中央研究院
生物多樣性研究中心
Biodiversity Research Center, Academia Sinica

TIGP-BIODIV Lecture, 3/31/2016

NGS: DNA/RNA preparation & different sequencing technologies

Mei-yeh Lu 呂美曄

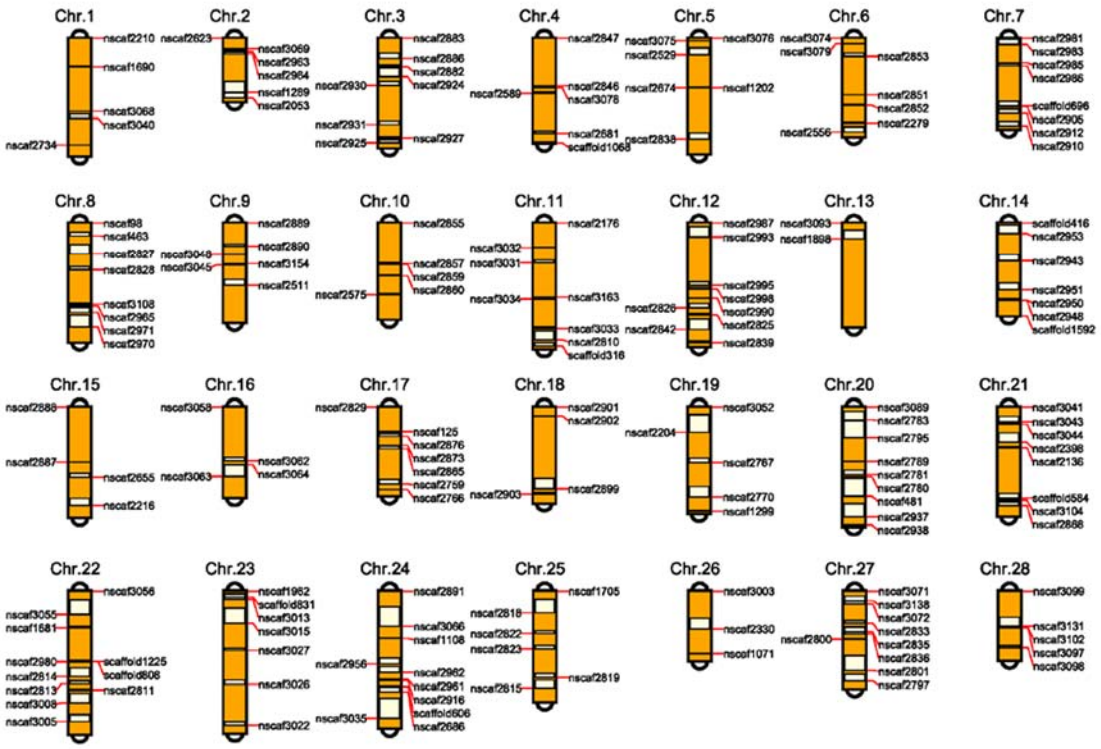
Associate Research Specialist
High Throughput Genomics Facility Manager
Biodiversity Research Center
Academia Sinica

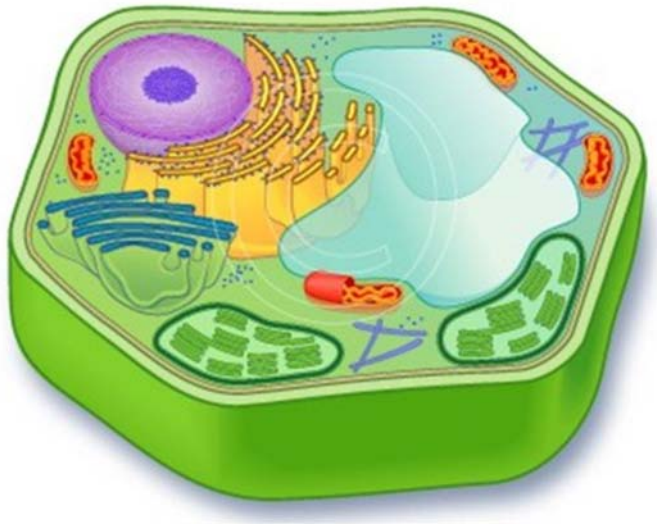
Outlines

- Evolution of sequencing technologies
- NGS platforms and comparisons
- Project considerations & Sequencing plan
- Data: types, preprocessing, & assessment
- Applications

Genome

- a full haploid set of chromosomes with all its gene; the total genetic constitution of a cell or organism.





- **Nuclear (chromosome)**
- **Plasmid (bacterial)**
- **Viral**

- **Mitochondrial**
- **Chloroplast**

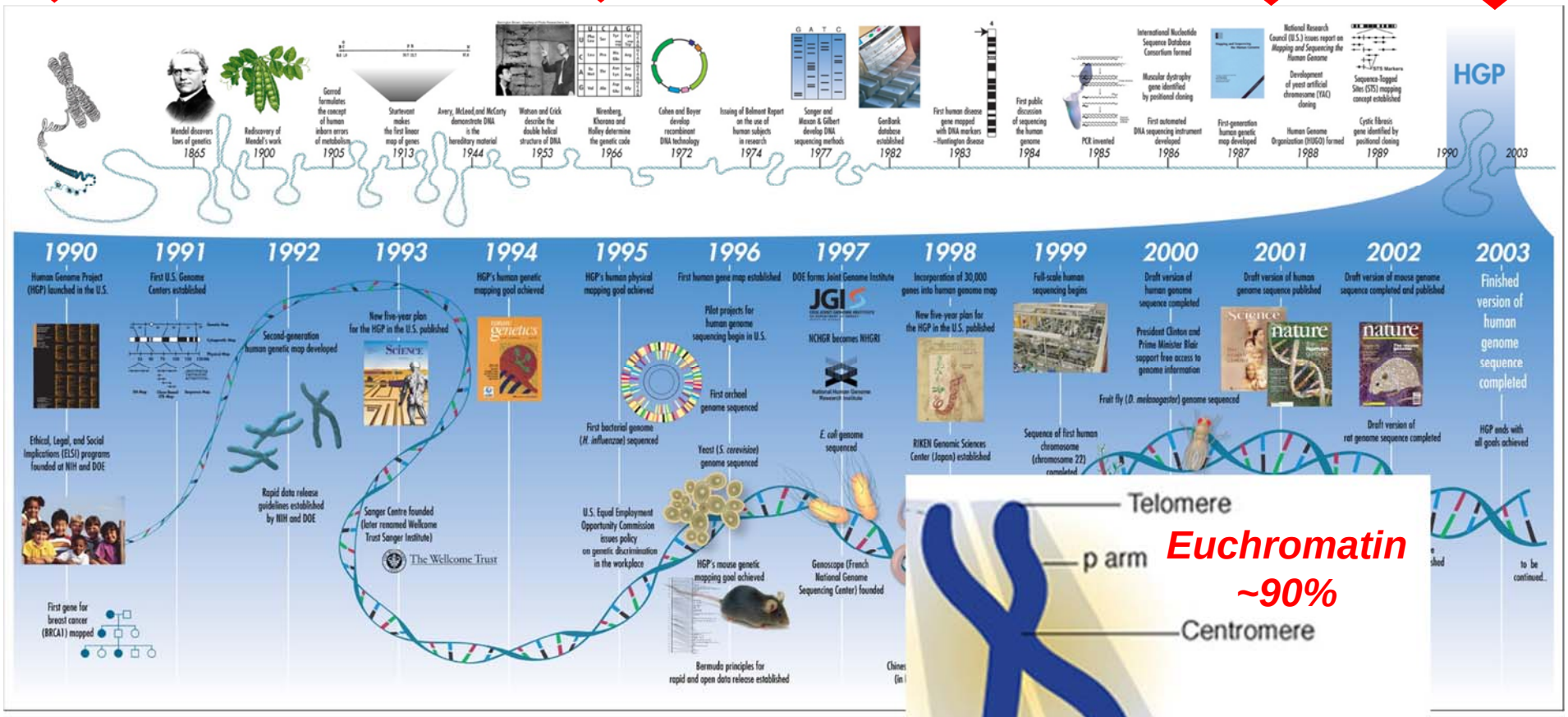
Timeline of the Human Genome Project

Project onset

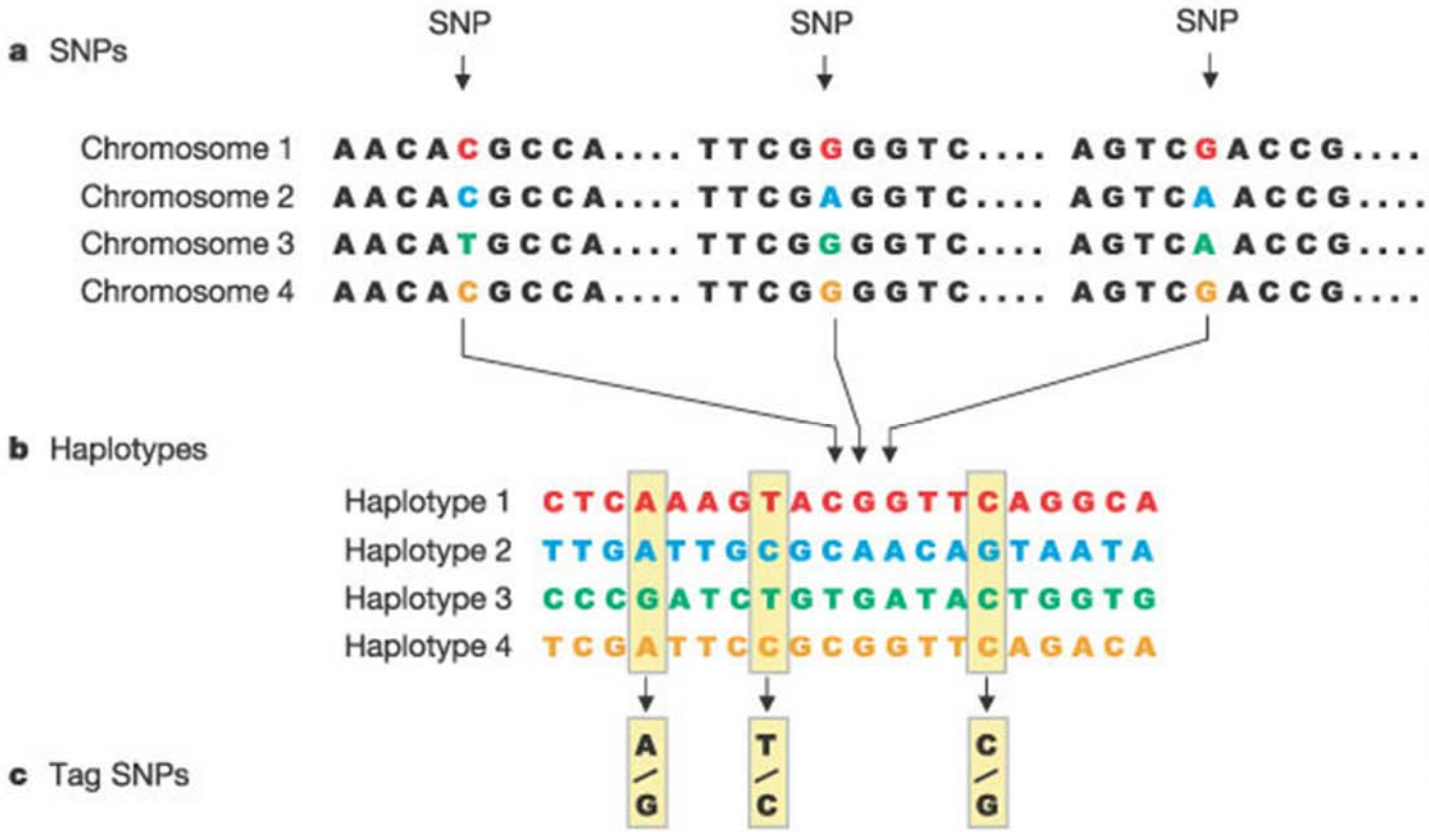
1st Genome: influenza

1st draft

Complete




International
HapMap
Project
2002~

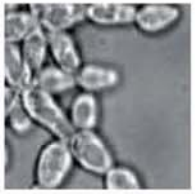

a SNPs
SNP
SNP
SNP
Chromosome 1 AACACGCCA.... TTCGGGGTC.... AGTCGACCG....
Chromosome 2 AACACGCCA.... TTCGAGGTC.... AGTCAACCG....
Chromosome 3 AACATGCCA.... TTCGGGGTC.... AGTCAACCG....
Chromosome 4 AACACGCCA.... TTCGGGGTC.... AGTCGACCG....
b Haplotypes
Haplotype 1 CTCAAAGTACGGTTCAGGCA
Haplotype 2 TTGATTGCGCAACAGTAATA
Haplotype 3 CCCGATCTGTGATACTGGTG
Haplotype 4 TCGATTCCGCGGTTCAGACA
c Tag SNPs
A/G
T/C
C/G

