TIGP-BIODIV Lecture, 5/20/2020

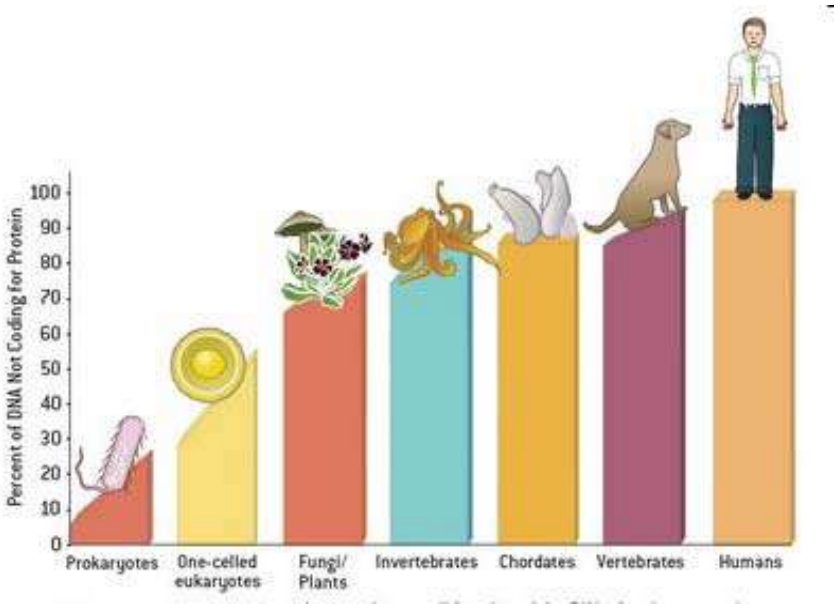
NGS: DNA/RNA preparation & different sequencing technologies

Mei-Yeh Jade Lu 呂美曄

Associate Research Specialist
High Throughput Genomics Core Manager
Biodiversity Research Center
Academia Sinica

HT Genomics Core in Academia Sinica

- 2008:Established for biofuel project
- 2013: Promoted as NGS service core on campus
- Internal & collaborative projects
  - Pathogenic bacteria
  - Pathogenic/medicinal fungi
  - worms
  - Insects
  - Evo-devo: avian species
  - C3/C4 plants
  - Marine animals
  - Metagenomes

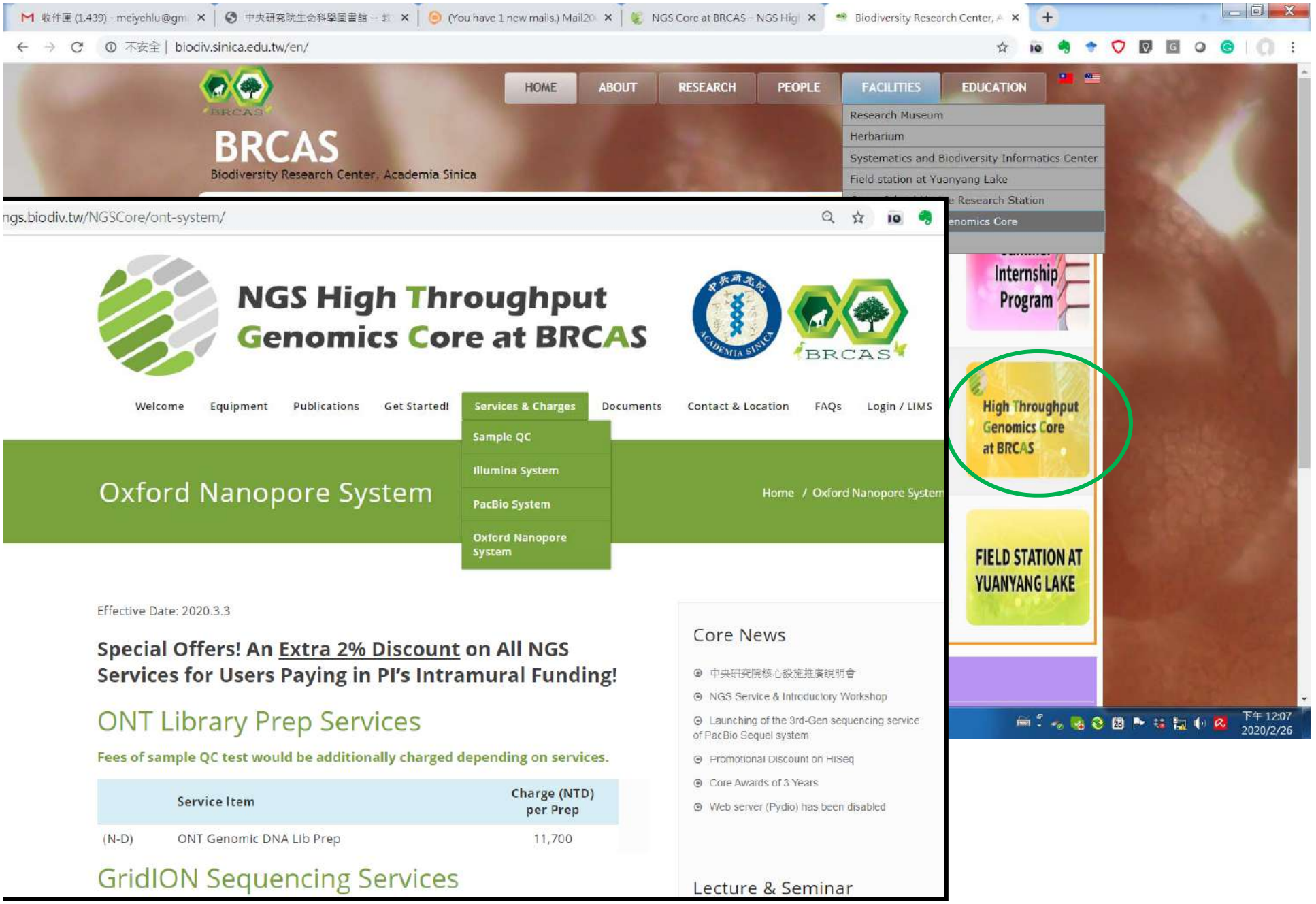


http://www.ashg.org/education/everyone_1.shtml

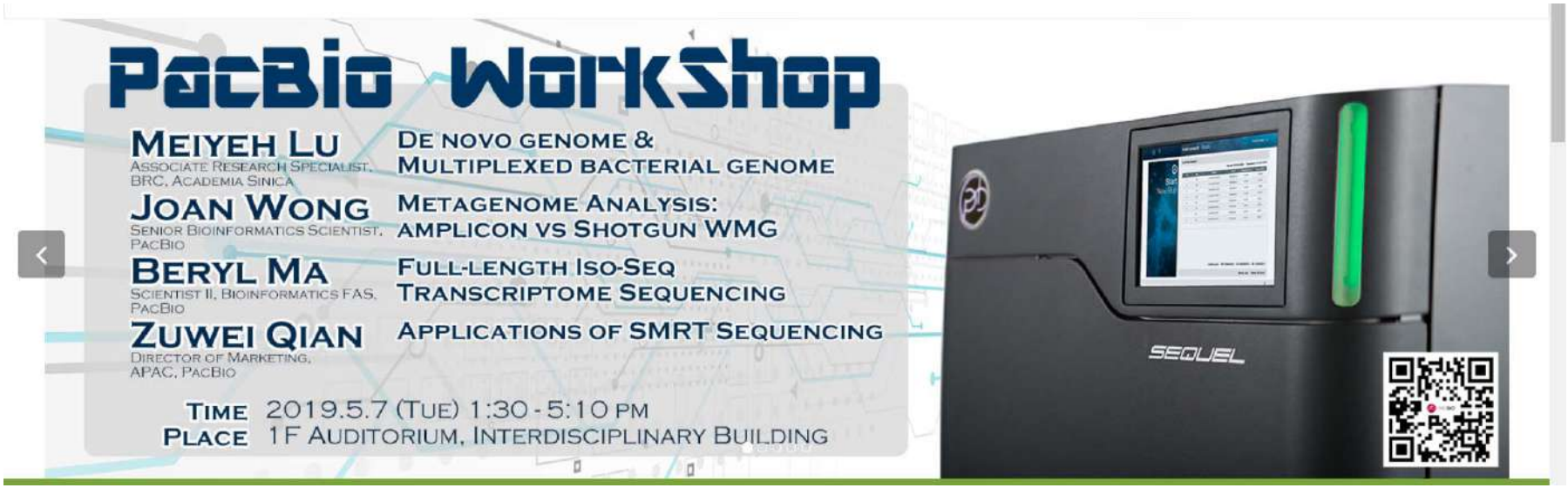
NGS Service at BRC core

- Current NGS lineup:
  - 3 Illumina (HiSeq2500 *2, MiSeq)
  - Roche 454 GS+
  - PacBio Sequel
  - Oxford NanoPore GridION *(new!!)*
- **SOPs:** established various NGS applications for NGS platforms
- Provides consultation on:
  - Project's need
  - suitable NGS experiment design
  - Sample preparation
  - Cost analysis

Where to Find Us?



Seminars /Workshops for advanced NGS Technologies

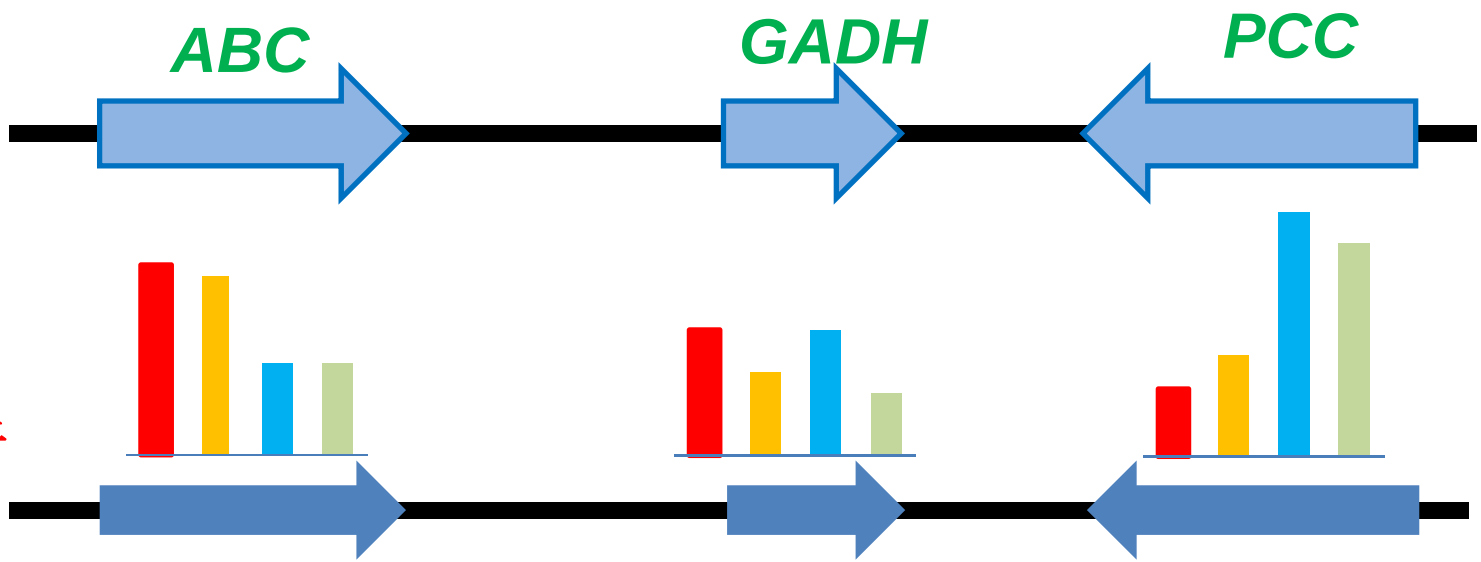


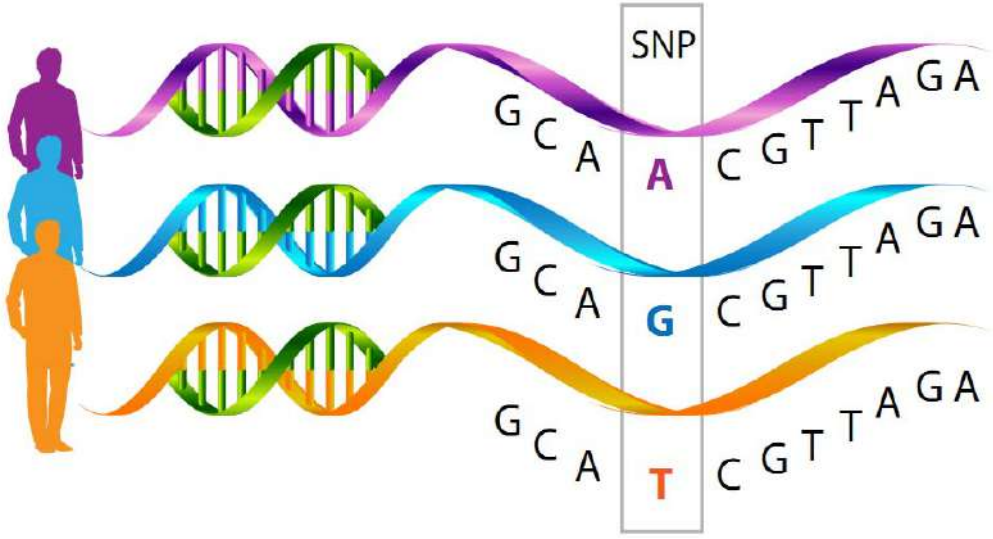
Outlines

1. Evolution of sequencing technologies
2. NGS platforms and comparisons
3. Project considerations & Sequencing plan
4. Good lab practice for NGS
5. Sample & library QC
6. Data types, preprocessing, and quality ctrl
7. Extended / Advanced NGS technologies

What can we learn from genome?

1. 基因體組序
2. 基因預測
3. 功能性註解
4. 基因表現量分析
5. 基因變異分析
基因調控 .....





I. Evolution of Sequencing Technologies

from Sanger to Next-Gen Seq.

Sanger:
ABI 3730

Single tube,
Di-deoxy termination

Roche 454

Illumina

Clonal Amplification
For signal enhancement

Ion Proton

S5

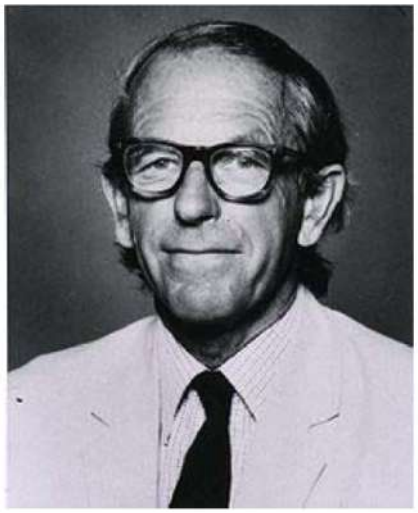
PacBio

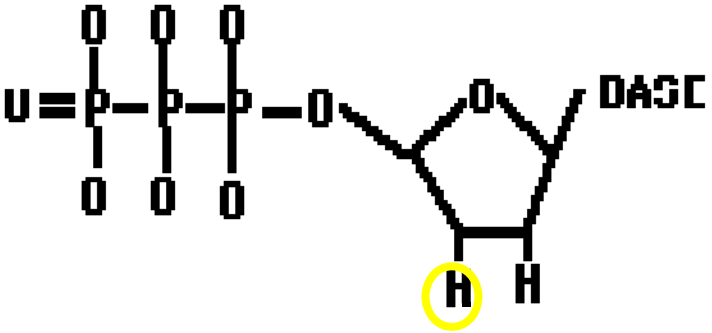
Oxford NANOPORE Technologies

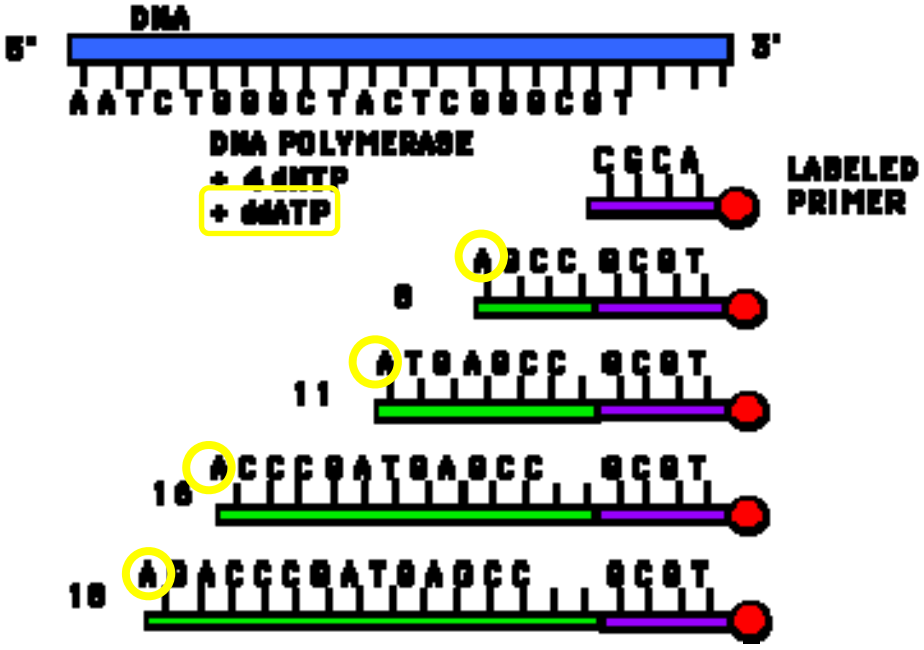
Single molecule sequencing

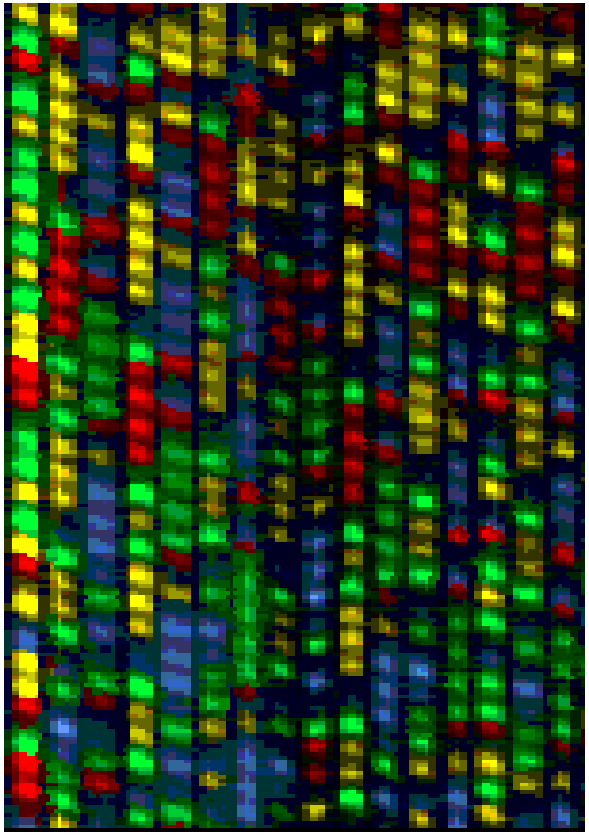
Sanger Seq. – dideoxy nucleotide termination

Frederick Sanger



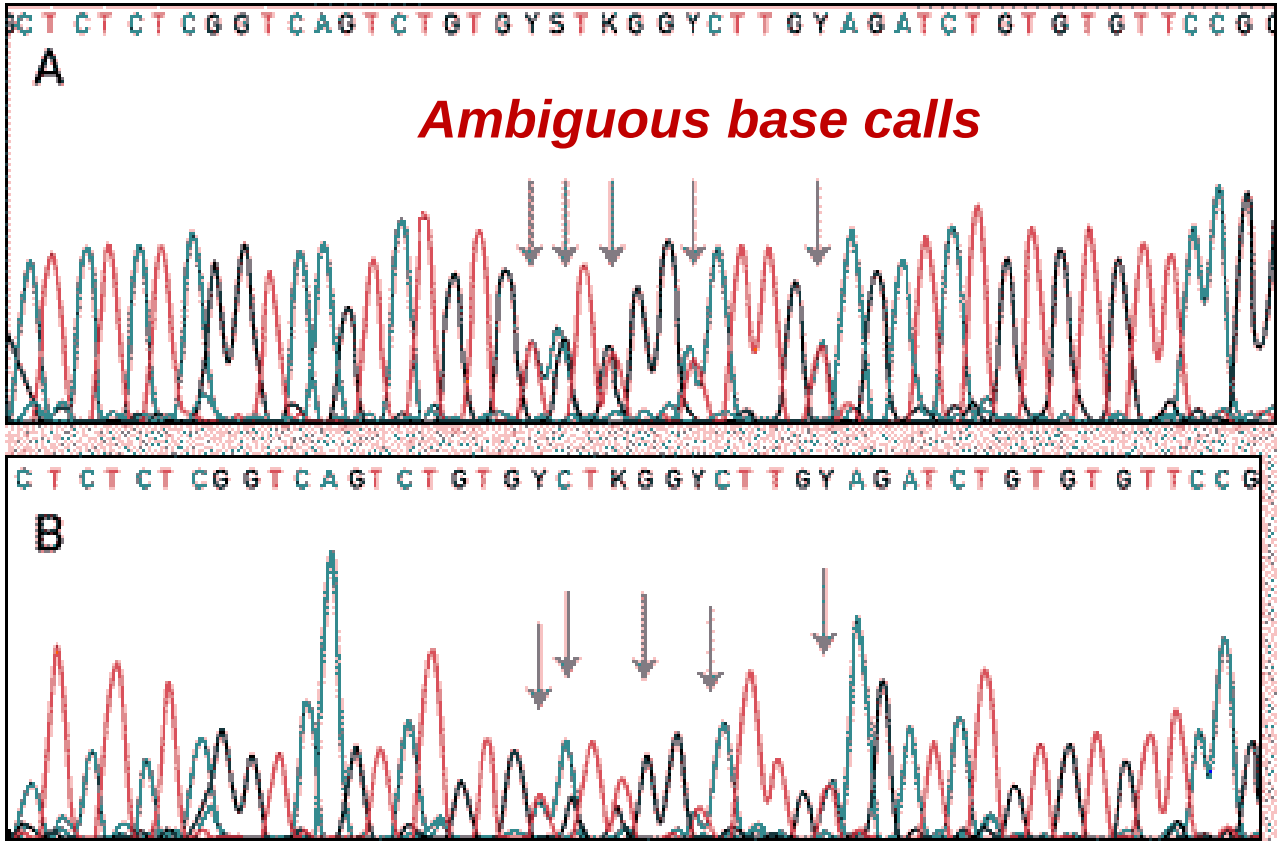


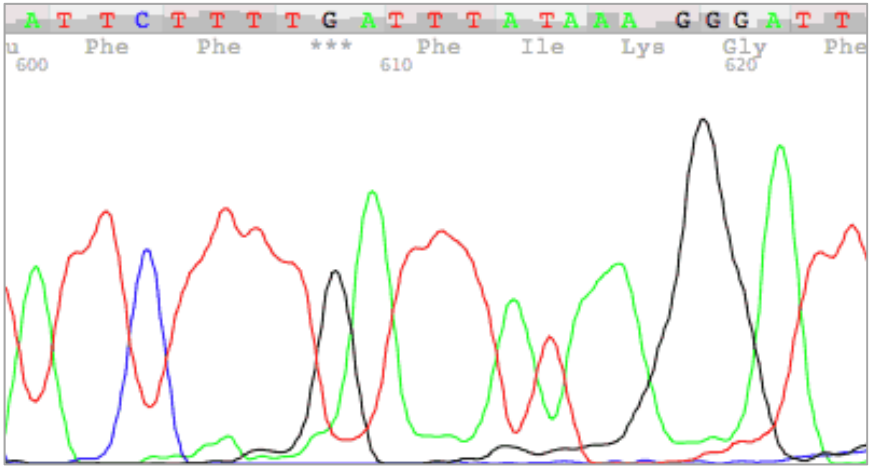




http://www.bio.davidson.edu/Bio111/seq.html

Fluorescent Dye-Terminator Cycle Sequencing





Blurry trace: Homopolymer

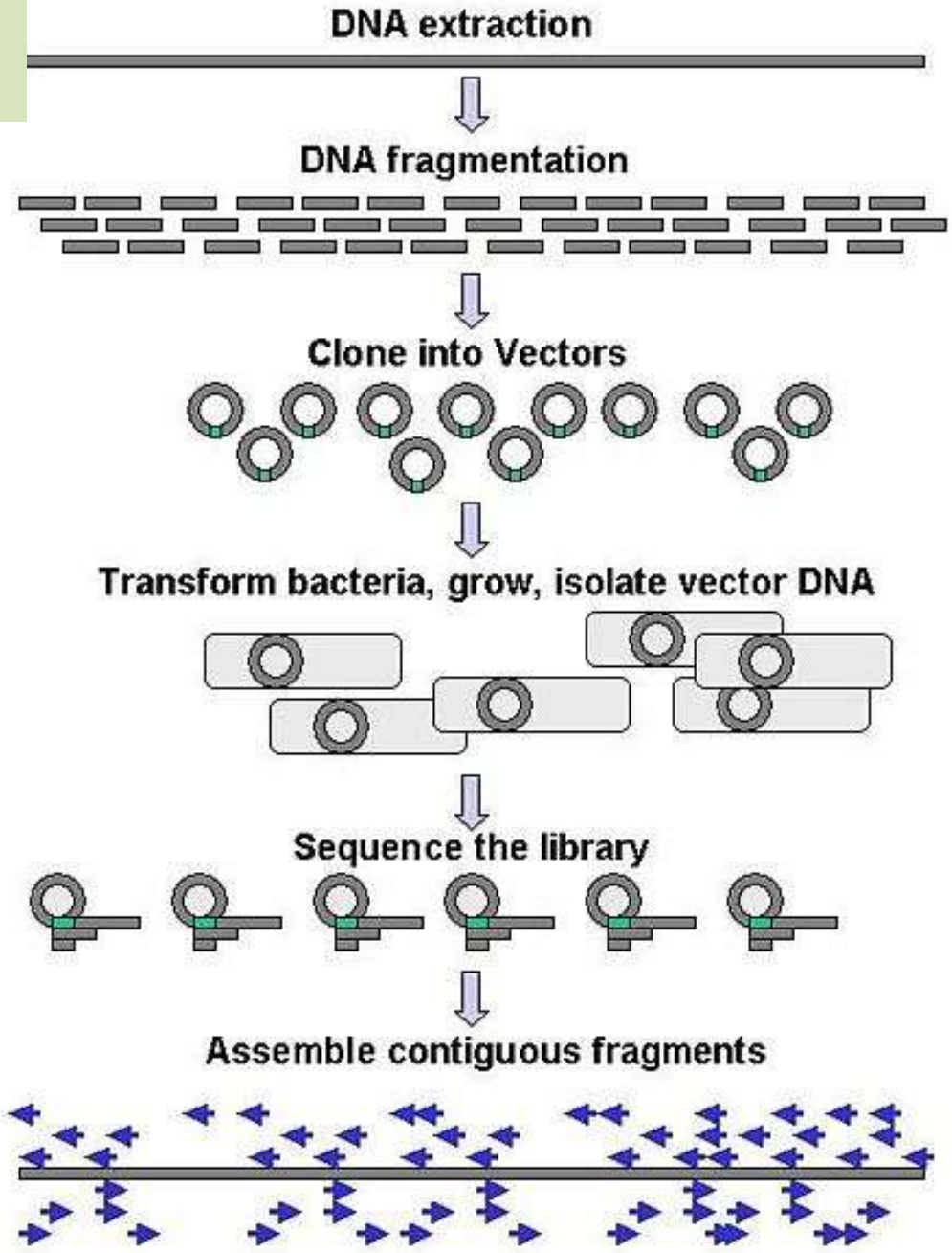
Genome Sequencing: Hierarchical cloning

BAC: 100-200kb

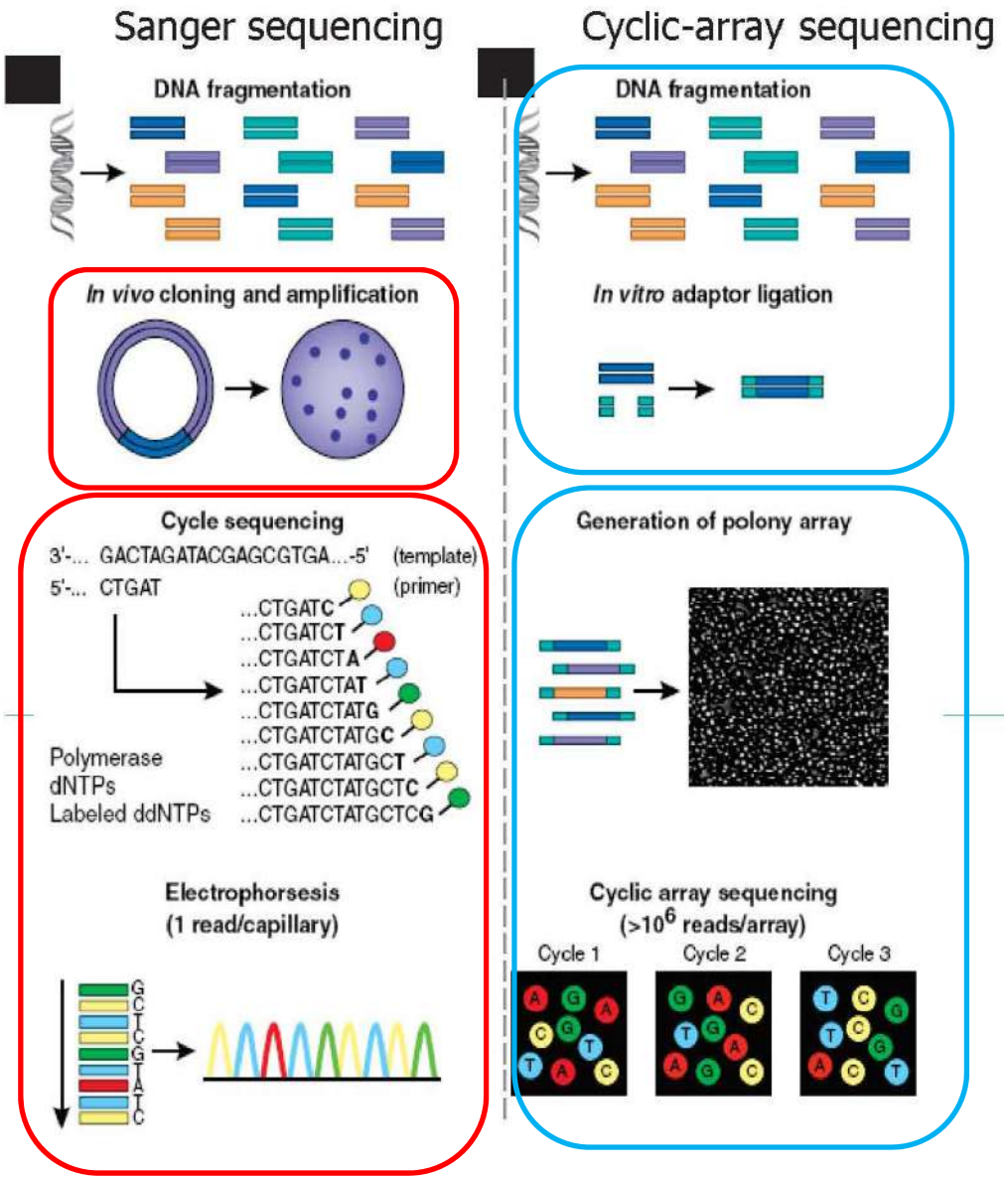
Cosmid

Fosmid: 30-40kb

Plasmid: 1-10kb



Next-generation DNA sequencing



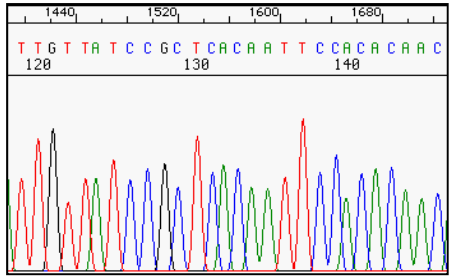
Advantages:

- adaptor-mediated library construction
- Clonal amplification to enhance signal intensity
- No bacterial cloning, colony picking, chr. Walking
- Array-based sequencing
- Massive parallel sequencing
- Much cheaper per *output unit*

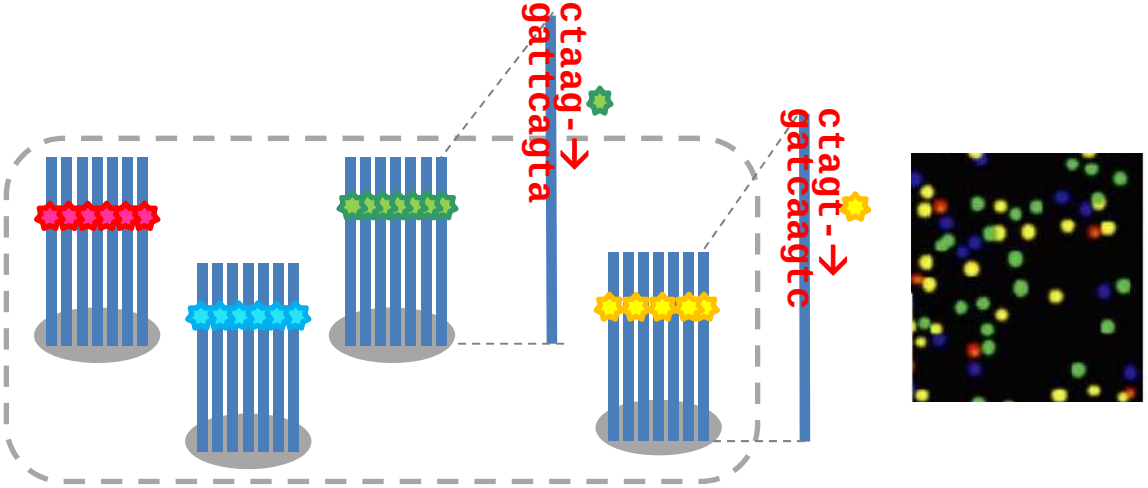
Evolution of Sequencing Technologies

Sanger
1 read/tube

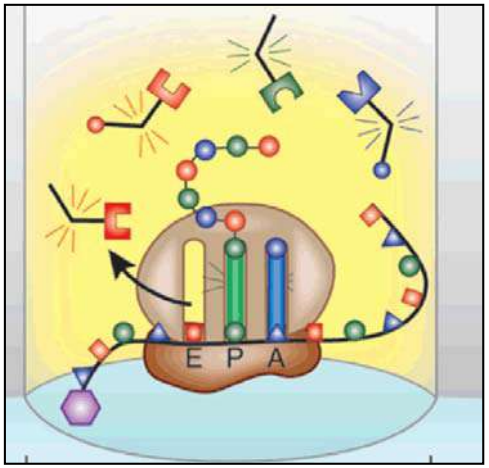
```
gctagttgaccttgaccaagcatggcgatcgat
|||||||
cgatca--->
```

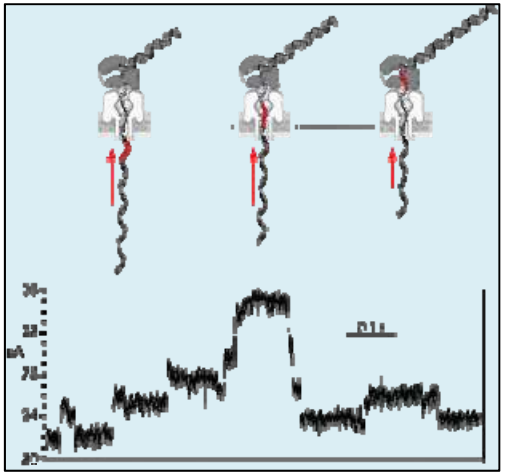


2nd-Gen
Clonal amplification



3rd-Gen
Single mol. Seq.





