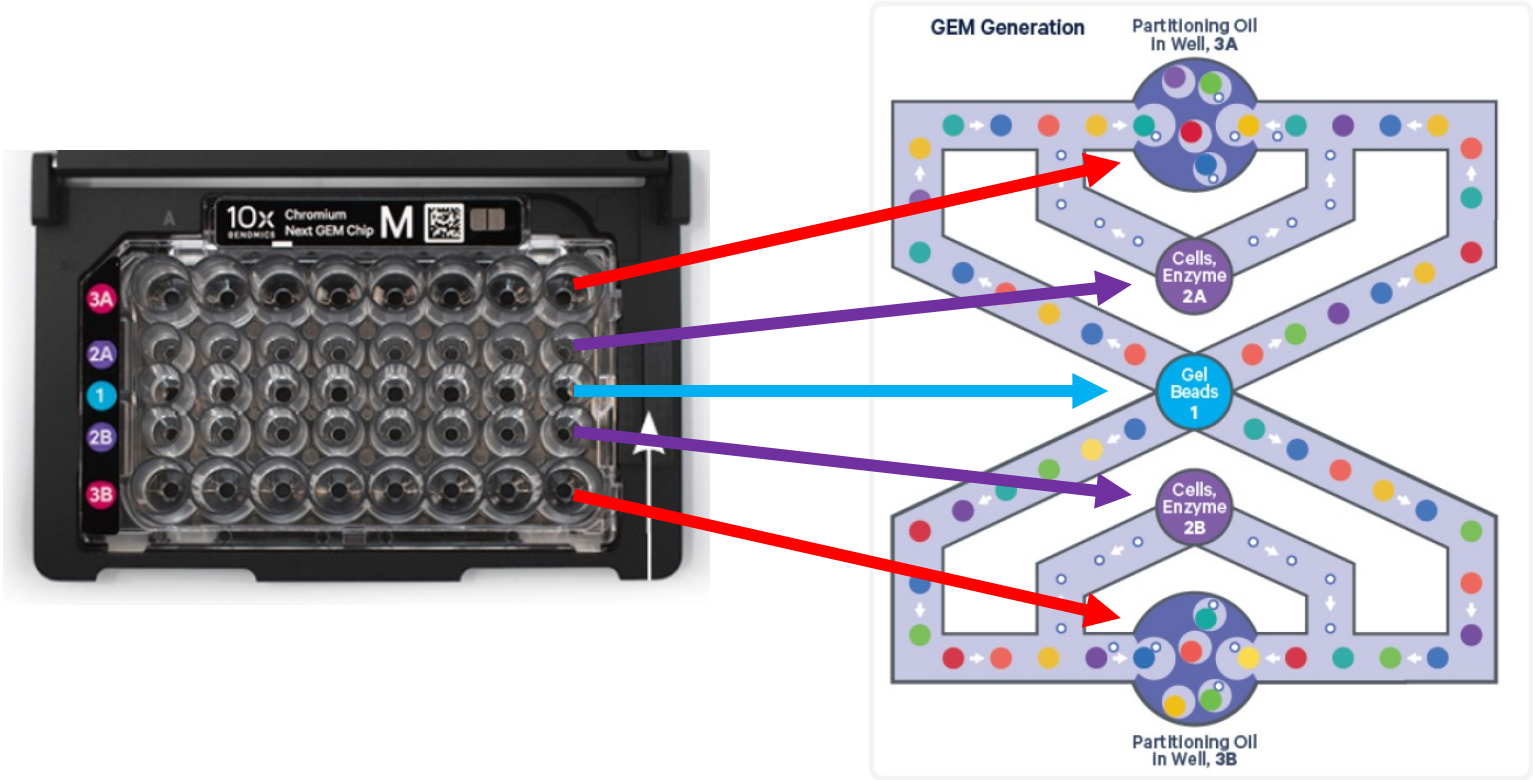
sc-RNA-seq (Droplet based)





Slides courtesy of Katie Miller

NATIONWIDE CHILDREN'S
When your child needs a hospital, everything matters.