Chr. 7

Vall d'Hebron
Institut de Recerca

Any DNA can be sequenced

M Tuberculosis



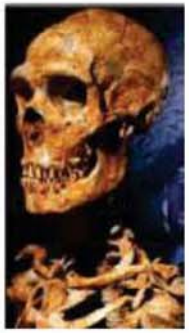
S. cervisiae

C. elegans

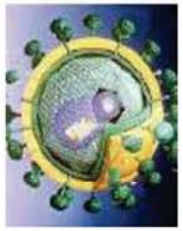
Barley

Arabidopsis

Maccaca



Neanderthal



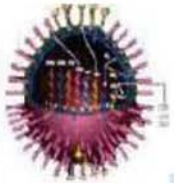
HIV

Tomato

Potato

Honeybee

Mammut



H5N1

James Watson

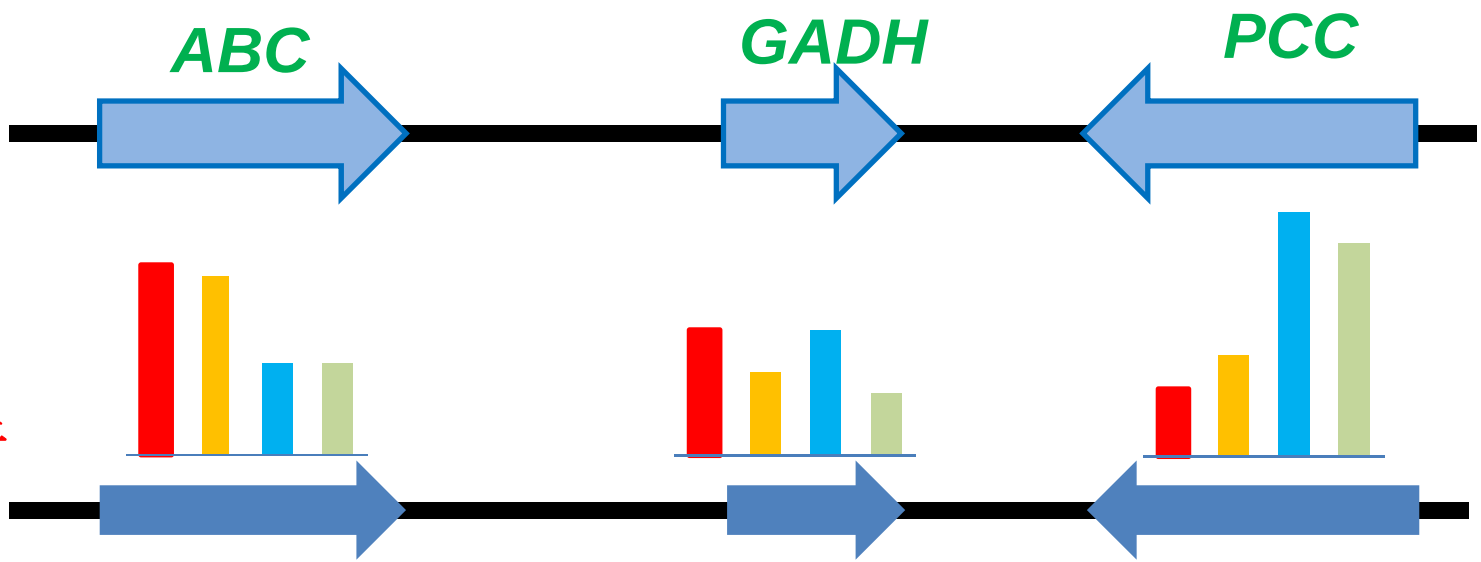
Grape wine

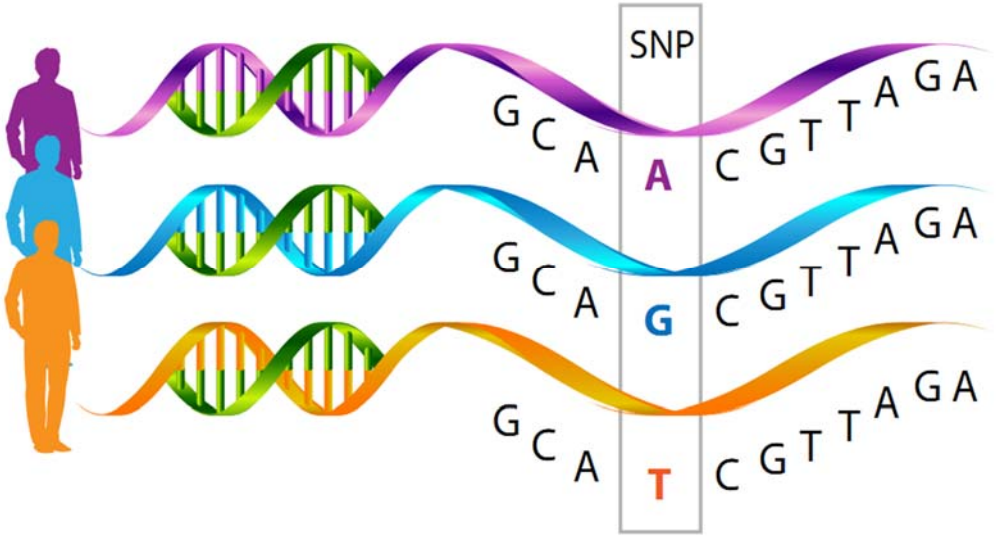
UNIVERSITAT DE BARCELONA

Vall d'Hebron
Institut de Recerca

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 .....