NGS – massive parallel sequencing

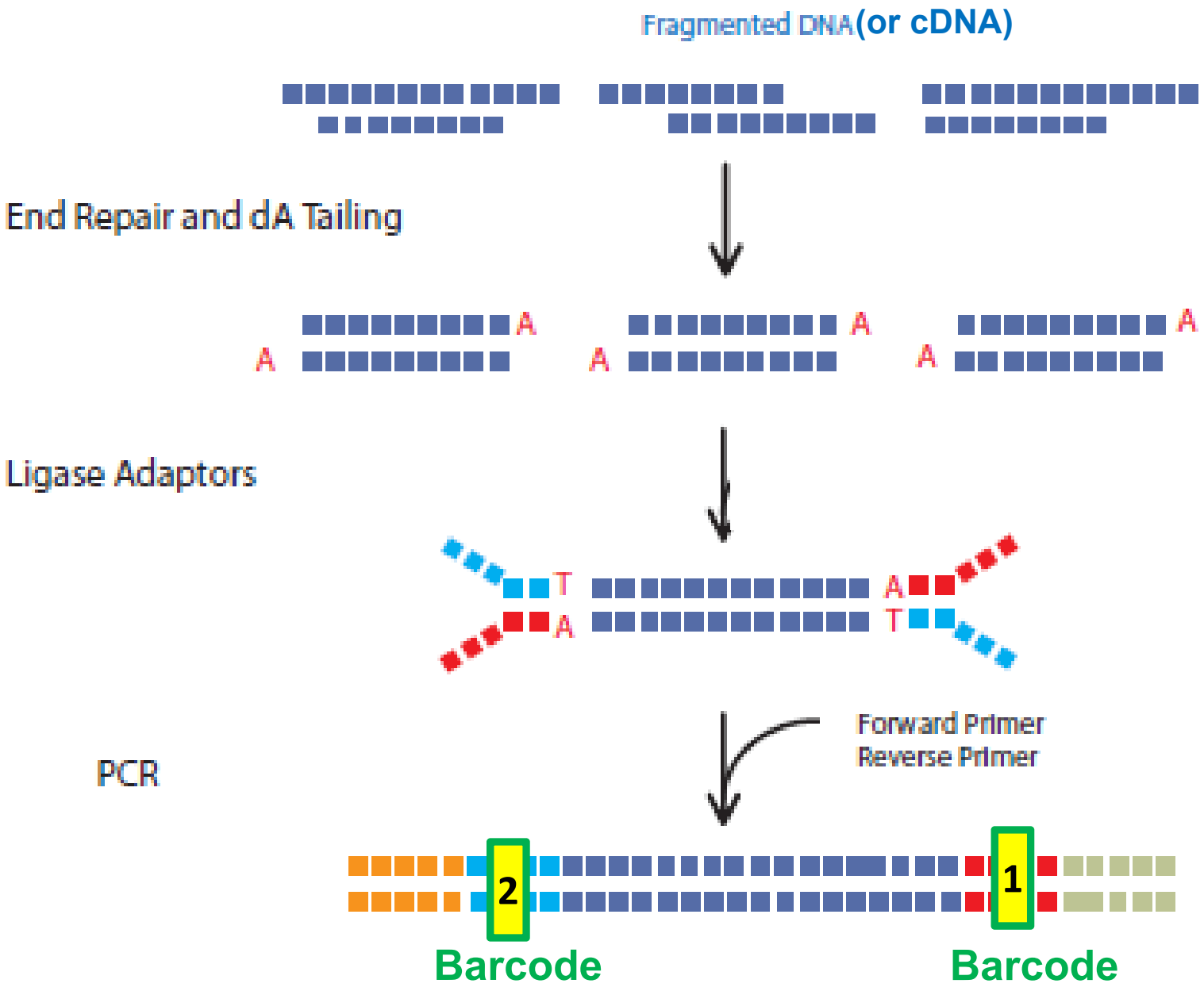
Current Popular platforms:

- **2nd-Gen: clonal amplification**
  - Roche 454: GS FLX, , 454 Jr., 454 XL+, 454 Jr.
  - Illumina: GA, Miseq, HiSeq, NovaSeq
  - Life Technologies: SOLiD, Ion Torrent, Ion Proton

- **3rd-Gen: single molecule sequencing**
  - Pacific Biosciences: PacBio RS II, Sequel
  - Oxford Nanopore Technologies

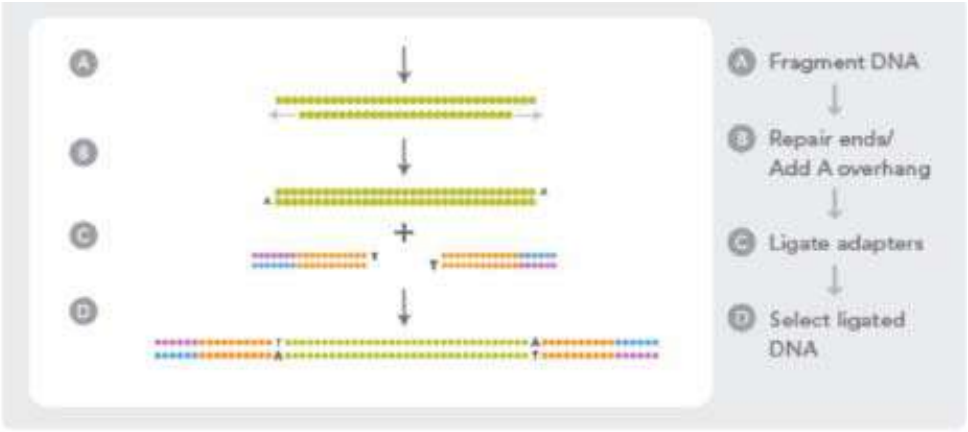
II. NGS platforms and comparisons

NGS Library Preparation Workflow

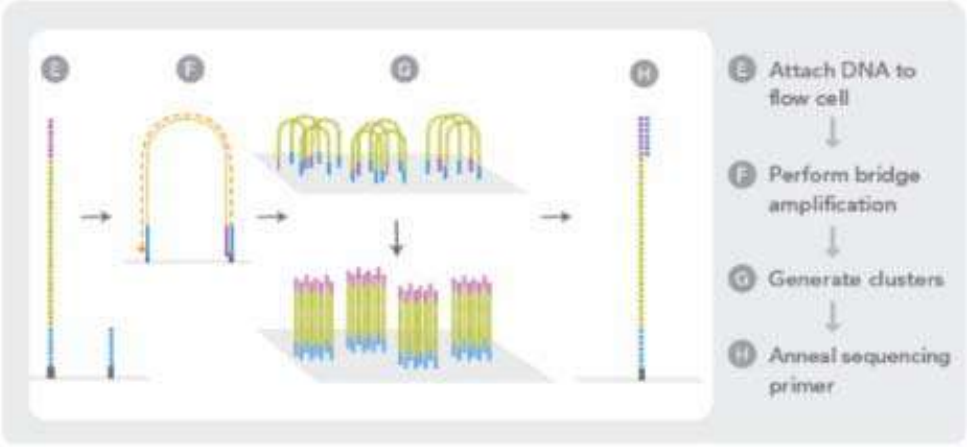


Illumina/Solexa: Cyclic Reversible Terminator

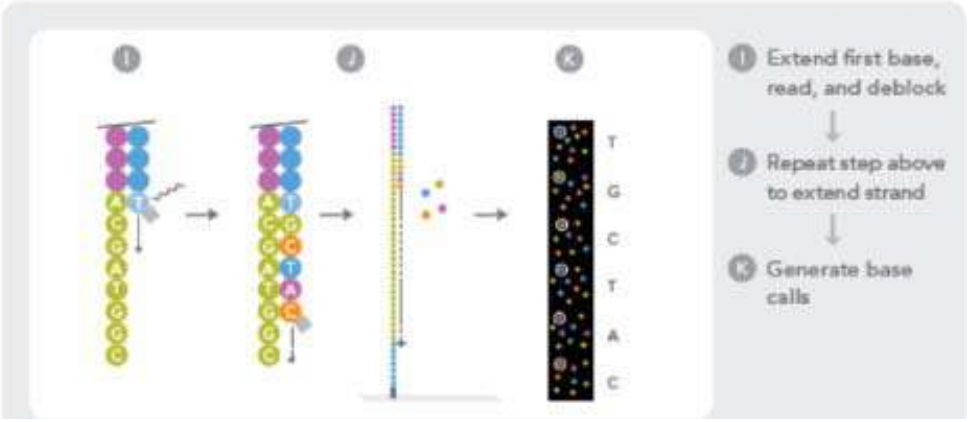
I.



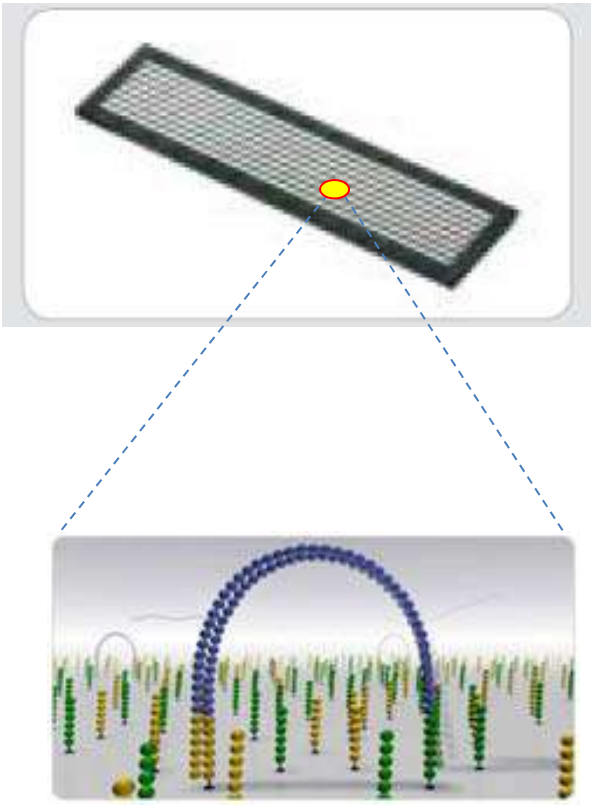
II.



III.



Flow Cell



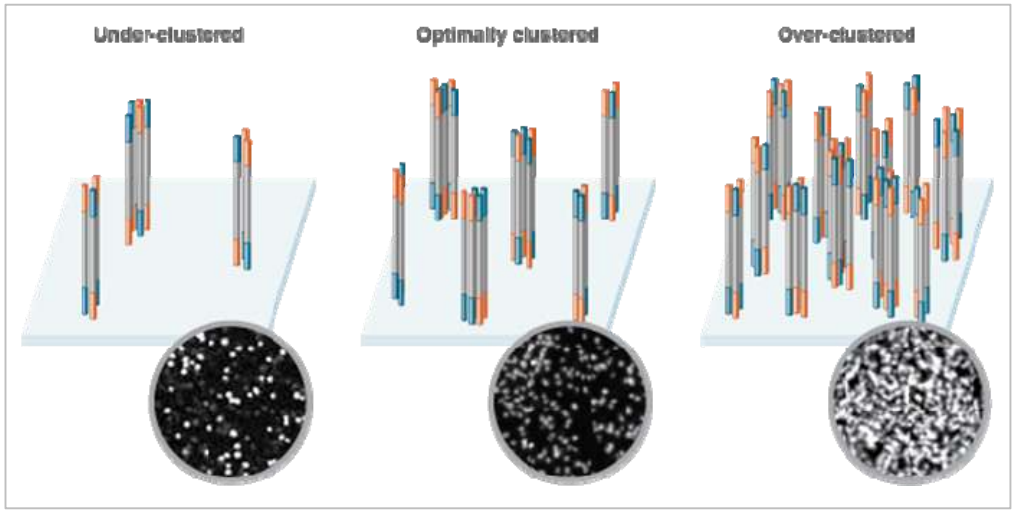
Illumina – Flow cell imaging

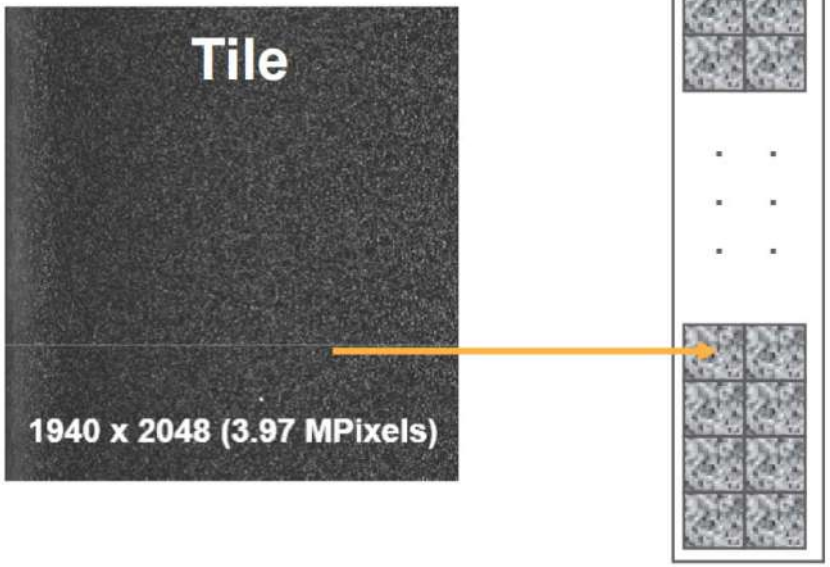
GA IIx

HiSeq 2500
(HT*8 / Rapid*2)

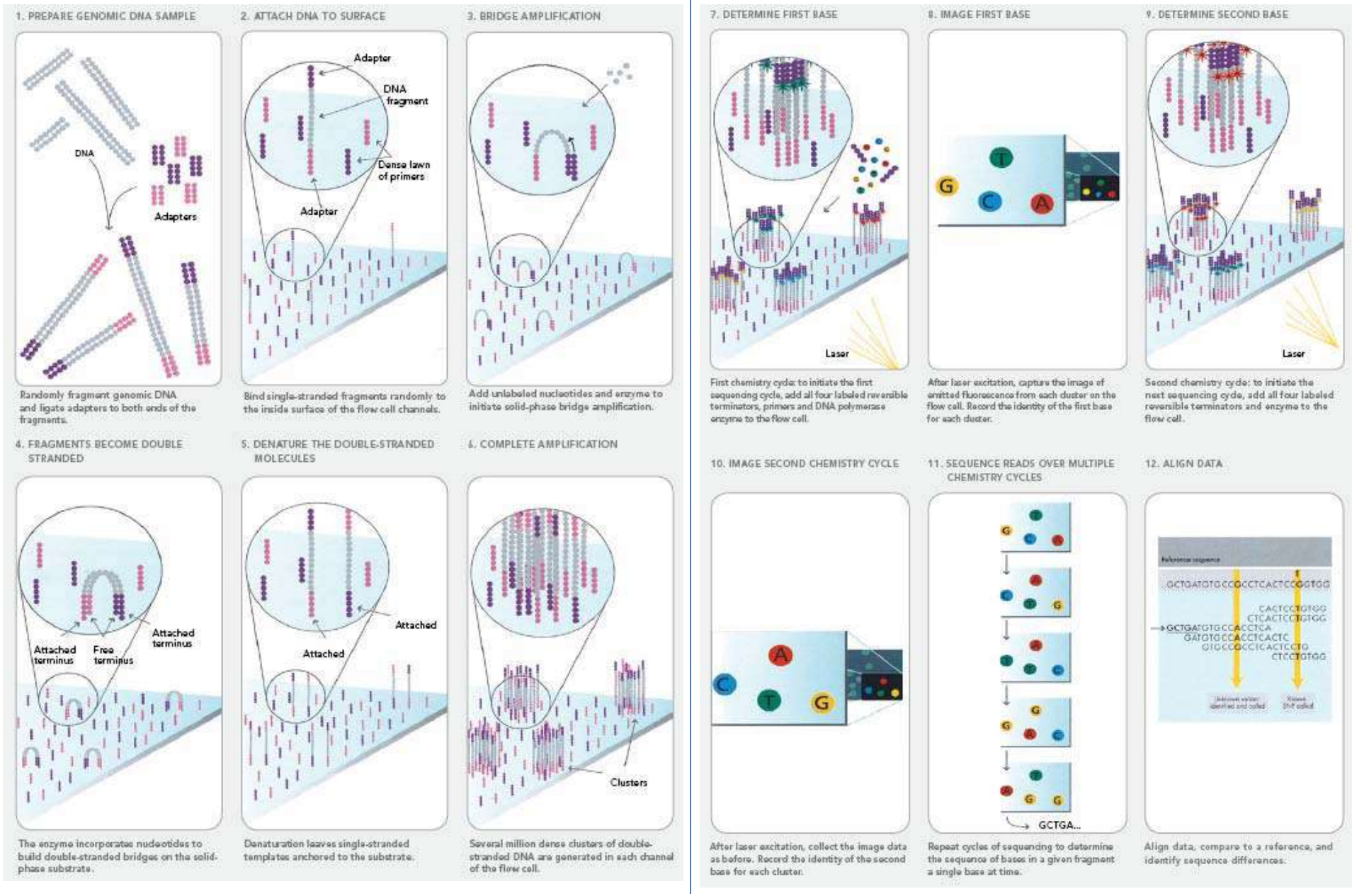
NextSeq 500

MiSeq v2

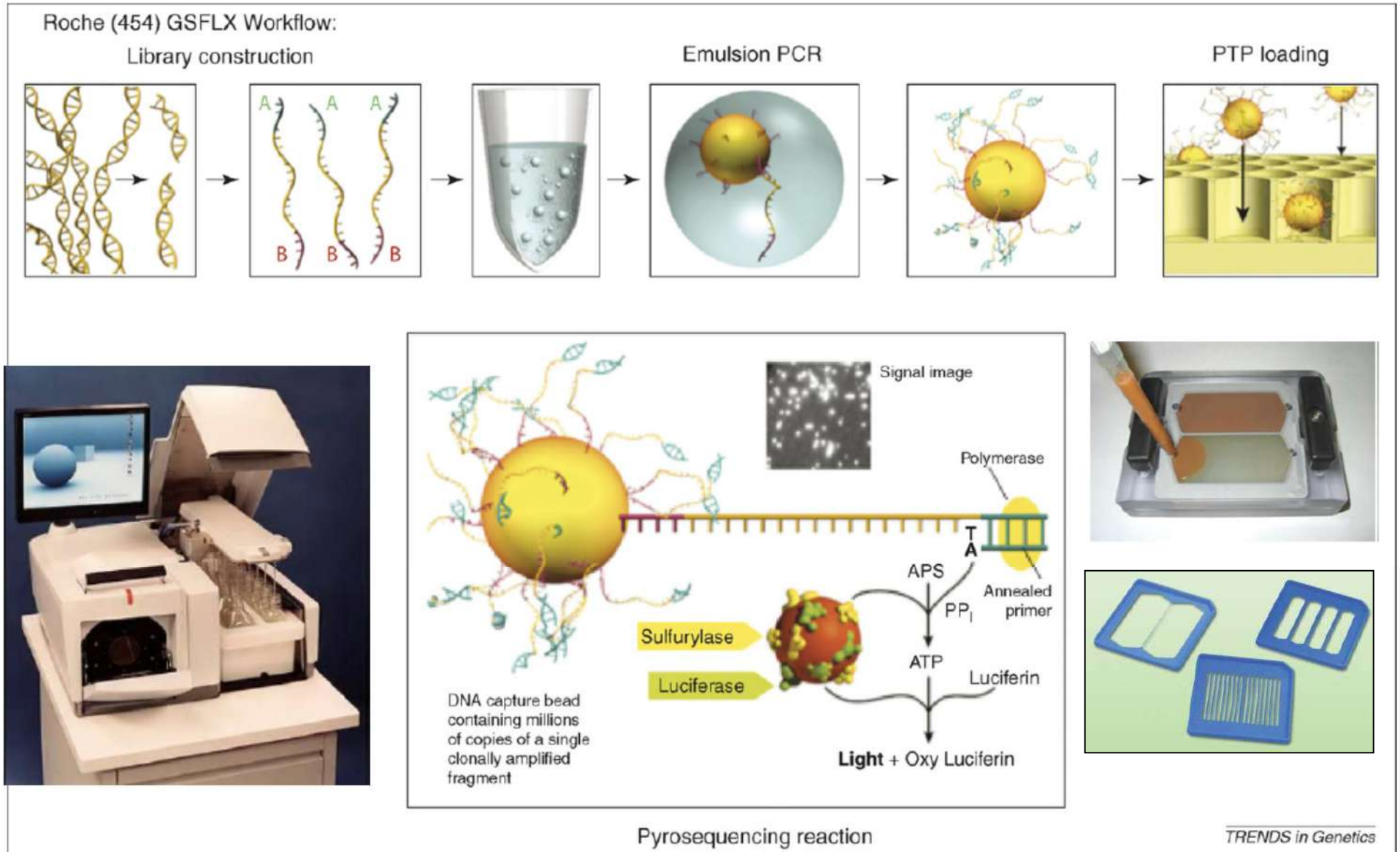




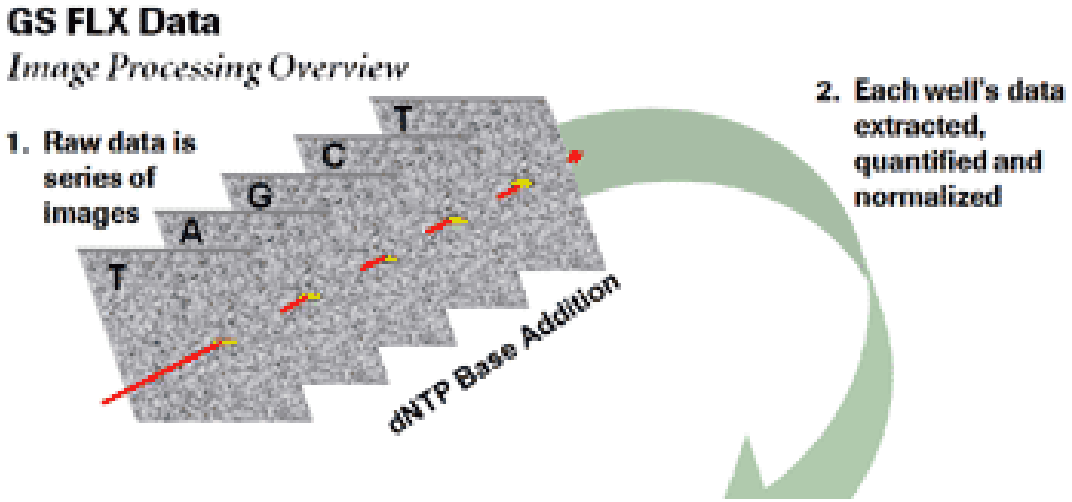
Illumina/Solexa: Cyclic Reversible Terminator

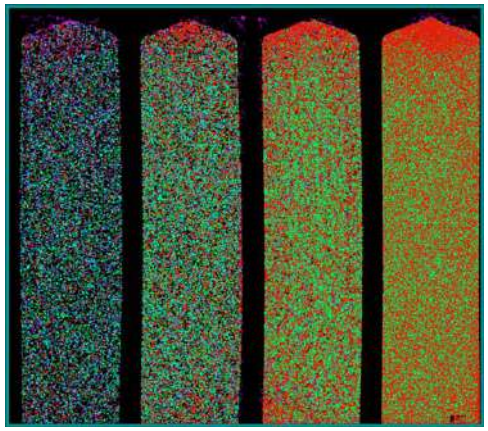


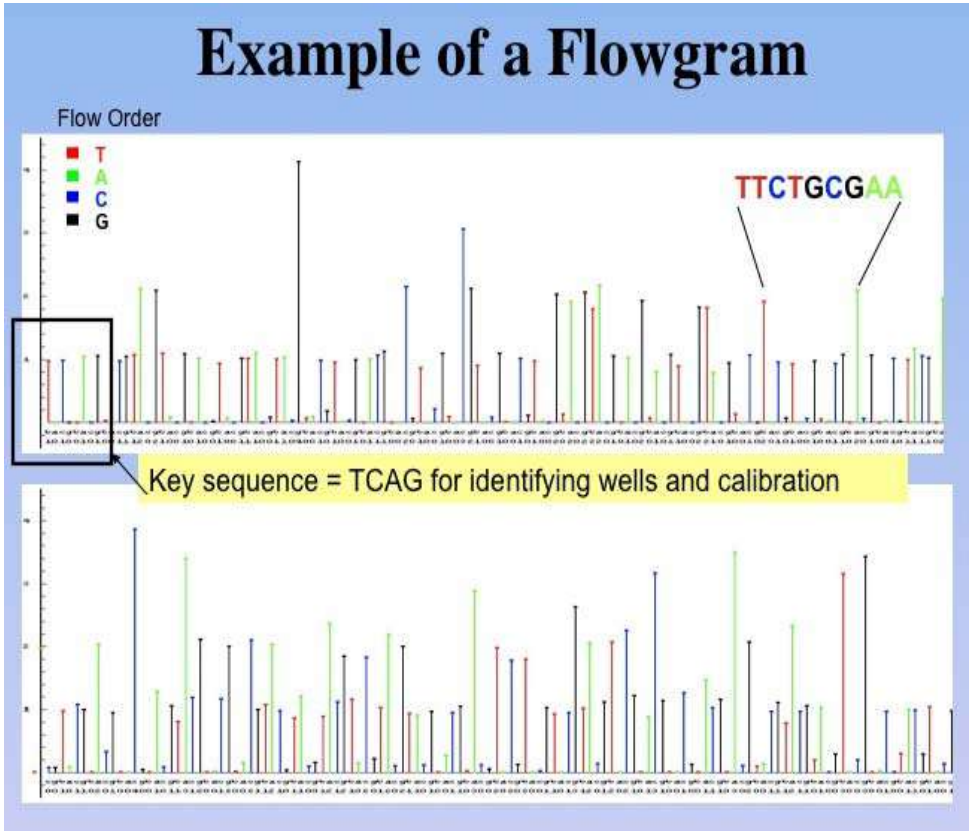
454: emPCR & pyrosequencing

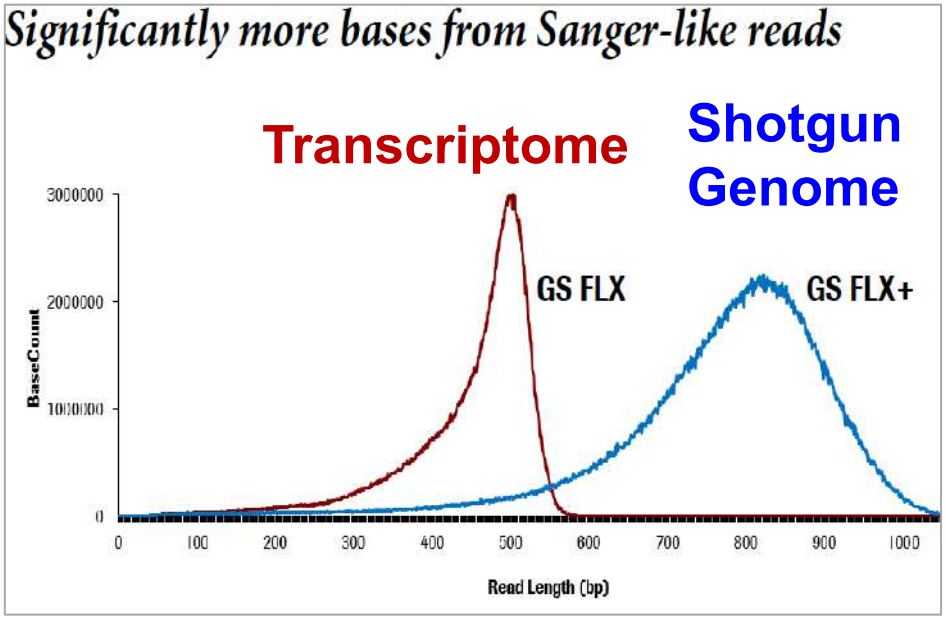


454 flowgram and read length profile

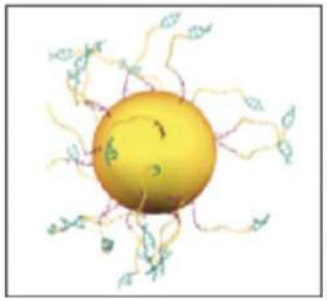






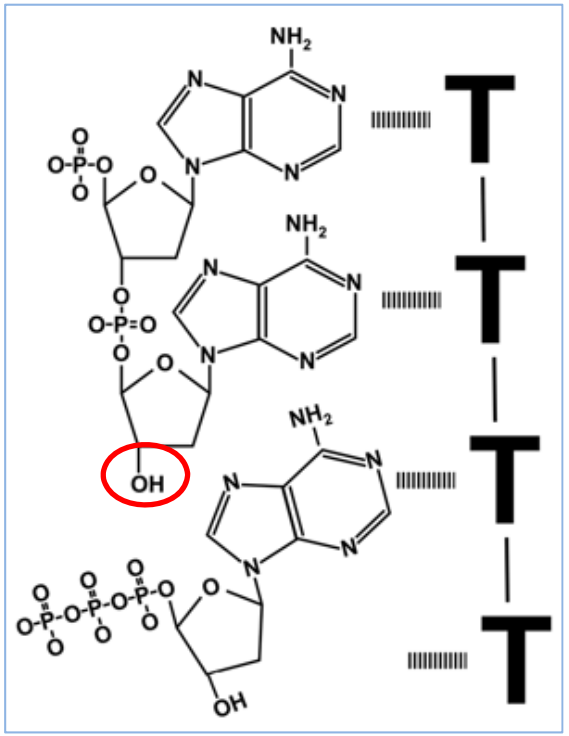


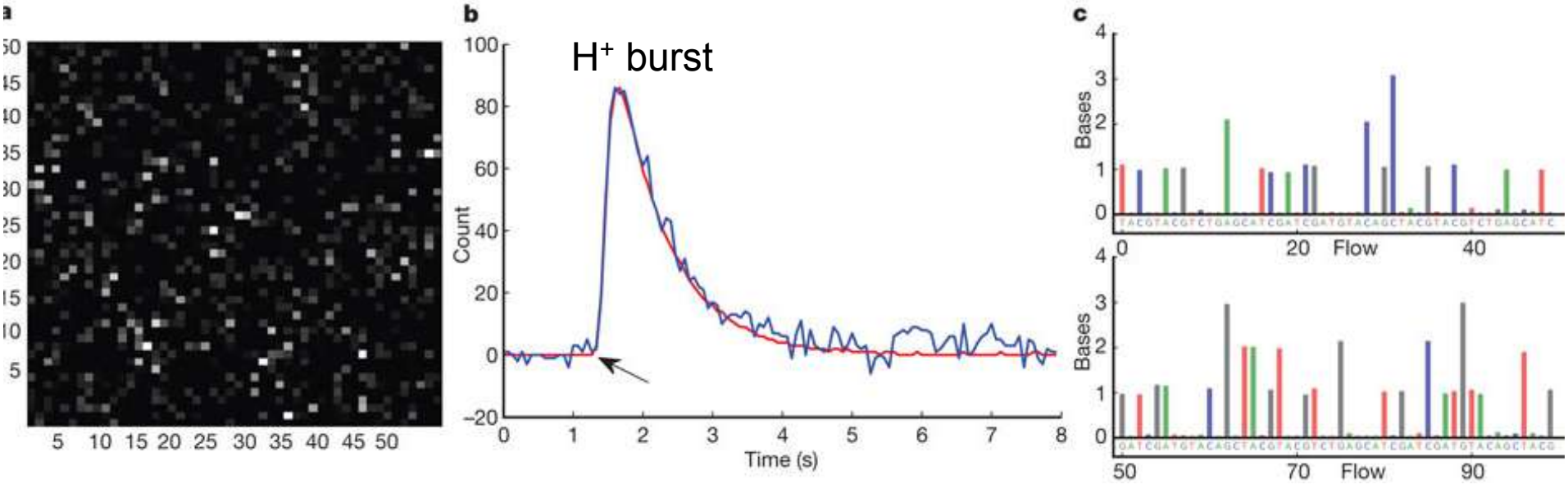
Ion Torrent/Proton:
Sensing bulk release of H^+



emPCR

Semi-conductor

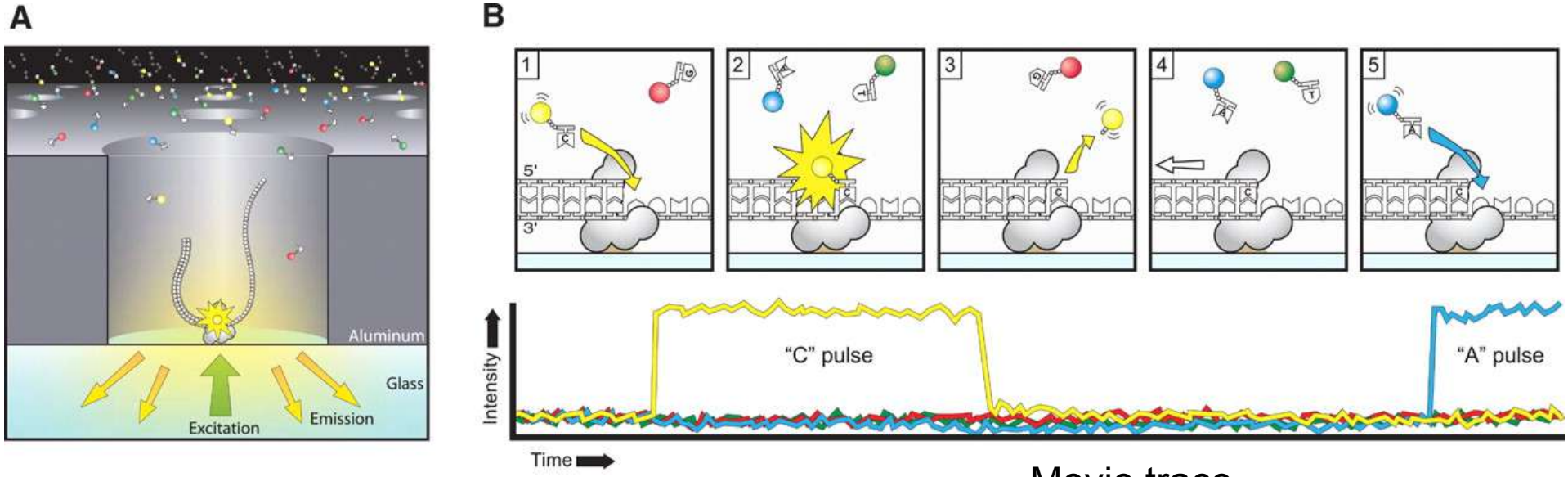




JM Rothberg *et al. Nature* **475**, 348-352 (2011) doi:10.1038/nature10242

PacBio: 3rd-Gen SMRT Sequencing

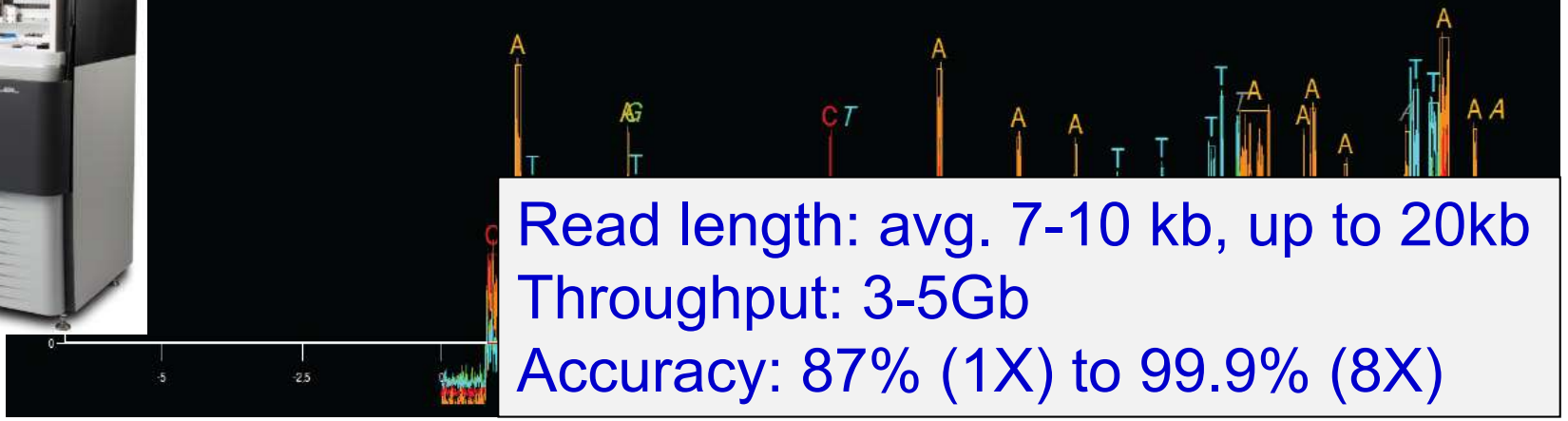
- Single Molecular Real Time (SMRT) real-time technology
- ZMW (zero-mode waveguides), a 100-nm hole with DNA/Polymerase complex immobilized at the bottom; recording fluorescence released from P-dNTP upon incorporation