Slides courtesy of Jason Navarro

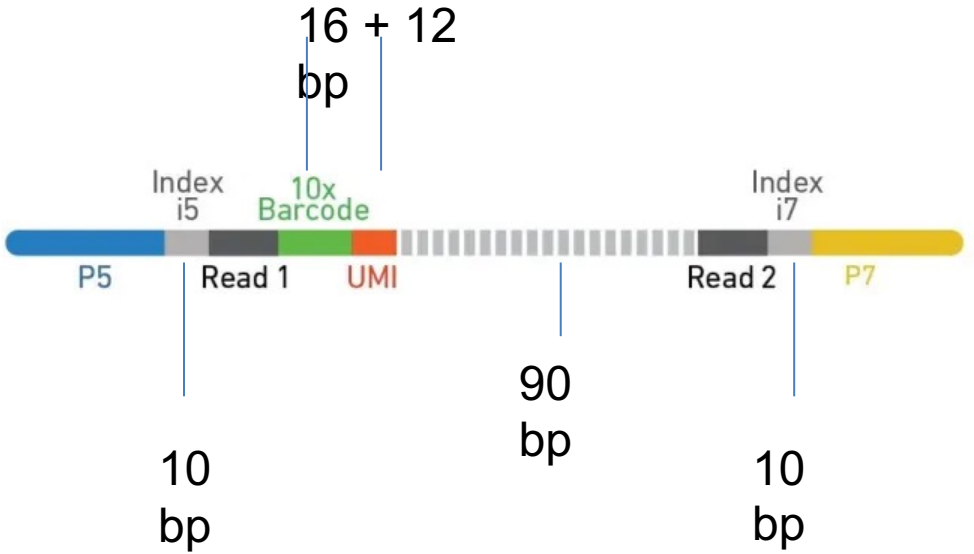
Chromium controller

Slides courtesy of Matt Cannon

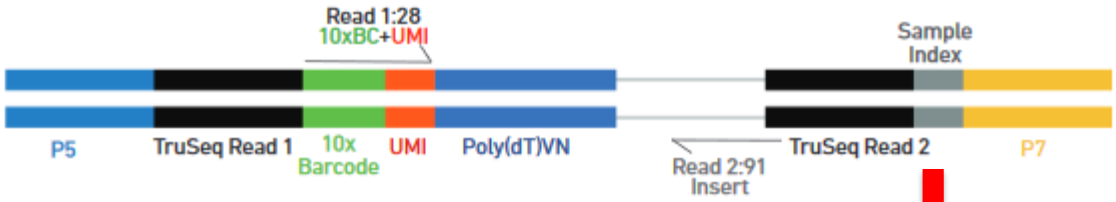
NATIONWIDE CHILDREN'S
When your child needs a hospital, everything matters.

Sequence data

"Run recipe" = 28 + 10 + 10 + 90
This gets sequenced on standard Illumina sequencer (about 150bp per read)



Slides courtesy of Katie Miller

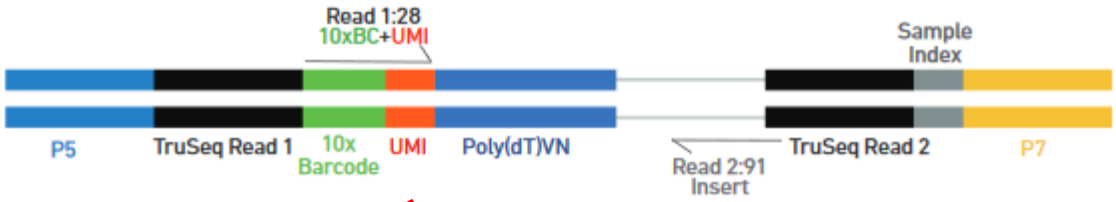


Sample index

```
@A00498:356:H53CMDRXY:1:2101:1371:1016 1:N:0:CATGCGAT
CATGCGAT
+
FF:FFFFF
@A00498:356:H53CMDRXY:1:2101:1443:1016 1:N:0:CATGCGAT
CATGCGAT
+
FFFFFFFF
@A00498:356:H53CMDRXY:1:2278:31349:37059 1:N:0:ACCCGACG
ACCCGACG
+
FFFFFFFF
```

Slides courtesy of Katie Miller

NATIONWIDE CHILDREN'S
When your child needs a hospital, everything matters.