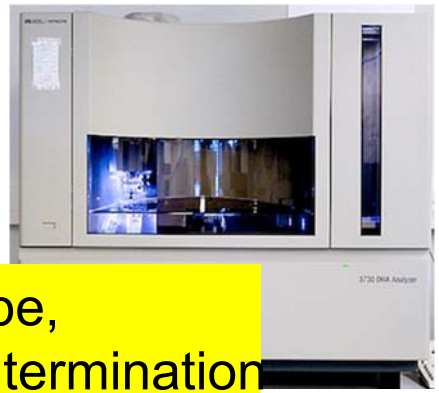




I. Evolution of Sequencing Technologies

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

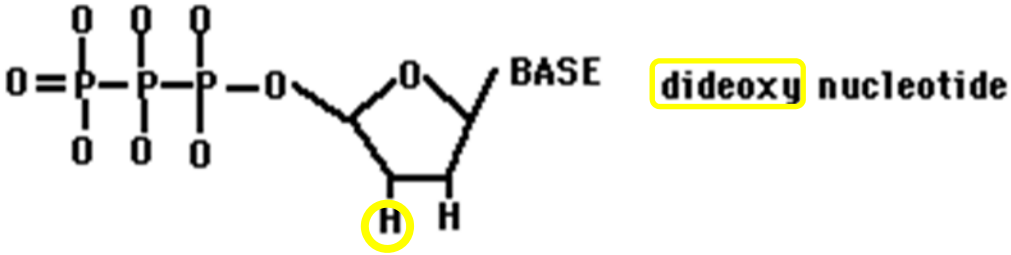
PacBio

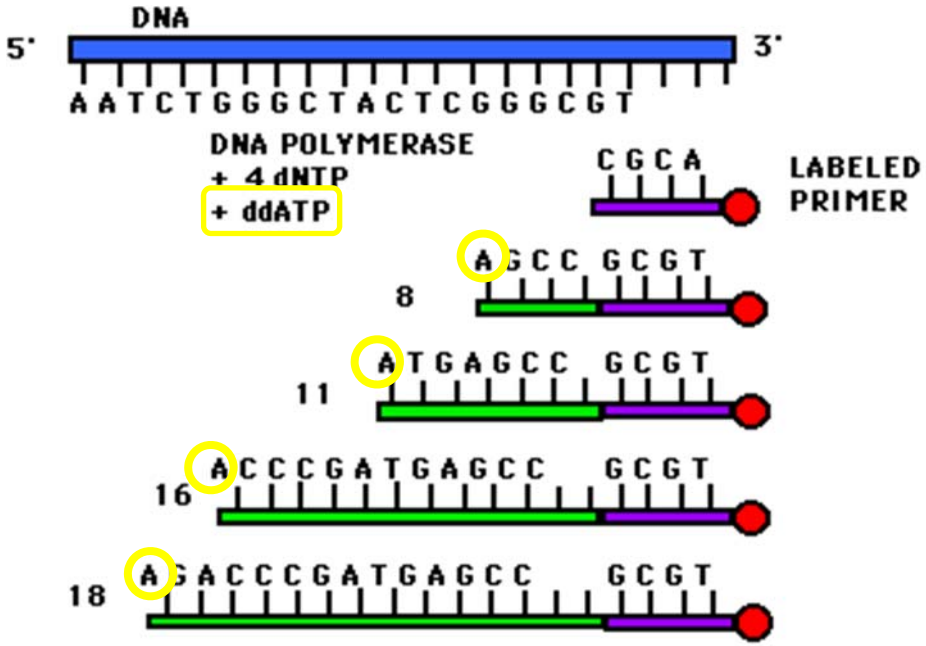
Oxford NANOPORE Technologies

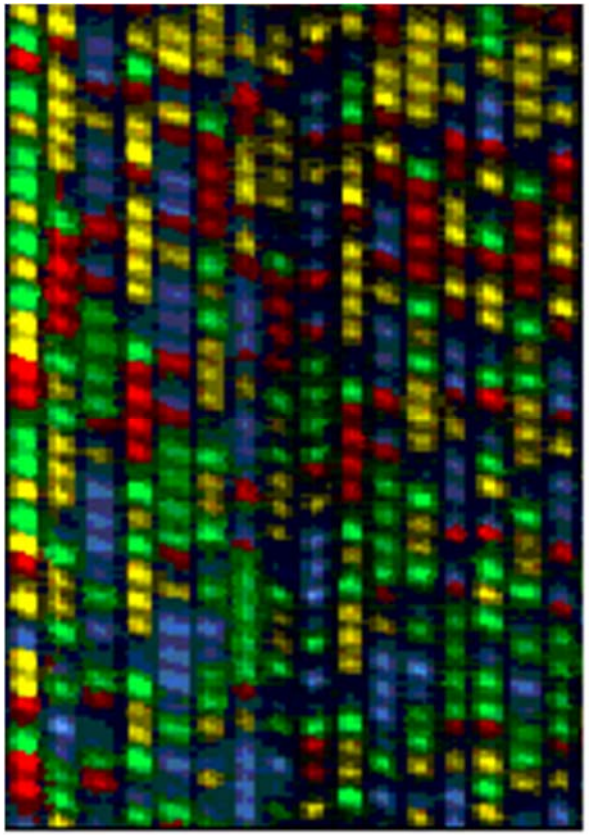
Single molecule sequencing

Sanger Seq. – dideoxy nucleotide termination

Frederick Sanger

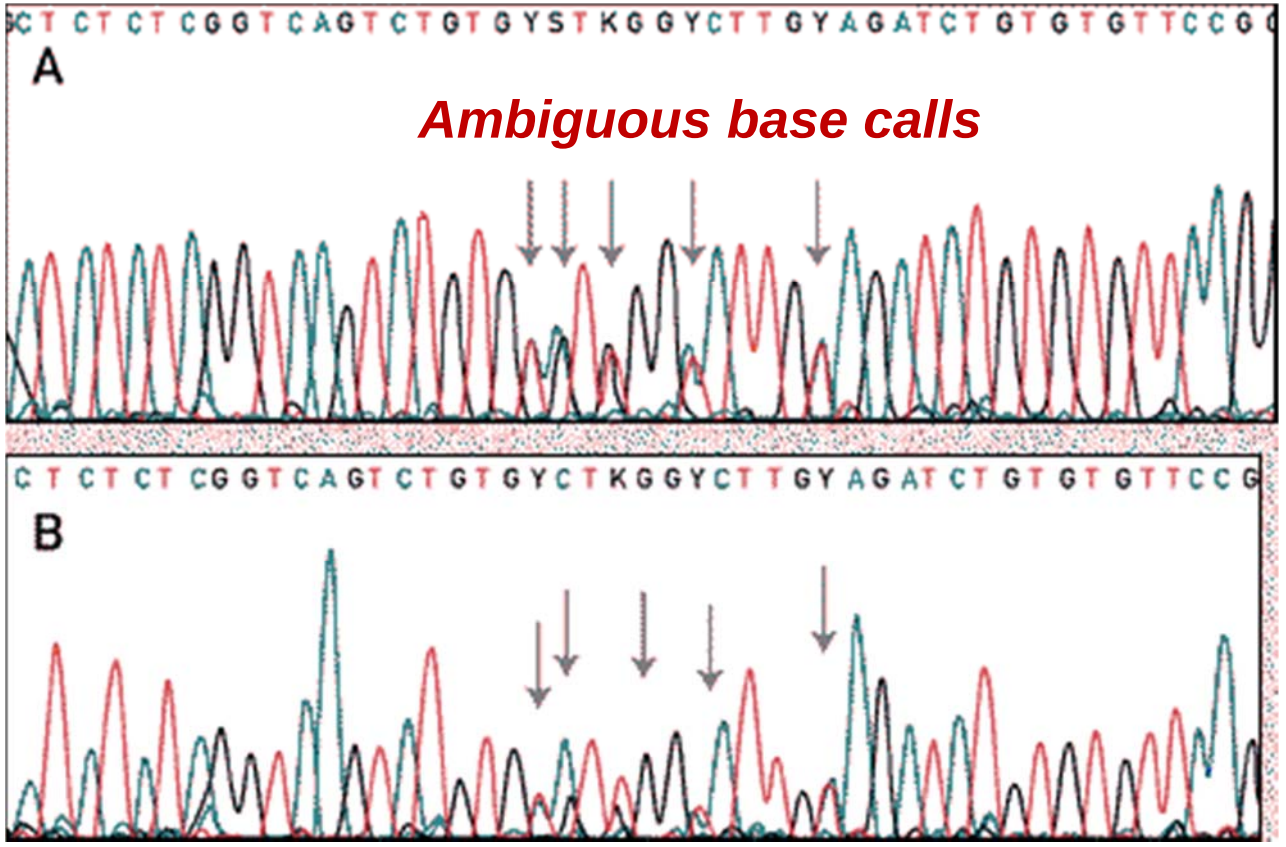


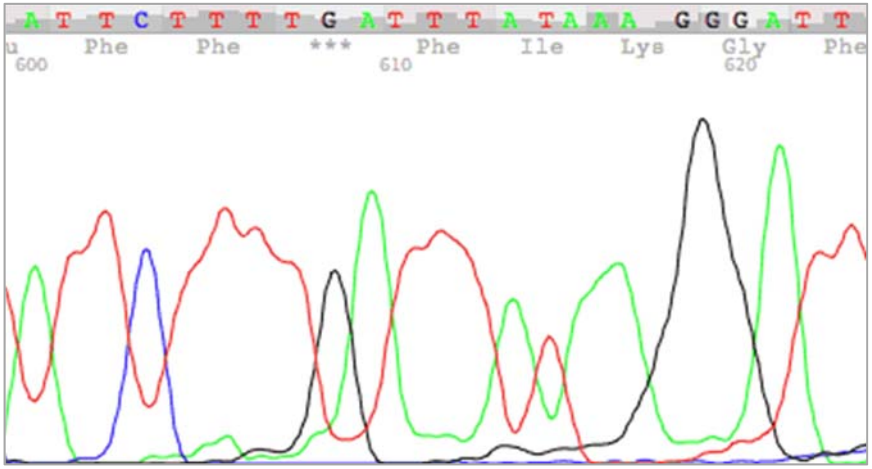




http://www.bio.davidson.edu/Bio111/seq.html

Fluorescent Dye-Terminator Cycle Sequencing





Blurry trace: Homopolymer

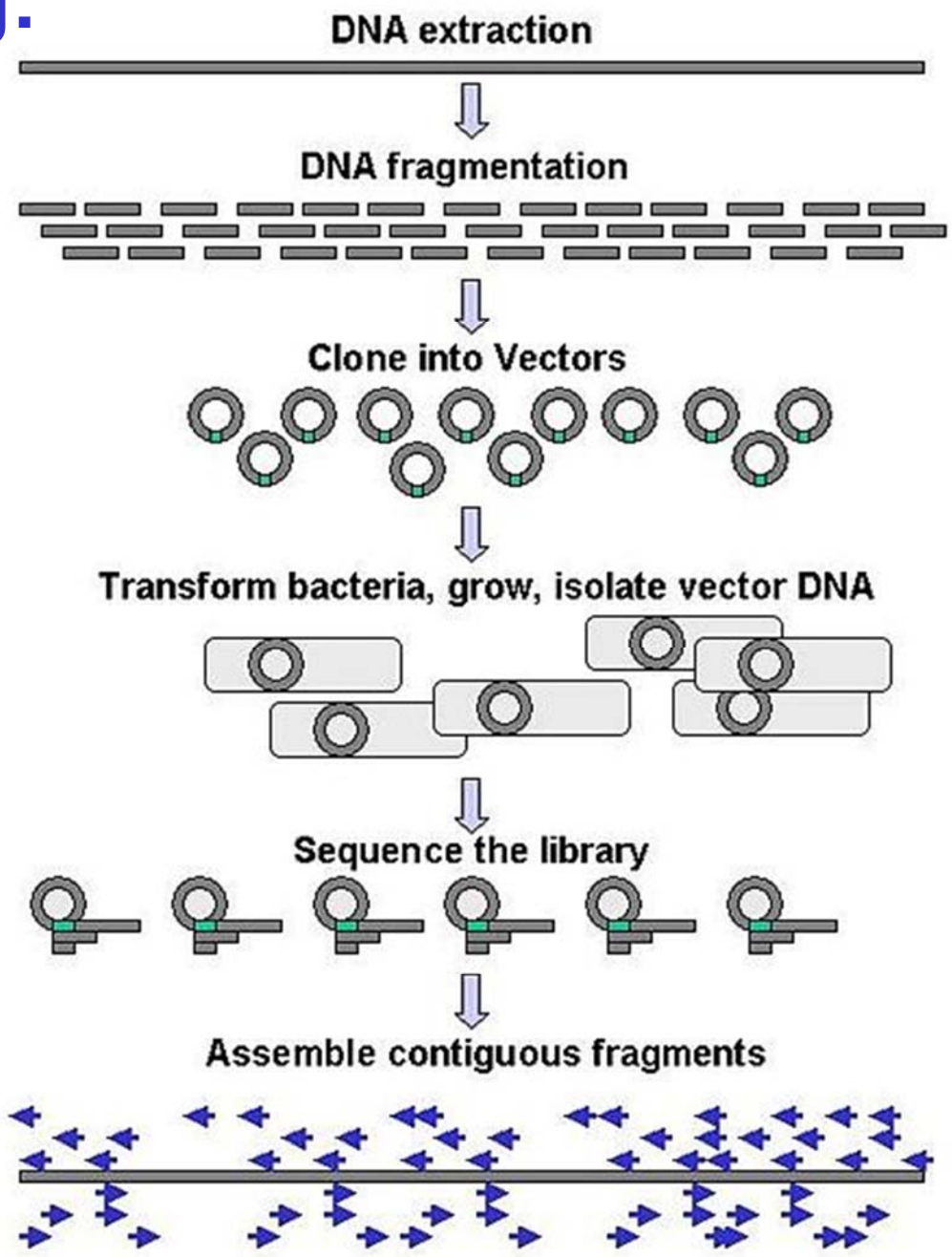
Genome Sequencing: Hierachical cloning

BAC

Cosmid

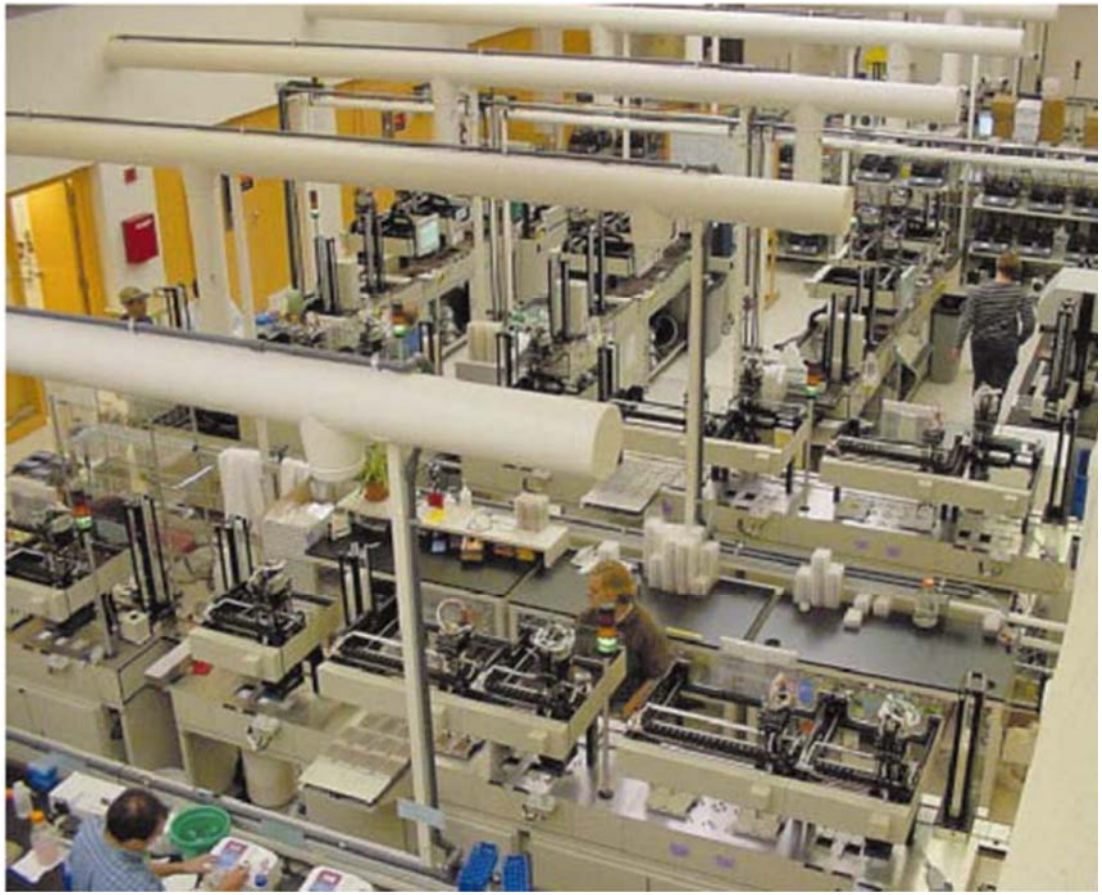
Fosmid

Plasmid



Large scale Cappilary Sequencing

Library factory - Whitehead Institute



Sequencing factory - Sanger Institute

Nature 409, 860-921(15 February 2001)

NGS–benefits over Sanger sequencing

- Massive parallel sequencing of shotgun libraries
- Use universal primer on adaptors
  - No need for prior sequence knowledge (good for non-model organisms)
- No bacterial cloning (less representation bias) and seq. walking
- High throughput (great coverage depth)
- More cost-effective per unit output
- Diversified applications
- Various analysis tools available
- Higher sensitivity than array-based detection
- Fast evolving for even greater performance

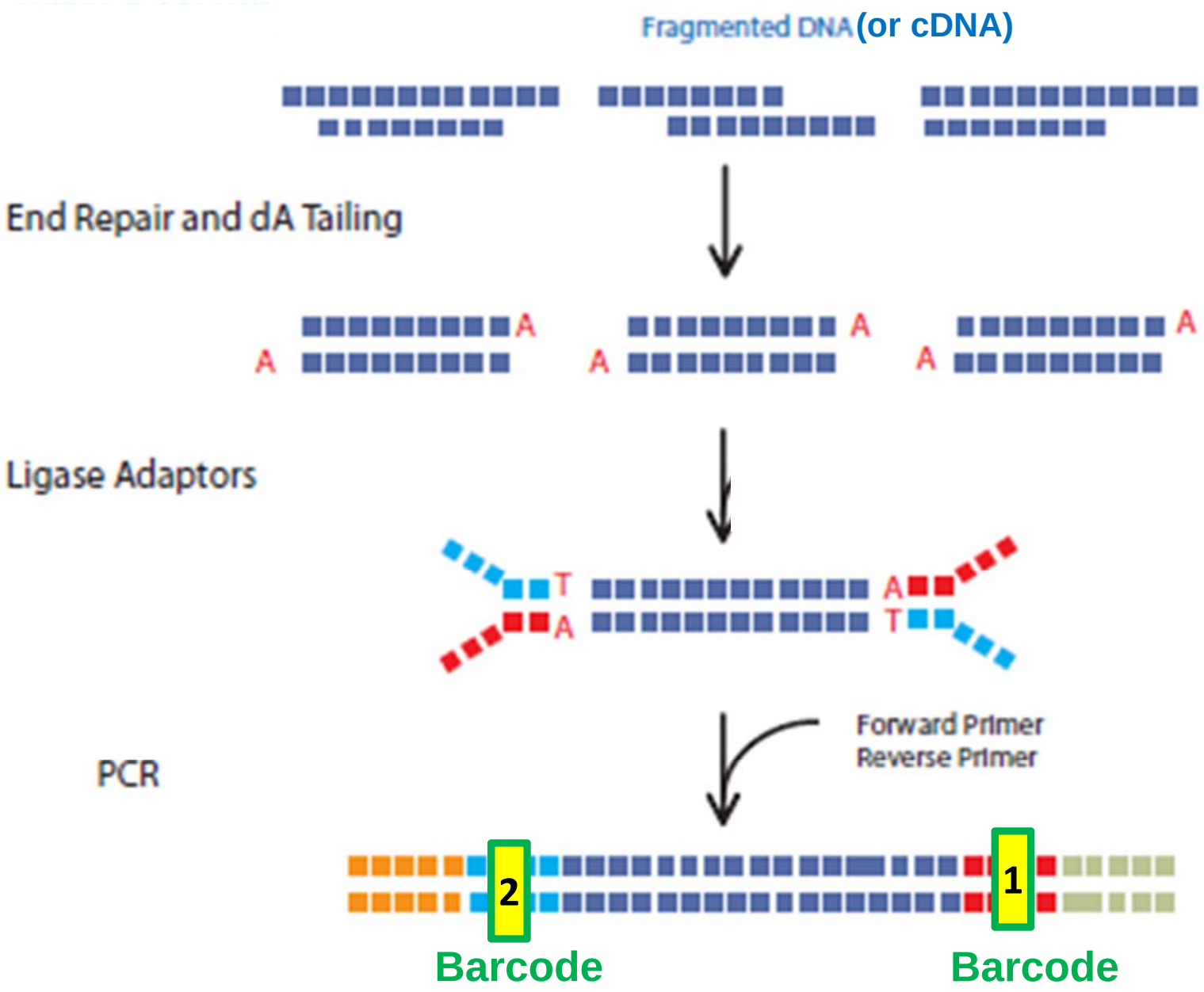
NGS – massive parallel sequencing

Current Popular platforms:

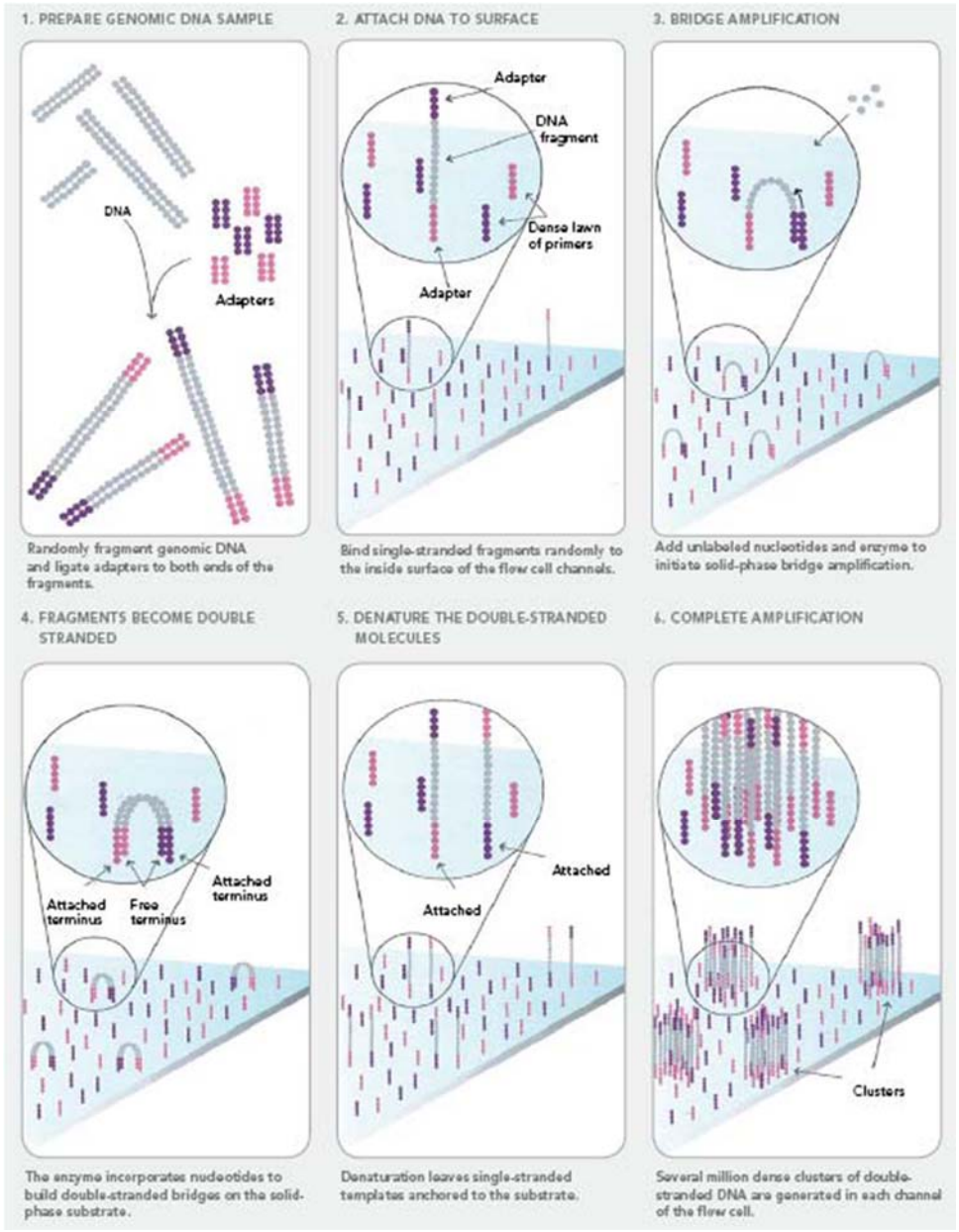
- **2nd-Gen: clonal amplification**
  - Roche 454: GS FLX, , 454 Jr., 454 XL+, 454 Jr.
  - Illumina: GA, Miseq, HiSeq2500, MiSeq
  - Life Technologies: SOLiD, Ion Torrent, Ion Proton