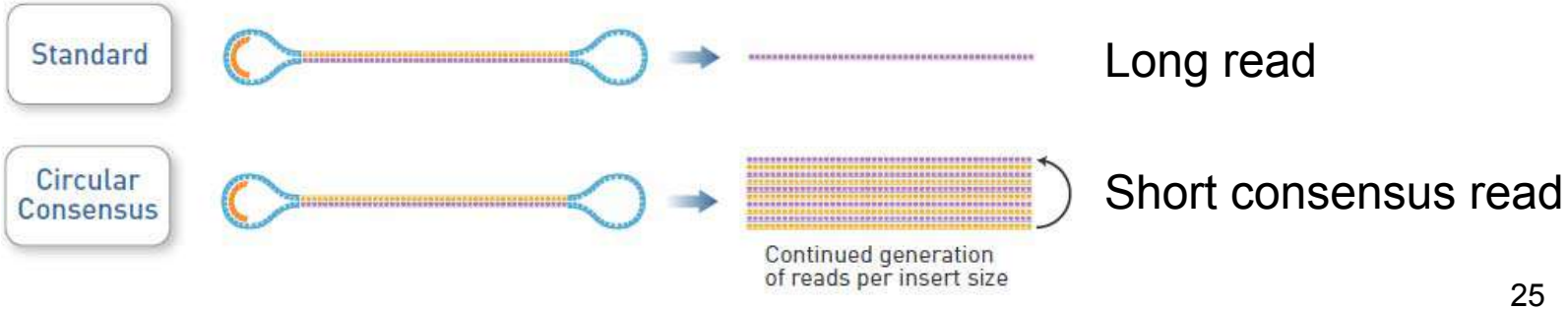
Movie trace

Signal Processing and Base Calling

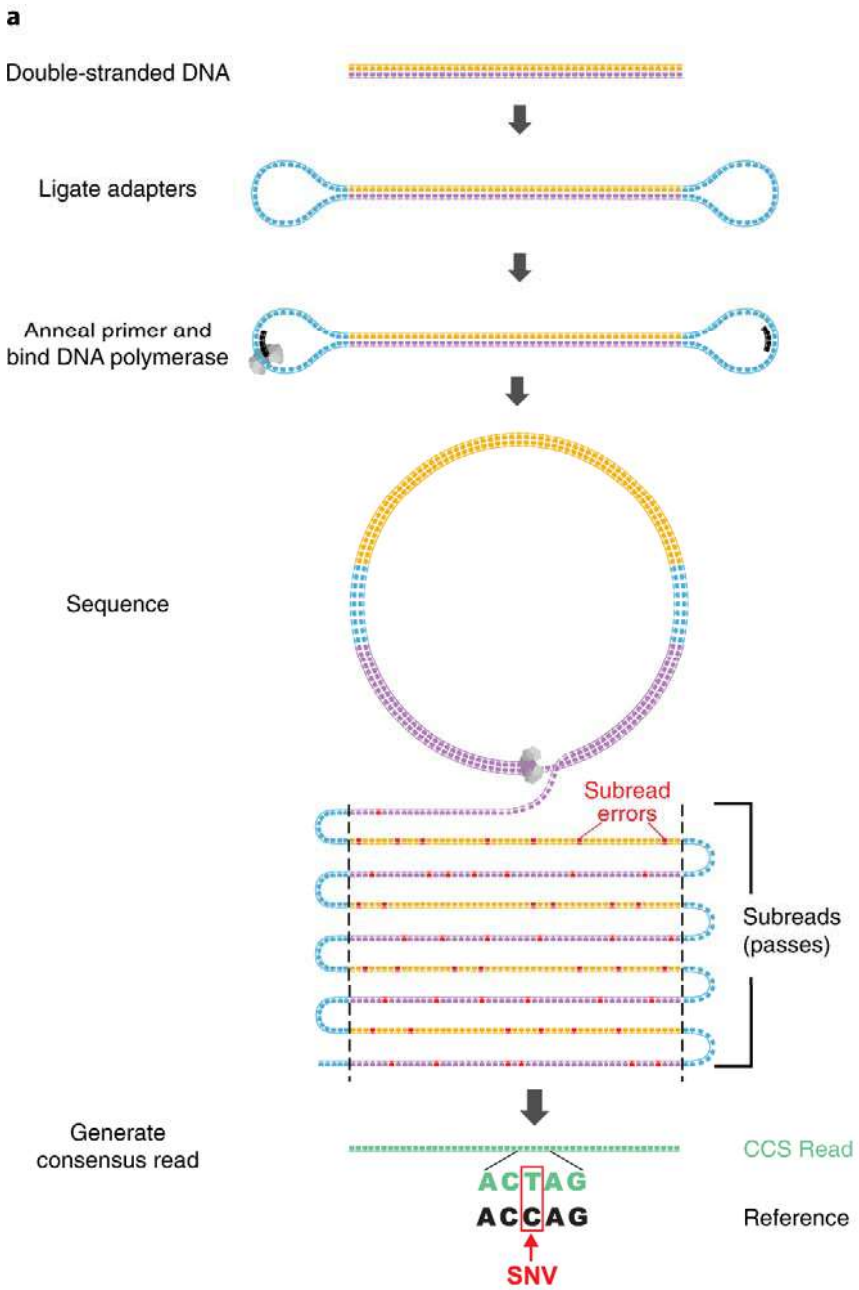
Converting pulses of light into DNA bases and kinetic measures

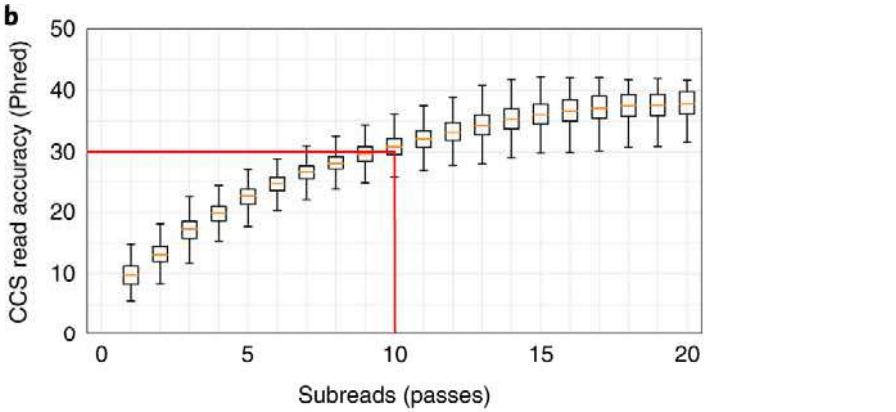


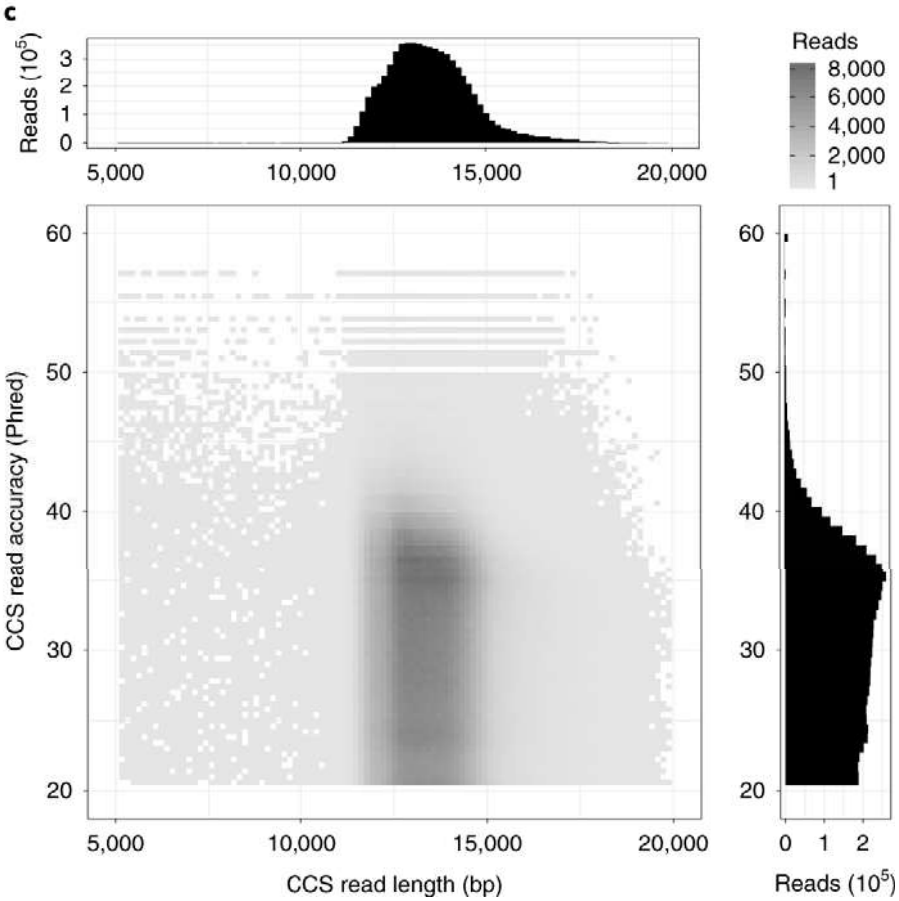
Read length: avg. 7-10 kb, up to 20kb
Throughput: 3-5Gb
Accuracy: 87% (1X) to 99.9% (8X)



PacBio: Circular Consensus Sequencing

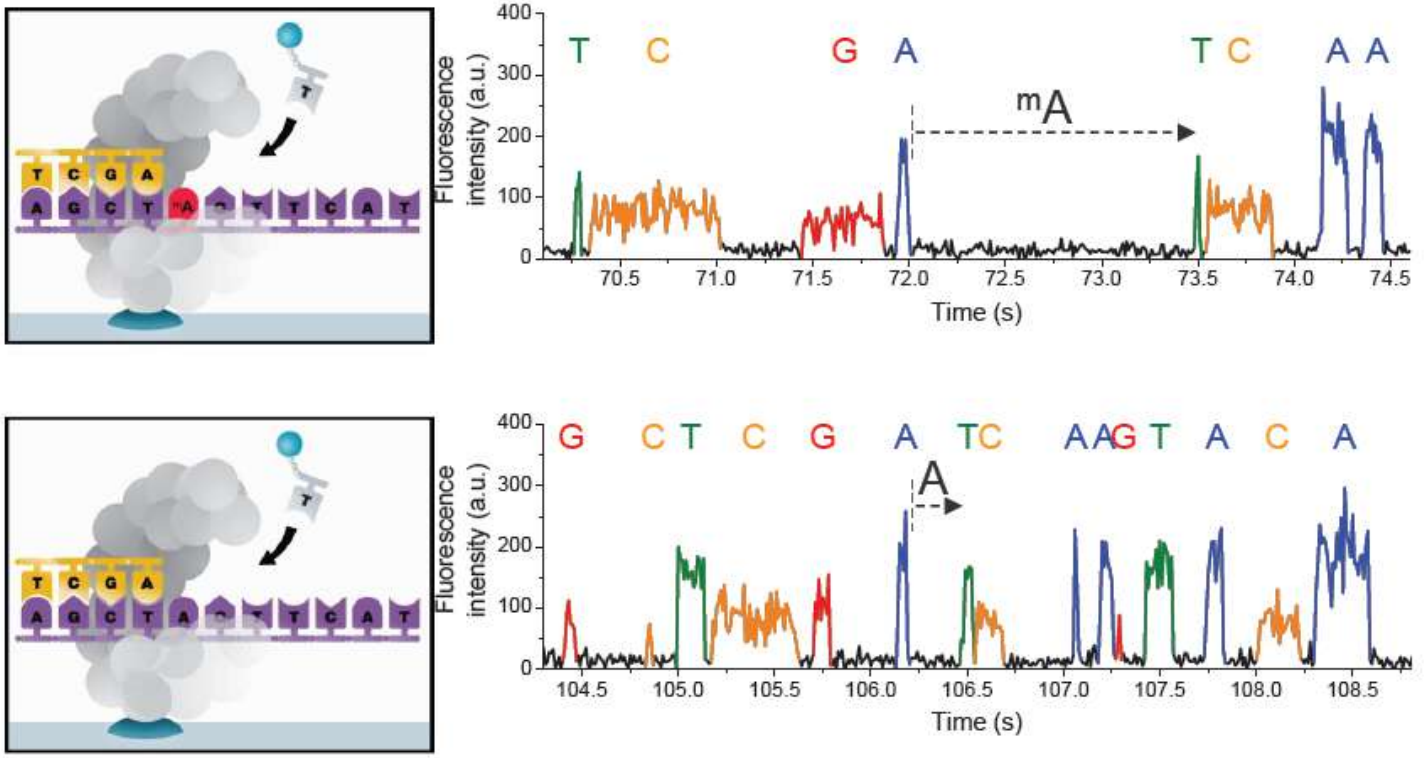






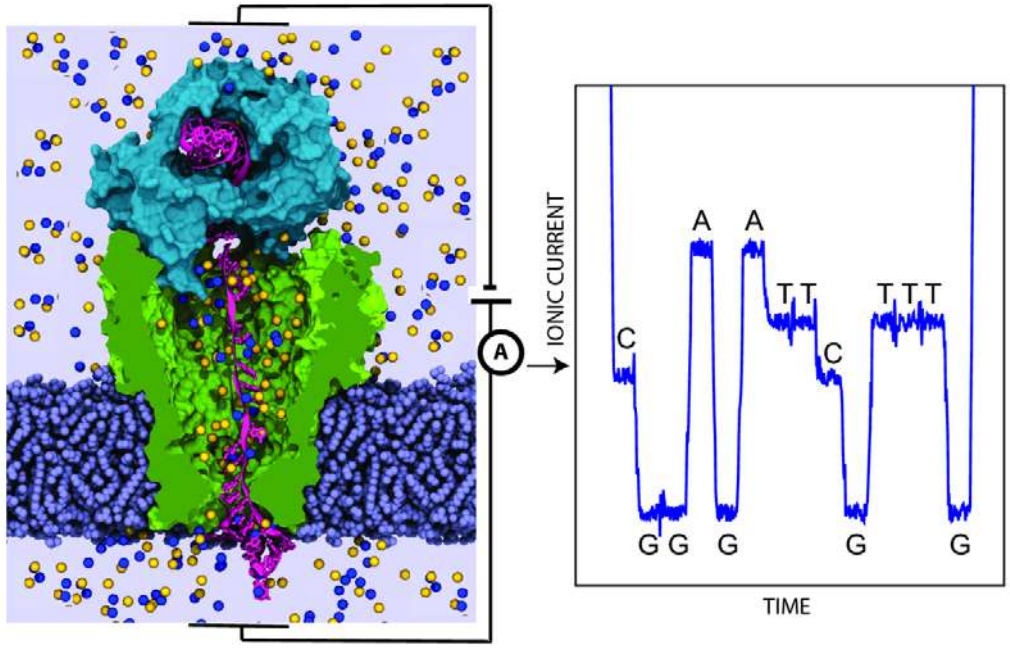
https://www.nature.com/articles/s41587-019-0217-9

Key Feature: Kinetic Information

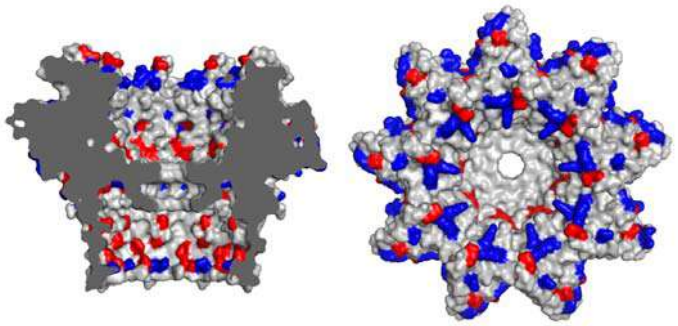


- Differentiation between modified and non-modified bases
  - Epigenetics, DNA damage, New, novel modifications
- Direct observation (*e.g.* no bisulfite)

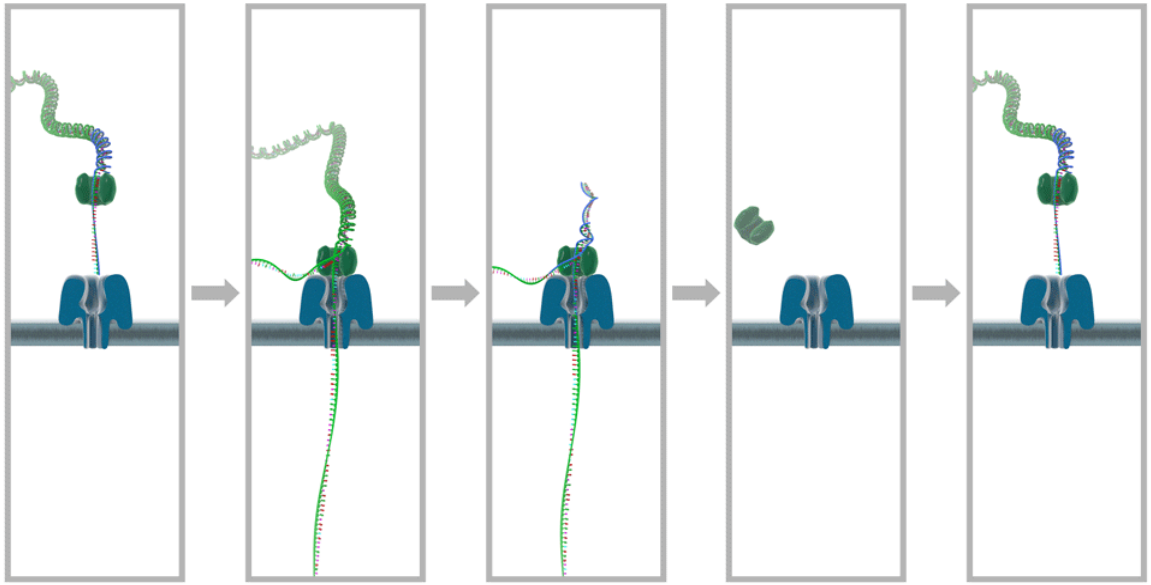
NanoPore Sequencing Technology

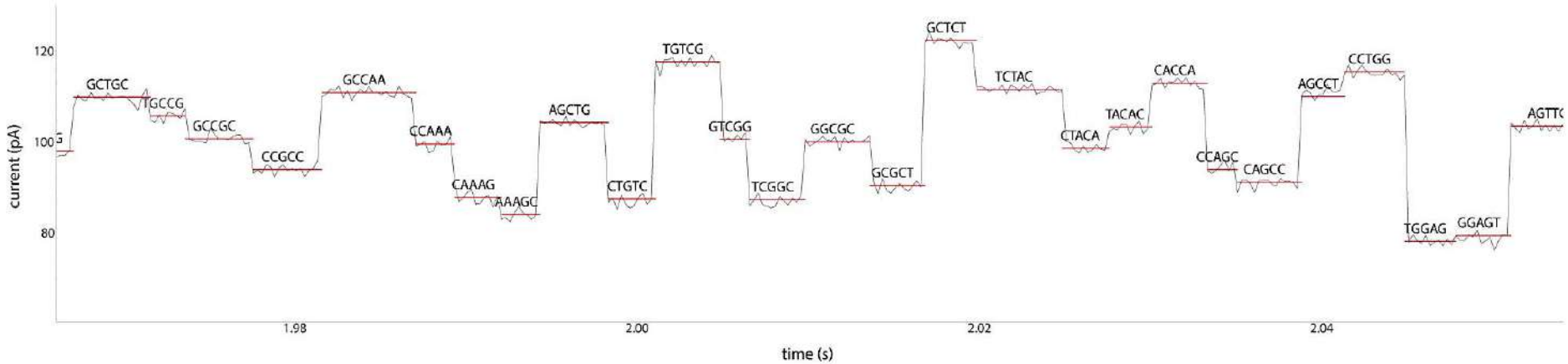


3rd-Gen: Oxford Nanopore (DNA, RNA, protein)



CsgG pore protein complex

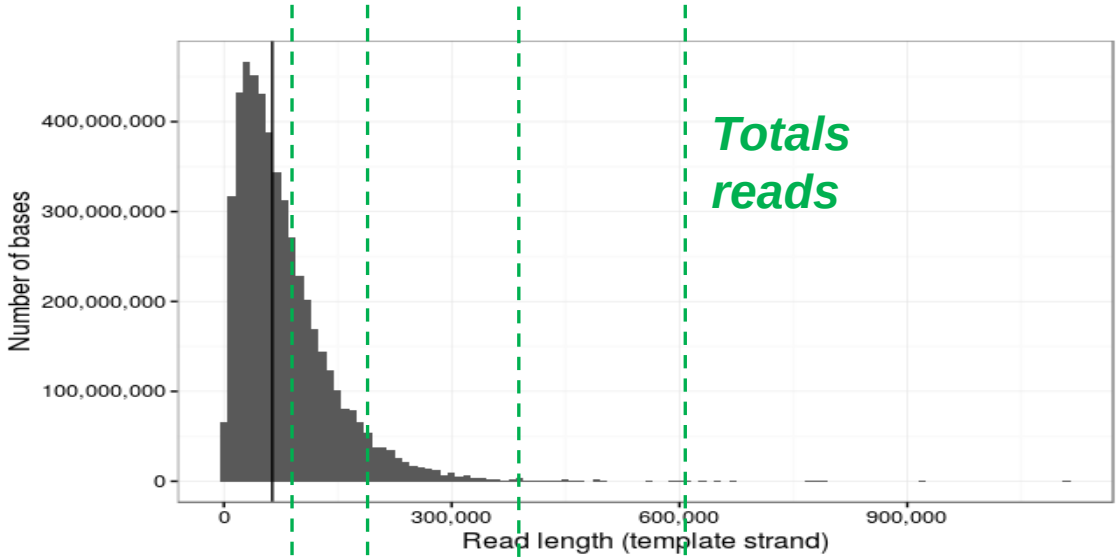




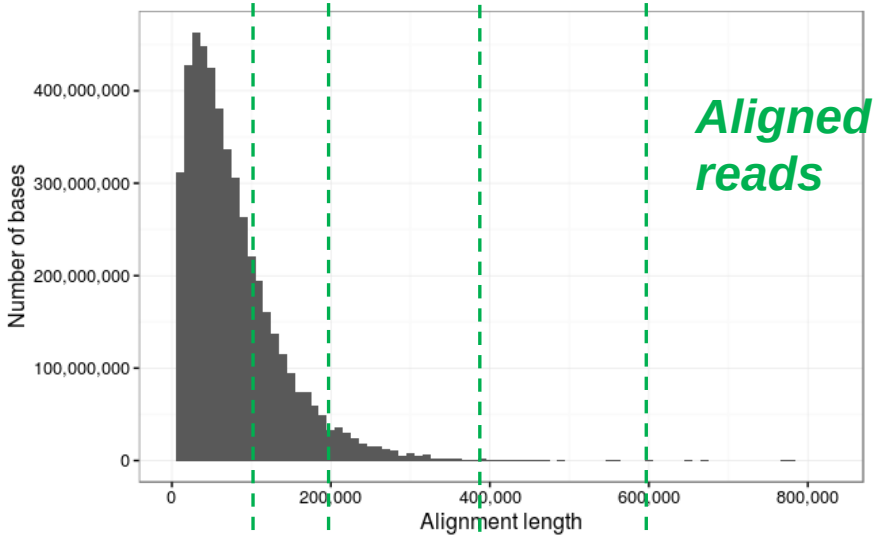
genome assembly using MinION reads [version 1]. F1000Research 2017, 6:1083 (doi: 10.12688/f1000research.12012.1)

F1000Research

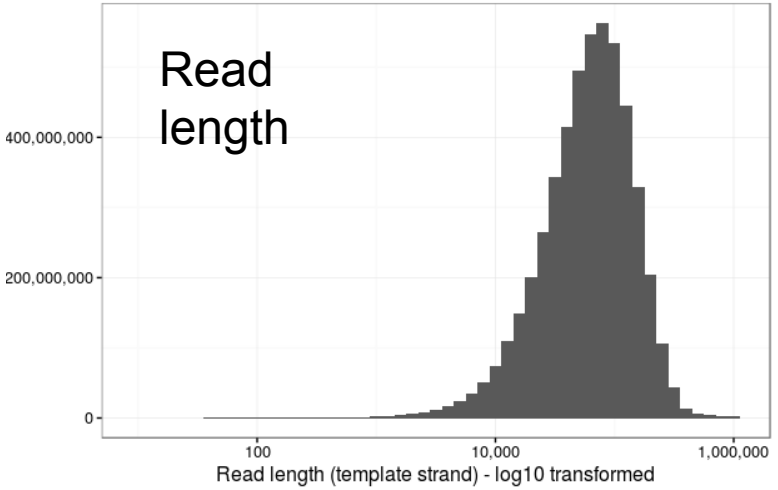
E. coli: on MonION flowcell v9.4

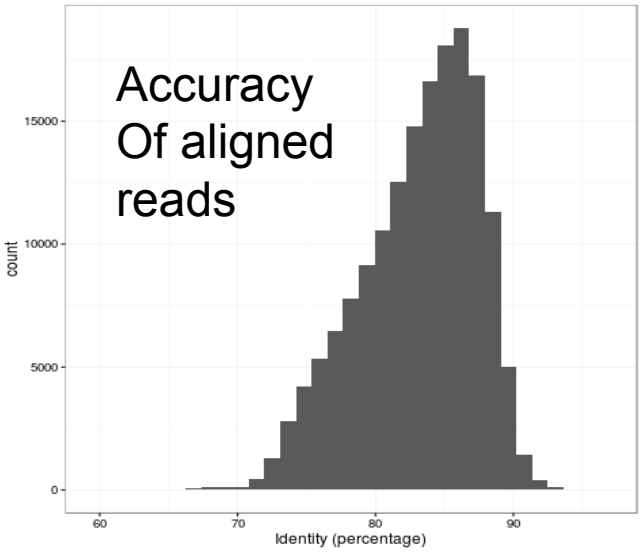


Totals reads



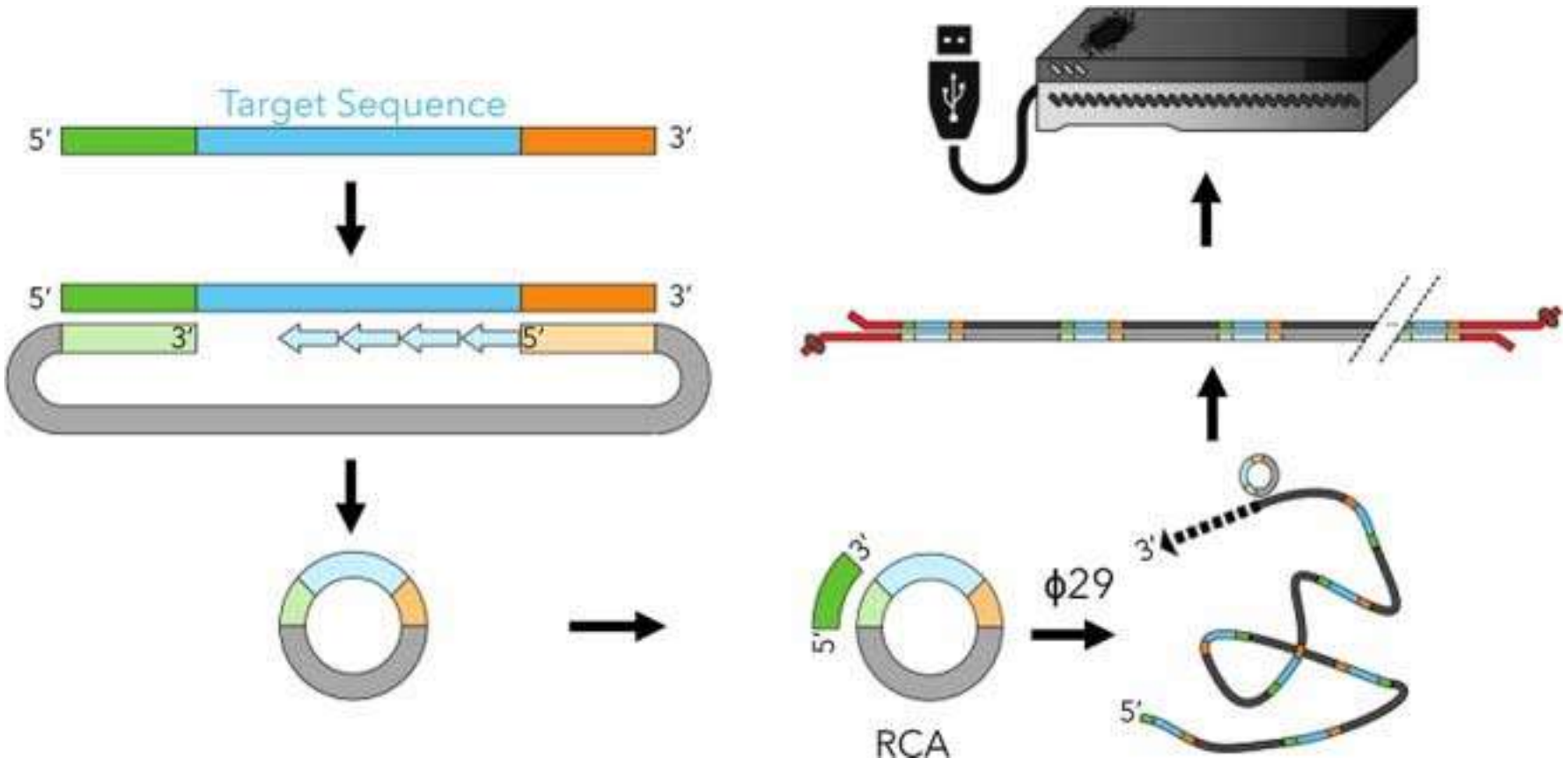
Aligned reads





Source: Loman Lab

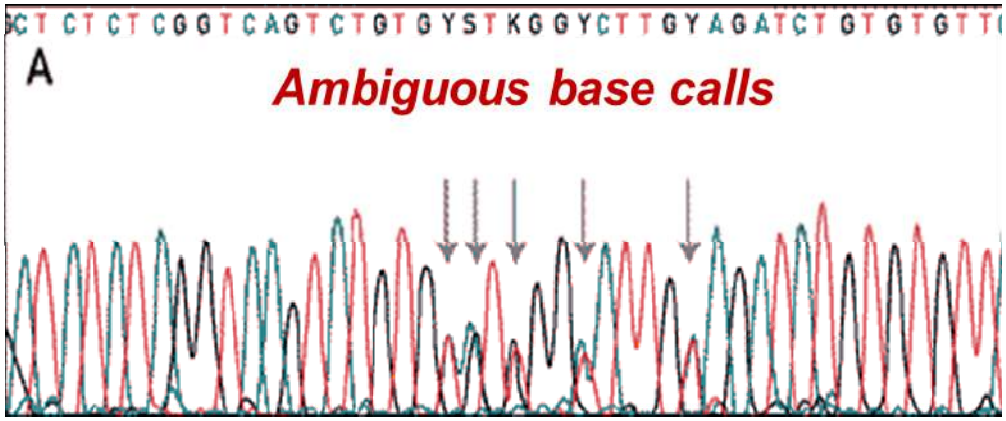
Nanopore: Rolling Circle Amplification proto

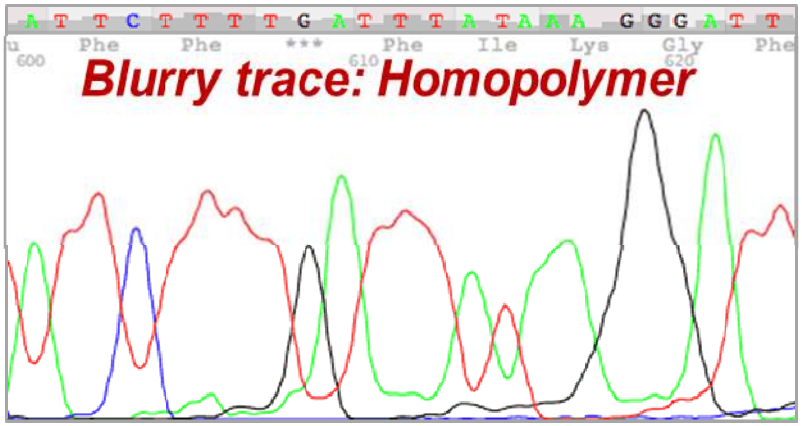


https://europepmc.org/article/pmc/pmc6533607

Problems of Sequencing

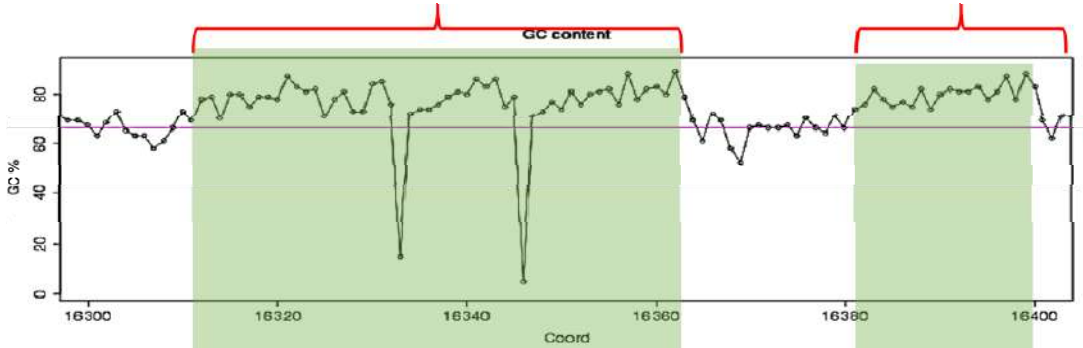
I. Sequencing errors: wrong calls, INDELs



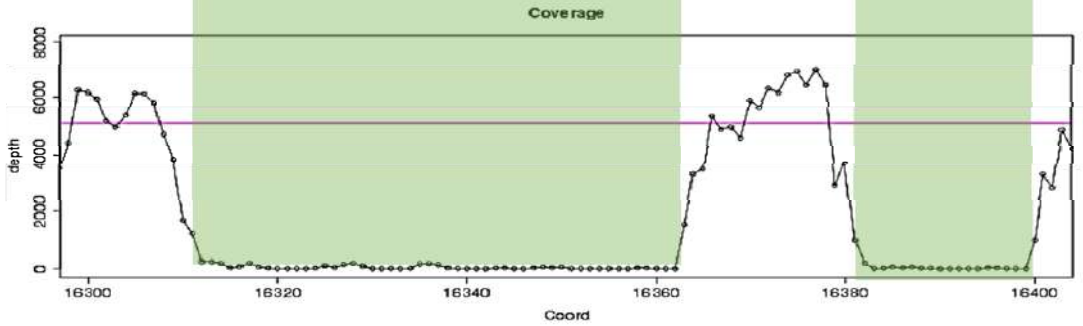


II. Low coverage at high GC%

GC% plot

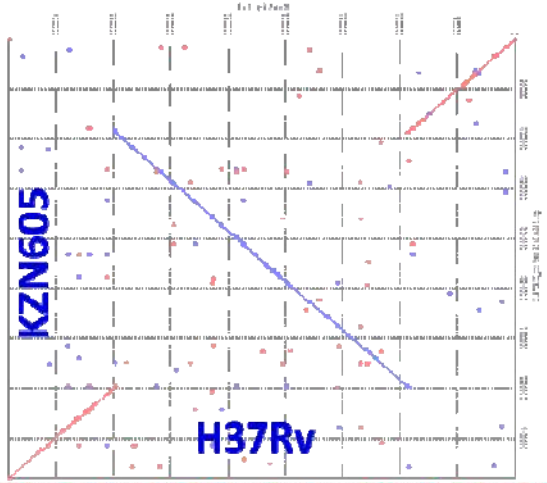


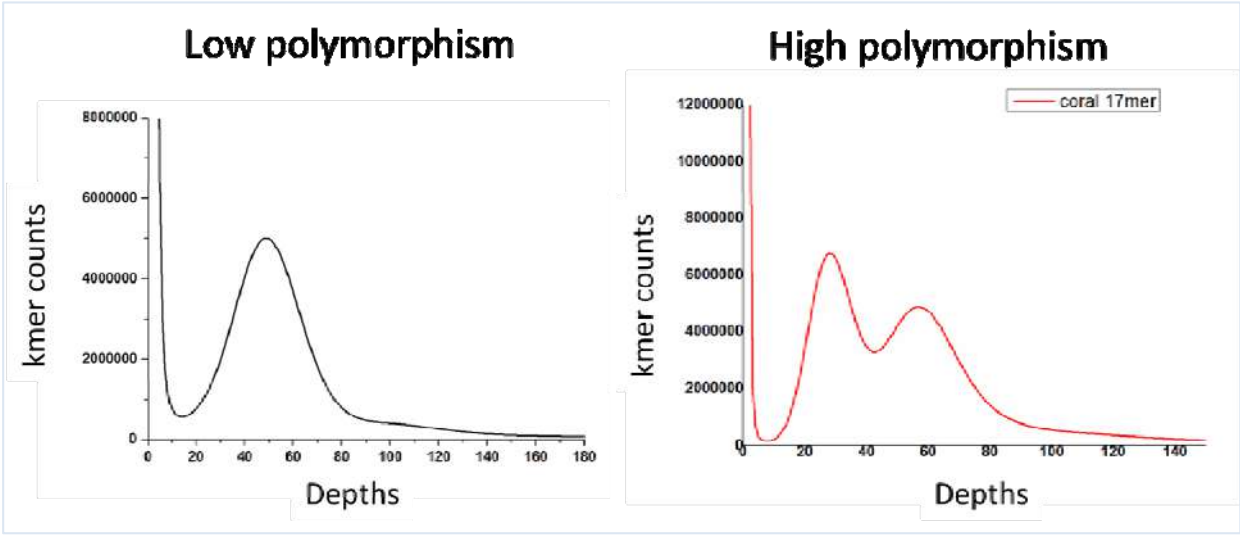
HISeq Read depth



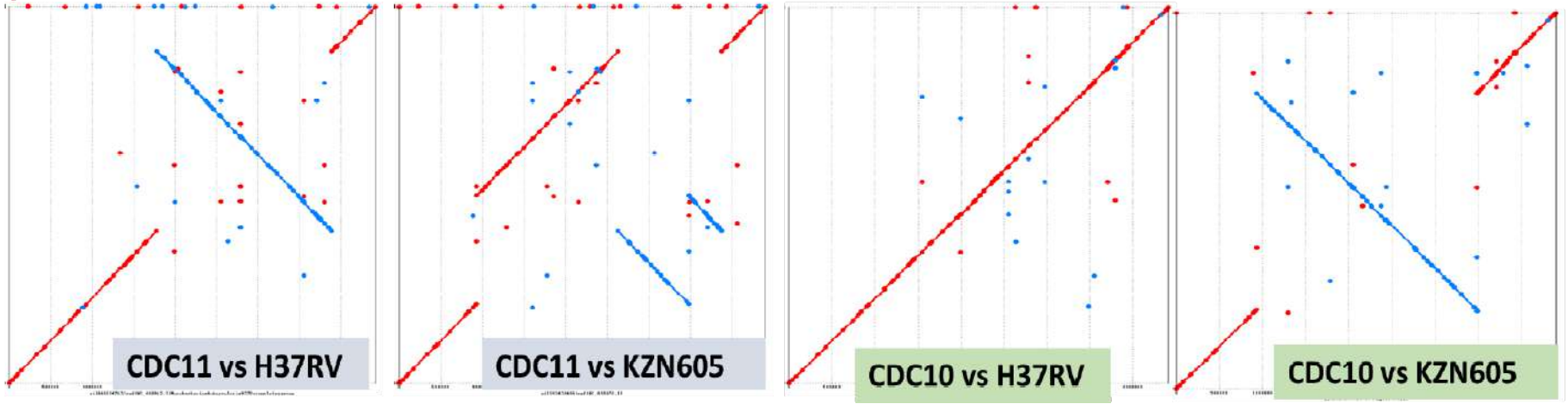
Genome assembly hurdles

- Repeats
- Heterozygosity



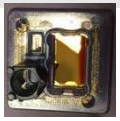


Ambiguous inversion



Current NGS Platforms & Features




	Illumina HiSeq 2500 (*2)	Illumina MiSeq	PacBio Sequel, Sequel II	Oxford ONT MinIon, Promethion
Chemistry	Cyclic reversible terminator Of amplified DNA clusters		SMRT-tech; DNA polymerization	Electrical current passing through a nanopore channel
Chip format				
Output/run	HT mode: 1.2 Tb Rapid mode: 150 Gb	up to 15 Gb	Current: 5-30 Gb	Current: 5-30 Gb
Read length	**PE 50-250 nt**	**PE 50-300 nt**	**1-20 kb (max>100kb)**	**1-50 kb (max>200kb)**
# Fragments /lane	150-180 M (Rapid) 200-250M (HT)	12-15 M (v2) 20-25M (v3)	350-700 K / SMRT cell	30-300 K / chip
Data quality	> 99.9%; Tolerate homopolymer; sensitive to high GC	> 99.9%; Tolerate homopolymer; sensitive to high GC	Raw 85-89%; HiFi ~99.9%; Random homopolymeric errors; tolerate high GC%	Raw 80~94%; Systematic homopolymeric errors; tolerate high GC%
Application	De novo assembly; Re-sequencing; RNA-seq	De novo assembly; Re-sequencing; amplicon	Genome assembly; structural variation; phasing; Iso-Seq	Genome assembly; structural variation; phasing; RNA/DNA-seq

III. NGS Project considerations

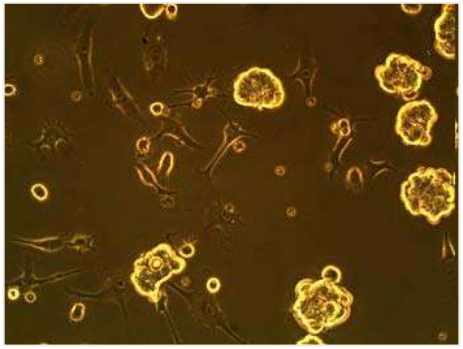
- Purpose: mapping vs de novo
- NGS platform?
- Data format & scale?
- **Sample issues**
- **Genome issues**

NGS project considerations (1)

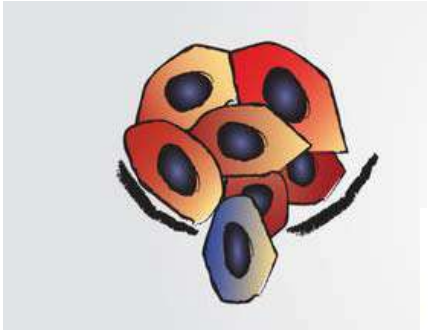
- New genome (de novo): assembly
  - High continuity and accuracy
  - Whole genome annotation: high quality (continuity and accuracy)
  - Phylogeny: diverged/low quality reference; guided assembly
- Re-sequencing: sensitivity & scale
  - Variation discovery: SNP, INDELs, Structural variations
  - Population sequencing &Genotyping
  - Comparative genomics of closely related species
  - RNA-seq:
    - Assembly vs DGE
    - Prokaryotes vs Enkaryotes; polyA-tailed vs none
    - Regulation? Network?