Read 1

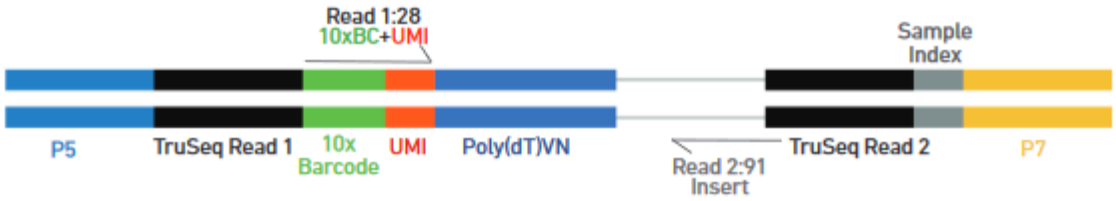
```
@A00498:356:H53CMDRXY:1:2101:4390:1266 1:N:0:CATGCGAT
AATAGAGAGTCTGTACTTTGACAACCGT
+
FFFFFFFFFFFFFFFFFFFFFFFFFFF
@A00498:356:H53CMDRXY:1:2101:15130:1141 1:N:0:CATGCGAT
ATCTCTATCCAACTGACAGTTAACTGGT
+
:FFFFFFF,FFFFFFFFFFFFFFFFFF
@A00498:356:H53CMDRXY:1:2101:15167:1141 1:N:0:CATGCGAT
GATGGAGTCGTGGACCACTCGGGCAGCC
+
FFFFFFFFFFFF:FFFF:FFFFFFFFF
@A00498:356:H53CMDRXY:1:2101:15528:1141 1:N:0:CATGCGAT
TCTACATAGTCTACCATCCAAACGAACT
+
FFFFFFFFFFFFFFFFFFFFFFFFFFF
```

10X barcode (cell)

Unique molecular identifier (UMI)

Slides courtesy of Katie Miller

NATIONWIDE CHILDREN'S
When your child needs a hospital, everything matters.