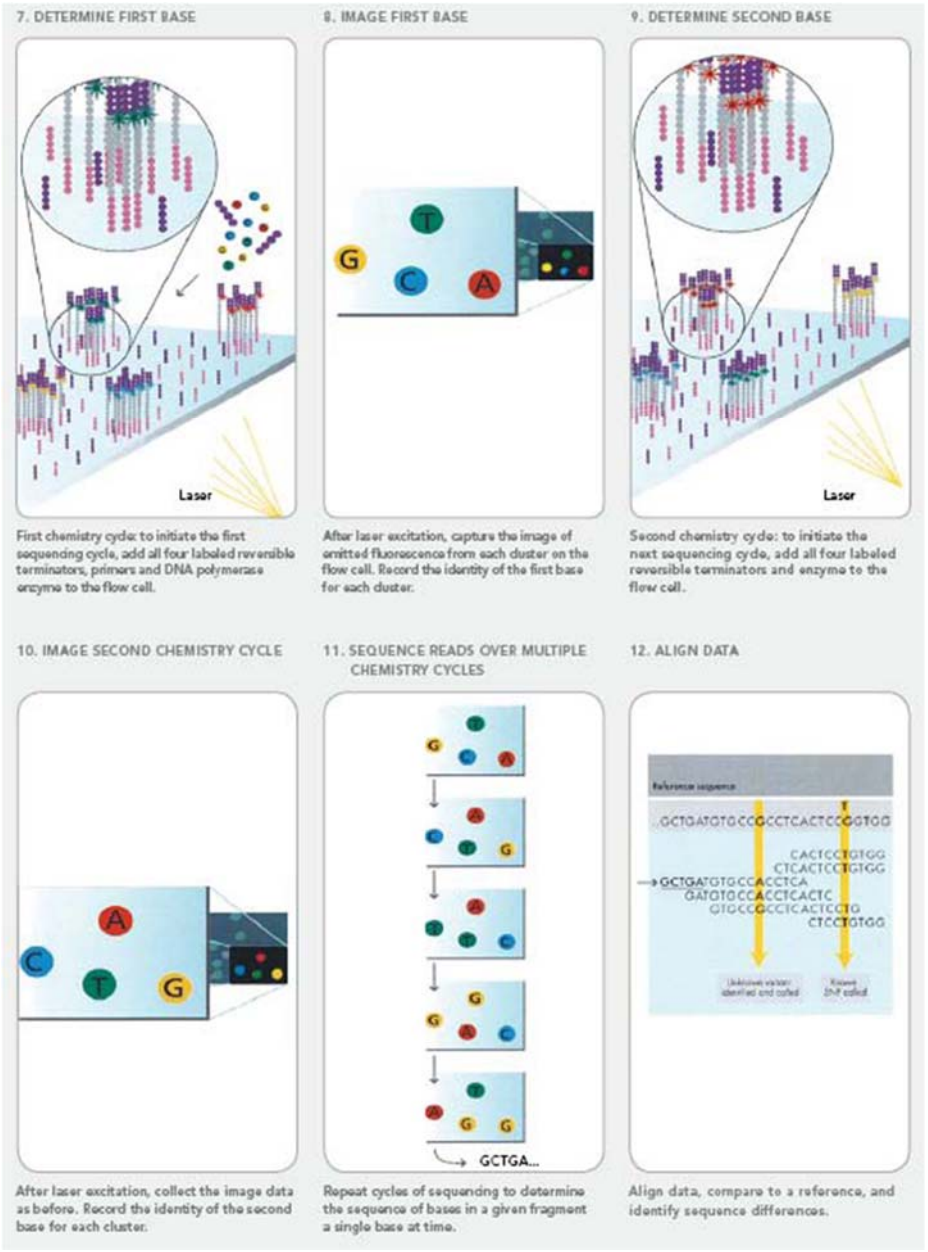
- **3rd-Gen: single molecule sequencing**
  - Pacific Biosciences: PacBio RS II
  - Oxford Nanopore Technologies

General workflow for DNA library prep

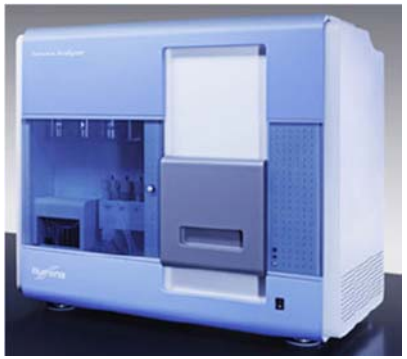


Illumina/Solexa: Cyclic Reversible Terminator





Illumina – Flow cell imaging

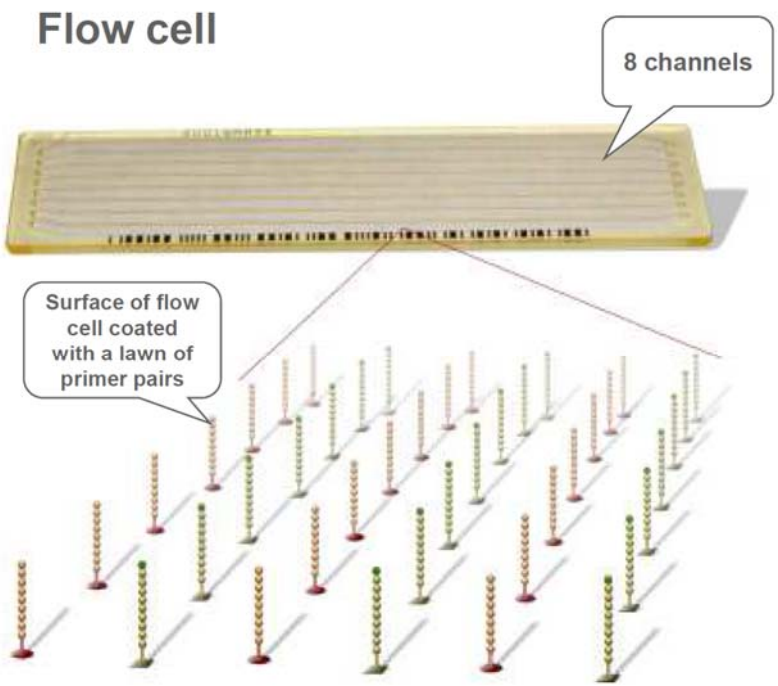


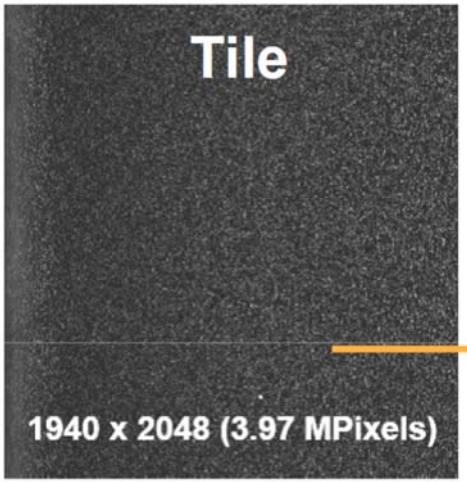
GA IIx

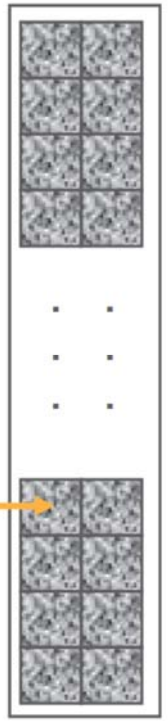
HiSeq 2500
(HT*8 / Rapid*2)