Sample considerations

- Pure strain?

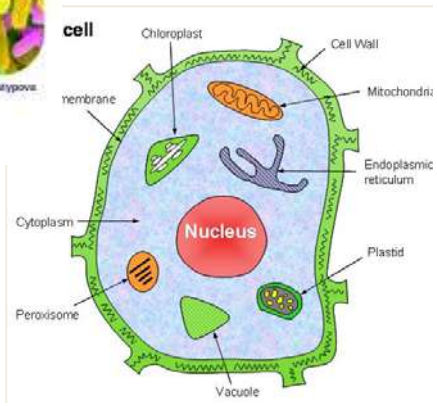


- Cell heterogeneity?



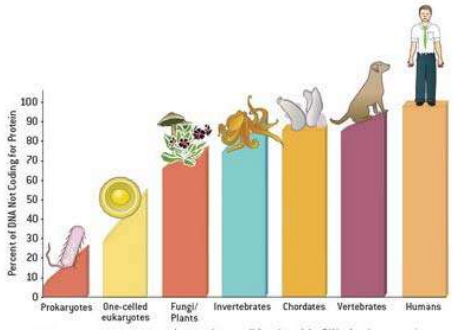
- Metagenome?

- Plastids (mitochondria, chloroplast)：

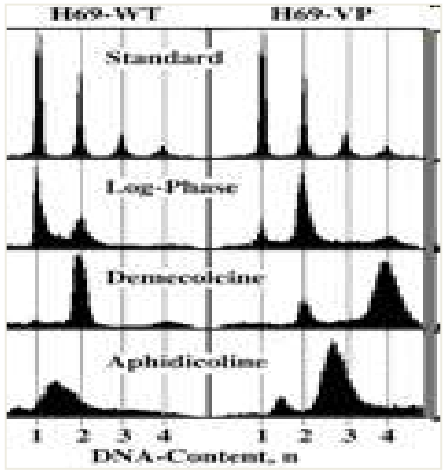


Genome consideration

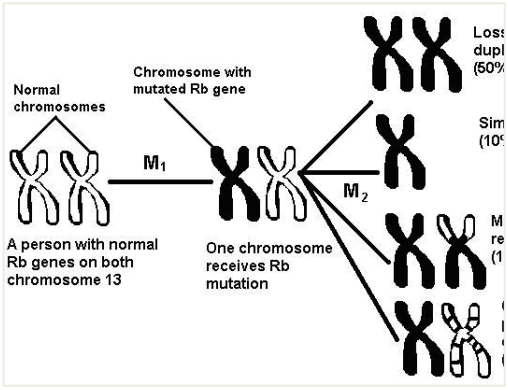
- Genome size



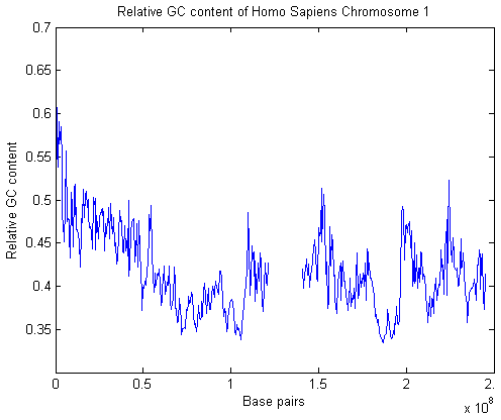
- Genome ploidity?



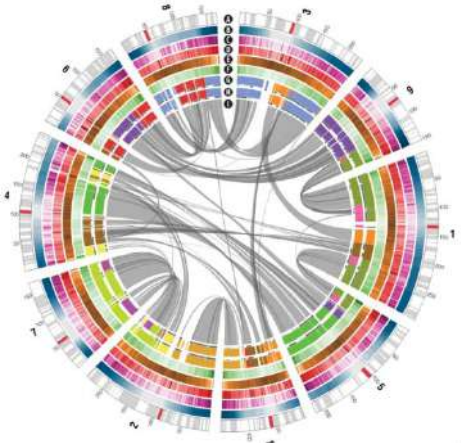
- Heterozygosity?



- GC%



- Genome complexity



- repeats, duplications…

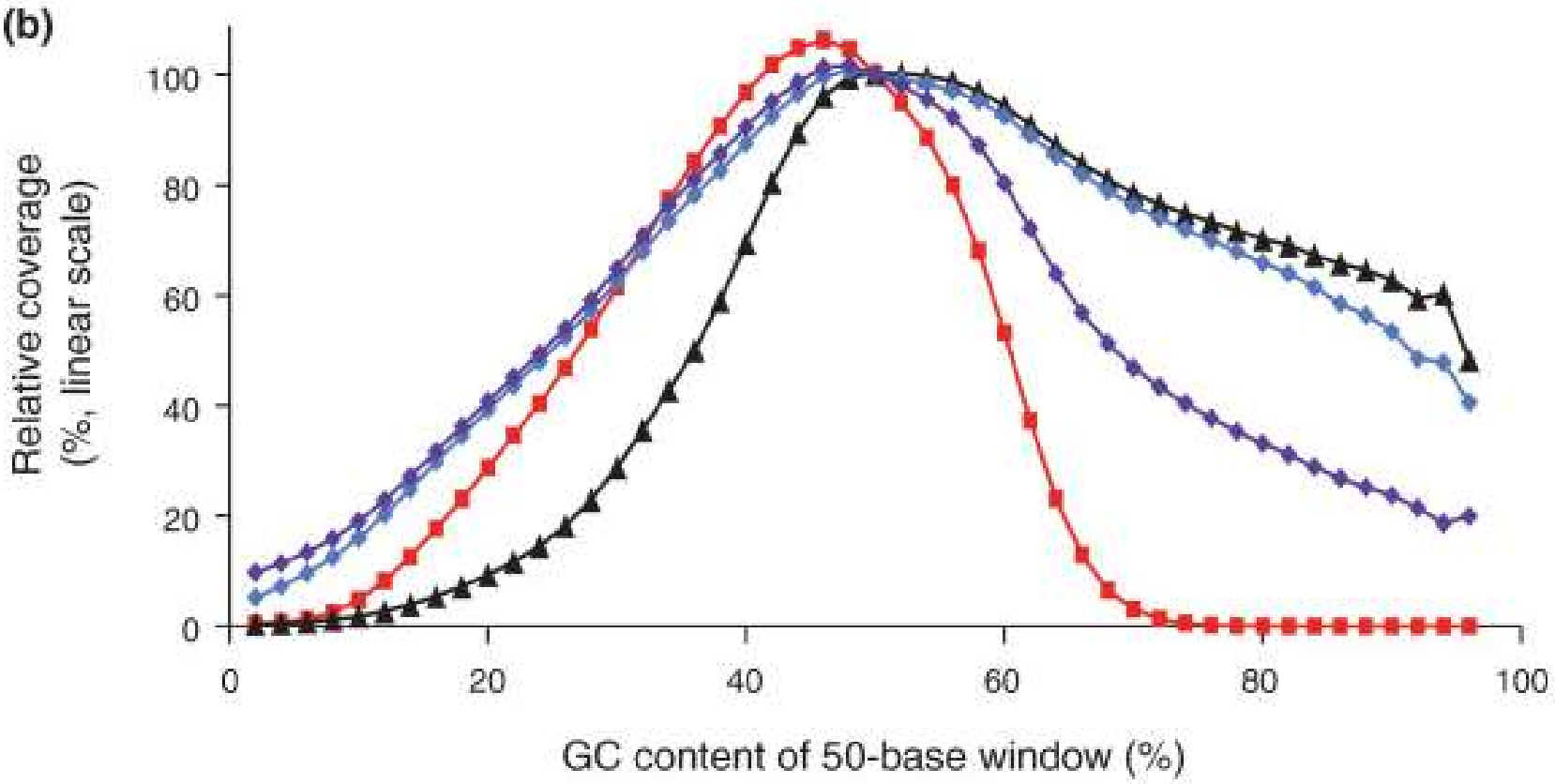
NGS project considerations (2)

- **Sample issues:**
  - Purity: chemical, environmental, endogenous (DNA/RNA)
  - Quality: integrity? processability (over-dried; inhibitors)?
  - Quantity: depends on application type; need spare amount for validation (prior RT-qPCR)?

- **Experimental design:**
  - Controls? Test (treatment? mutants? time points?)
  - Biological replicate: n = 3 (simple/homogeneous) to n=50 (single cell)
  - Barcodes for multiplexed sequencing?
  - Repeat content? Repeat sizes?
  - Huge family of highly conserved genes?
  - GC%?

Uneven presentation due to PCR bias:

1. PCR optimum at ~50% GC
2. Seq. with exterme GC (>80%) are under-represented



Red, Illumina PCR protocol
Others, modified protocols

Aird et al., Genome Biology (201

NGS project considerations (3)

- NGS prep issues:
  - Sample input amount (normal vs low input)
  - PCR amplification (sample, target, library?)
  - Multiple size range required?
- Sequencing concerns:
  - Data: read length, SR vs PE, base accuracy
  - Platform: strength vs weakness
  - Template bias from sequencing/imaging?
- Data scale: coverage depth
  - Genome ploidy
  - pure vs population
  - Expression level or detection sensitivity

Data requirement (assembly consideration)

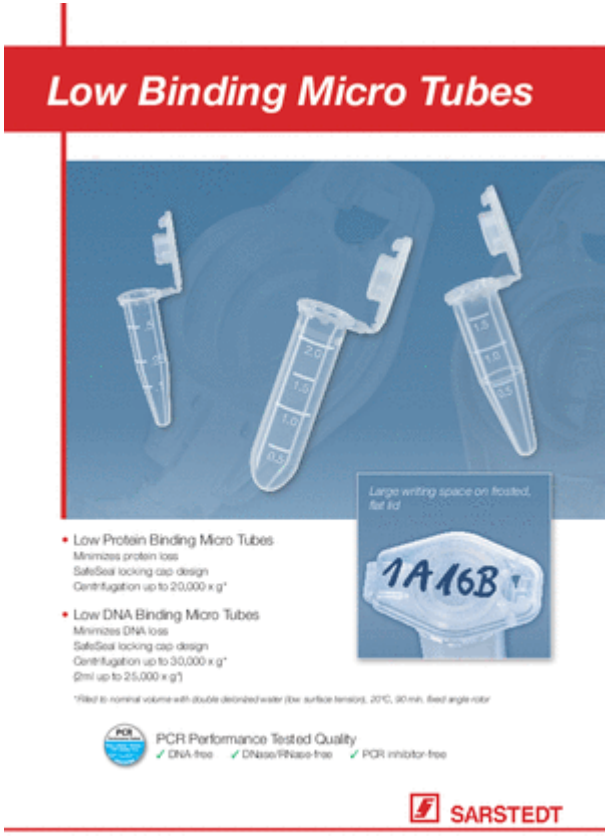
- NGS platform
- Read length
- Sequencing depth (Fold coverage)
- Single Read vs Paired-end

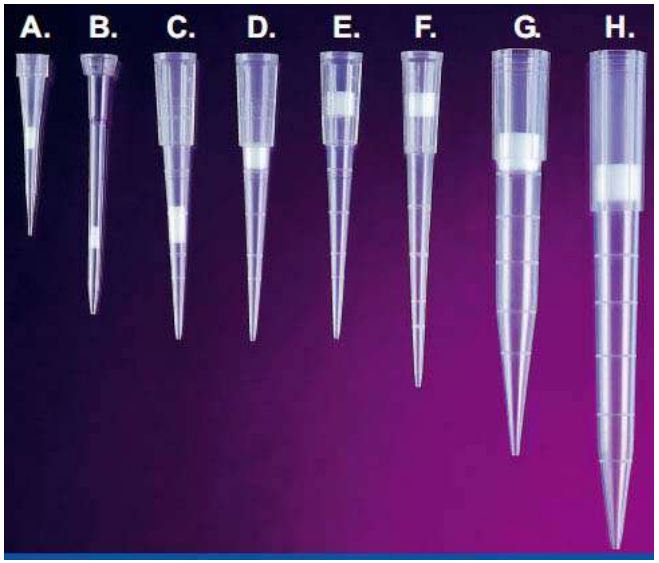
IV. Good Lab Practices & DNA/RNA preparation

Lab wear and clean bench

Plastic wares

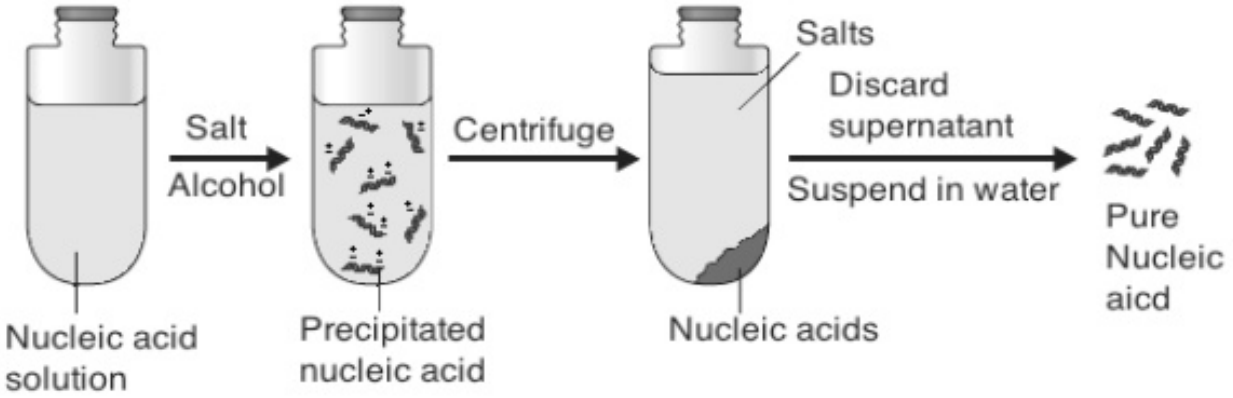
1. Dnase/Rnase free (pre-sterilized)
2. non-sticky / Low-bind

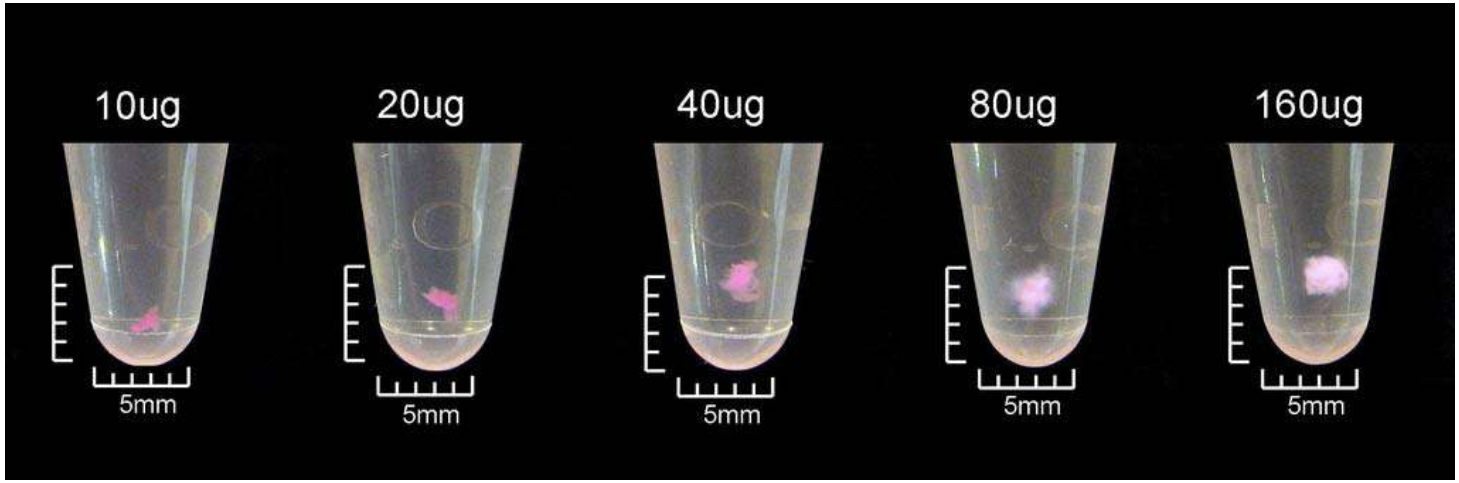




Inverting vs Vortexing

- Ethanol Precipitation is a method for purifying or concentrating DNA/RNA from aqueous solutions using Ethanol as anti-solvent.
- In presence of monovalent cations (eg: Na+) ethanol efficiently precipitates nucleic acids. The precipitate can be collected by centrifugation.





V. Sample & Library QC

General Sample Requirements

DNA:

- RNase-treated and purified
- Submission amount > 3X of library input
- High purity (NanoDrop ratios, BioA, gel)
- Long integrity (>23~48kb)
- Low in inhibitors and contaminants (eg. EDTA, CTAB)
- Concentration: 200-800 ng/ul

RNA:

- DNase-treated and purified
- Submission amount >2X of library input
- High purity (NanoDrop ratios, BioA, gel)
- High in rRNA ratio, RIN
- Low in inhibitors and contaminants
- Concentration:
  - mRNA 200-1000 ng/ul

Auxiliary equipments

Qubit Fluorometer

BioAnalyzer
(up to 11 samples)

Fragment Analyzer
(up to 96well plates*3)

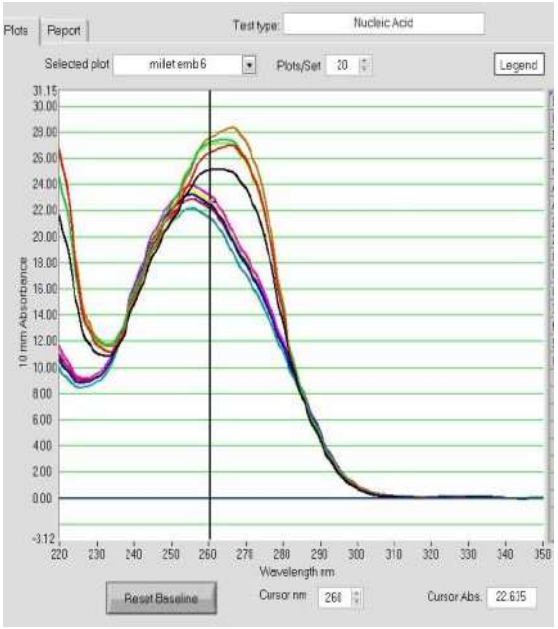
Covaris

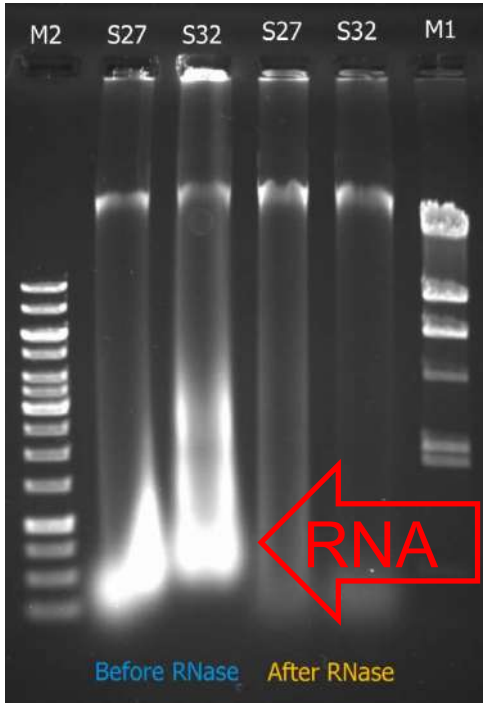
(DNA shearing)
0.2~10kb

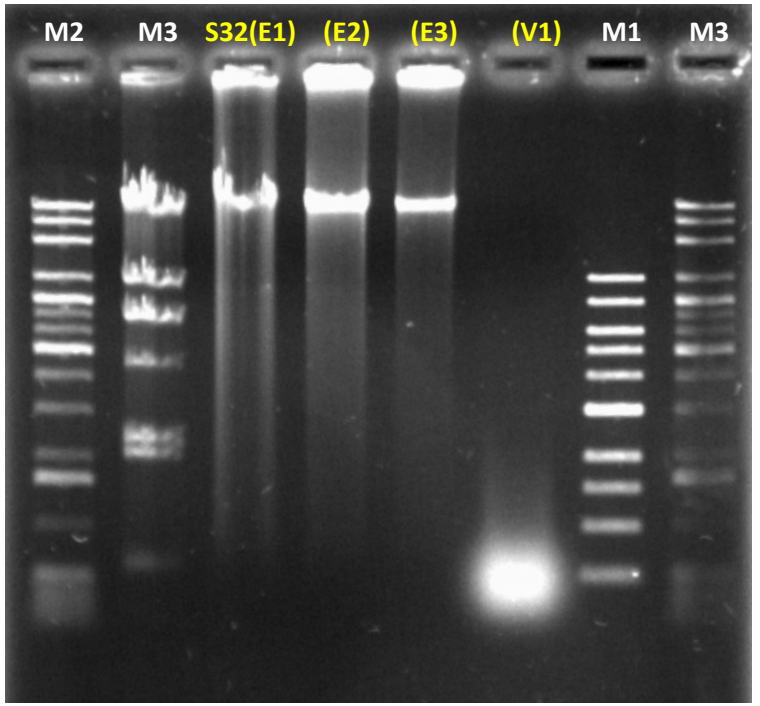
BluePippin
(gel size selection)

LC480
qPCR

Genomic DNA QC

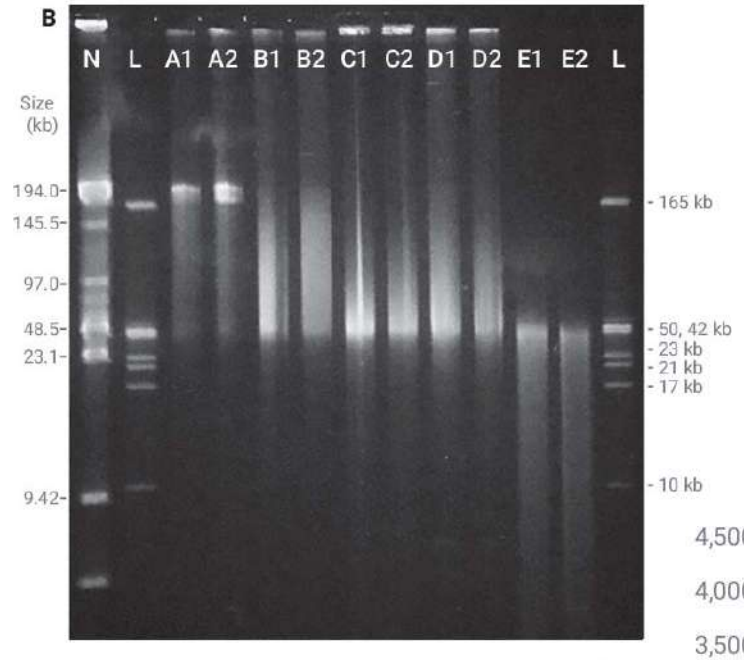


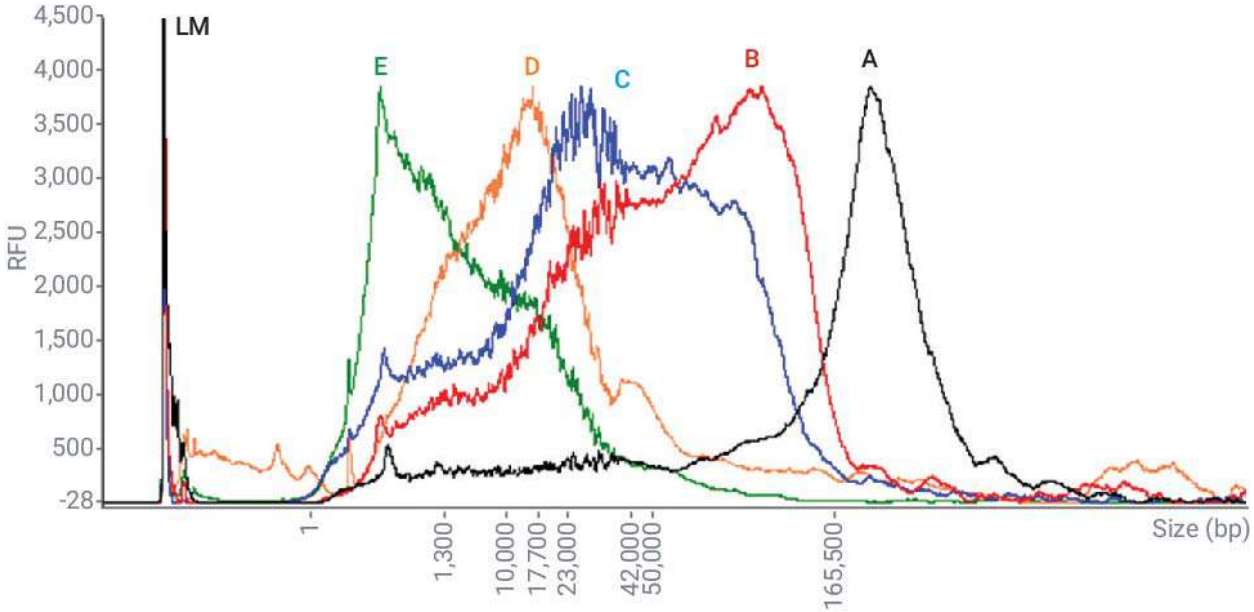




	OD 260/280	OD 260/230	NanoDrop (ng/uL)	Qubit DNA (ng/uL)	RNA Carry over (NanoDrop/Qubit)
S32-original	2.04	2.05	2250.8	56.0	40.19 X
S32_V1	2.16	2.52	1207.40	7.29	165.62 X
S32_E1	1.77	0.94	394.50	131.00	3.01 X
S32_E2	1.67	0.82	45.06	14.10	3.20 X
S32_E3	1.75	0.75	11.48	4.49	2.56 X

Genomic DNA assessment on Fempto Pulse





https://www.semanticscholar.org/paper/Genomic-DNA-Extractions-Compared-with-the-Agilent-Pocernich-Uthe/aad08d7899bac06a4af39c88f02016df9b741e6f

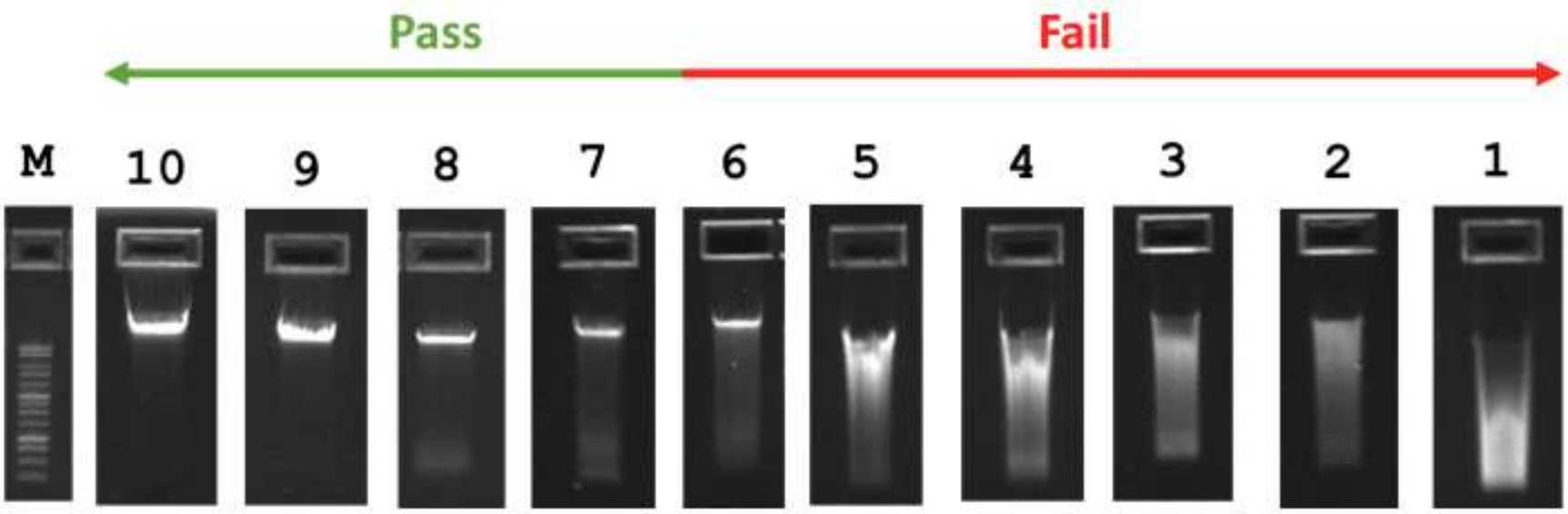
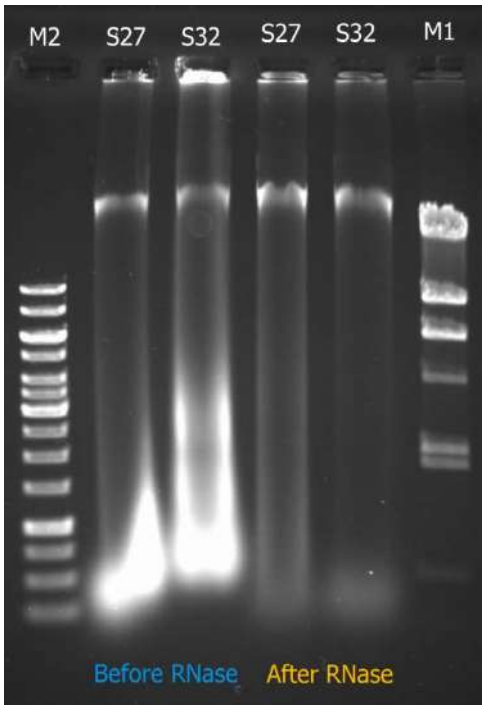
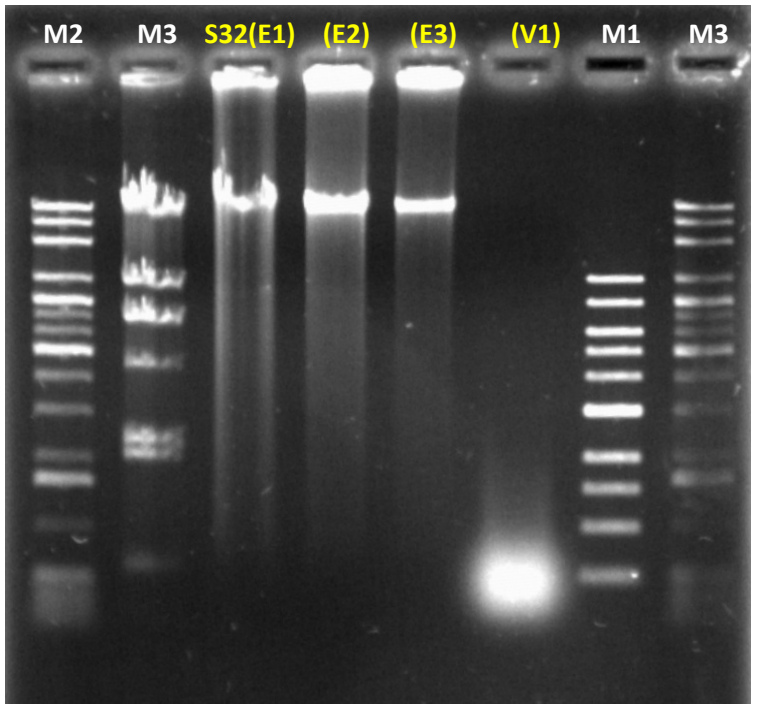


Figure 1: The marker is a 1Kb ladder (Promega, G5711) with DNA fragments ranging from 250bp to 10Kbp. The gDNA samples range from very good quality HMW gDNA with very little degradation or RNA contamination (Image 10) to extremely degraded gDNA (Image 1). gDNA samples in images 10 to 7 would pass sample QC. gDNA samples in images 6 or less would fail sample QC showing greater levels of RNA contamination or DNA degradation.