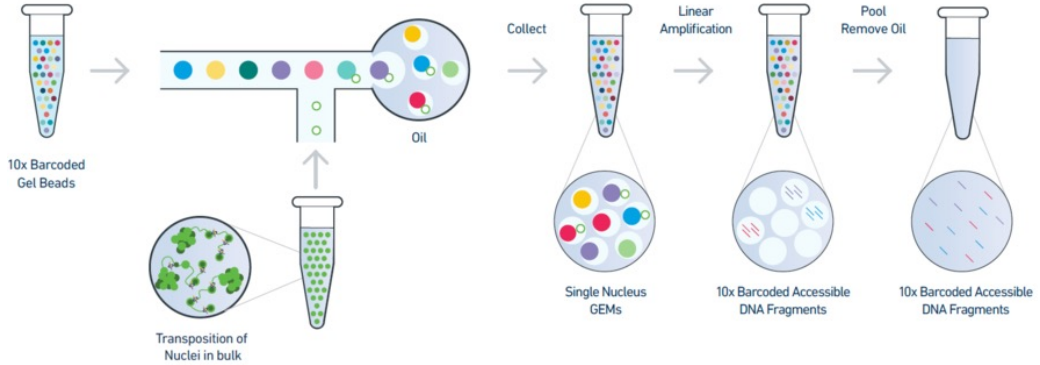
Read 2 = insert = transcript of interest

```
@A00498:356:H53CMDRXY:1:2101:28221:1204 2:N:0:CATGCGAT
AGATGATCTGTTCAGTGCTAAGCGTAGGATGTTGAAGTCCCCAACTGTTATTGTGTTGGAGTCTATTTGTCTCTTTAGGTTTAATAATATT
+
FFF:FFFFFFFF:FFFFF:FFF,F,FFFF,F,:FFFFFFFFFFFFF:FFF,FFFFFFFFFF:,FFFF,FF:FFFFFFFFFFFFFF,FFFF
@A00498:356:H53CMDRXY:1:2101:28583:1204 2:N:0:CATGCGAT
ATGCCCTAGCCCACTTCTTACCACAAGGCACACCTACACCCATTATCCCCATACTACTTATAATCGAAACCATCAGACTACACATTCAACC
+
F::FFF,F,FFFFFFFF,,FF:FFFFFF,FFF:FFF,FFFF,:FF::,F:FF,::F,,F,,,F:FFFFF:FFFFF,,FFFF,:F,FF:FF,
@A00498:356:H53CMDRXY:1:2101:28673:1204 2:N:0:CATGCGAT
GTGAAGAGGATCTTGAATTTCTTAATGCATCCAAAGCCTTCTGGCAAACCATTGCCGAAATCTAAAAAAACTTGTAGCAAAGGCAGTAAA
+
FFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF:FFFF,FFFFFFFFFFF:FFF:FFFFFFFFFFF:FFFFFFFFFFFF
@A00498:356:H53CMDRXY:1:2101:28709:1204 2:N:0:CATGCGAT
CGCGAGGTGGGGGCGTCGTGTAAGCAGCGGAGGATGGGGGGGCGGTGCACGTGGGTGGGCGTGGCTGAGATCTAAGTGTCCTGCAGCTGTG
+
FFFFFFFFFFFFFFFFFFFFF:FFFFFF:FFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF
```

Slides courtesy of Katie Miller

scATAC-seq

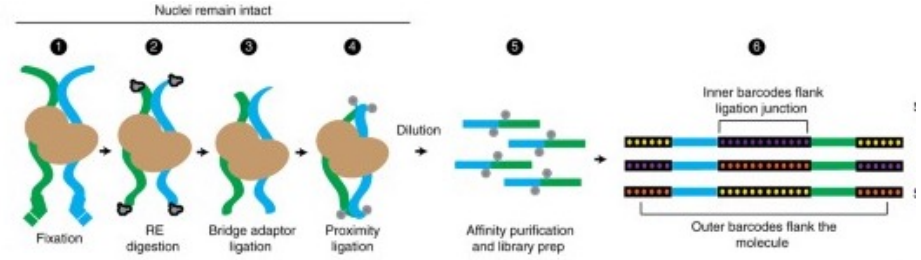
- Droplet based.
- tn5 in oil.



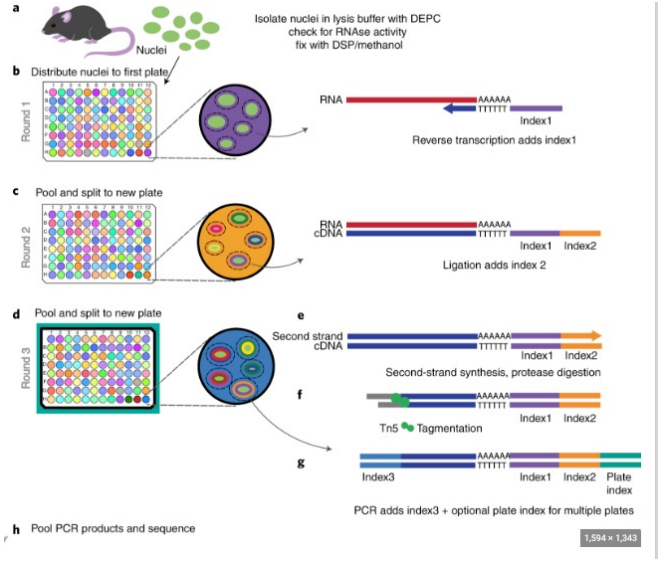
scChIP-seq

- Droplet based
- Uses cut-and-tag (oil contain tn5+antibody)
- Work somewhat decent for K27ac (easiest to ChIP)