NextSeq 500

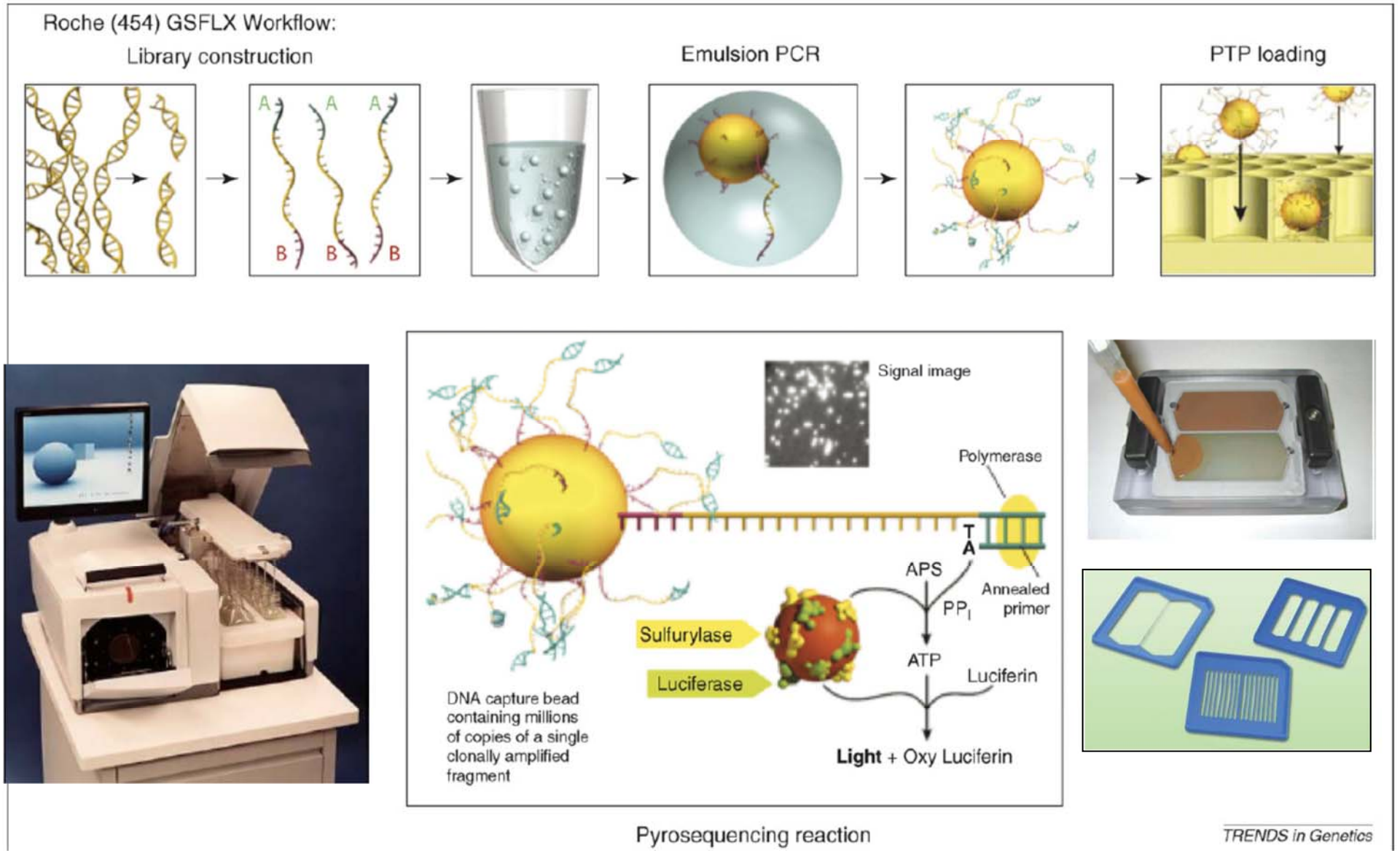
MiSeq v2



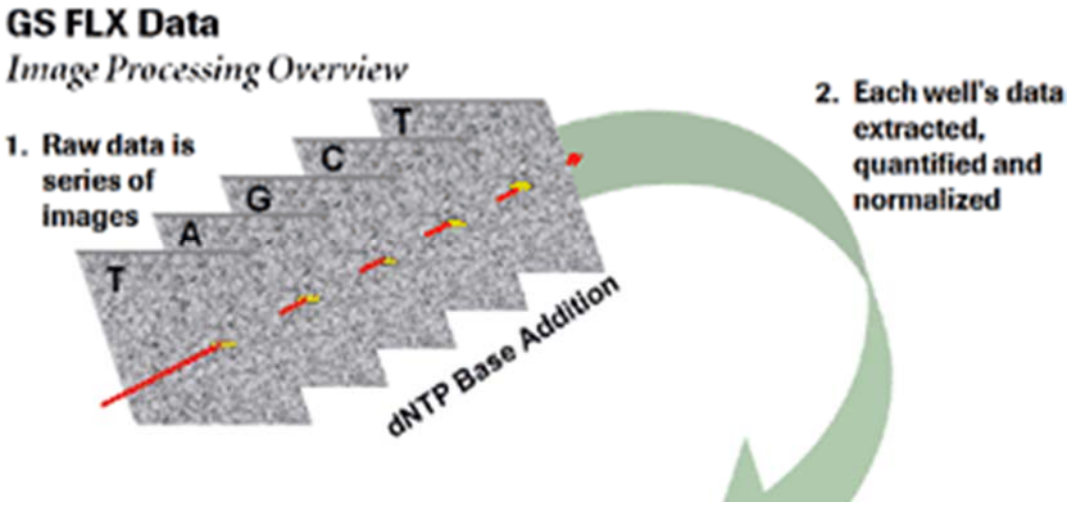


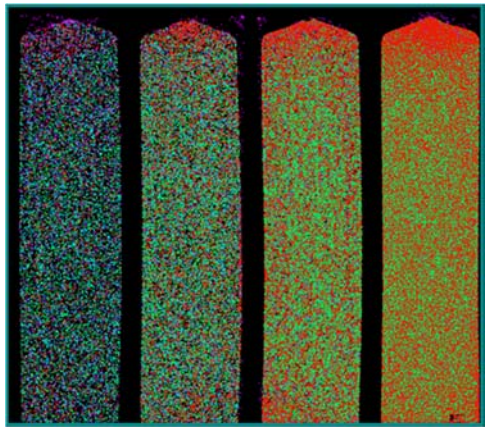


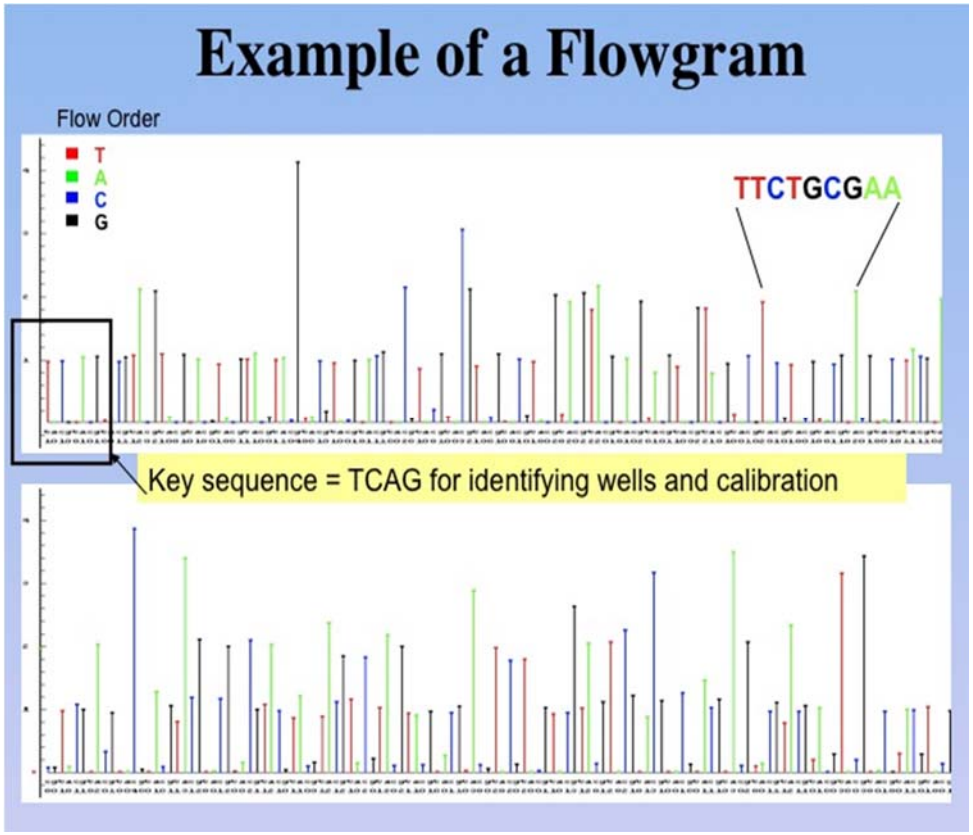
454: emPCR & pyrosequencing

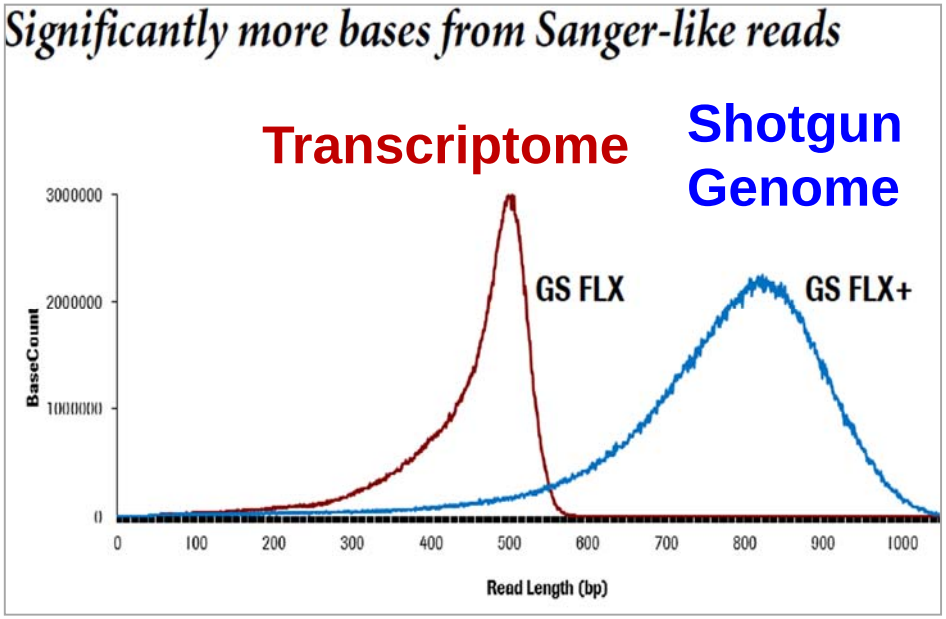


454 flowgram and read length profile

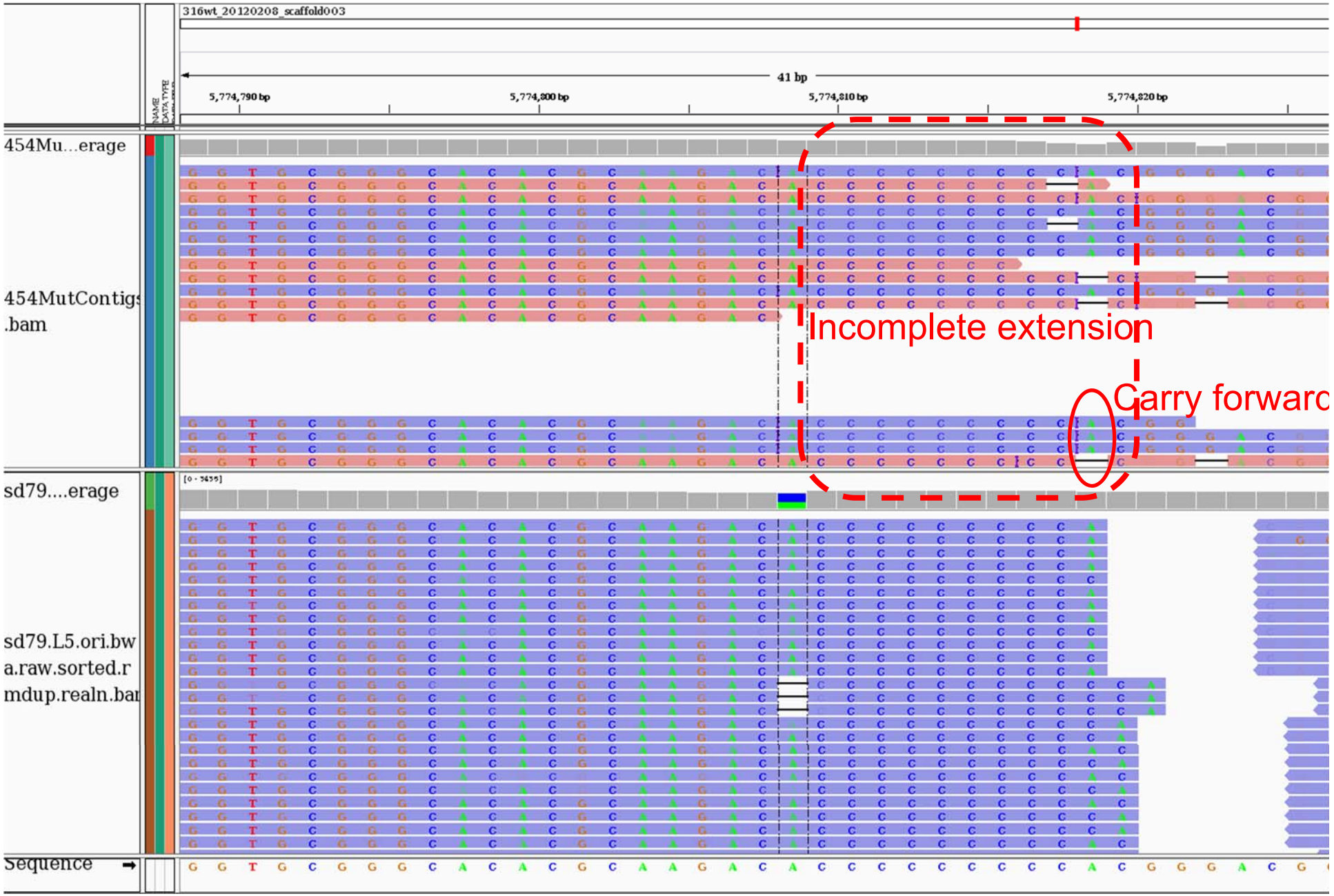




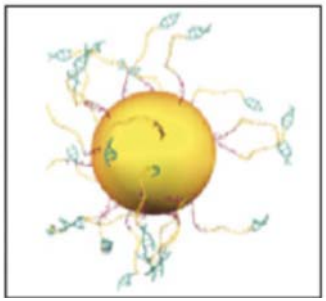




Homopolymer errors – 9 C's

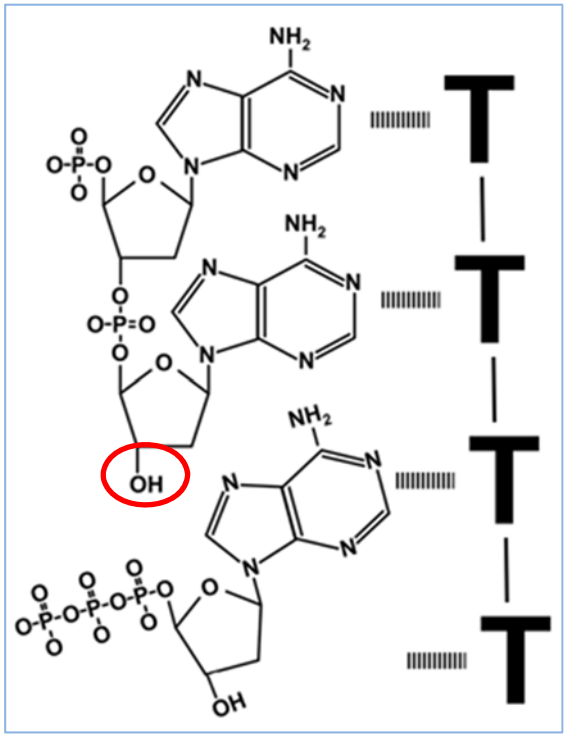


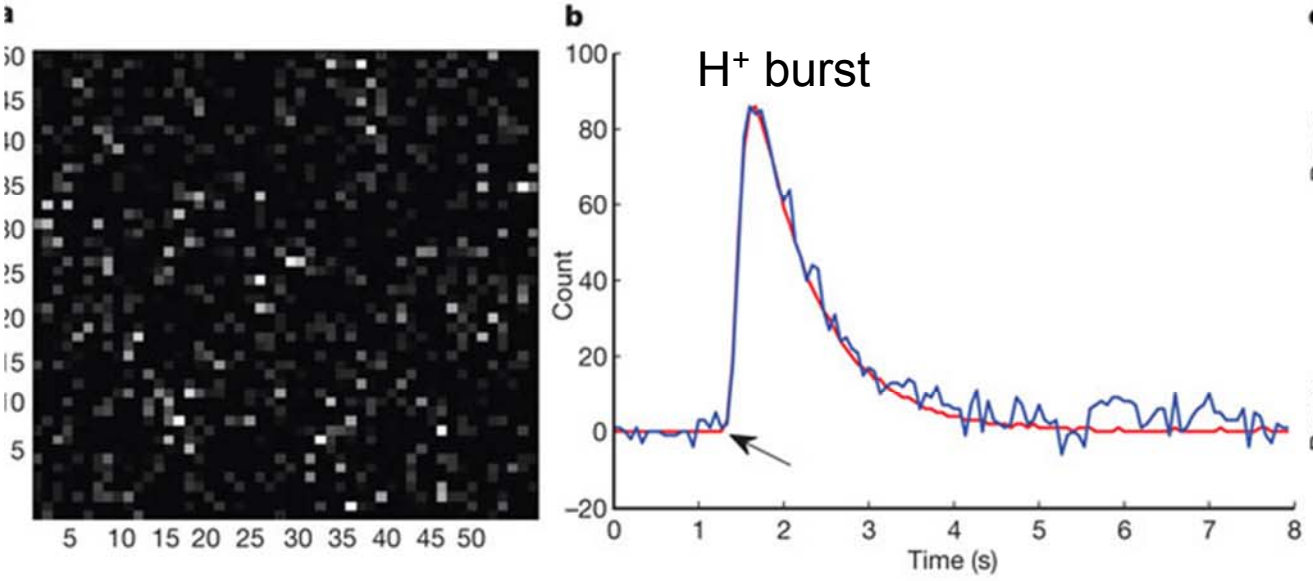
Ion Torrent/Proton:
Sensing bulk release of H^+