Genomic DNA QC





	OD 260/280	OD 260/230	NanoDrop (ng/uL)	Qubit DNA (ng/uL)	Carry over (NanoDrop/Qubit)
S32-original	2.04	2.05	2250.8	56.0	40.19
S32_V1	2.16	2.52	1207.40	7.29	165.62
S32_E1	1.77	0.94	394.50	131.00	3.01
S32_E2	1.67	0.82	45.06	14.10	3.20
S32_E3	1.75	0.75	11.48	4.49	2.56

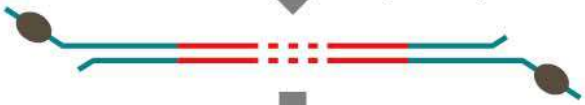
Bead Free Long Fragment LSK109 Library Prep

Genomic DNA Fragmentation

End-Prep

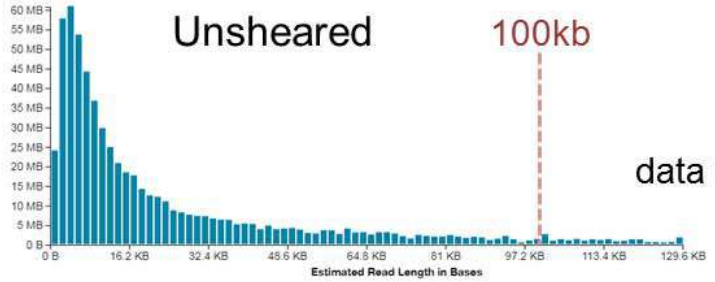
Add Equal volume (9% PEG8000 (w/v), 1M NaCl)
Incubate 30mins, Spin 13K rpm 30mins, 2x 70% EtOH Wash
Resuspend in EB

Adapter Ligation

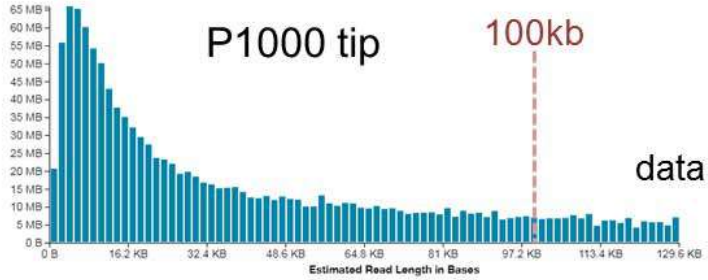


Add NaCl to 400mM
Incubate 30mins, Spin 13K rpm 30mins
2x Wash (4.5% PEG8000, 500mM NaCl)
Resuspend in EB

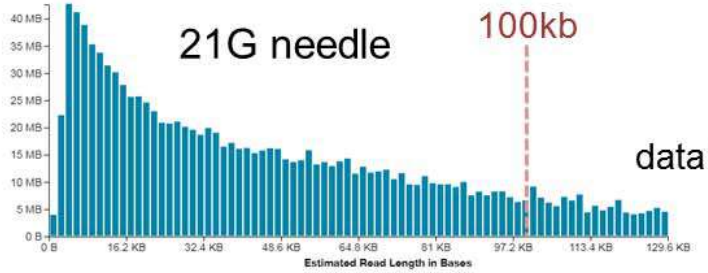
Load



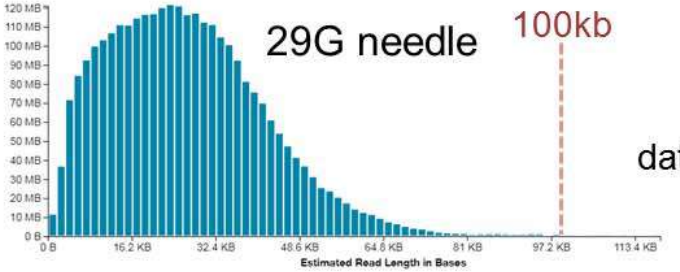
6%
data in reads >100Kb



21%
data in reads >100Kb



15%
data in reads >100Kb

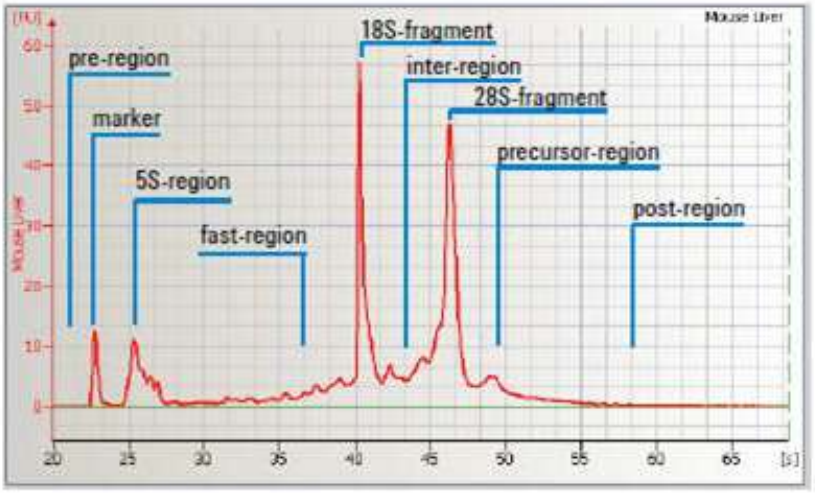


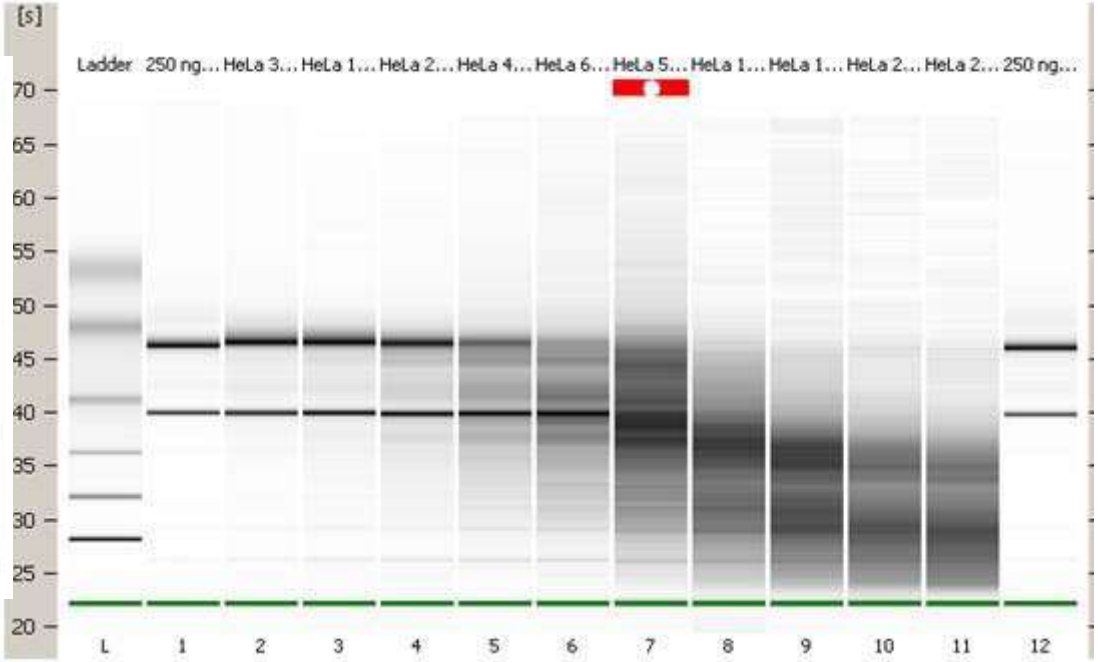
0.03%
data in reads >100Kb

(~12Gb of sequence collected from screening different library preparations on a single MinION flowcell with intervening DNaseI resets. Libraries prepared using intermediate fractions from a sequential shearing series performed in the same tube on genomic DNA (Phenol Chloroform and spooled out, ~200ng/ul, 10ug input per library))

https://borevitzlab.anu.edu.au/projects-overview/long-read-sequencing-of-eucalyptus-tree-genomes/

RNA integrity - BioAnalyzer





RNA integrity - BioAnalyzer

BioAnalyzer RNA ladder

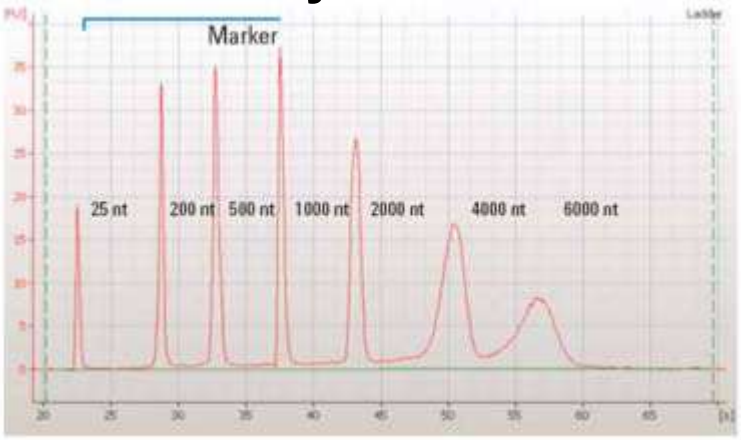
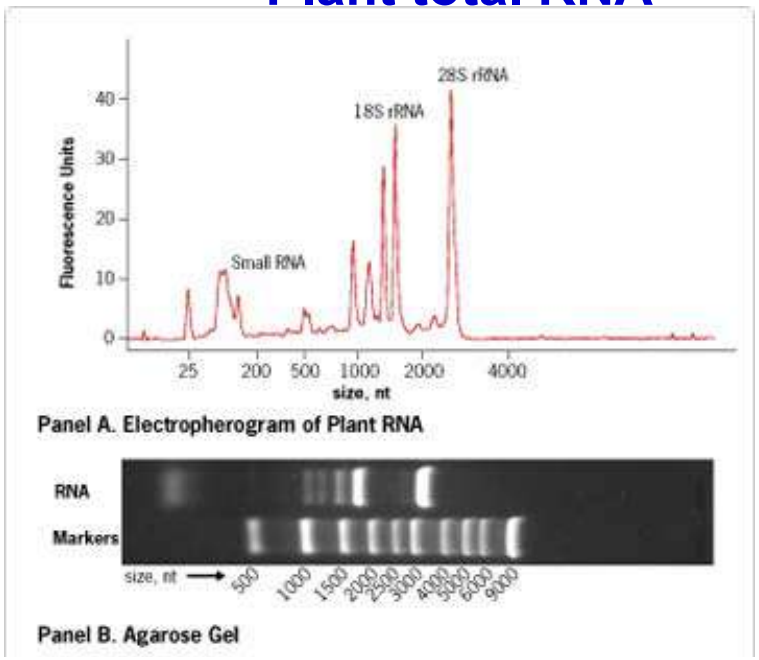
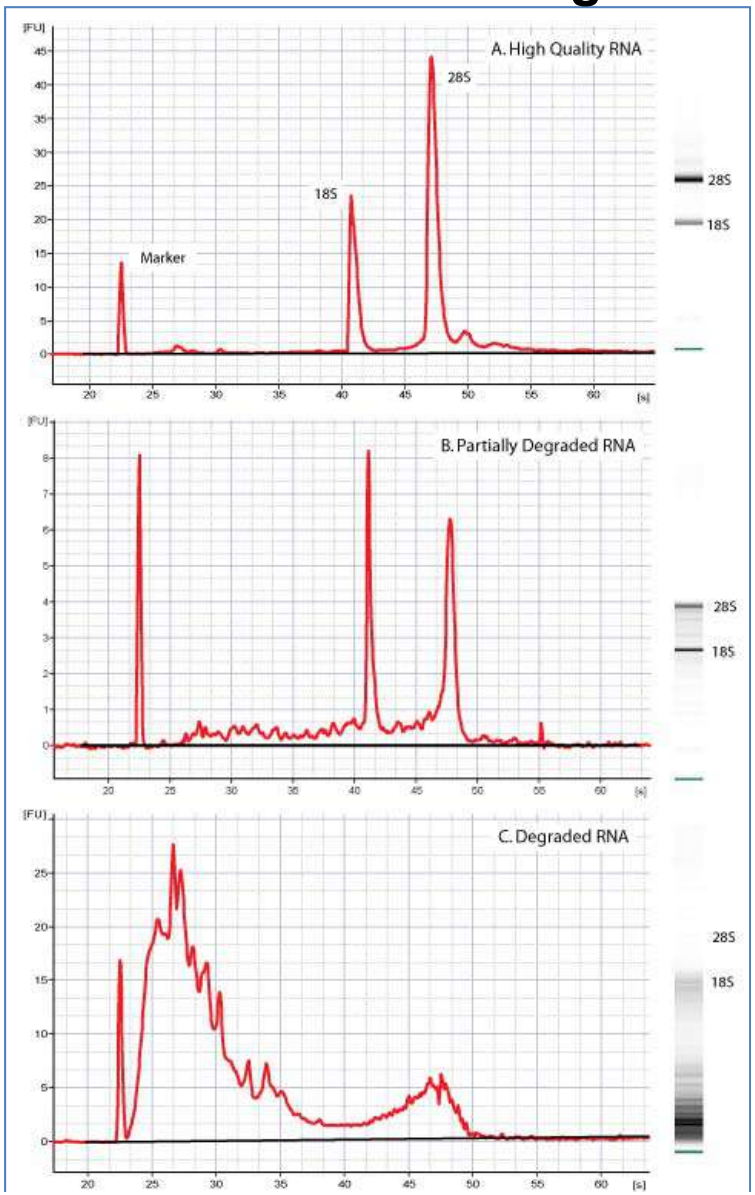


Figure 1 RNA 6000 Nano ladder

Plant total RNA

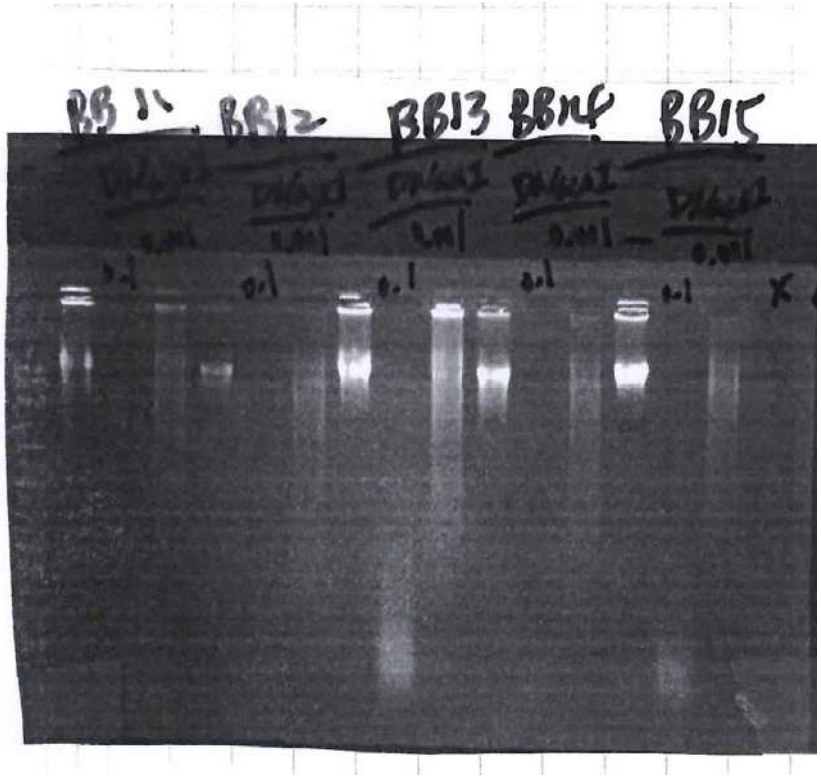


Human RNA – various degradation

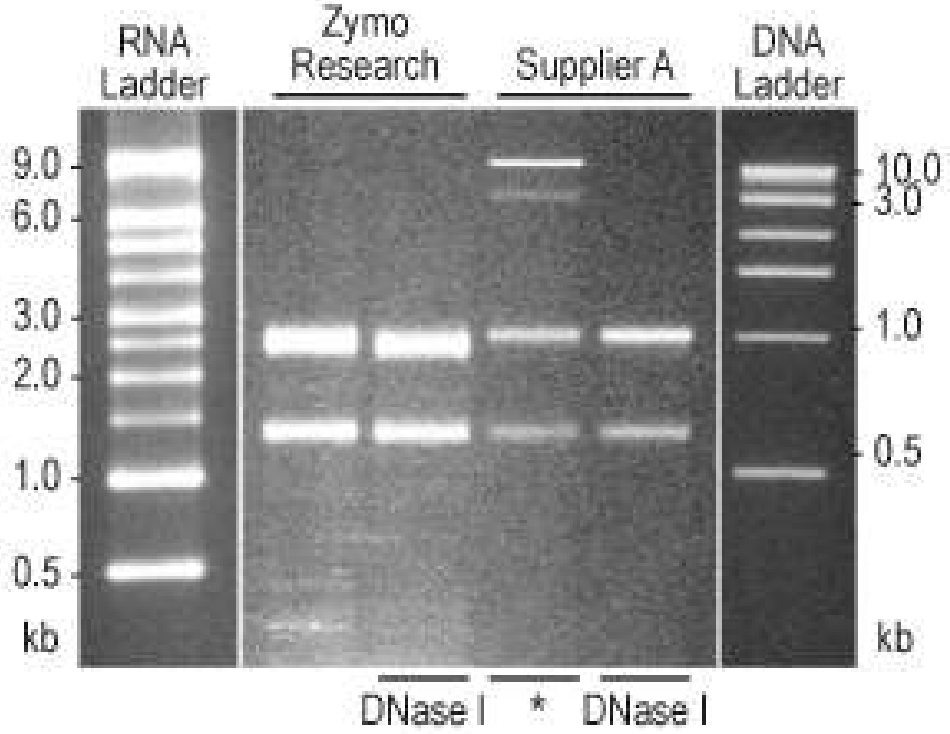


DNase I treatment

gDNA sample

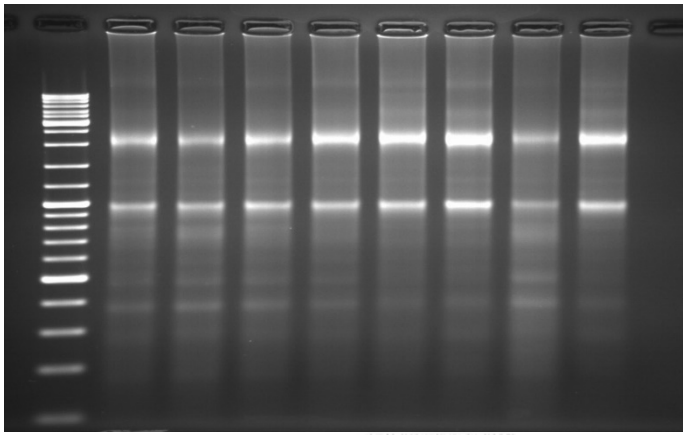


RNA sample

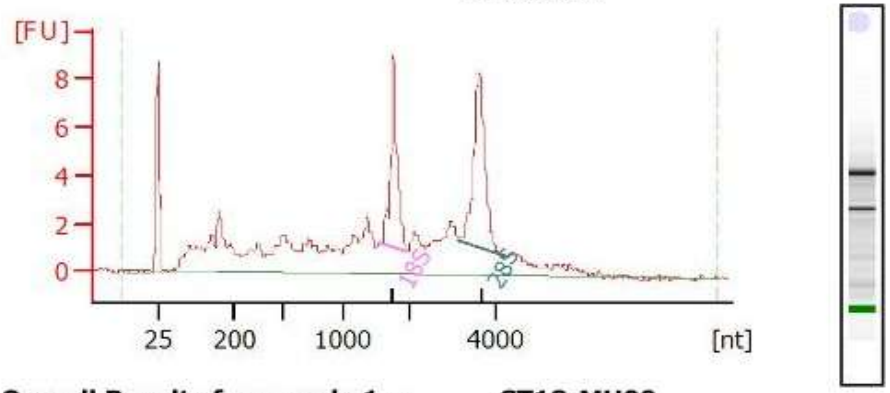


Different methods of sample collection and RNA extraction

Liquid N2- snap freeze
RNAzol + PCI



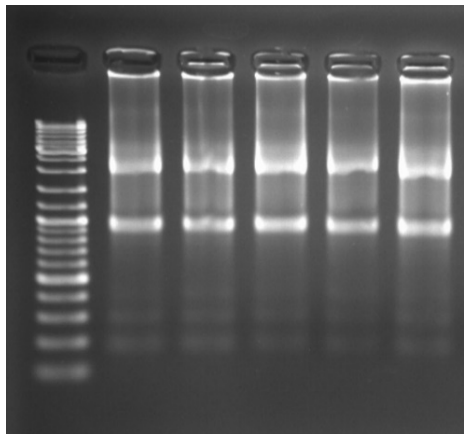
ST18-MU09

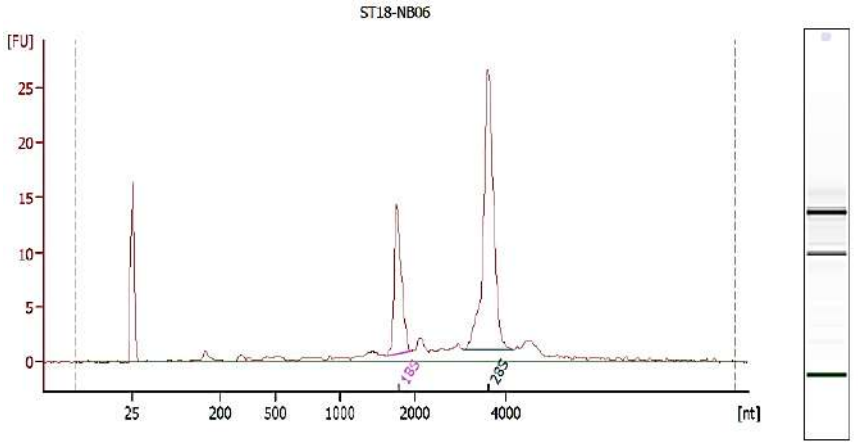


Overall Results for sample 1 : ST18-MU09

RNA Area:	124.6
RNA Concentration:	59 ng/µl
rRNA Ratio [28s / 18s]:	1.3
RNA Integrity Number (RIN):	7.2 (B.02.07)
Result Flagging Color:	
Result Flagging Label:	RIN: 7.20

RNALater
RNeasy-Tissue kit

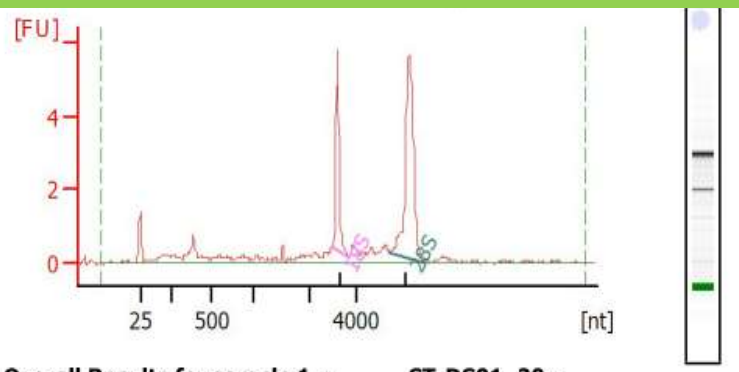




Overall Results for sample 6 : ST18-NB06

RNA Area:	110.1	RNA Integrity Number (RIN):	9.6 (B.02.07)
RNA Concentration:	61 ng/µl	Result Flagging Color:	
rRNA Ratio [28s / 18s]:	2.4	Result Flagging Label:	RIN: 9.60

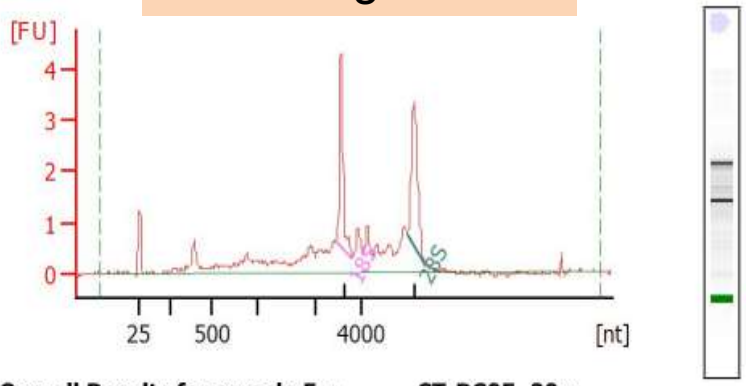
Good RNA: rRNA ratio>1.8, RIN>8



Overall Results for sample 1 : ST-DS01_20x

RNA Area:	29.4
RNA Concentration:	70 ng/µl
rRNA Ratio [28s / 18s]:	2.1
RNA Integrity Number (RIN):	8.8 (B.02.07)
Result Flagging Color:	
Result Flagging Label:	RIN: 8.80

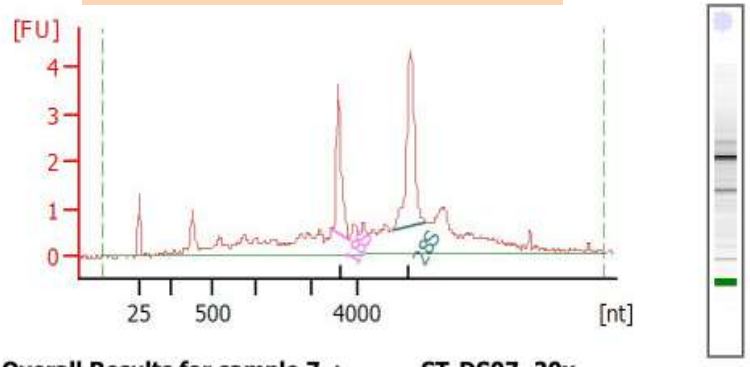
Some degradation



Overall Results for sample 5 : ST-DS05_20x

RNA Area:	28.1
RNA Concentration:	67 ng/µl
rRNA Ratio [28s / 18s]:	1.2
RNA Integrity Number (RIN):	6.9 (B.02.07)
Result Flagging Color:	
Result Flagging Label:	RIN: 6.90

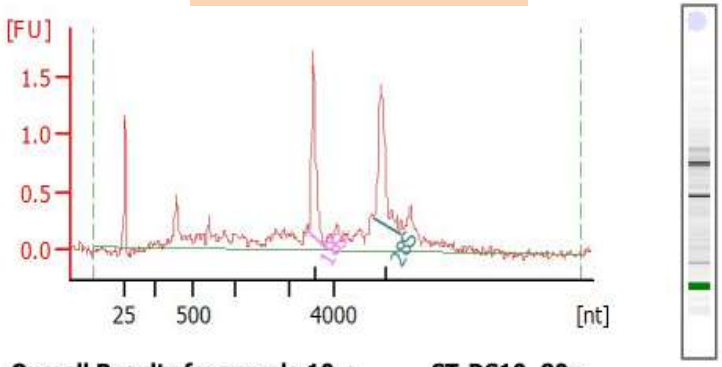
gDNA contamination



Overall Results for sample 7 : ST-DS07_20x

RNA Area:	42.7
RNA Concentration:	101 ng/µl
rRNA Ratio [28s / 18s]:	1.7
RNA Integrity Number (RIN):	7.4 (B.02.07)
Result Flagging Color:	
Result Flagging Label:	RIN: 7.40

Too much salt



Overall Results for sample 10 : ST-DS10_80x

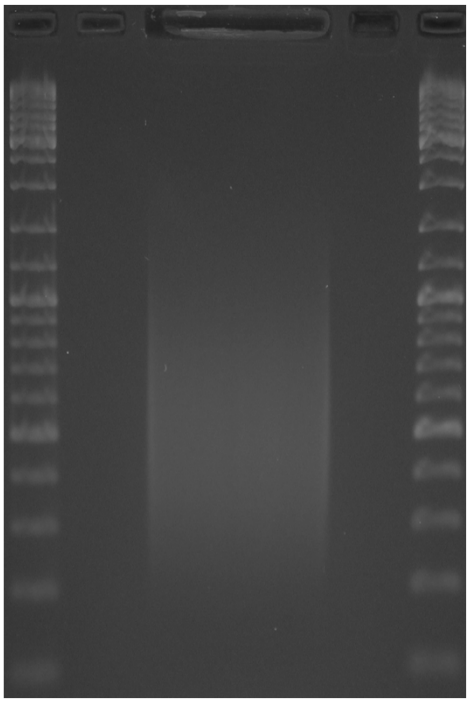
RNA Area:	14.3
RNA Concentration:	34 ng/µl
rRNA Ratio [28s / 18s]:	1.0
RNA Integrity Number (RIN):	7.1 (B.02.07)
Result Flagging Color:	
Result Flagging Label:	RIN: 7.10

Library QC

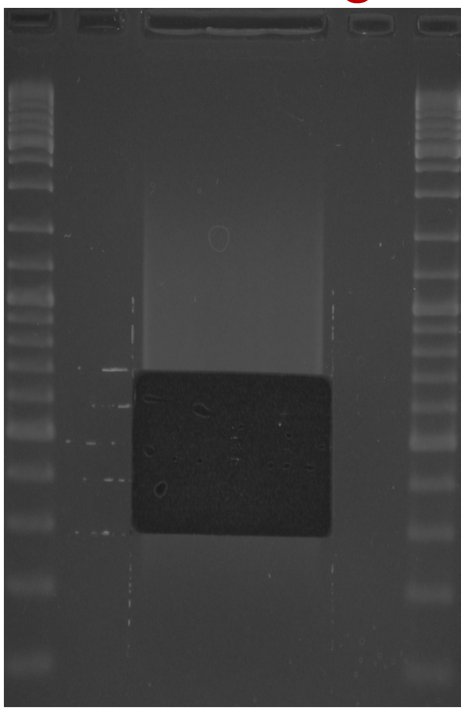
- Gel check / size selection
- BioAnalyzer
- Qubit quantification
- qPCR normalization

Example: Shotgun gDNA library

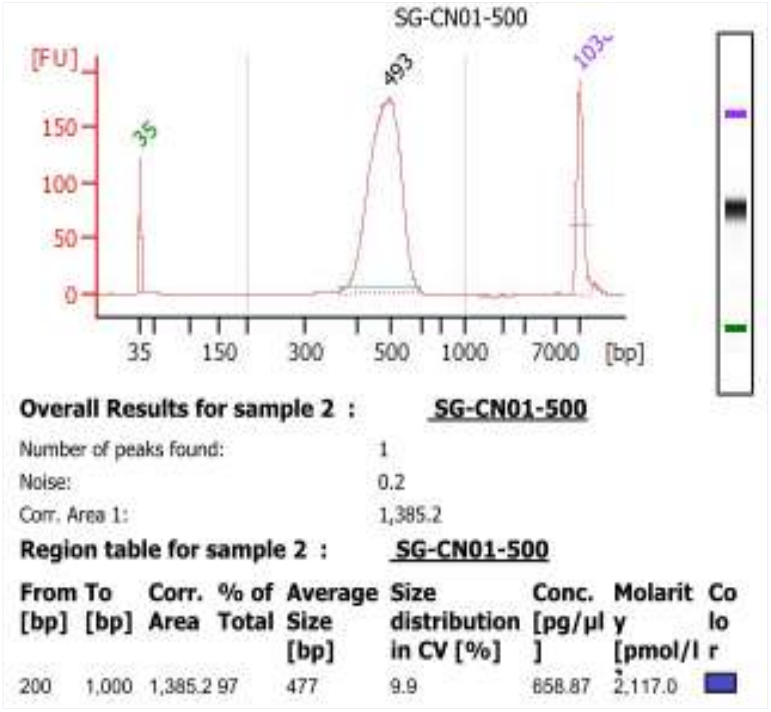
Total profile



Gel sizing



Final PCR library



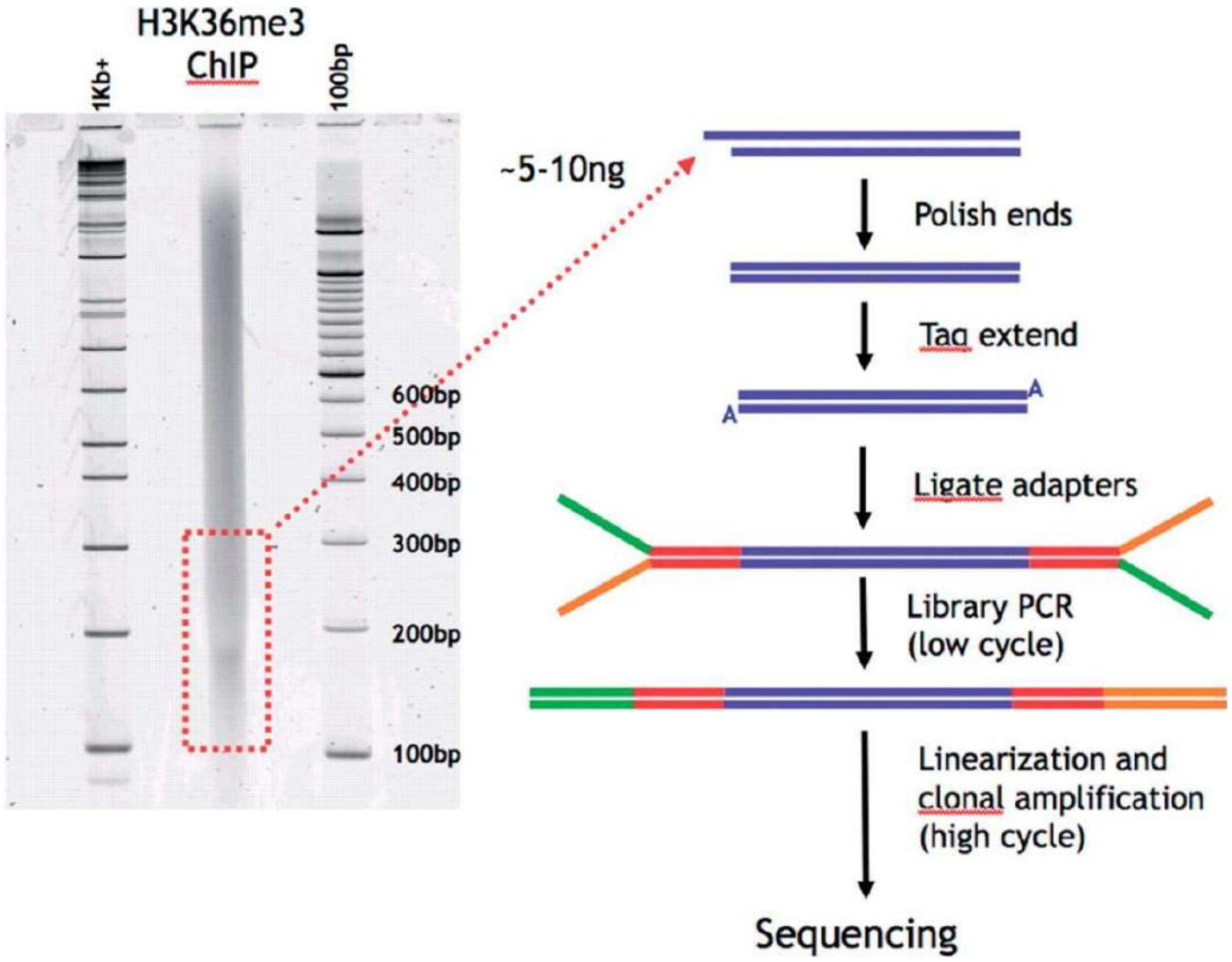
Overall Results for sample 2 : **SG-CN01-500**

Number of peaks found:	1
Noise:	0.2
Corr. Area 1:	1,385.2

Region table for sample 2 : **SG-CN01-500**

From [bp]	To [bp]	Corr. Area	% of Total	Average Size [bp]	Size distribution in CV [%]	Conc. [pg/µl]	Molarity [pmol/l]	Color
200	1,000	1,385.2	97	477	9.9	858.87	2,117.0	