sc-Hi-C



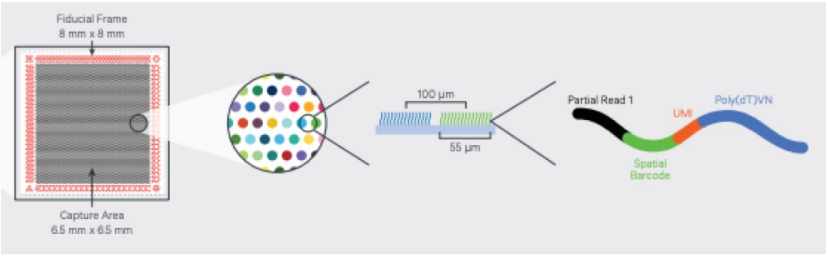
Proximity ligation with nuclei intact

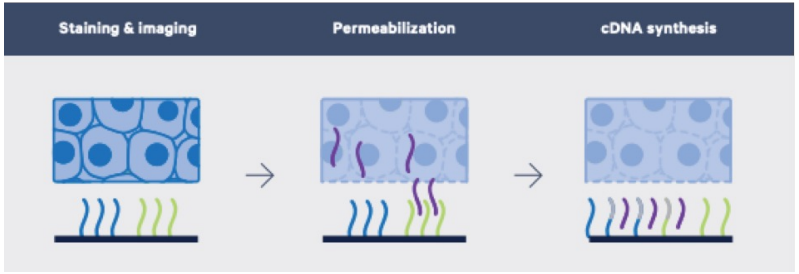


Combinatorial Indexing

Spatial Transcriptomics

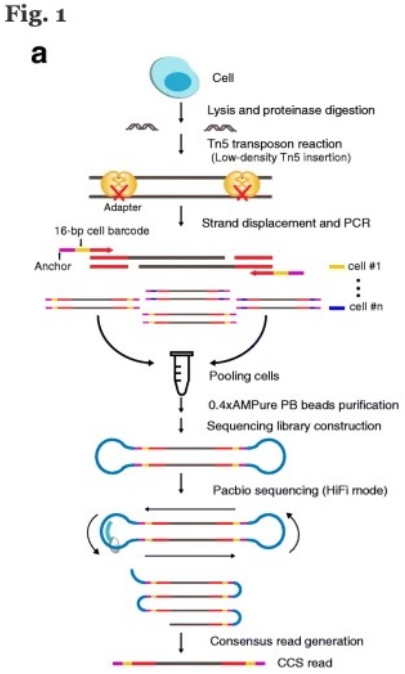
- Each microscope slide includes ~5000 “spots”.
- Each spot is about ~55µm (1-10 cells).
- mRNAs released from cell,
- mRNAs bind to spatially barcoded oligos,
- RT to produce cDNA.
- Not exactly single cell.





sc-PacBio

sc-WGS and sc-RNA have both been done and reported on PacBio HiFi platform.



Costs

- Bulk: ~$200-$500 per sample (including reagents but not time cost)
- Hi-C is more expensive due to sequencing depth.
- Pacbio $3000 per sample for bulk.
- sc-RNA-seq could cost over $5000 per sample.

In the next 5 years…according to Katie

- Freedom from <expensive> hardware (e.g., Parse and Fluent Biosciences)
- Decreasing sequencing costs
  - …therefore increase number of cells to be analyzed
- Enhanced/more-sensitive mRNA capture rates
- More user-friendly computational pipelines
- Multi-omics approaches
  - Proteomics
  - CRISPR
  - Non-coding RNA
  - Epigenomics
  - Long-reads
- Clinical applications

Special Thanks to…

Ben Stanton
PI, CCCR

Ryan Roberts,
PI, CCCR

Matt Cannon
bioinformatician

Katherine Miller
PI, IGM

Anthony Miller
Director of TechDev, IGM

NATIONWIDE CHILDREN'S®
When your child needs a hospital, everything matters.