emPCR

Semi-conductor

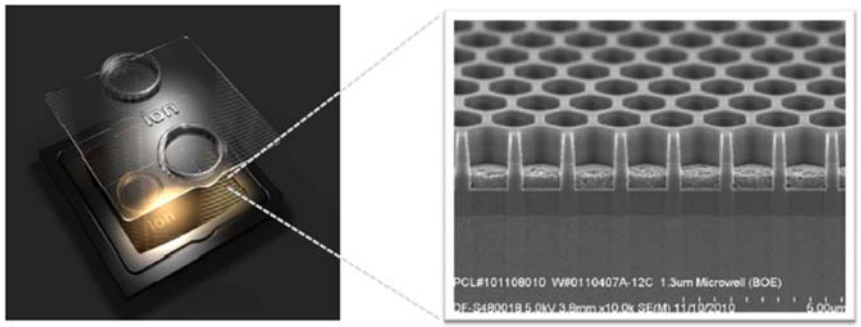


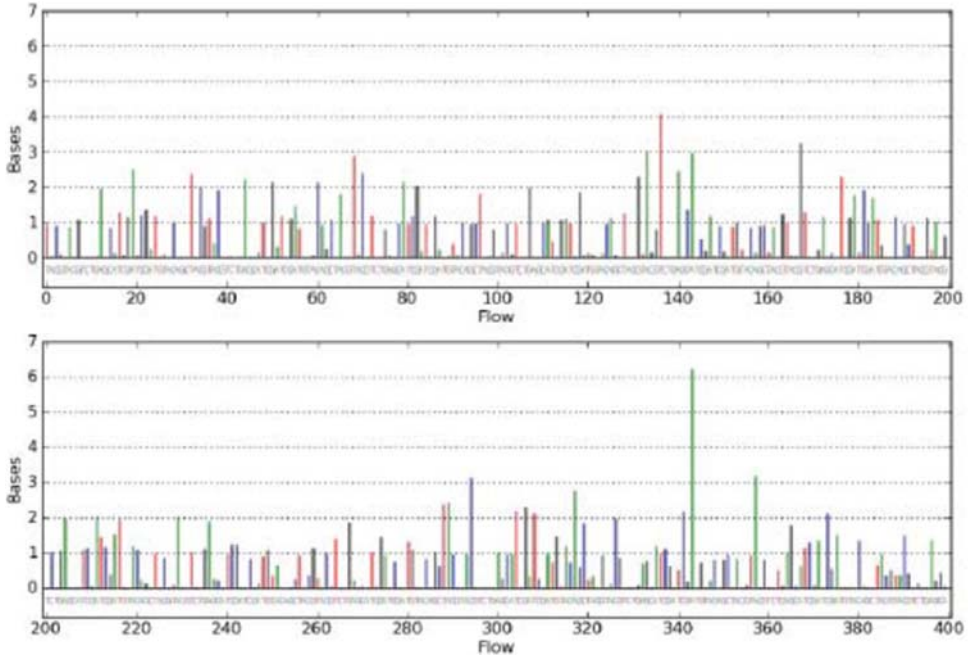


JM Rothberg *et al. Nature* **475**, 348-352 (2011) doi:10.1038/nature10242

Life Technologies – proton sensing

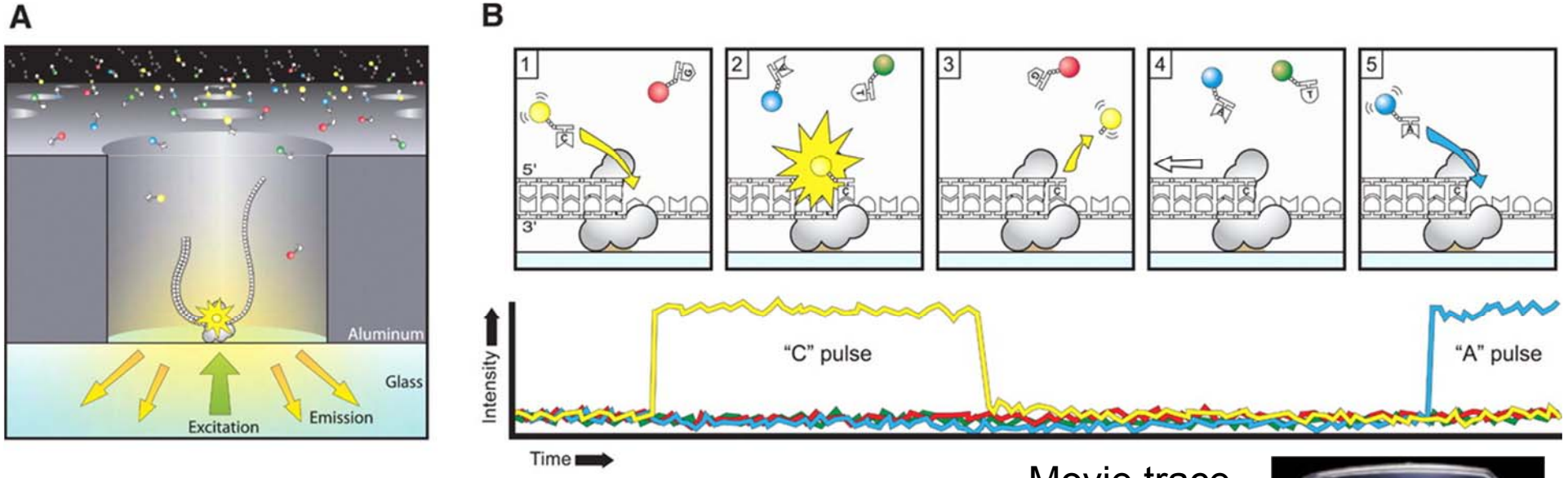
PGM™ for genes.
Proton™ for genomes.
Sequencing for all.





PacBio: 3rd-Gen SMRT Sequencing

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

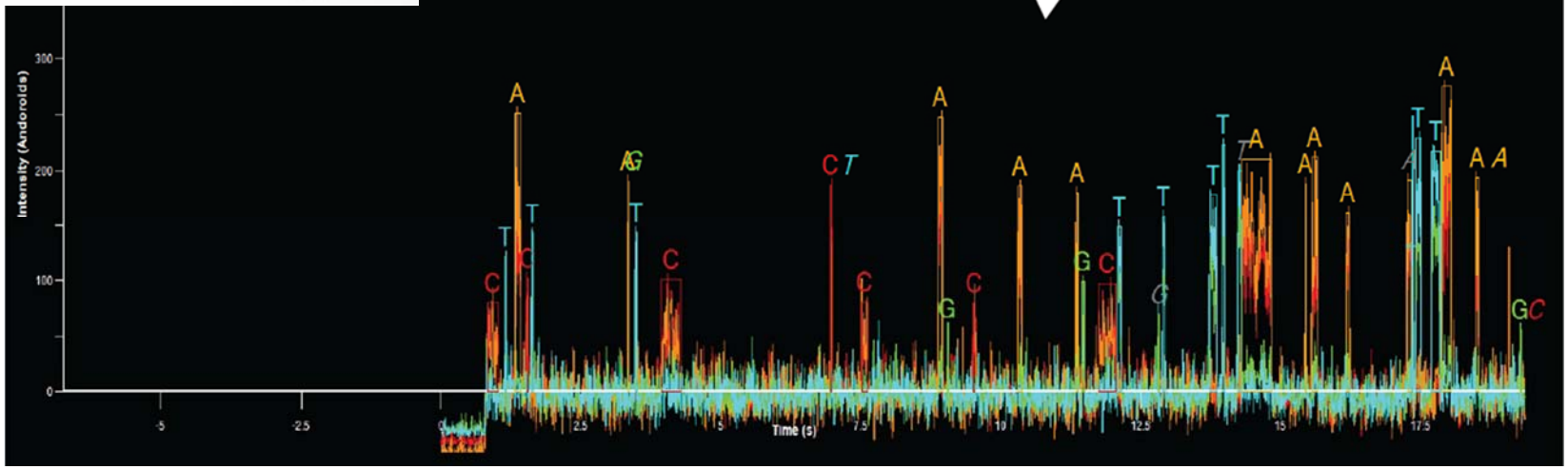


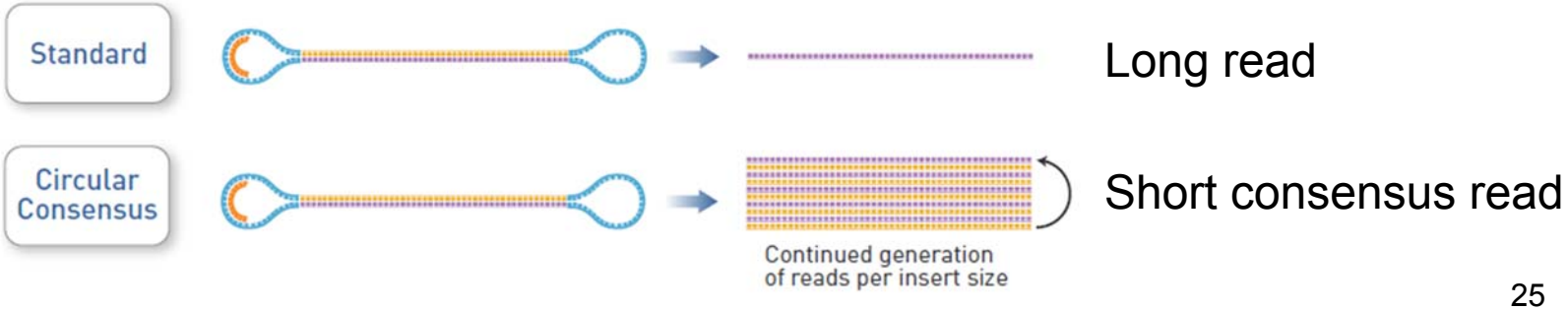
Movie trace

Read length: avg. 1.3 kb, up to 3.4kb
Throughput: 45 MB
Accuracy: 85% (1X) to 99.99% (30X)