Overview of ChIP-seq construction

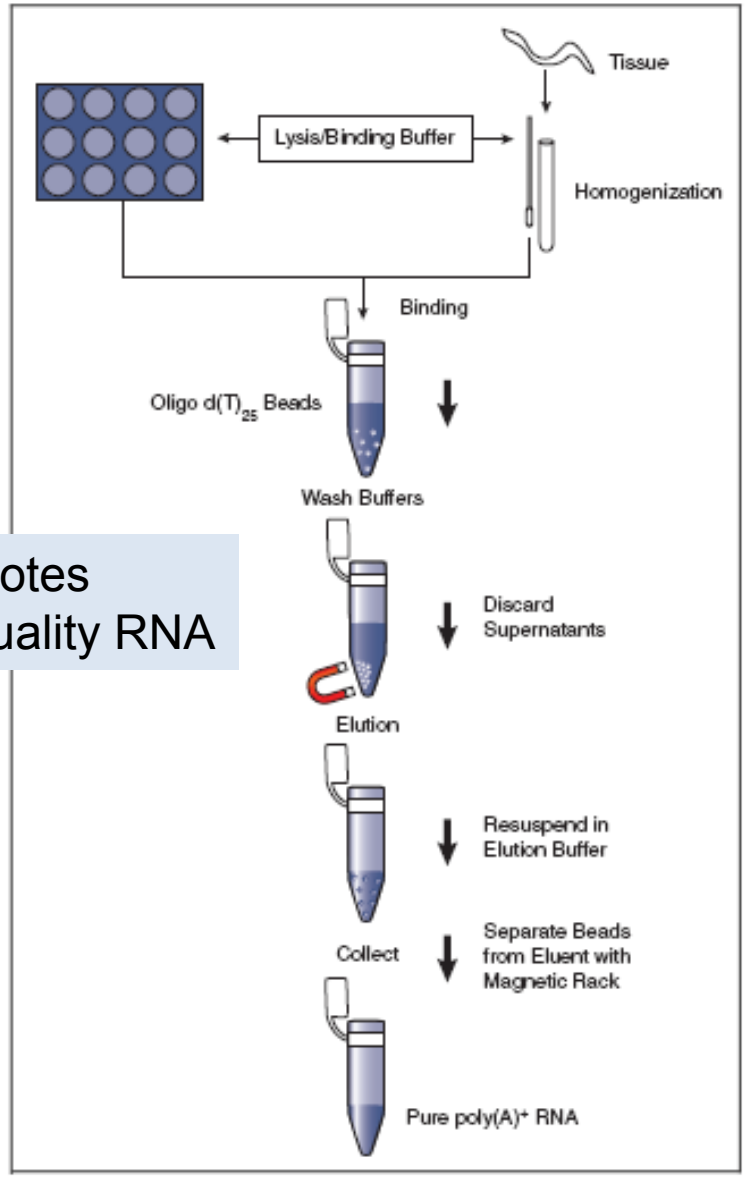


Hirst M , Marra M A Briefings in Functional Genomics 2010;9:455-465

Briefings in Functional Genomics

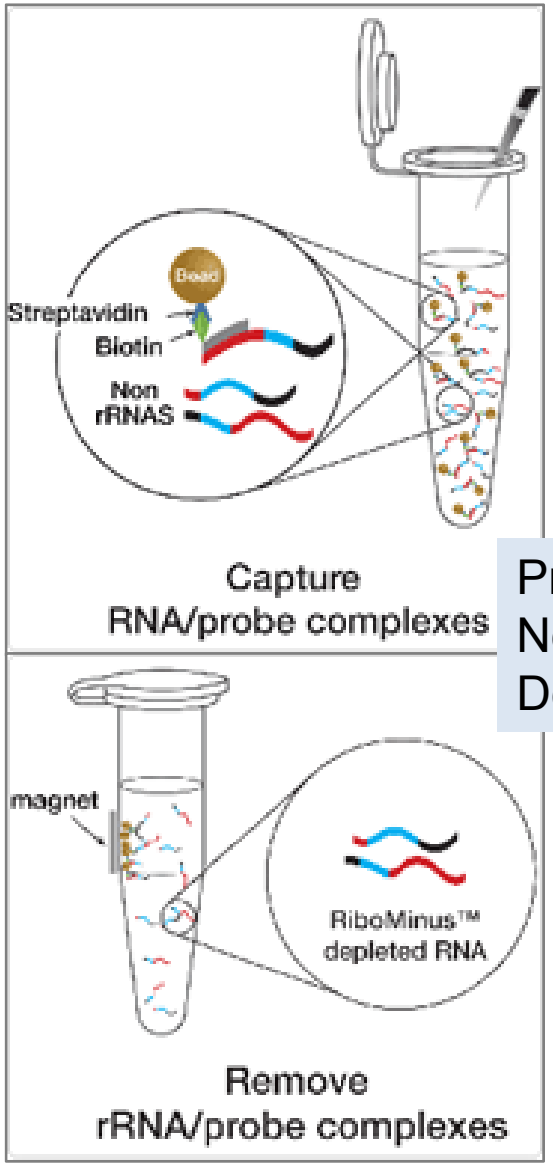
mRNA enrichment methods

Oligo-dT binding



Eukaryotes
High quality RNA

rRNA removal



Prokaryotes
Non-A-tailed RNA
Degraded RNA

Example of RiboZero purificatic

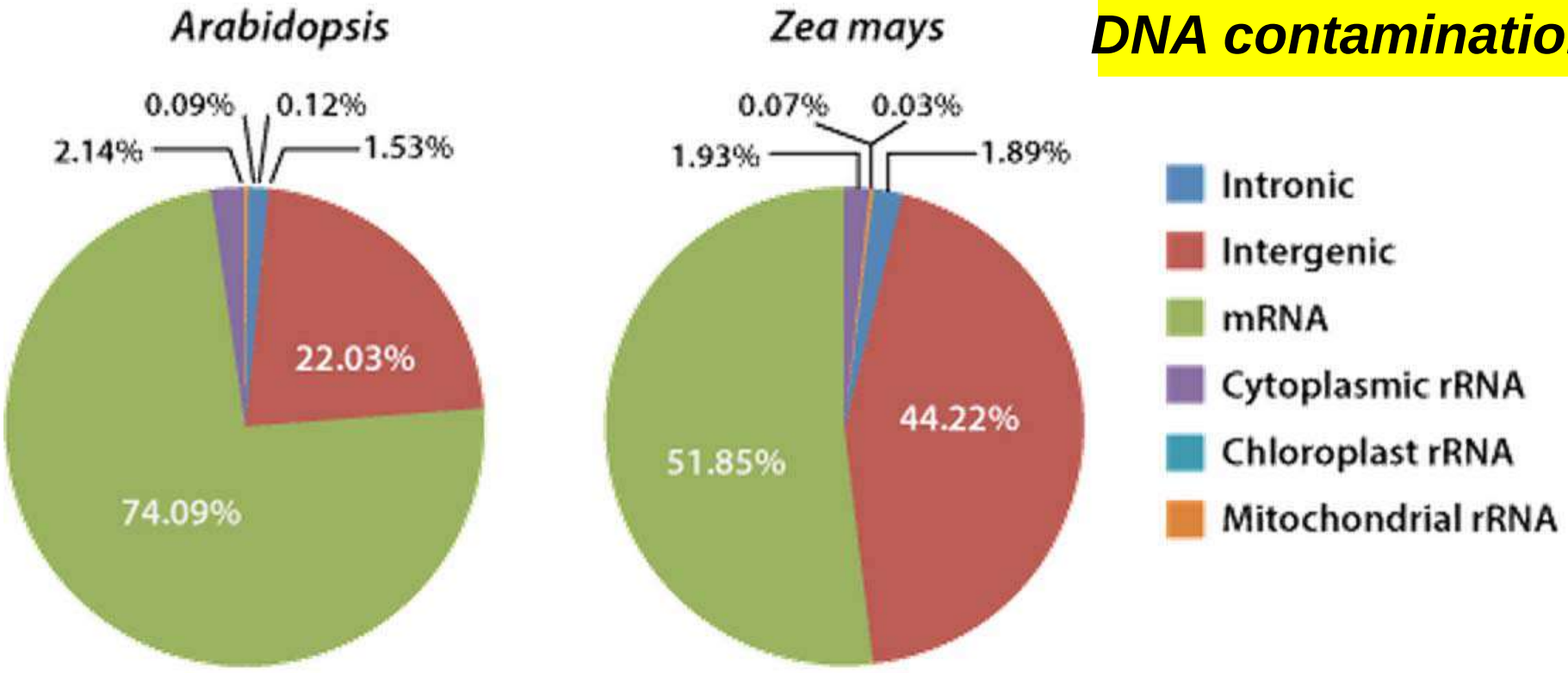
LARGE-SCALE BIOLOGY ARTICLE

RNA Sequencing of Laser-Capture Microdissected Compartments of the Maize Kernel Identifies Regulatory Modules Associated with Endosperm Cell Differentiation[OPEN]

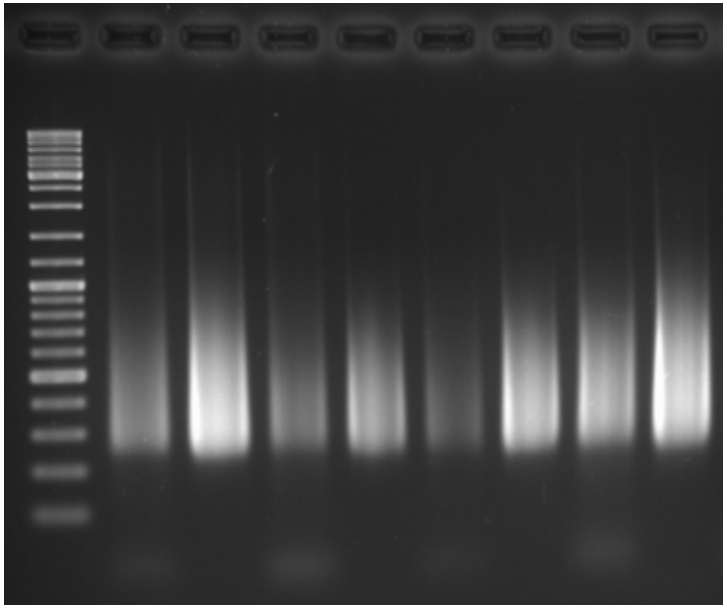
Junpeng Zhan,[a,1] Dhiraj Thakare,[a,1] Chuang Ma,[a,2] Alan Lloyd,[b] Neesha M. Nixon,[b] Angela M. Arakaki,[b] William J. Burnett,[b] Kyle O. Logan,[b] Dongfang Wang,[a,3] Xiangfeng Wang,[a,4] Gary N. Drews,[b] and Ramin Yadegari[a,5]

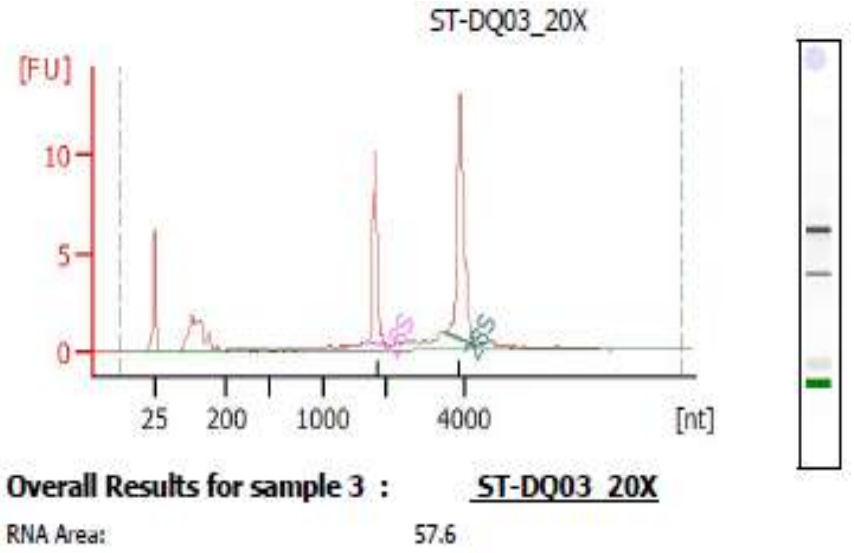
[a] School of Plant Sciences, University of Arizona, Tucson, Arizona 85721
[b] Department of Biology, University of Utah, Salt Lake City, Utah 84112

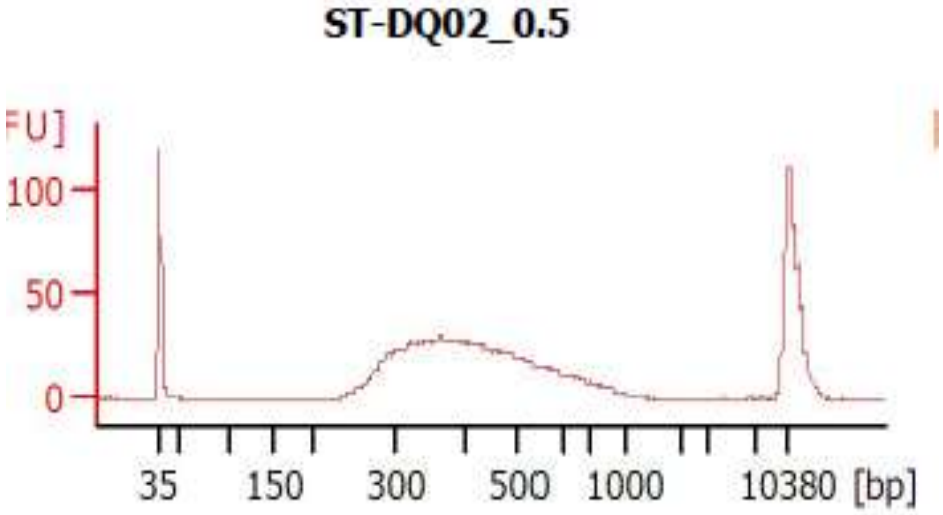
Genome divergence
RNA integrity
DNA contamination



mRNA-seq prep
(no smRNA)



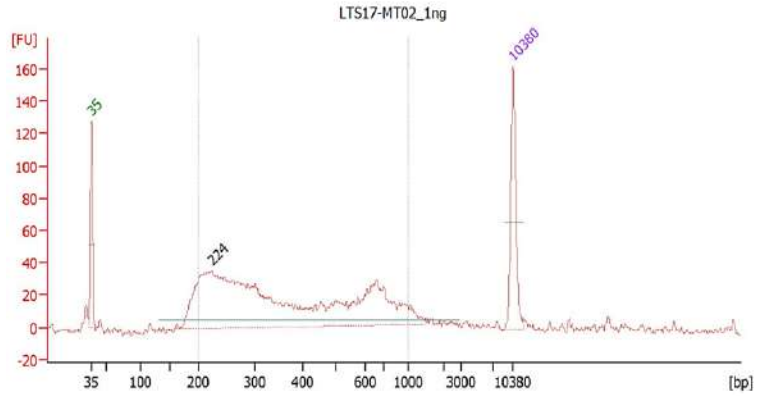




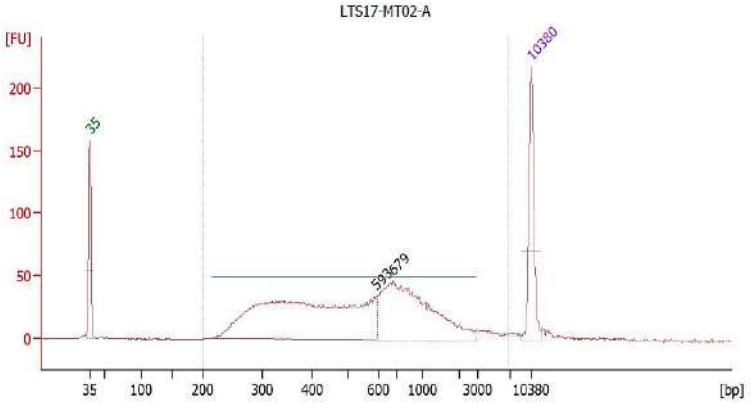
Problematic libraries

Different library output from various prep methods

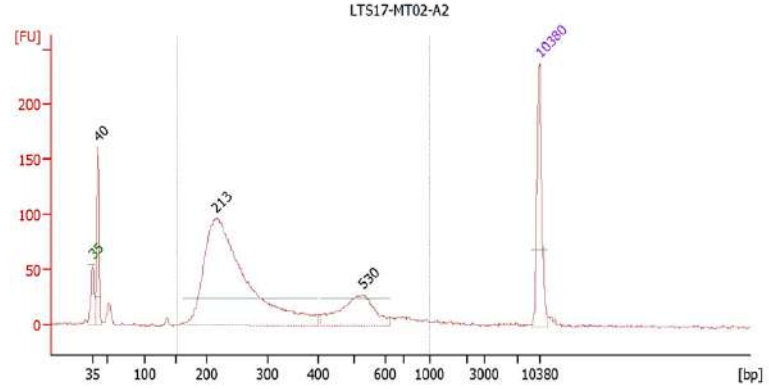
1.KAPA kit (polyA)



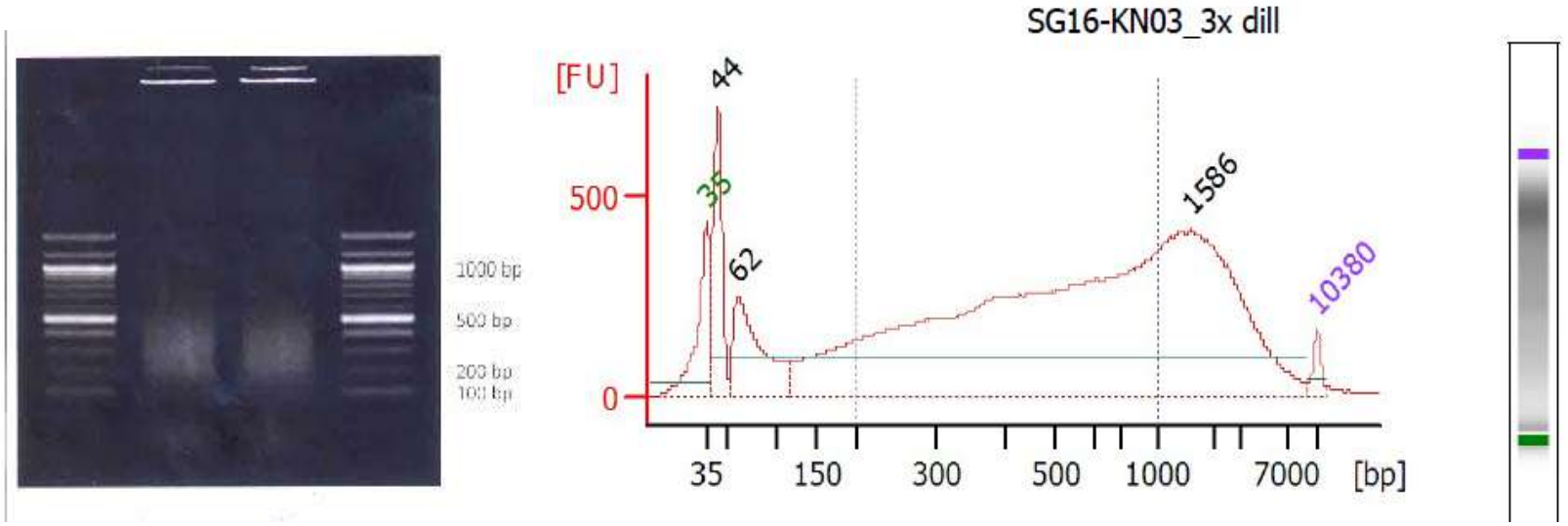
2. Ribozero + PolyA

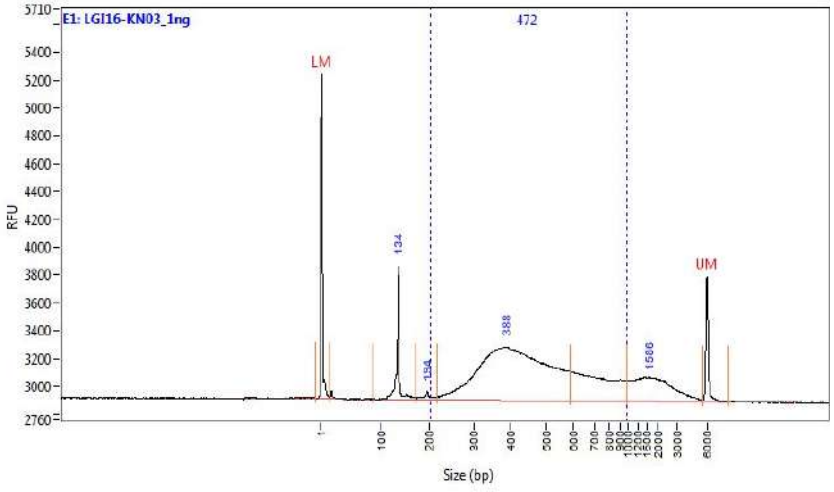


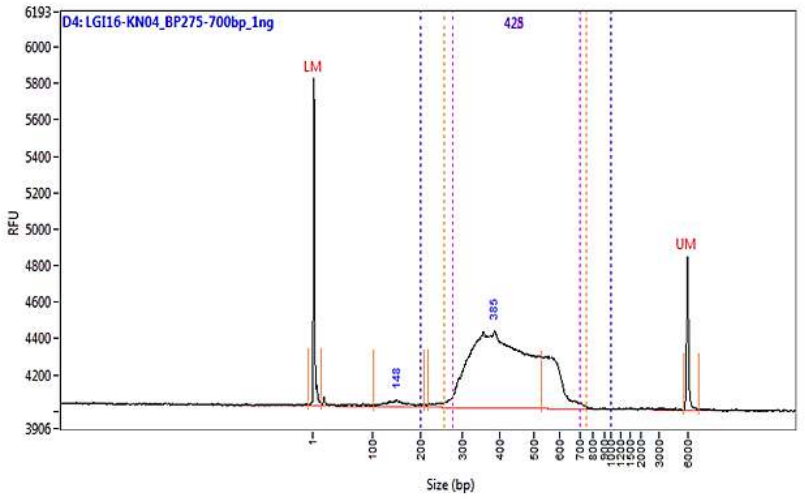
3. DNaseI treatment + PolyA



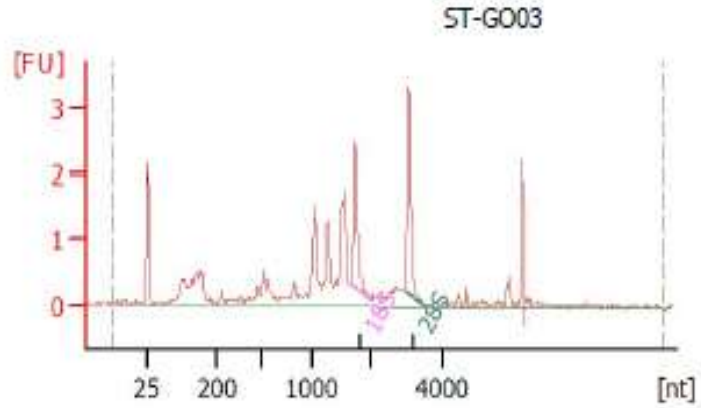
1. Adapter Dimer, large size fragments







2. rRNA contamination in RNA-seq

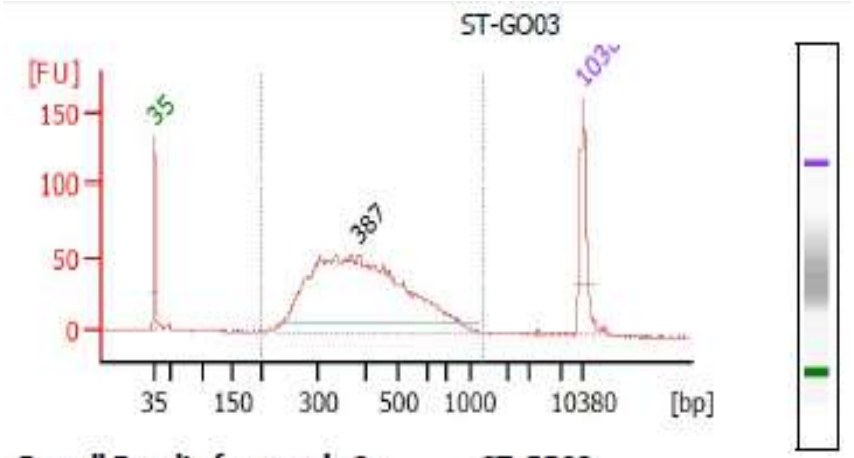


Overall Results for sample 3 : ST-GO03

RNA Area:	25.1
RNA Concentration:	38 ng/µl
rRNA Ratio [28s / 18s]:	1.4
RNA Integrity Number (RIN):	6.9 (B.02.07)
Result Flagging Color:	
Result Flagging Label:	RIN: 6.90

Fragment table for sample 3 : ST-GO03

Start Size [nt]	End Size [nt]	% of total Area	Area	Name
1,665	2,020	8.9	2.2	18S
2,922	3,472	12.4	3.1	28S



Overall Results for sample 3 : ST-GO03

Number of peaks found:	1
Noise:	0.6
Corr. Area 1:	1,452.2

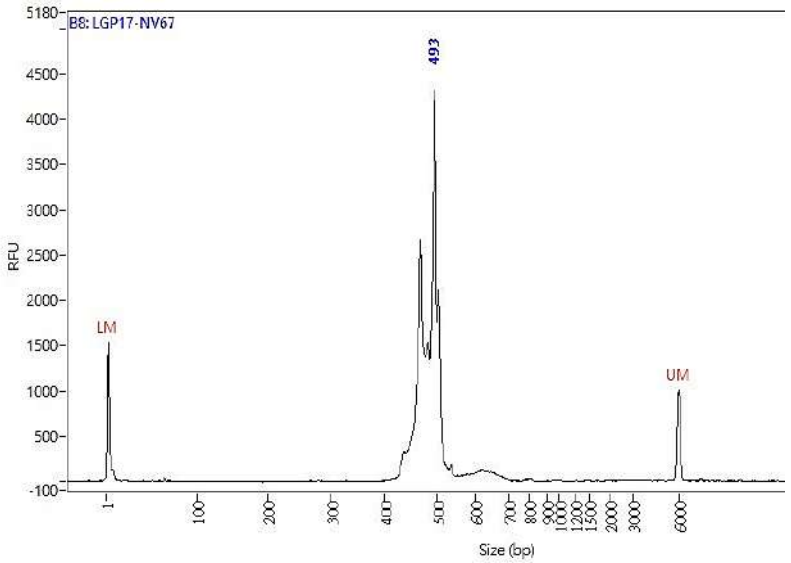
Region table for sample 3 : ST-GO03

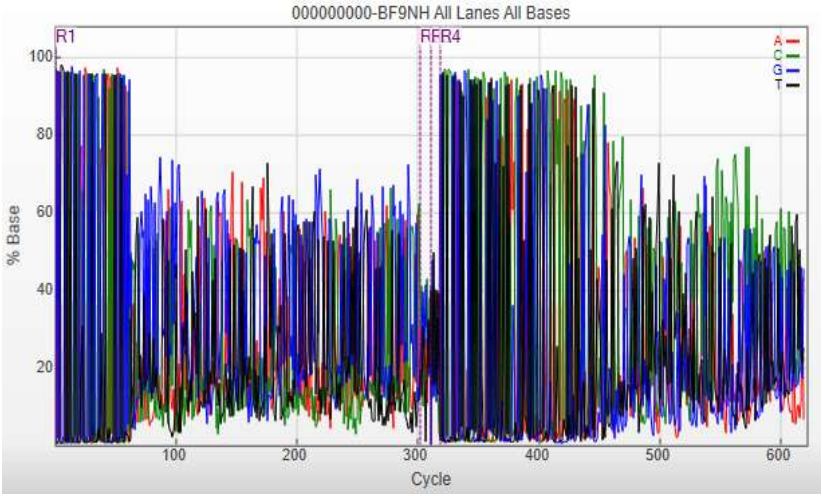
From [bp]	To [bp]	Corr. Area	% of Total	Average Size [bp]	Size distribution in CV [%]	Conc. [pg/µl]	Molarity [pmol/l]	Color
200	1,358	1,452.2	93	425	34.1	951.93	3,858.3	

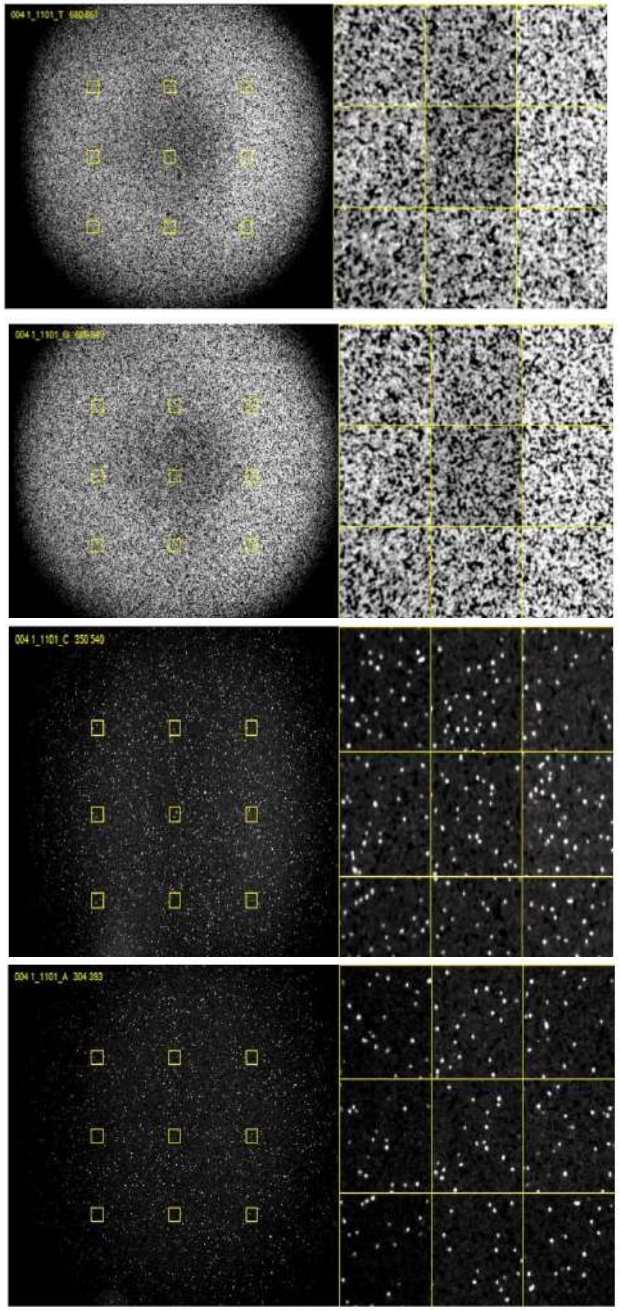
3. biased imaging – amplicon

(over-crowded loading);

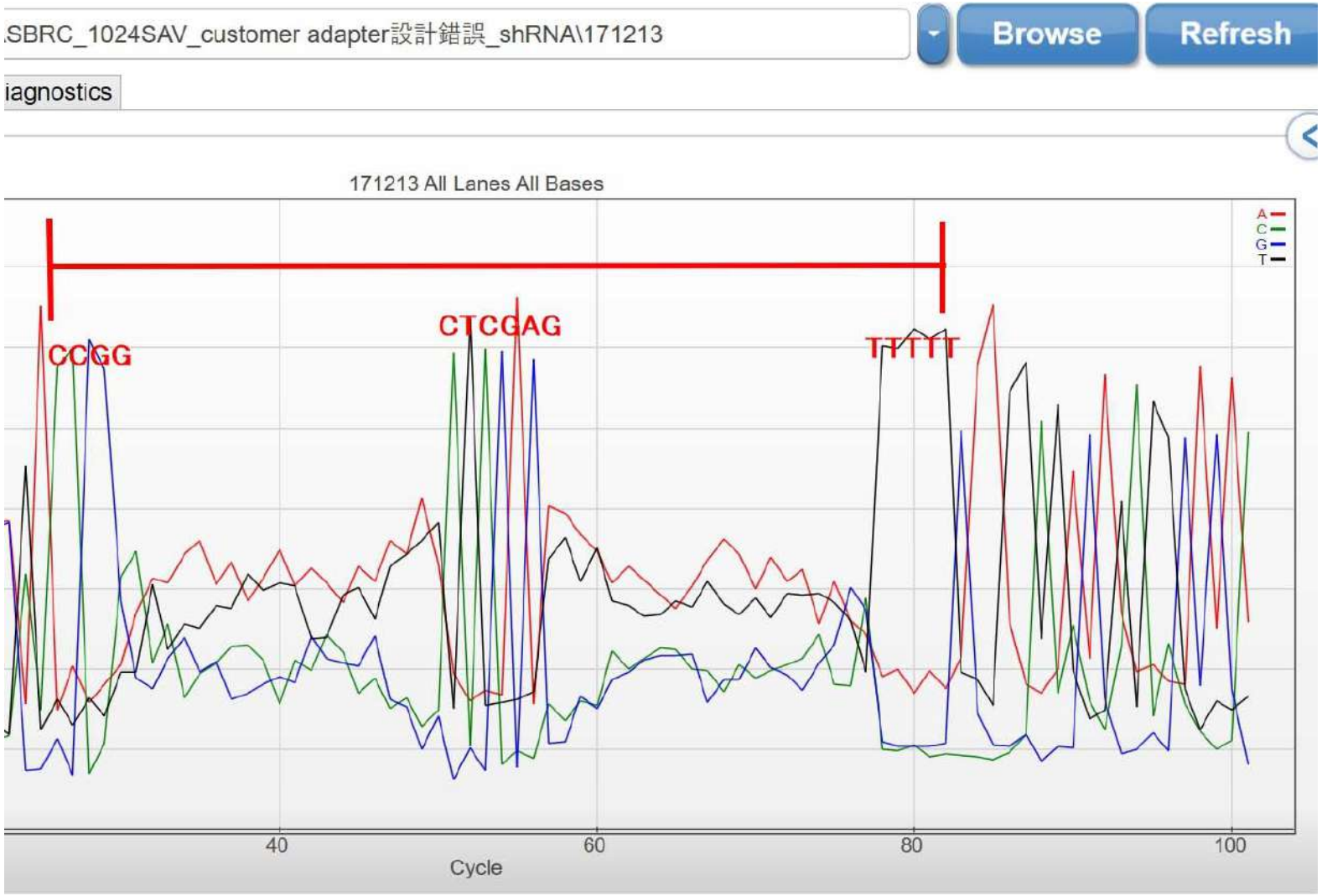
need to spike-in PhiX library







4. wrong primer design



5. barcode mixture with only 1 difference (result in data loss by default)

Primers	Barcodes
[I7]	
N701	TCGCCTTA
N702	CTAGTACG
N703	TTCTGCCT
N704	GCTCAGGA
N705	AGGAGTCC
N706	CATGCCTA
N707	GTAGAGAG
N708	CCTCTCTG
N709	AGCGTAGC
N710	CAGCCTCG
N711	TGCCTCTT
N712	TCCTCTAC
N714	GCTCATGA
N715	ATCTCAGG
[I5]	
N/S/E501	TAGATCGC
N/S/E502	CTCTCTAT
N/S/E503	TATCCTCT
N/S/E504	AGAGTAGA
N/S/E505	GTAAGGAG
N/S/E506	ACTGCATA
N/S/E507	AAGGAGTA
N/S/E508	CTAAGCCT

If feeding total barcode list, demultiplex would fail!!

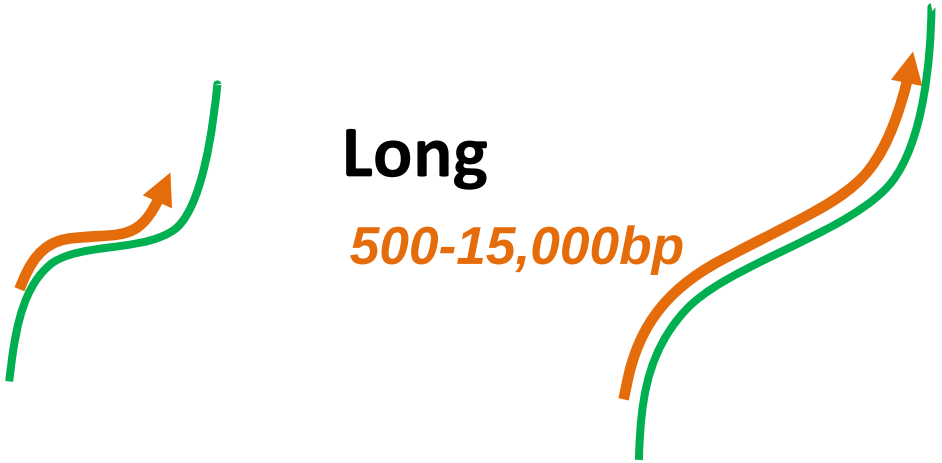
Demultiplex	Read	Clusters PF (%)	Yield (Mb)	# of Reads
perfect match	Read 1	76.41 +/- 2.04	7,878	26,190,976
	Read 2	76.41 +/- 2.04		

Demultiplex	Read	Clusters PF (%)	Yield (Mb)	# of Reads
One mismatch	Read 1	76.41 +/- 2.04	6,353	21,104,720
	Read 2	76.41 +/- 2.04		

VI. Data types, preprocessing, and quality control

Types and Characteristics of NGS Reads

- **Read length:** **Short** *50-300bp* **Long** *500-15,000bp*

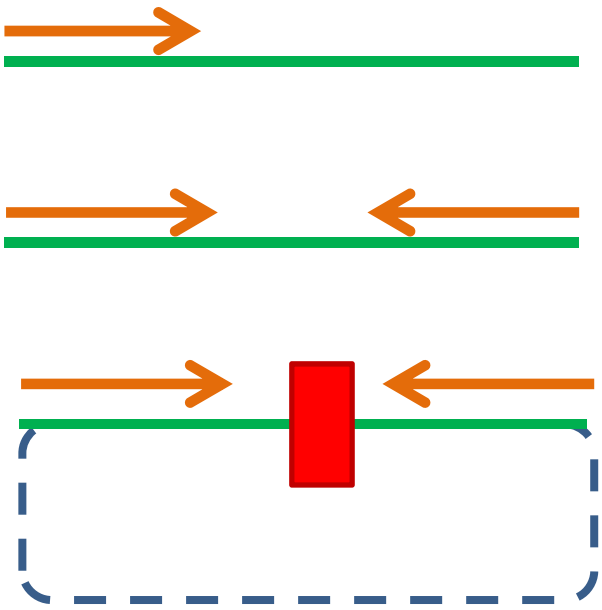


- **Read types:**

SR (single end) 50bp-20kb

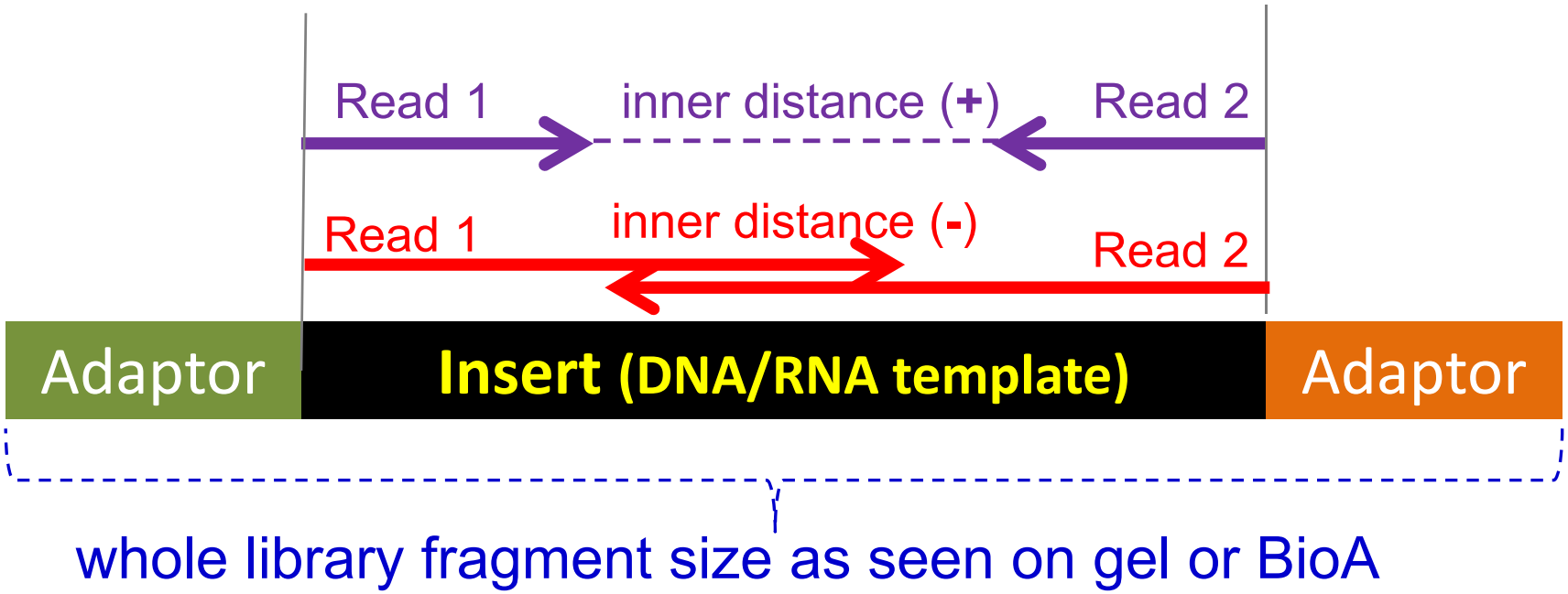
PE (paired-end) 50-300 bp; 1~1.5 kb jump