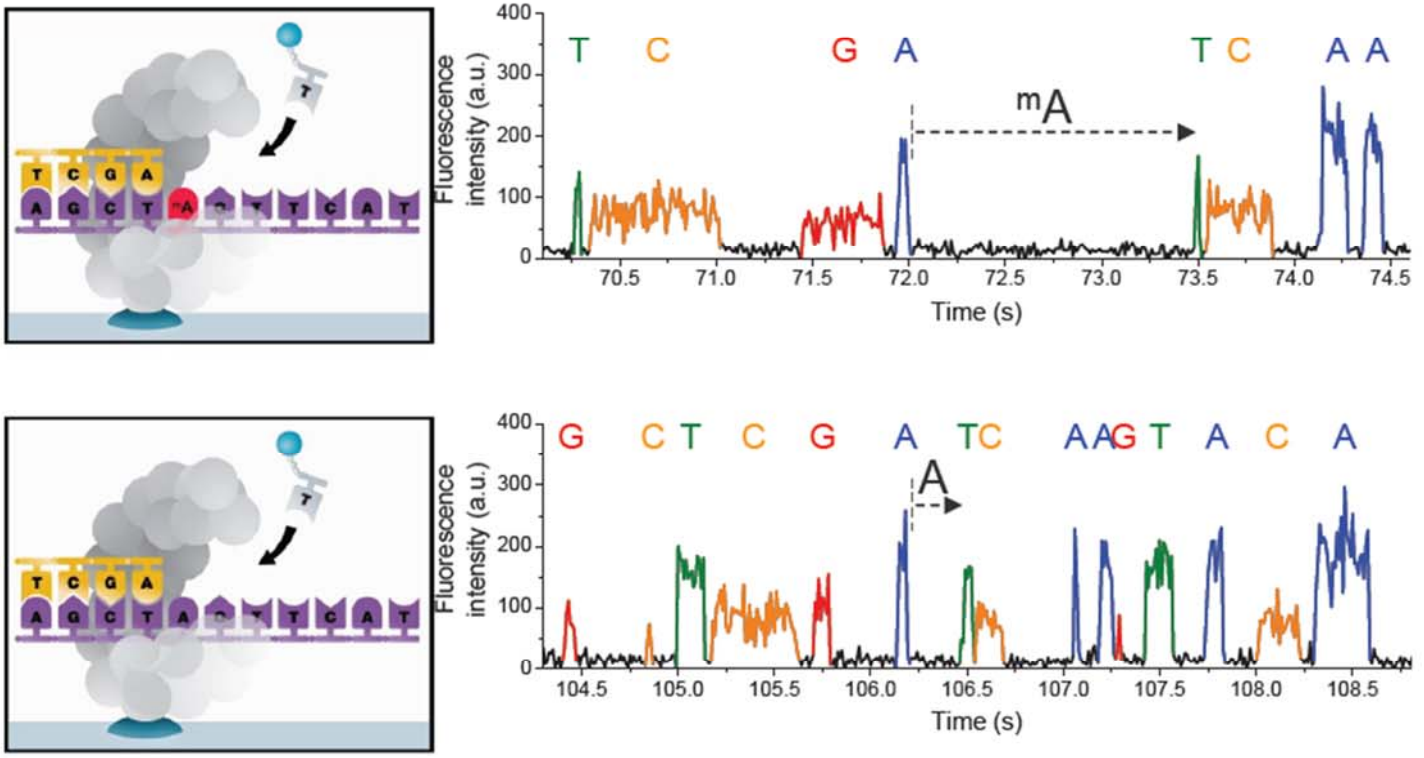
Signal Processing and Base Calling

Converting pulses of light into DNA bases and kinetic measures



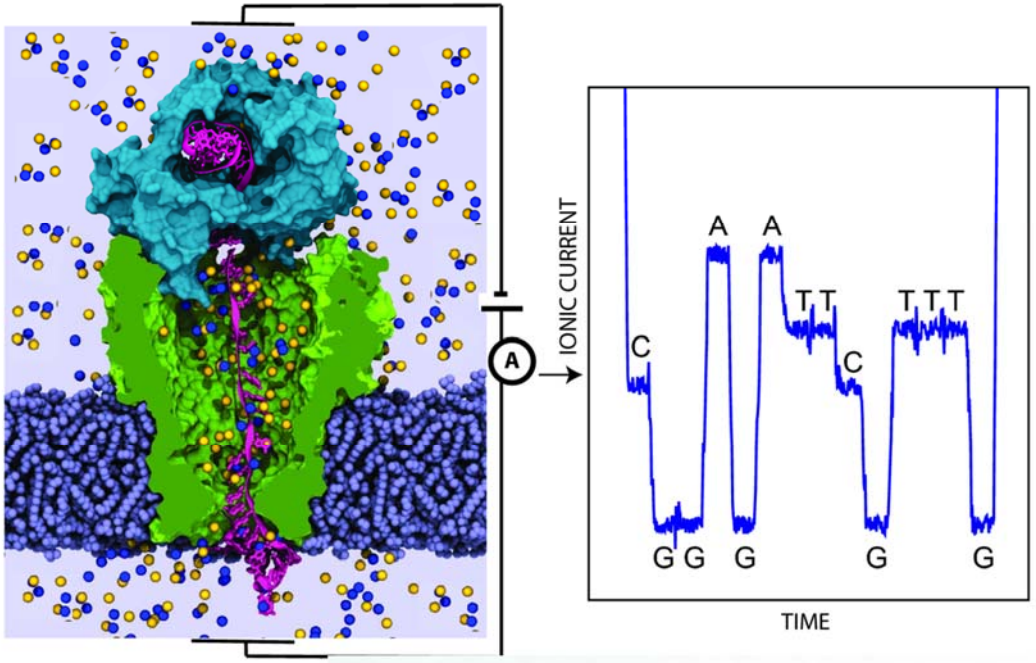


Key Feature: Kinetic Information



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

NanoPore Sequencing Technology

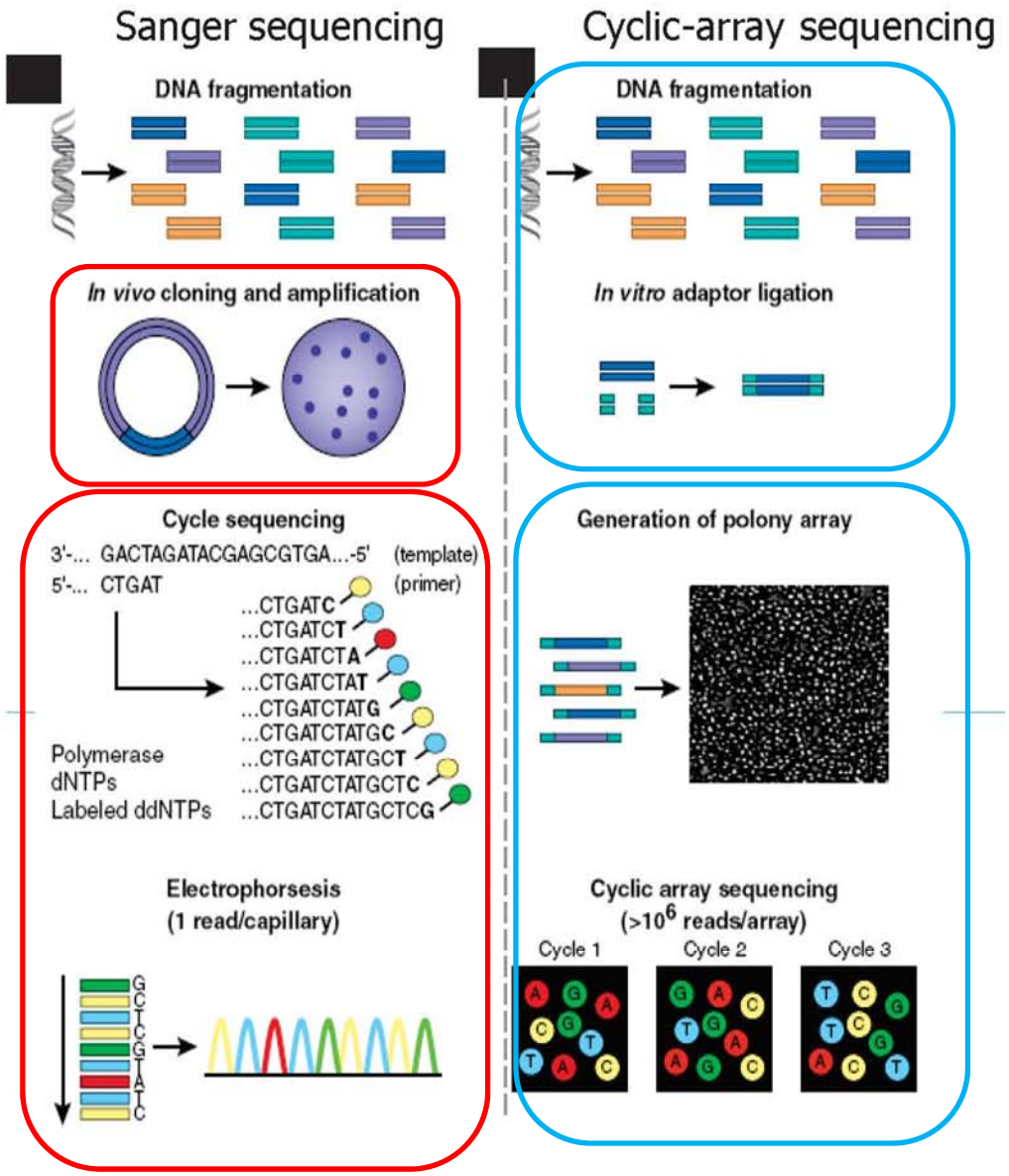


Read length comparisons of NGS platforms




	Roche 454 +	Illumina	Ion Torrent/H^+	PacBio
Chemistry	Pyrosequencing	Cyclic reversible terminator	Proton sensing by semi-conductor	Real-time Fluor. DNA polymerization
Throughput Per run	Jr.: 60Mb 454+: **800 Mb**	MiSeq: **15 Gb** HiSeq: **1 Tb**	Torrent: 700 Proton: **10 Gb**	C4 P6: 800 Mb Sequel: 7 Gb
Read length	**avg. 400-1000 nt**	**50-300 nt; SLR @ 5-15kb**	**200-400 nt**	**5-15 kb**
Data quality	>99.9% homopolymeric errors; tolerate high GC regions	>99.99% No homopolymer concern; more susceptible to high GC	>99.9% homopolymeric errors	Raw ~85% homopolymeric errors; single base INDELs
Application	Large De novo assembly; Long amplicons	De novo assembly; Re-sequencing	Re-sequencing	Genome assembly /scaffolding; structural variation

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*

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

II. Project considerations & Sequencing plan

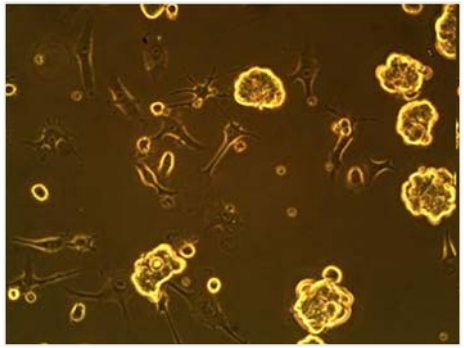
- Factors
- Sample QC
- Library QC

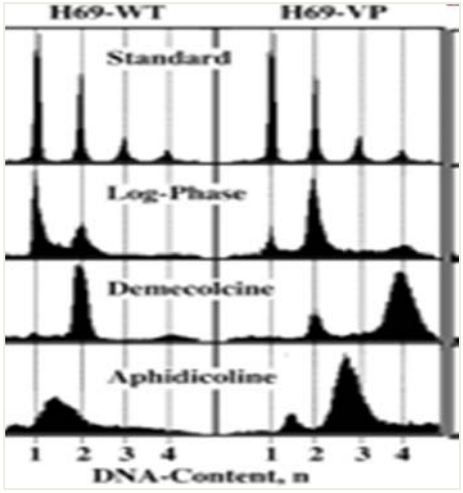
NGS project considerations (1)

- 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

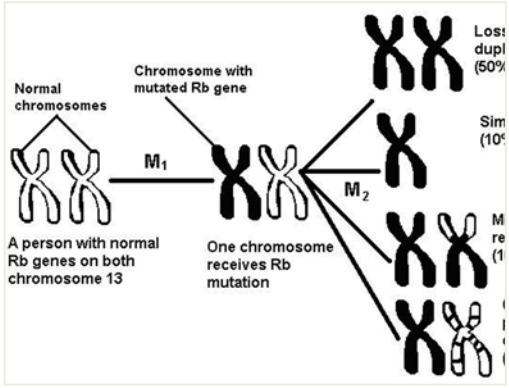
- Pure strain?

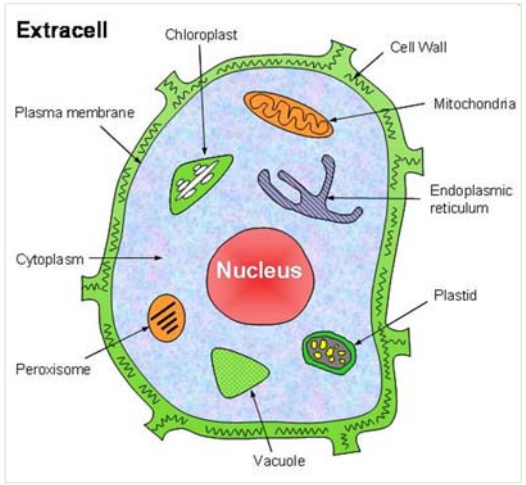




- Genome ploidity?

- Heterozygosity?

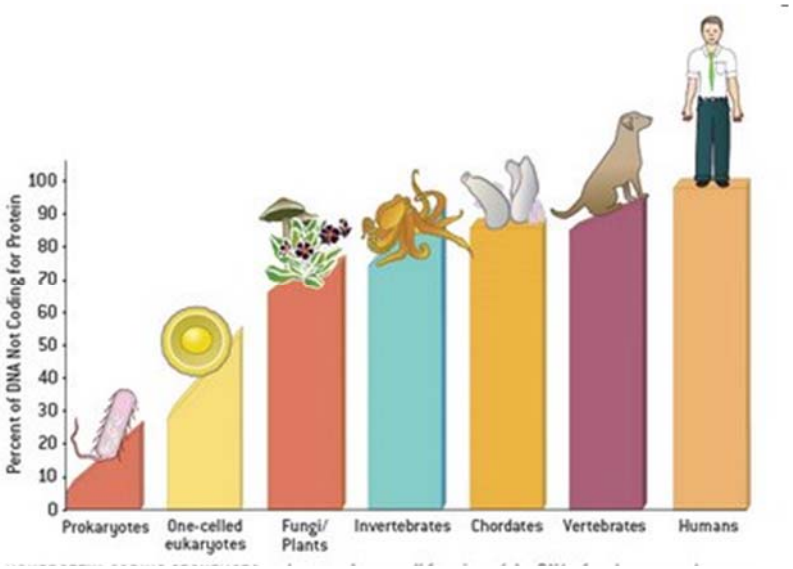




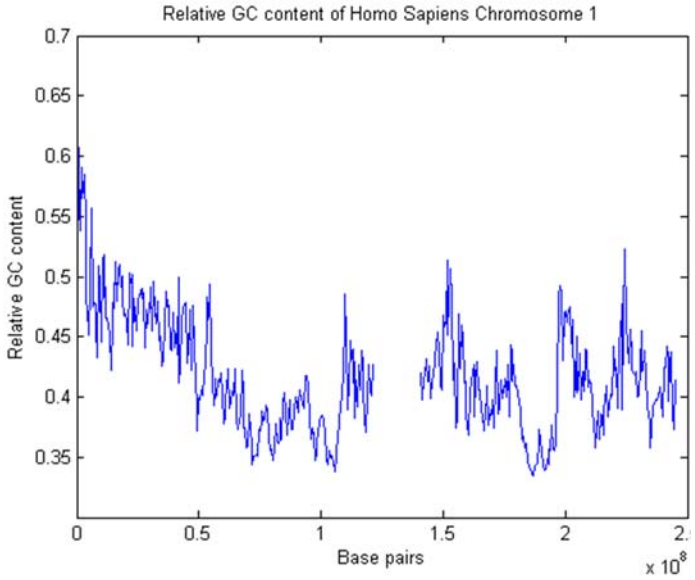
- Plastids (mitochondria, chloroplast)?