MP (mate-pair) 50-300bp; 2~15kb jump



Fragment v. Insert v. Inner Distance

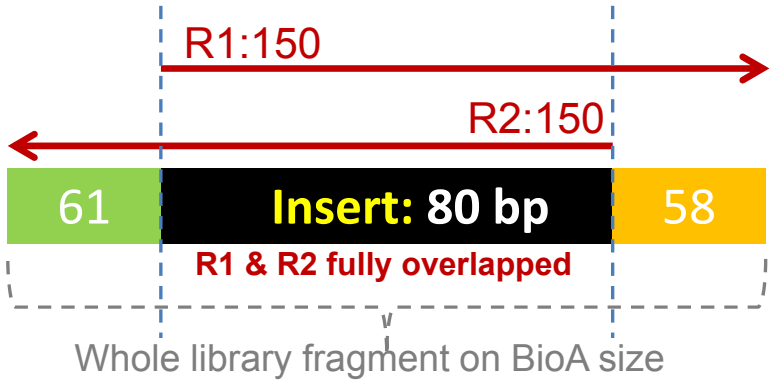
A. **Library fragment** = length detected by BioA/agarose gel
B. **Insert size** = DNA or RNA template; no adaptors included
C. **Inner distance** = distance b/w the end base of R1 and R2
    1. Positive distance = gapped ends
    2. Negative distance = overlapped ends



Insert size vs Library Fragment Size: PE2*150

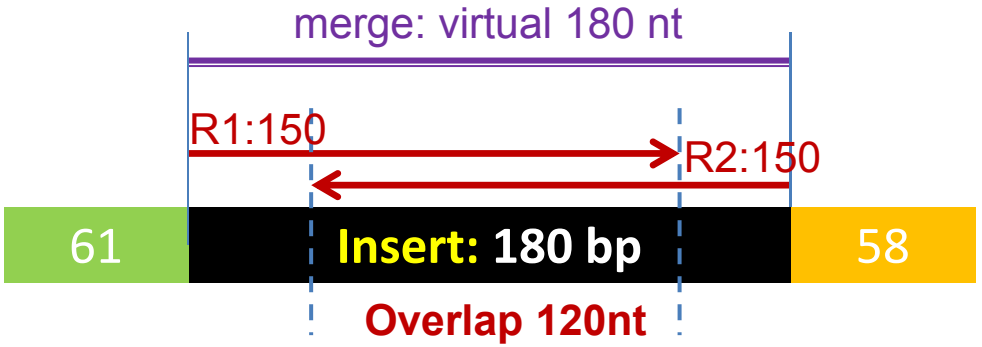
200-bp fragment (BioA):

- *RNA Insert = 80 bp*
- *R1 & R2 fully overlapped*
- *Seq. runs into adaptor (adaptors fully covered)*



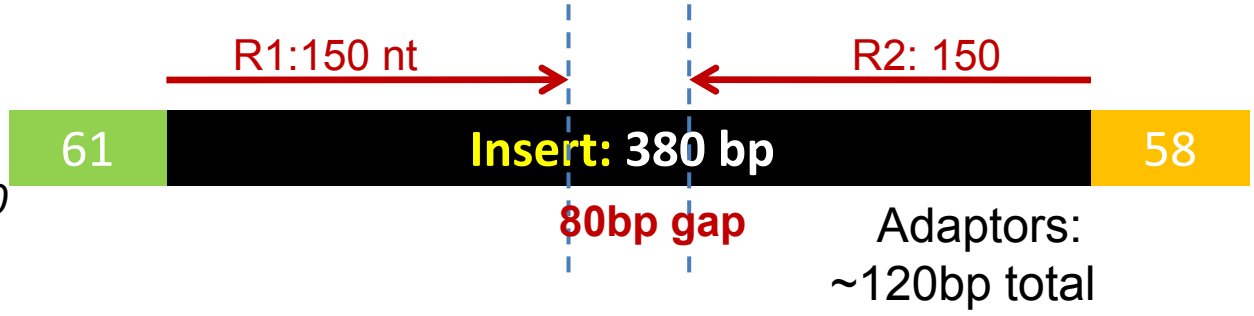
300-bp fragment (BioA):

- *RNA Insert = 180 bp*
- *Read end overlapped 120bp*
- *No reading of adaptors*



500-bp fragment:

- *RNA Insert = 380 bp*
- *End gapped 80 bp by PE2*150*
- *No reading of adaptors*



Illumina Read – fastQ

Index sequence

no control

Y/N: failing PF or not

Sequence head *Machine ID, FC ID* *Lane ID* *Read1 or Read2*

```
@HWI-D00368:32:H8R31ADXX:2:1101:2034:2140 1:N:0:CAGATC
TTTGNCGAGAACTGGAATTGAACCAATATTTAAGTCTTACAAGGAATTCGTTTTAAC
+
@@@F#2ADFDHHHJJJJJGHHIIJIIJJJIJGGJHEIIJIJIIJIIJJJIJJJJIGI
```

Q-score header

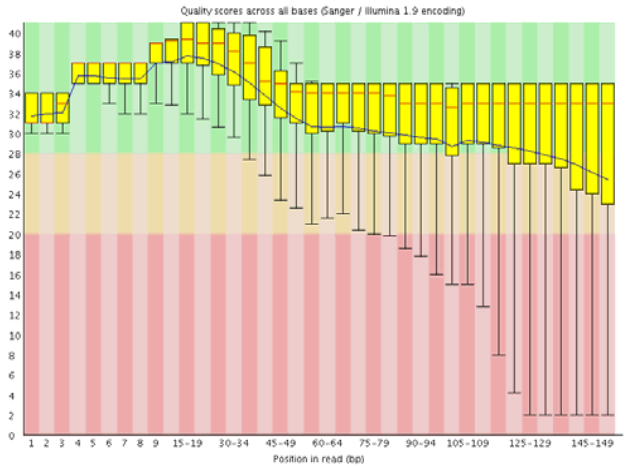
Base quality: error probability

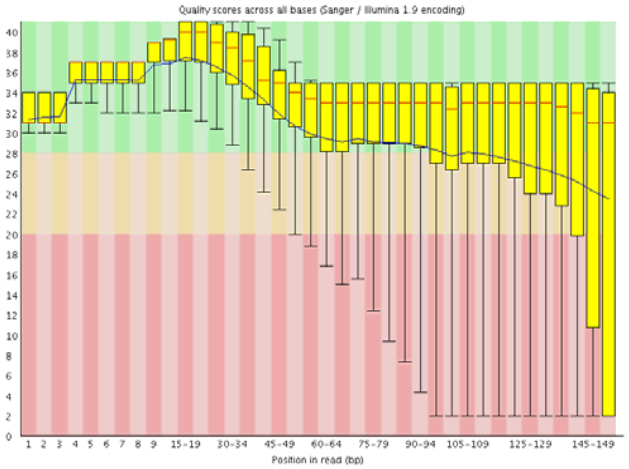
P by Q = [− 10 * $\log_{10}(P)$]

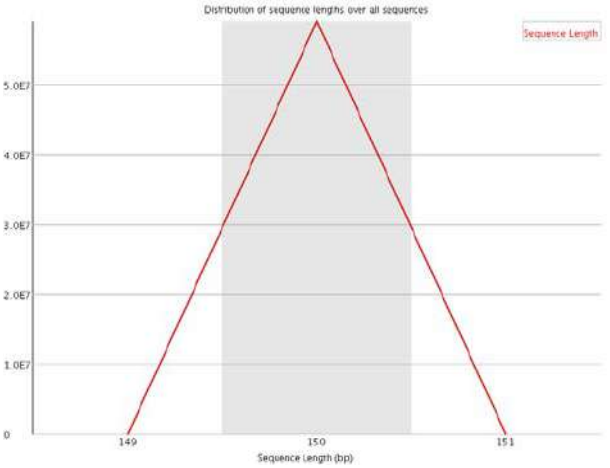
Phred Score Q	Error probability
10	1 in 10
20	1 in 100
30	1 in 1,000
40	1 in 10,000

Adapter Trimming Result of HiSeq genomic PE Reads

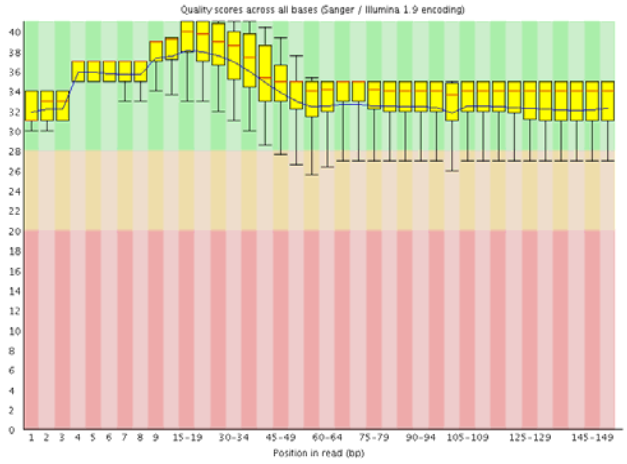
HiSeq.CDC10.raw150

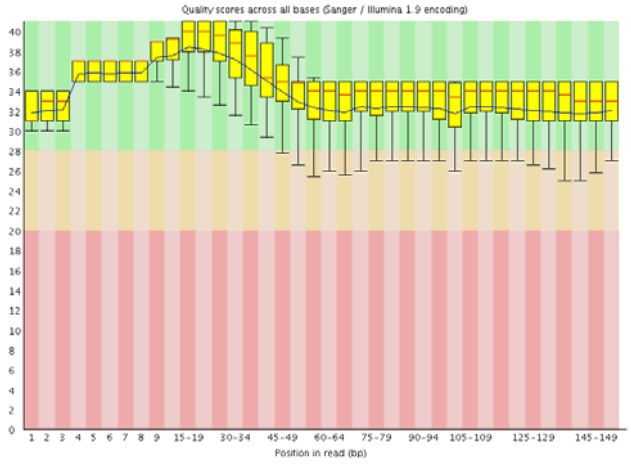


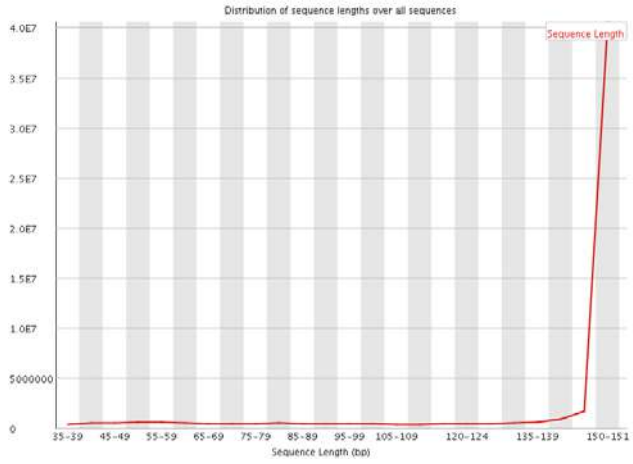




HiSeq.CDC10.raw150 (after adapter trimming by Trimmomatic)

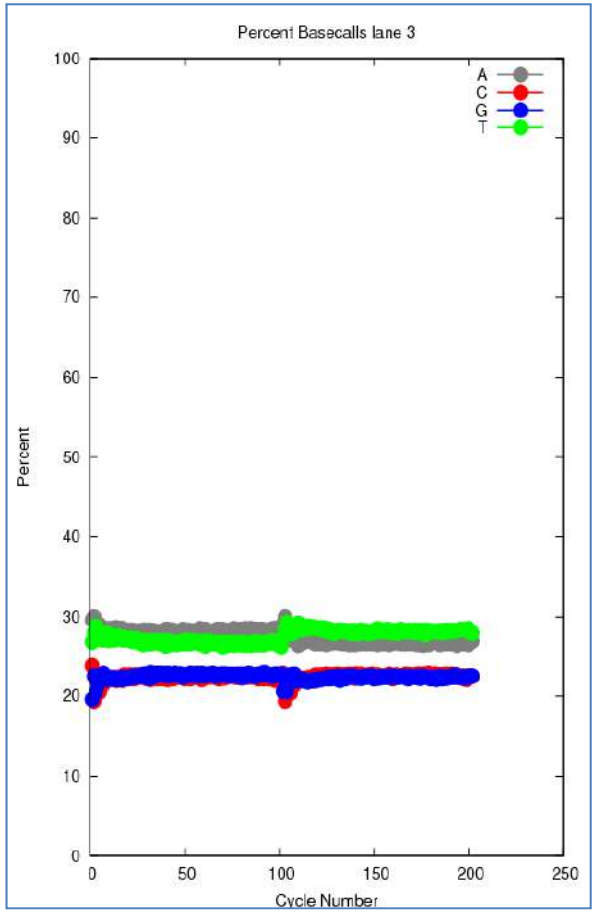




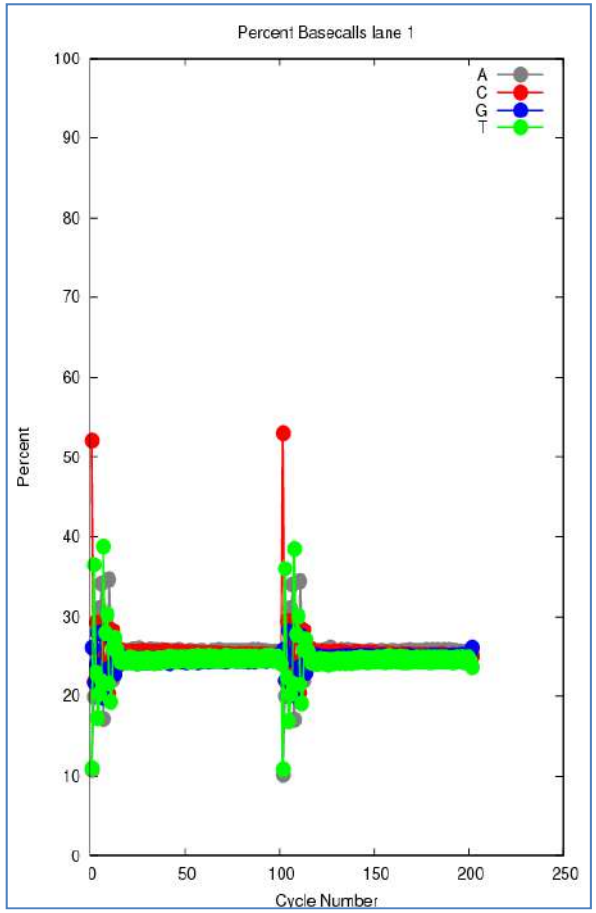


IVC plots (Intensity vs Cycle)

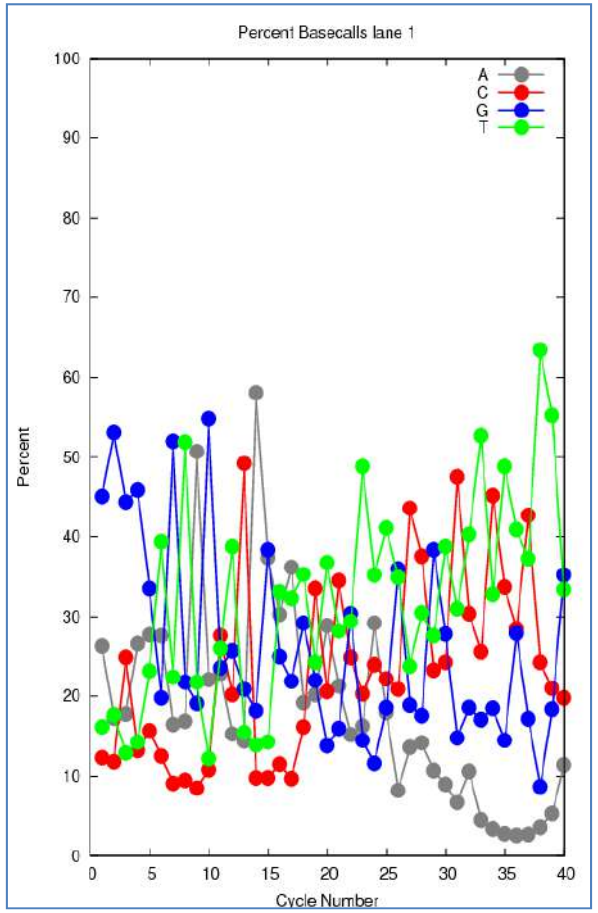
Normal GC%



mRNA-seq

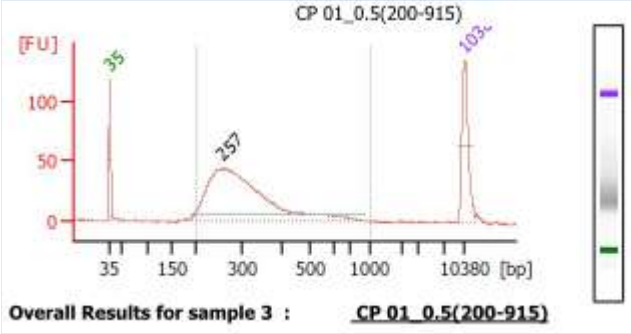


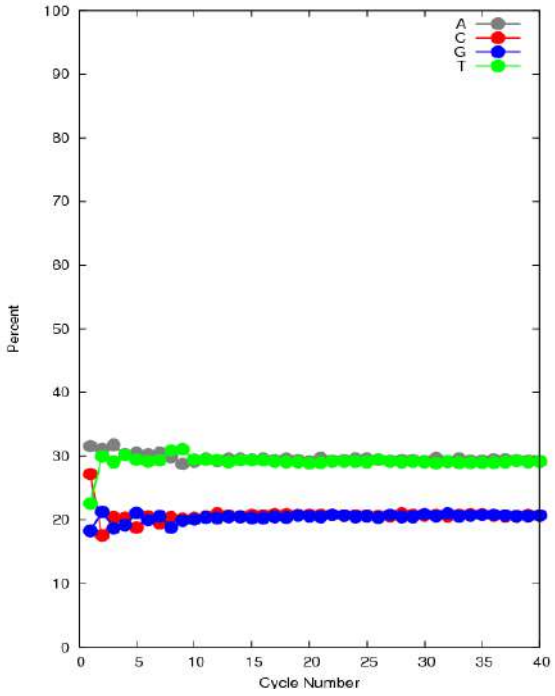
smRNA-seq



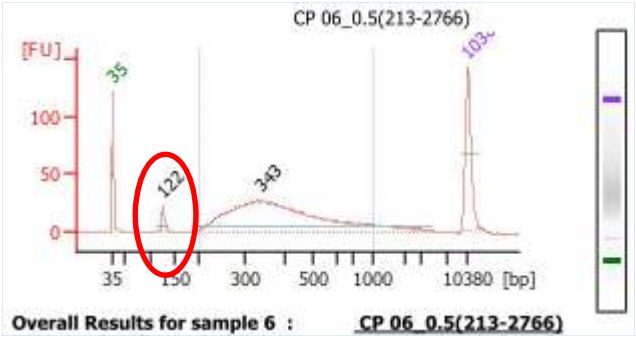
IVC plots (Intensity vs Cycle)

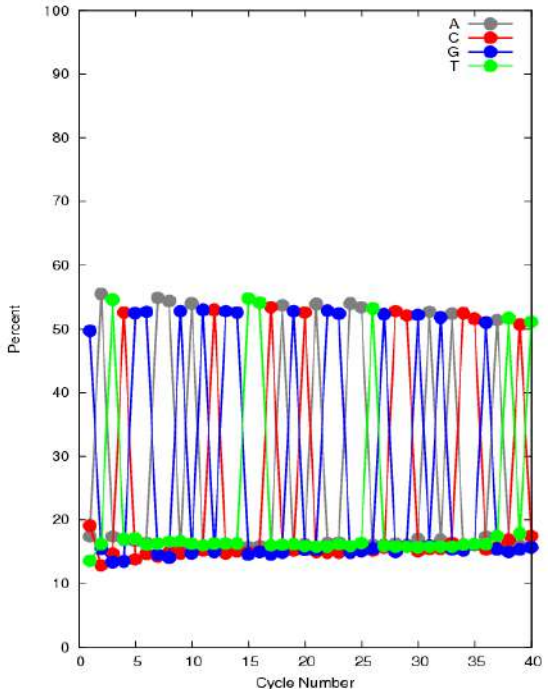
Normal input- Little bias





Low input - Strong bias





VII. Extended / Advanced NGS technologies

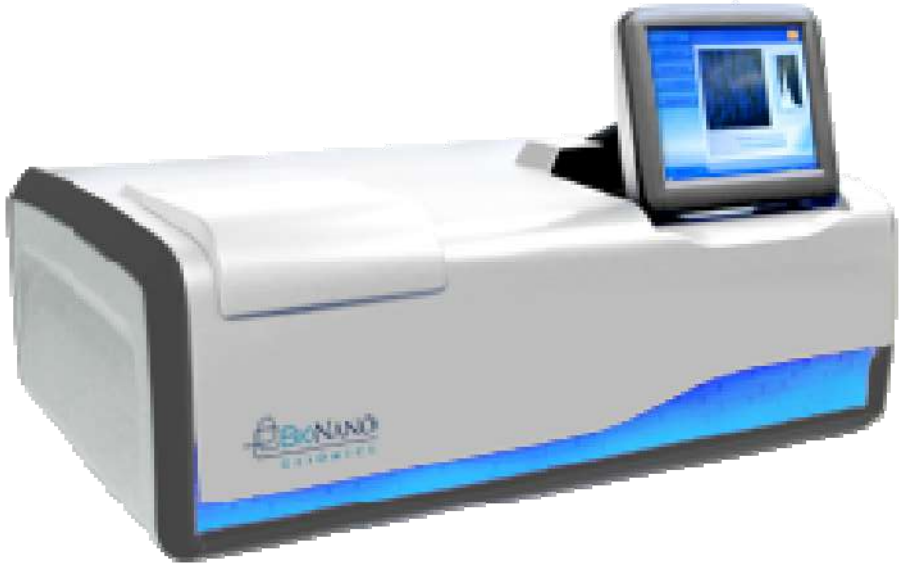
Various NGS approaches

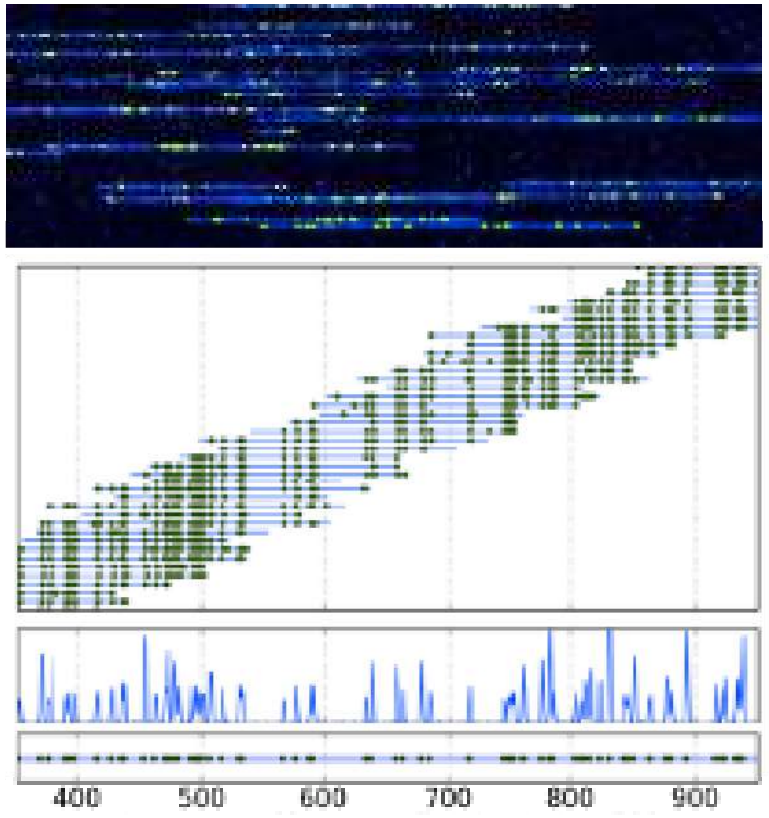
1. Hybrid NGS
   - BioNano (optical mapping)
   - Hi-C (high-order chromatin folding)

2. Single-cell technologies

BioNano:

Build reference genome
detect SV
visualize CNV



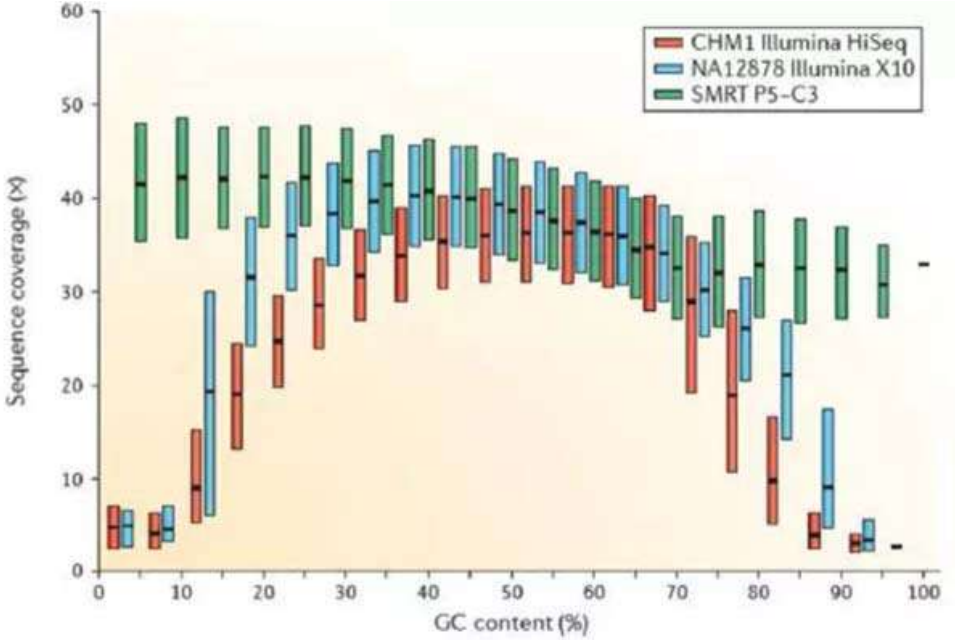


Assembly by chr RE fragment patterns.

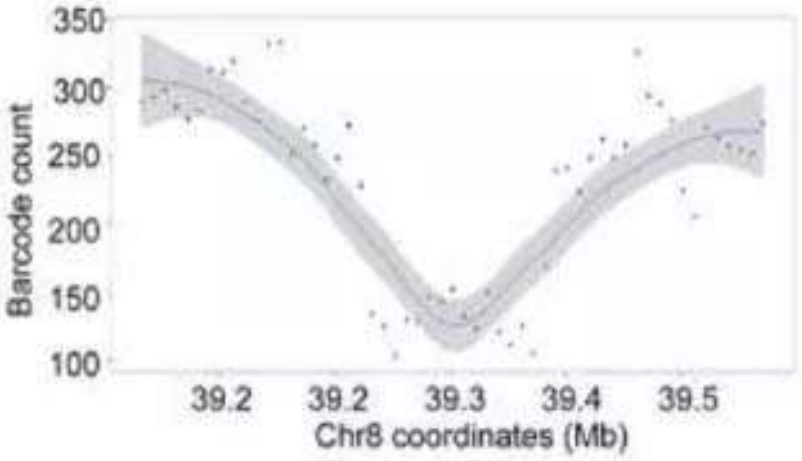
https://youtu.be/0-NFpOPADiQ

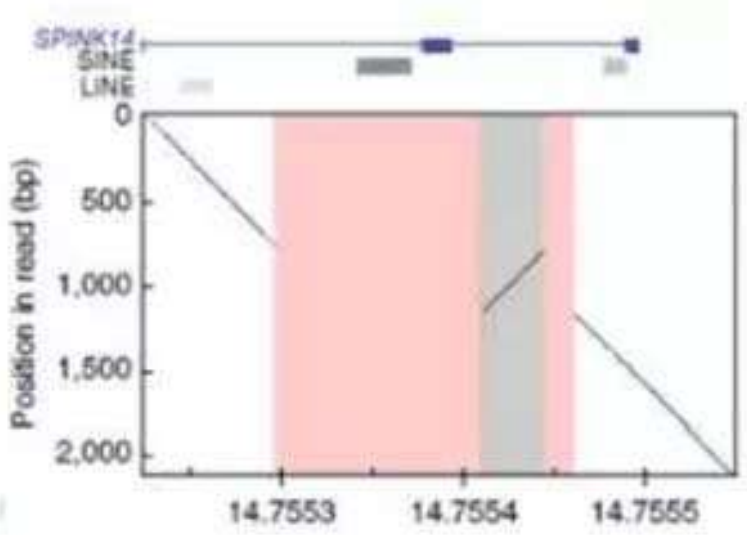
Long-read NGS for High GC%, long SV

Coverage at GC% biased regions

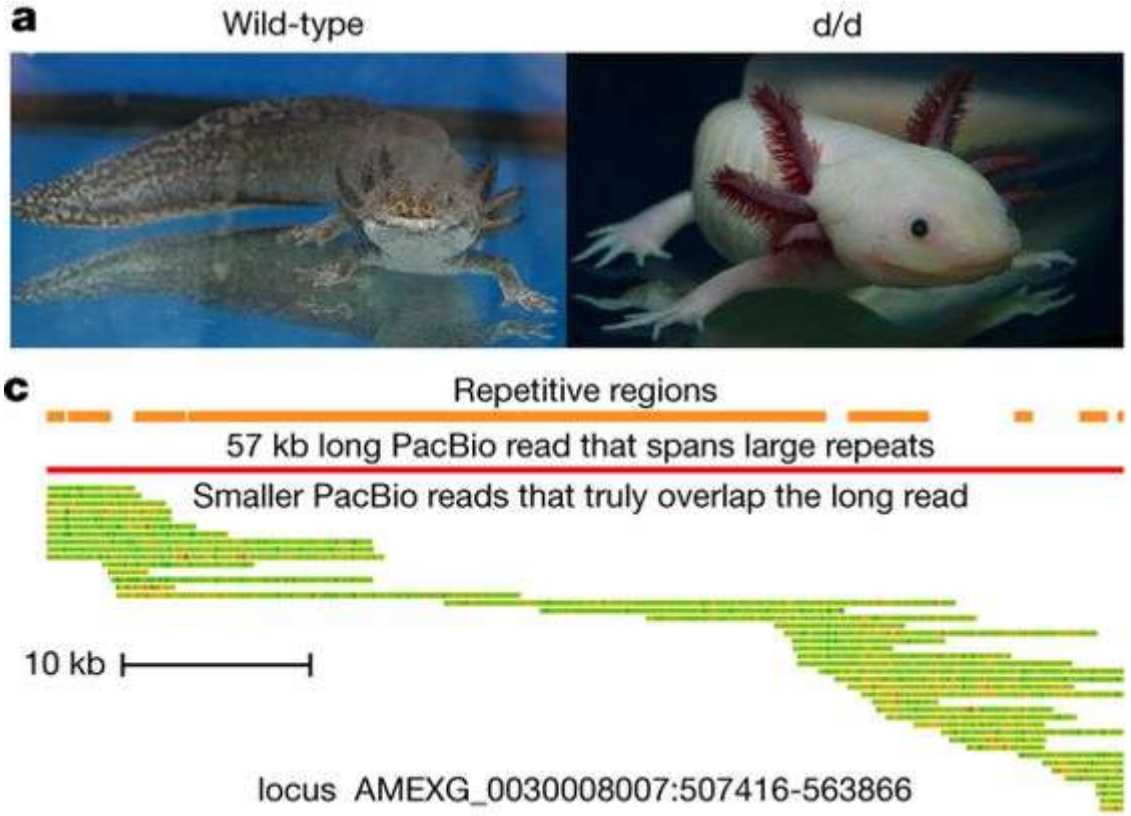


Resolution of SV analysis





PB - *de novo* large genome: Axolotl

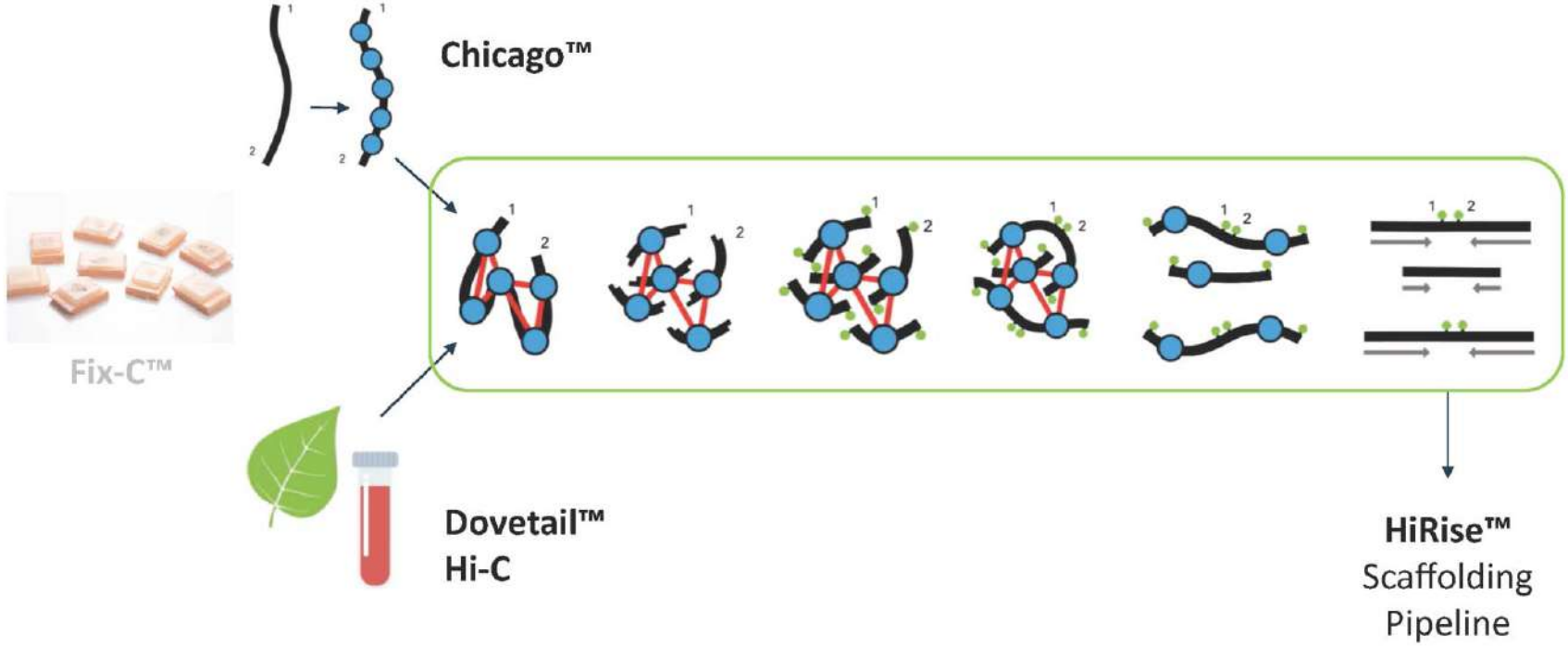


Genome assembly strategy

Metrics	Axolotl (*A. mexicanum*)
Assembly size (Gb)	**32.4 Gb** (28.4 in contigs)
Genome size	32 Gb
Chromosomes	14
Sequencing technology	PacBio; Optical map
Coverage	**32**x
Assembler	MARVEL
Contig N50)	216,277 bp
Number of contigs	217,461
Scaffold N50	3,052,786 bp
# Scaffolds	**125,724**

Nature : The axolotl genome and the evolution of key tissue formation regulators
S Nowoshilow *et al. Nature* **554,** 50–55 (2018)

Hi-C: Chromosome Proximity Ligation



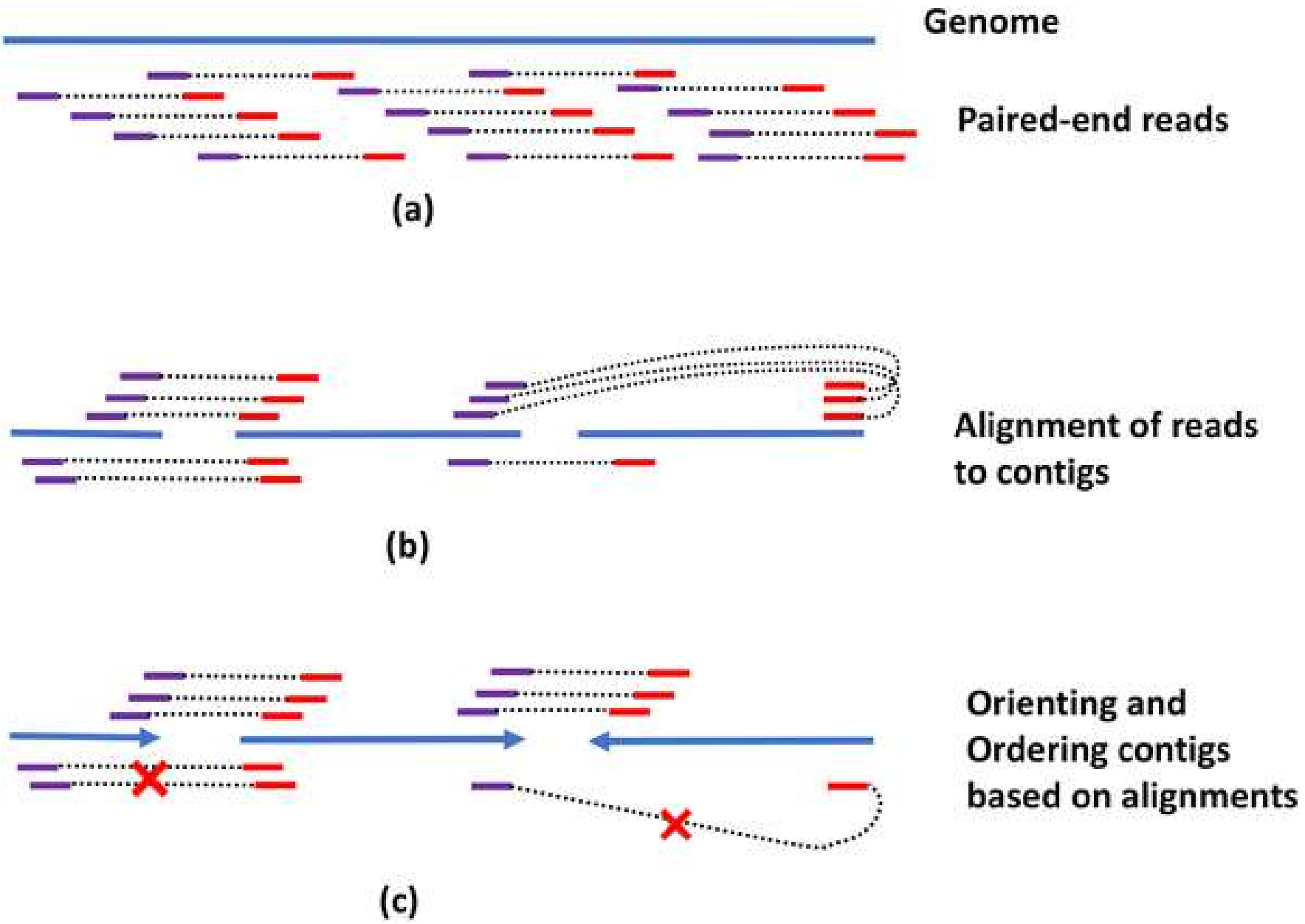
Chicago vs Hi-C

Chicago

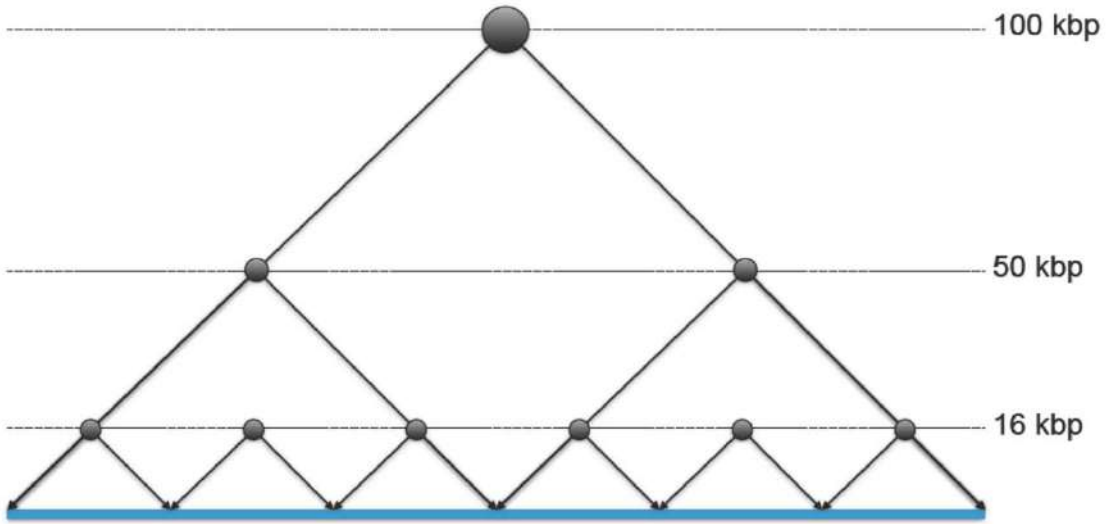
- In vitro crosslinking
- Scaffolding
- Order and orientation
- Denovo assembly with N50 < 1 Mbp (at least > 20kb)

Hi-C

- In vivo crosslinking
- Scaffolding
- Chromosome scaffolding
- Denovo assembly with N50 > 1 Mbp


Genome
Paired-end reads
(a)
Alignment of reads
to contigs
(b)
Orienting and
Ordering contigs
based on alignments
(c)

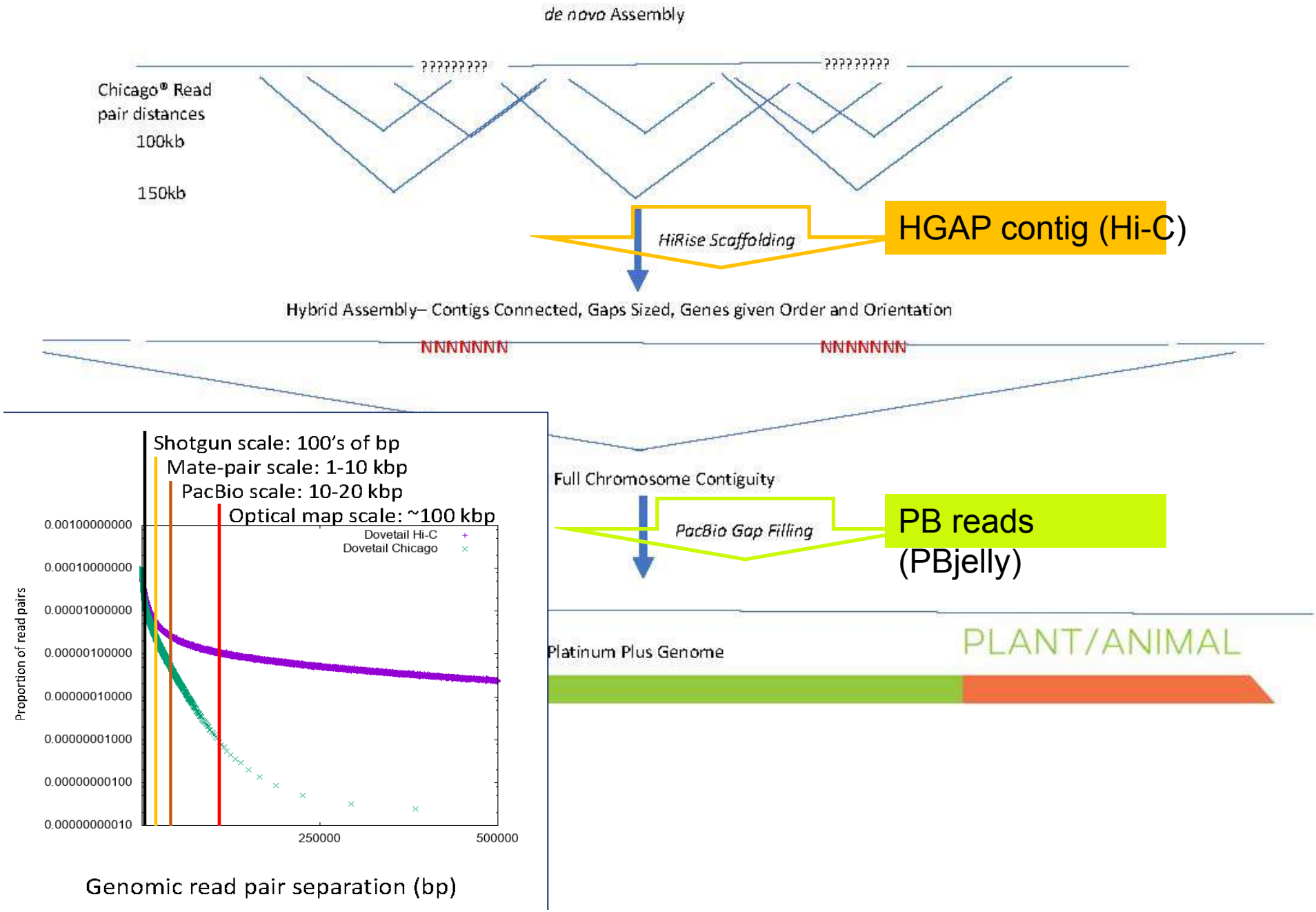
Methods:
Proximity Ligation Approaches



Read pairs of varying lengths allow you to create a scaffold of the genome

PLANT/ANIMAL

Improvements to achieve highest quality assembly

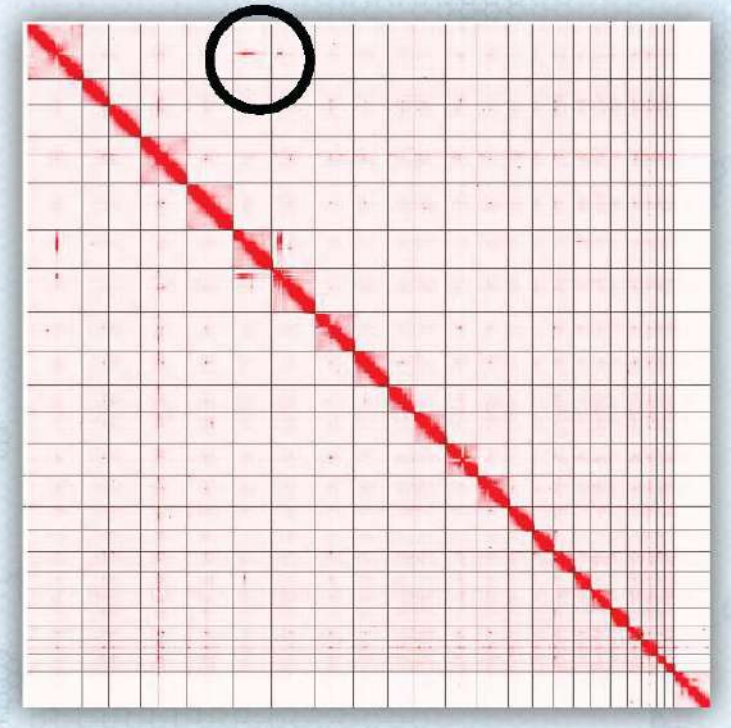


Hi-C: long range SV; detect assembly error

Flagging scaffolding issues with Hi-C

Clint the chimpanzee

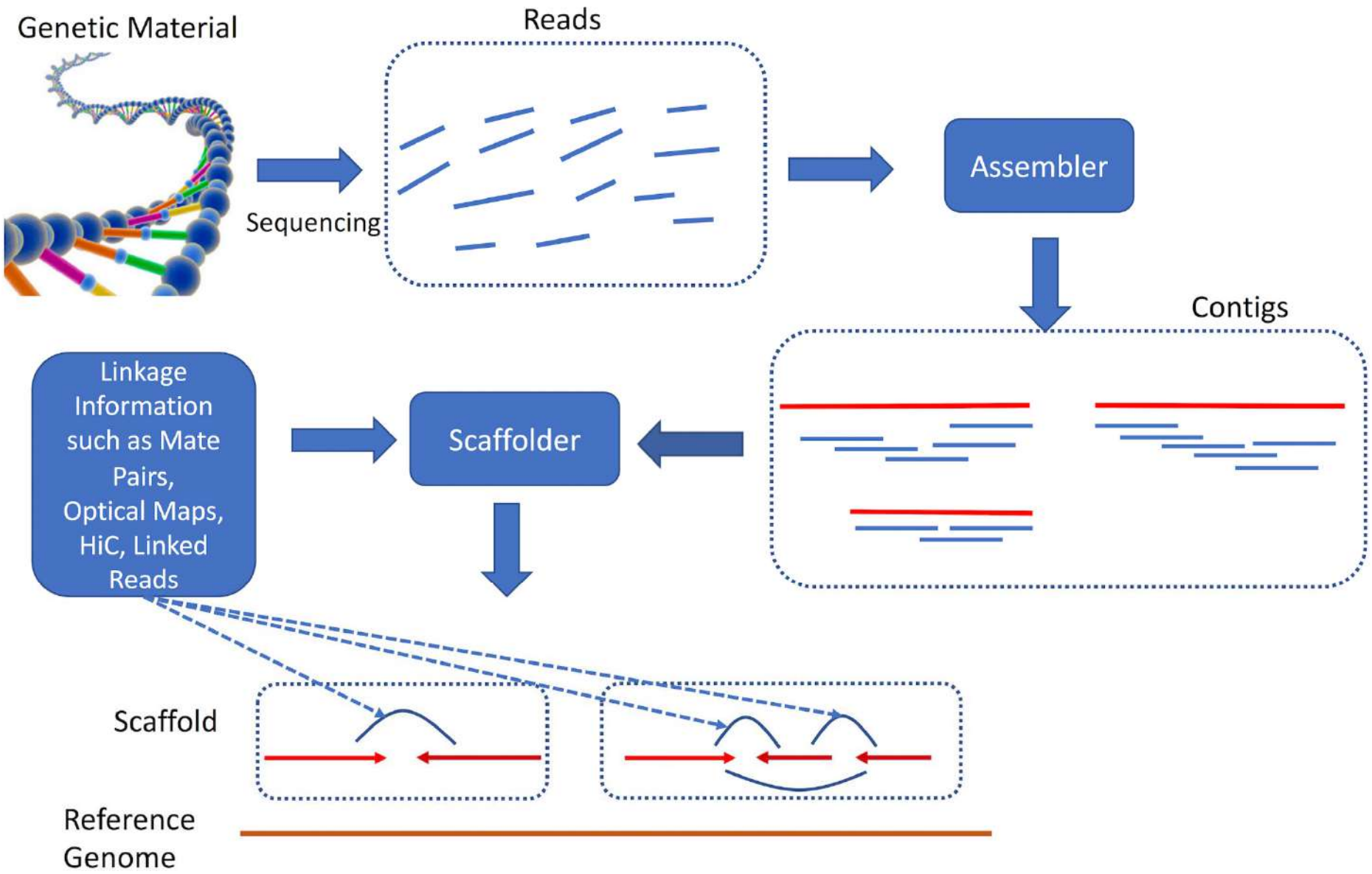
Clint scaffolds



Kronenberg et al., in review

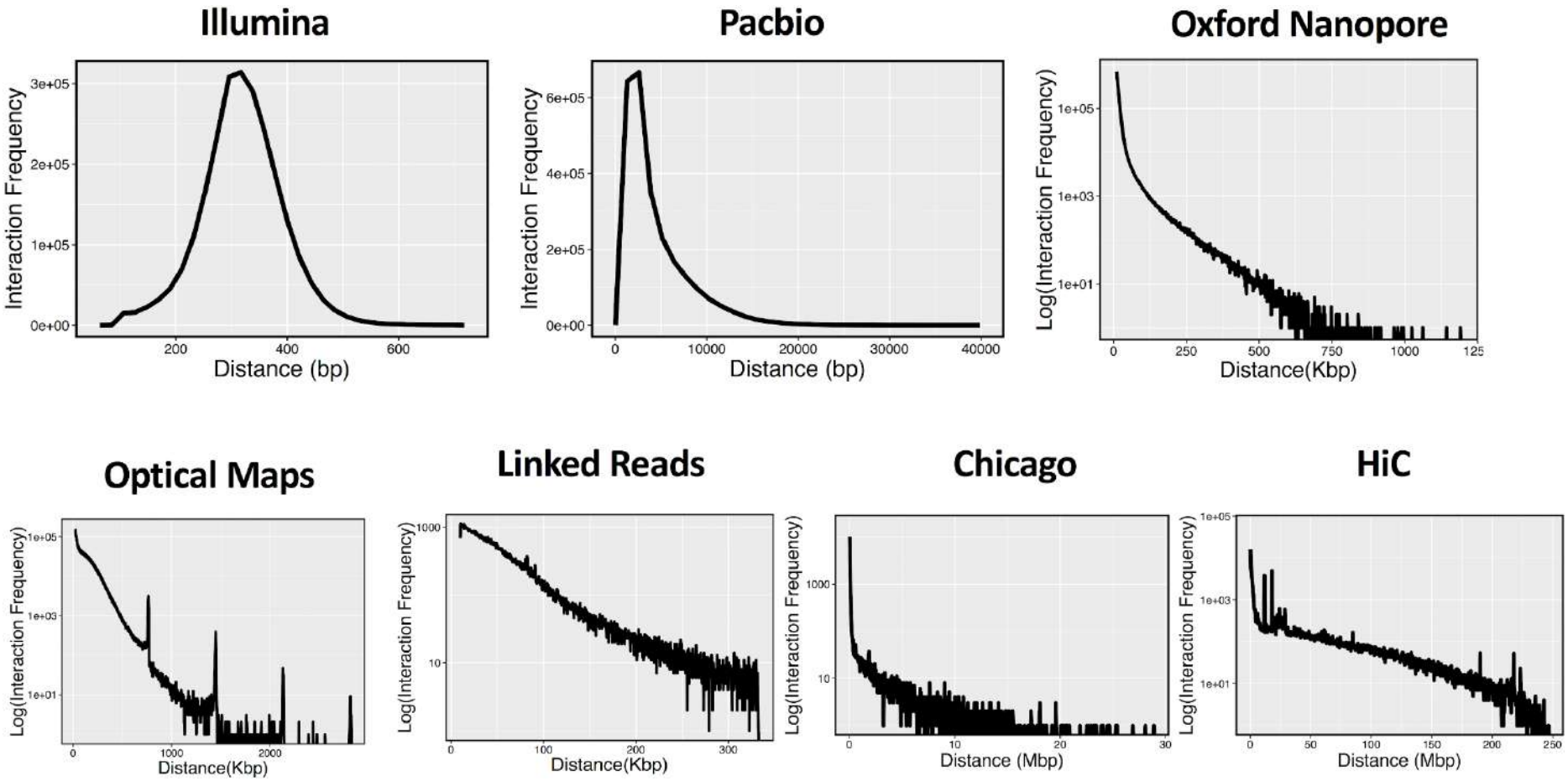
https://youtu.be/uzINKcj-p78

Modern technologies & algorithms for scaffolding assembled genomes



https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1006994

Modern technologies & algorithms for scaffolding assembled genomes



https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1006994

Modern technologies & algorithms for scaffolding assembled genomes

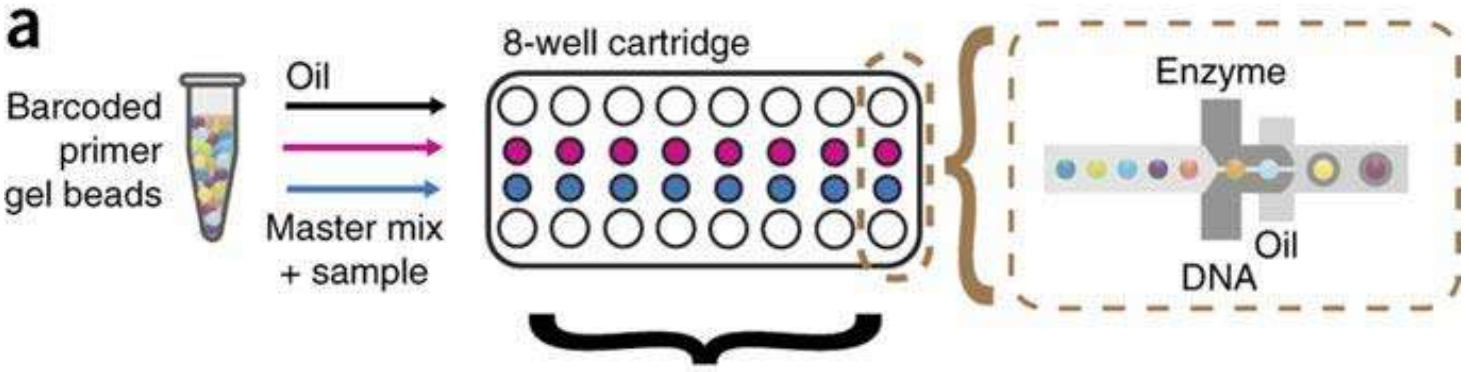
Category	Scaffolding Data	Separation on the Genome	Orientation	Ordering	Distance
Physical mapping	Restriction maps	10–100 Kb	Yes	No	Yes
	Optical maps	10–100 Kb	Yes	Yes	Yes
Subcloning	10x Genomics	100 Kb	Yes	Yes	Yes
	Illumina TSLR	100 Kbp	Yes	Yes	Yes
Long-read data	Pacific Biosciences	10–15 Kb	Yes	Yes	Yes
	Oxford Nanopore	15–20 Kb	Yes	Yes	Yes
Paired read	Paired-end reads	100–500 bp	Yes	Yes	Yes
	Mate pairs	1,000–10,000 bp	Yes	Yes	Yes
Chromosome conformation	Hi-C	30–100 Mb	Yes	Yes	No
	Chicago	3–100 Mb	Yes	Yes	No
Synteny	Reference genome(s)	Up to genome size	Yes	Yes	Yes

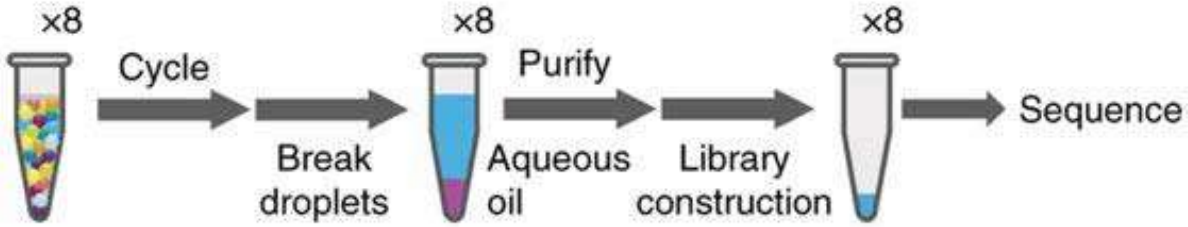
Abbreviation: TSLR, TruSeq Synthetic Long Read.

https://doi.org/10.1371/journal.pcbi.1006994.t001

https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1006994

10X Genomics – Linked Read Sequencing

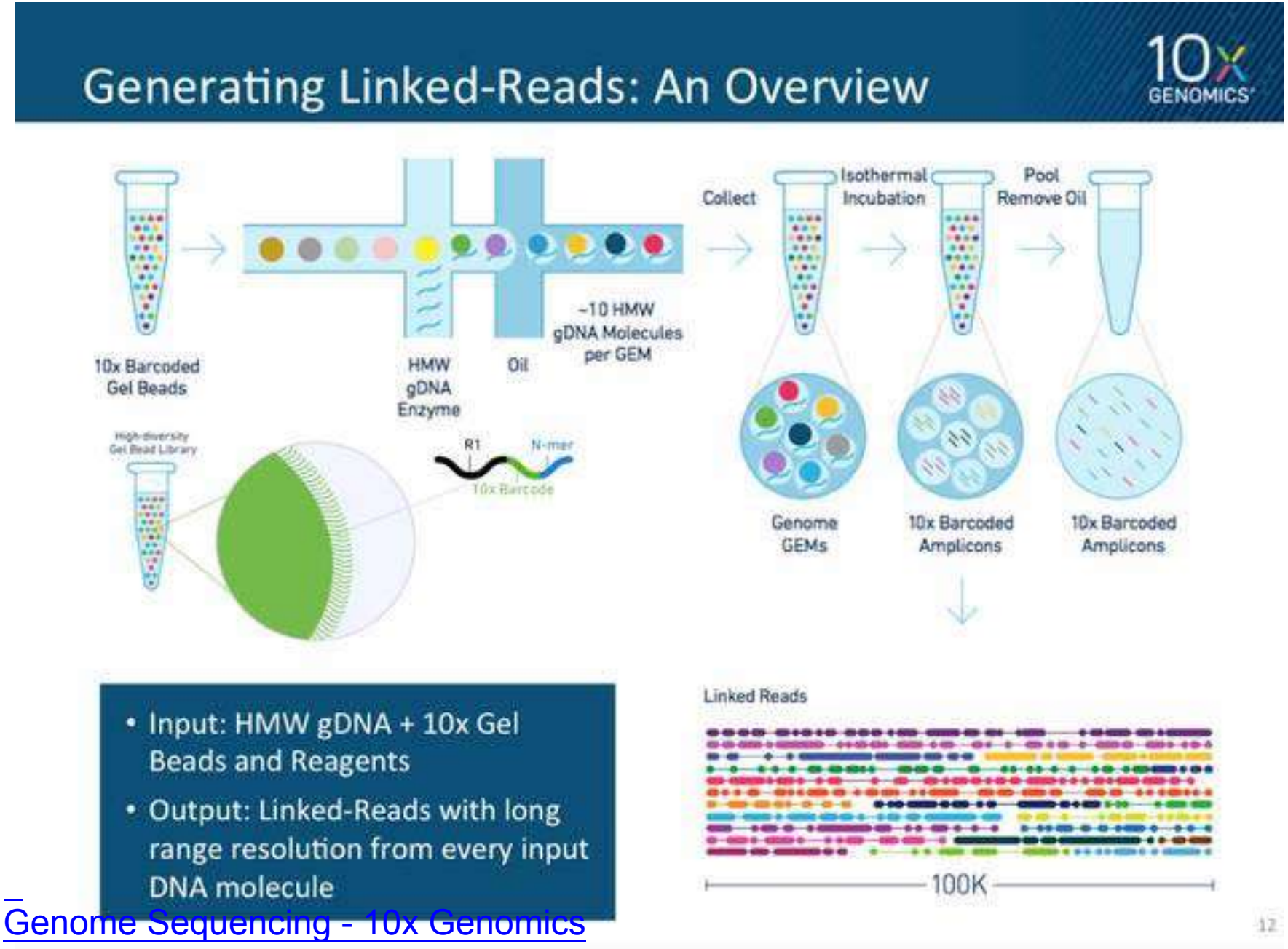




Input data	This study Illumina paired-end and mate-pair reads; 10XG reads; BNG genome maps	Ref. 21 PacBio reads; BNG genome maps	ALLPATHS-LG[22] Illumina paired-end, mate-pair, and fosmid-based short reads
Scaffold N50 (Mb)	33.5	31.1	11.5
Number of scaffolds	170	202	23,634
Assembly length (Gb)	2.86	2.76	2.78
Validity at 100 kb (%)	95.2	97.5	93.5
N content (%)	10.2	4.61	5.90
Phase block N50	4.7 Mb	145 kb	N/A
Phased SNVs	2,783,119	2,421,740	N/A

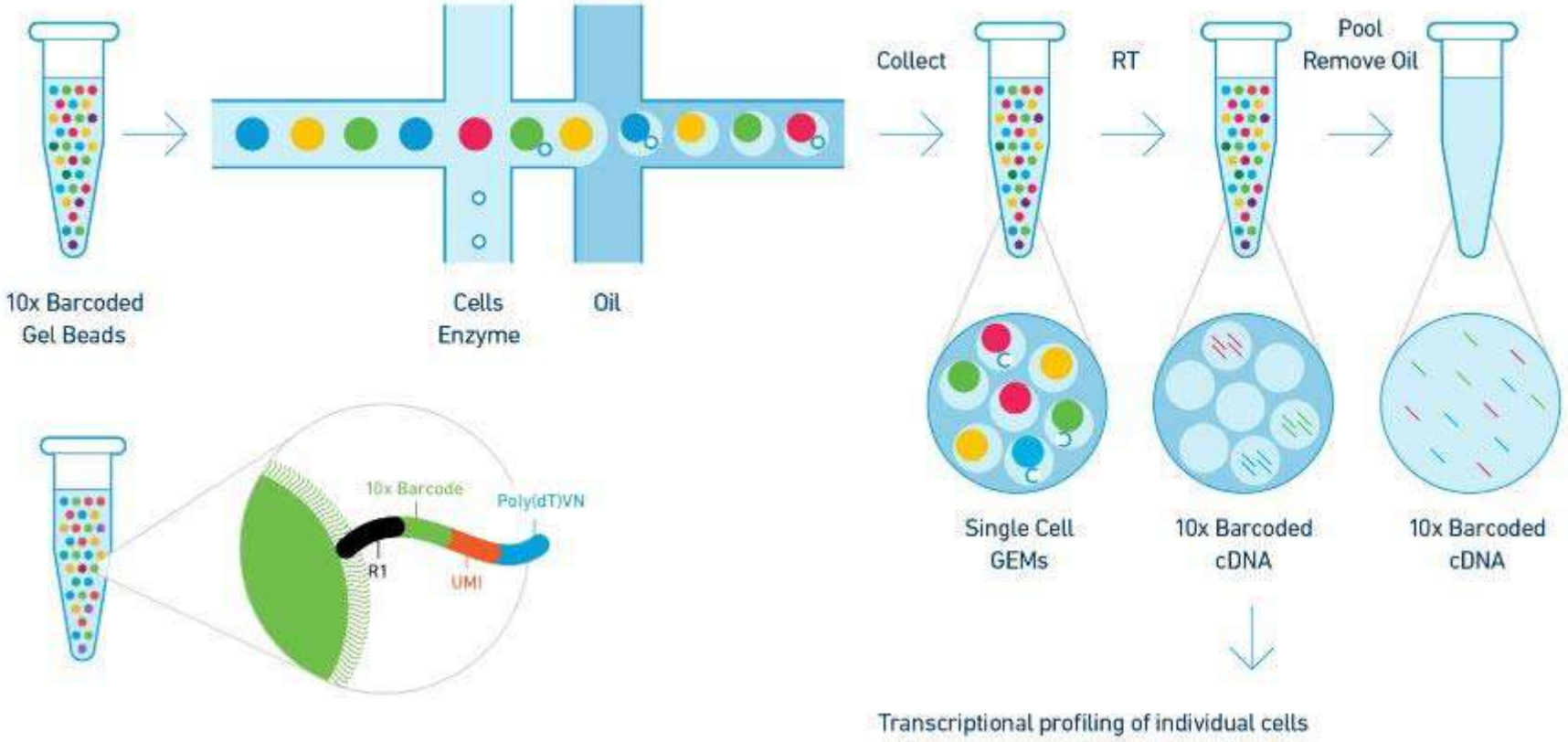
s303–311 (2016)

10X Genomics – Genome Sequencing



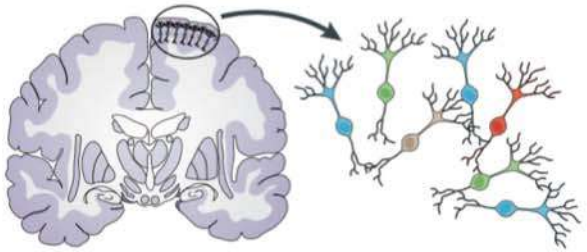
Genome Sequencing - 10x Genomics

Single-cell RNA-seq

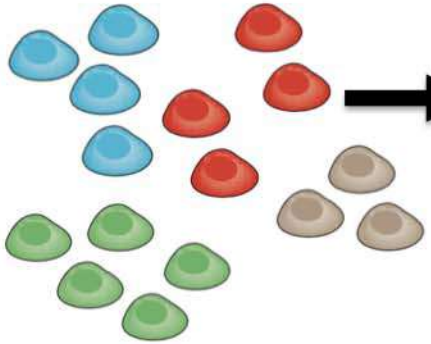


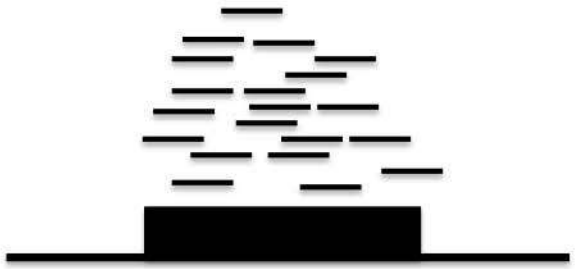
Single-cell RNA-Seq (scRNA-Seq)

Tissue (e.g. tumor)



Isolate and sequence individual cells



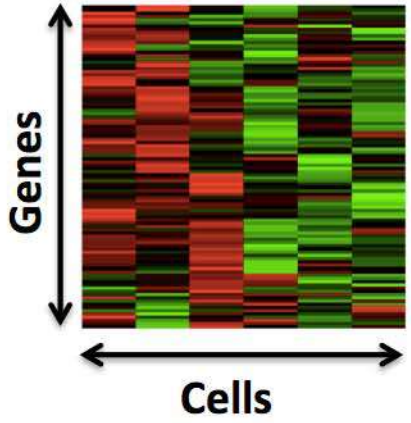


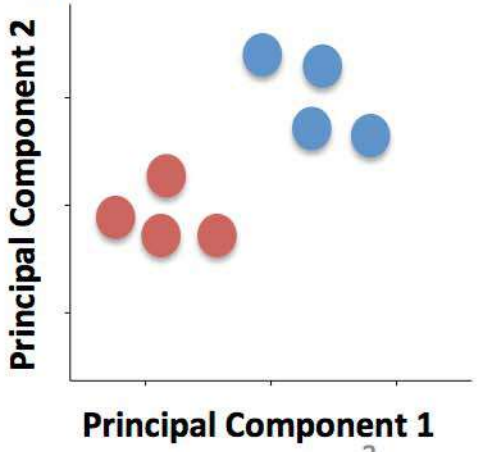
Gene 1
Cell 1

Read Counts

	Cell 1	Cell 2	…
Gene 1	18	0	
Gene 2	1010	506	
Gene 3	0	49	
Gene 4	22	0	
…			

Compare gene expression profiles of single cells





Next-generation DNA sequencing

Jay Shendure[1] & Hanlee Ji[2]

***Nature Biotechnology* 26**, 1135 - 1145 (2008)

APPLICATIONS OF NEXT-GENERATION SEQUENCING

Sequencing technologies — the next generation

*Michael L. Metzker**‡

***Nature Review Genetics* 11,** 31-46 (2010)

Published in final edited form as:
J Genet Genomics. 2011 March 20; 38(3): 95–109. doi:10.1016/j.jgg.2011.02.003.

The impact of next-generation sequencing on genomics

Jun Zhang[a,b,*], **Rod Chiodini**[c], **Ahmed Badr**[a], and **Genfa Zhang**[d]
[a] COE for Neurosciences, Department of Anesthesiology, Texas Tech University Health Sciences Center El Paso, TX 79905, USA

NGS Reviews

Next-generation DNA sequencing

Jay Shendure[1] & Hanlee Ji[2]

***Nature Biotechnology* 26**, 1135 - 11

APPLICATIONS OF NEXT-GENERATION SEQUENCING

Sequencing technologies – the next generation

*Michael L. Metzker**‡

***Nature Review Genetics* 11,** 31-46 (2010)

ADVANCES IN WOUND CARE

Mary Ann Liebert, Inc. publishers | Journals | Search | Alerts

Adv Wound Care (New Rochelle). 2015 Jan 1; 4(1): 50–58. PMCID: PMC4281878
doi: 10.1089/wound.2014.0542

Next-Generation Sequencing: A Review of Technologies and Tools for Wound Microbiome Research

Brendan P. Hodkinson and Elizabeth A. Grice*

Author information ▶ Article notes ▶ Copyright and License information ▶

nature REVIEWS GENETICS

nature.com ▸ journal home ▸ archive ▸ issue ▸ review ▸ full text

NATURE REVIEWS GENETICS | REVIEW

ARTICLE SERIES: Applications of next-generation sequencing

Coming of age: ten years of next-generation sequencing technologies

Sara Goodwin, John D. McPherson & W. Richard McCombie

Affiliations | Corresponding author

Nature Reviews Genetics **17**, 333–351 (2016) | doi:10.1038/nrg.2016.49
Published online 17 May 2016

NIH Public Access
Author Manuscript
J Genet Genomics. Author manuscript; available in PMC 2011 April 13.

Published in final edited form as:
J Genet Genomics. 2011 March 20; 38(3): 95–109. doi:10.1016/j.jgg.2011.02.003.

The impact of next-generation sequencing on genomi

Jun Zhang[a,b,*], **Rod Chiodini**[c], **Ahmed Badr**[a], and **Genfa Zhang**[d]

[a] COE for Neurosciences, Department of Anesthesiology, Texas Tech University Center El Paso, TX 79905, USA

Video Clips

- Sanger Sequencing of DNA [HD Animation]
  - https://www.youtube.com/watch?v=nudG0r9zL2M
- Pyro Sequencing
  - https://www.youtube.com/watch?v=nFfgWGFe0aA
- Illumina Sequencing Technology
  - https://www.youtube.com/watch?v=womKfikWlxM
- Ion Torrent™ next-gen sequencing technology
  - https://www.youtube.com/watch?v=WYBzbxIfuKs
- Single Molecule Real Time Sequencing - Pacific Biosciences
  - https://www.youtube.com/watch?v=v8p4ph2MAvI
- Oxford Nanopore Technologies
  - https://www.youtube.com/watch?v=3UHw22hBpAk
- Next-Generation Sequencing Technologies - Elaine Mardis (2014)
  - https://www.youtube.com/watch?v=6Is3W7JkFp8

- PCR (Polymerase Chain Reaction)
  - https://www.youtube.com/watch?v=iQsu3Kz9NYo
- Polymerase Chain Reaction [HD Animation]
  - https://www.youtube.com/watch?v=0HCWmD7Mv8U

High Throughput Genomics Core

NGS CORE ACADEMIA SINICA

Home 首頁 | Member 成員 | Instruments 儀器 | Service 服務 | Application Forms 表單下載 | Contacts 聯絡我們 | Other 其他

Instruments

The two NGS platforms have gone through timely upgrades and capacity expansion through new acquisition.

Illumina MiSeq

Illumina HiSeq-2500

- Illumina platform : the current models include two HiSeq2500 and one MiSeq sequencers. Sequencing can be single-end (SR) or paired-end (PE) format. Read length can be defined according to the length most suitable to the desired application. Mate-pair library is standard

Search Go

Sequencing Data Download

- Pydio
- sFTP

Related Web Links

- Illumina
- Roche 454
- NCHC NGS Software Platform (國家高速網路與計算中心)

http://ngs.biodiv.tw/NGSCore/

Thank you!