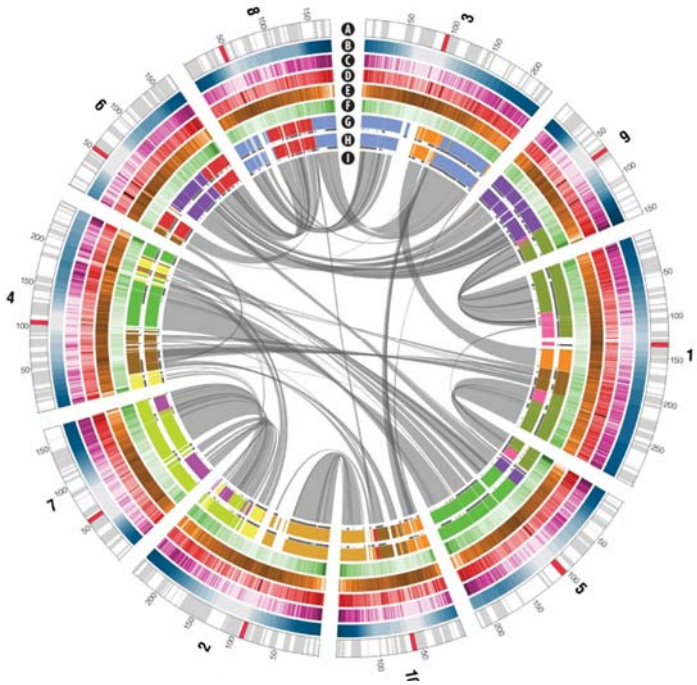
Genome consideration

- Genome size



- GC%





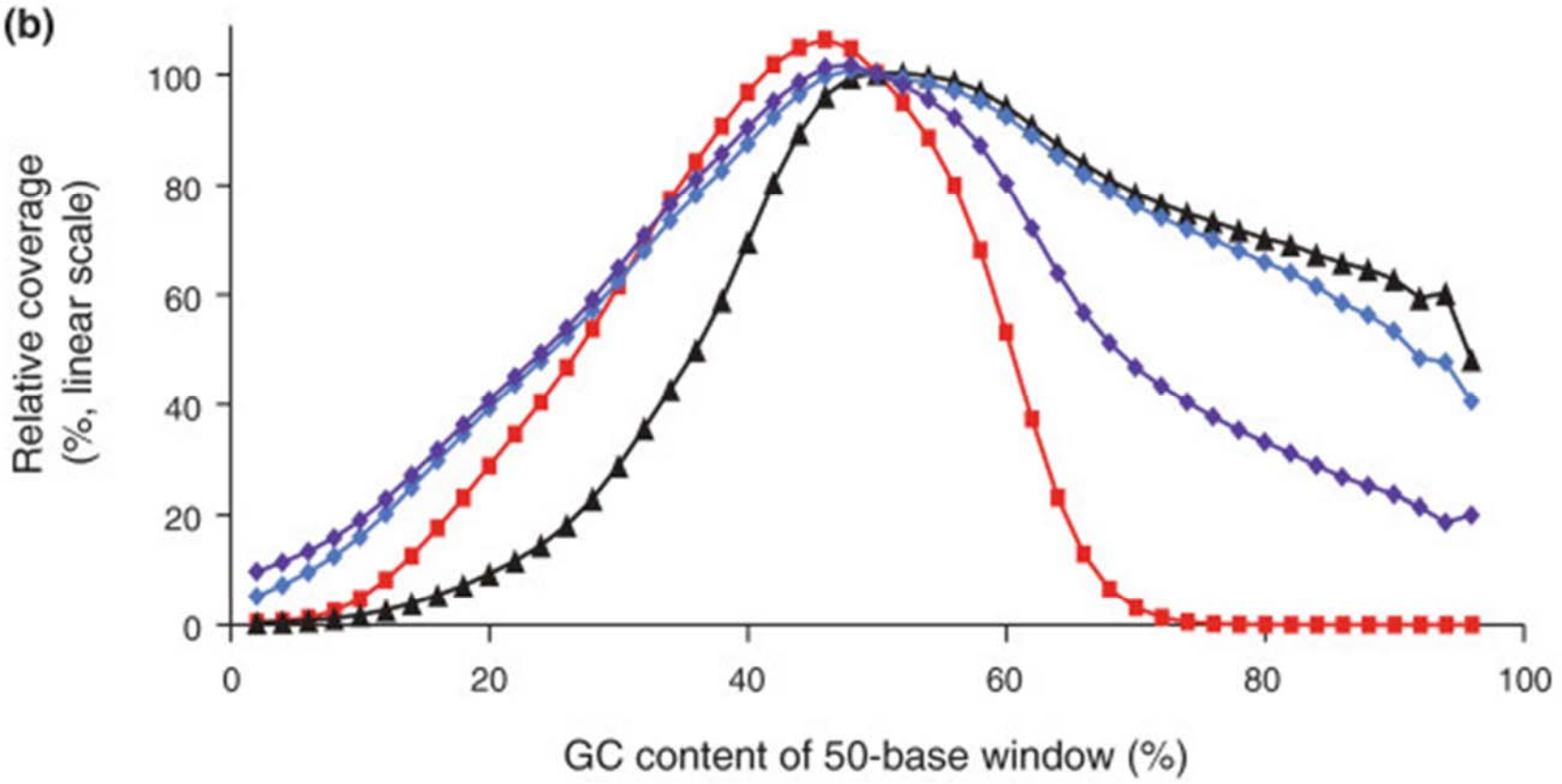
- Genome complexity
  - repeats, duplications…

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)?
  - 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



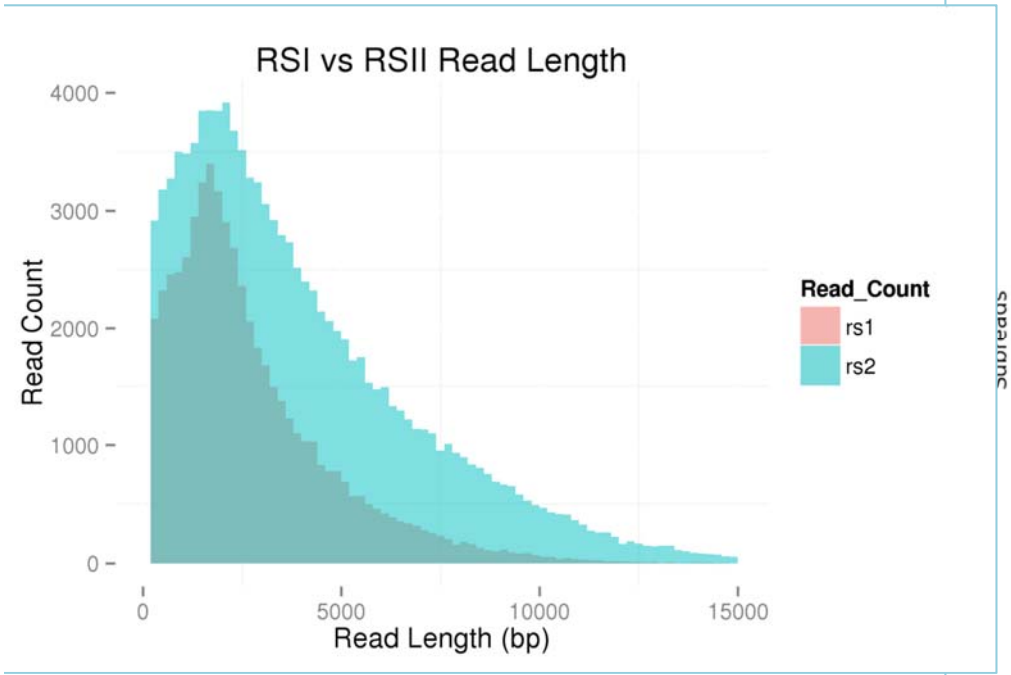
Red, Illumina PCR protocol
Others, modified protocols

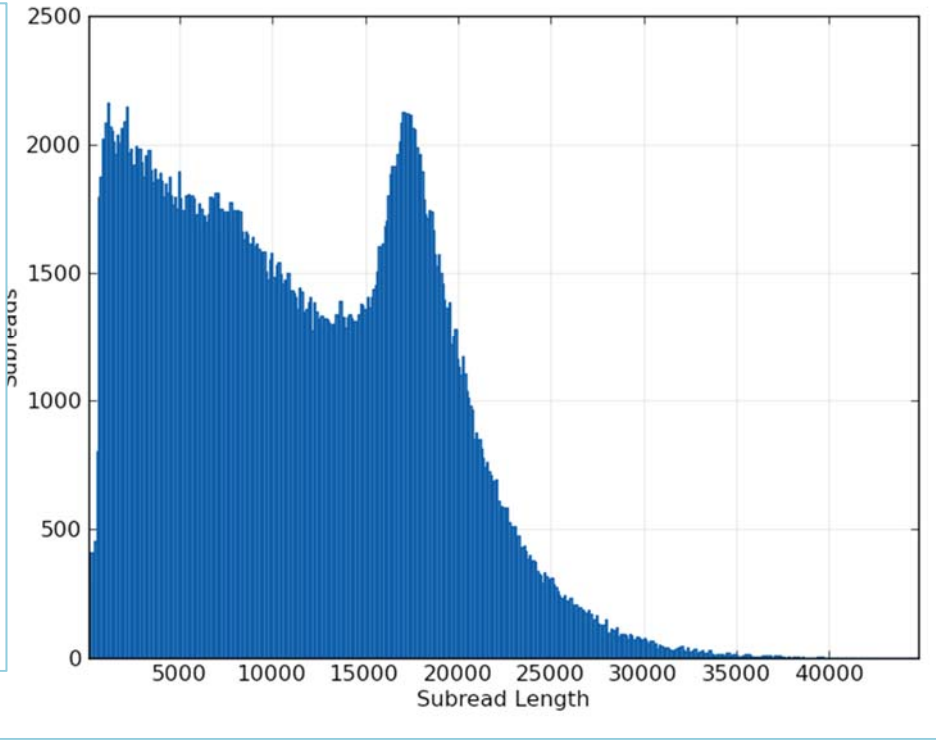
Aird et al., Genome Biology (201

NGS project considerations (3)

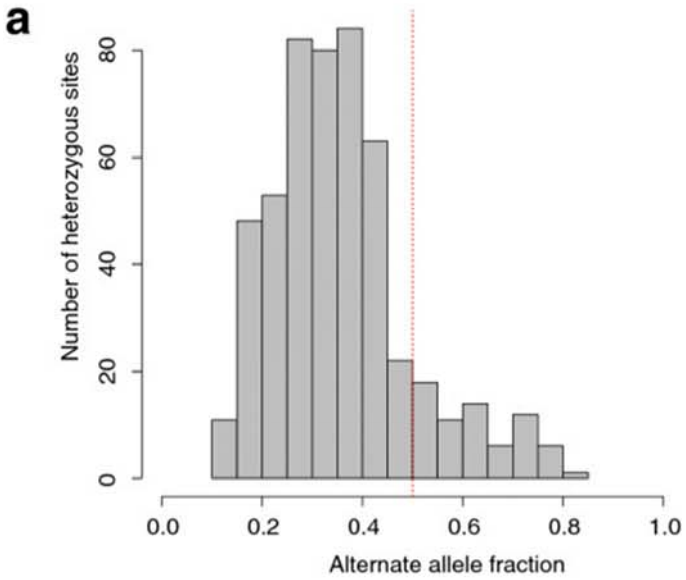
- NGS prep issues:
  - Sample input amount (normal vs low input)
  - PCR amplification (sample, target, library?)
  - Multiple size range required?
- Sequencing concerns:
  - Data: read length, type, base accuracy
  - Platform: strength vs weakness
  - Template bias from sequencing/imaging?
- Data scale: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

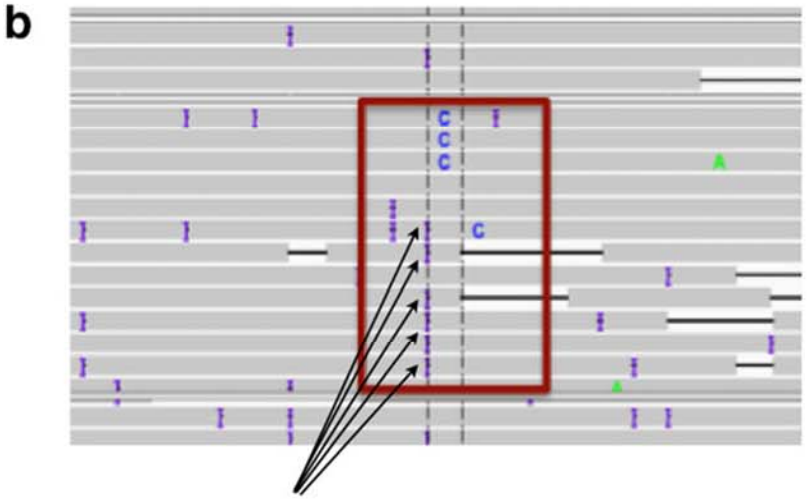
PacBio – current longest read lengths



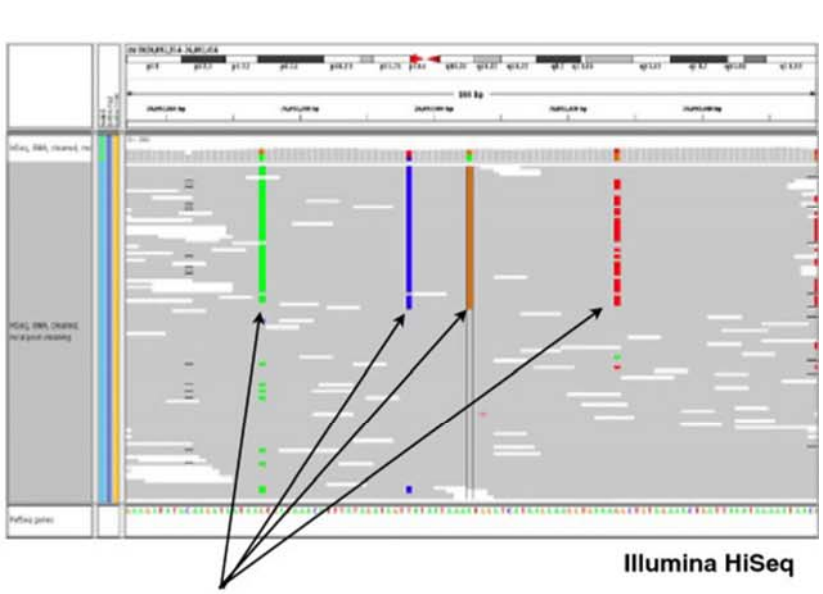


Error profile of Pacific Biosciences data

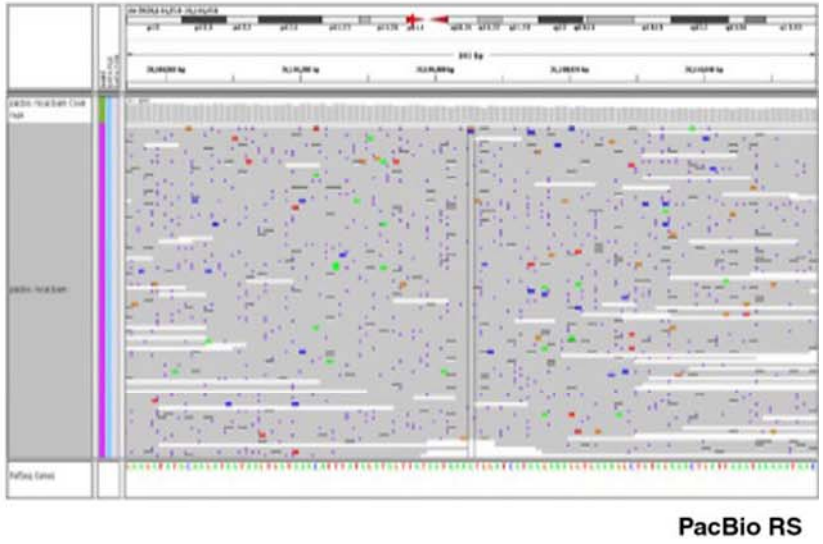




"C" INSERTIONS

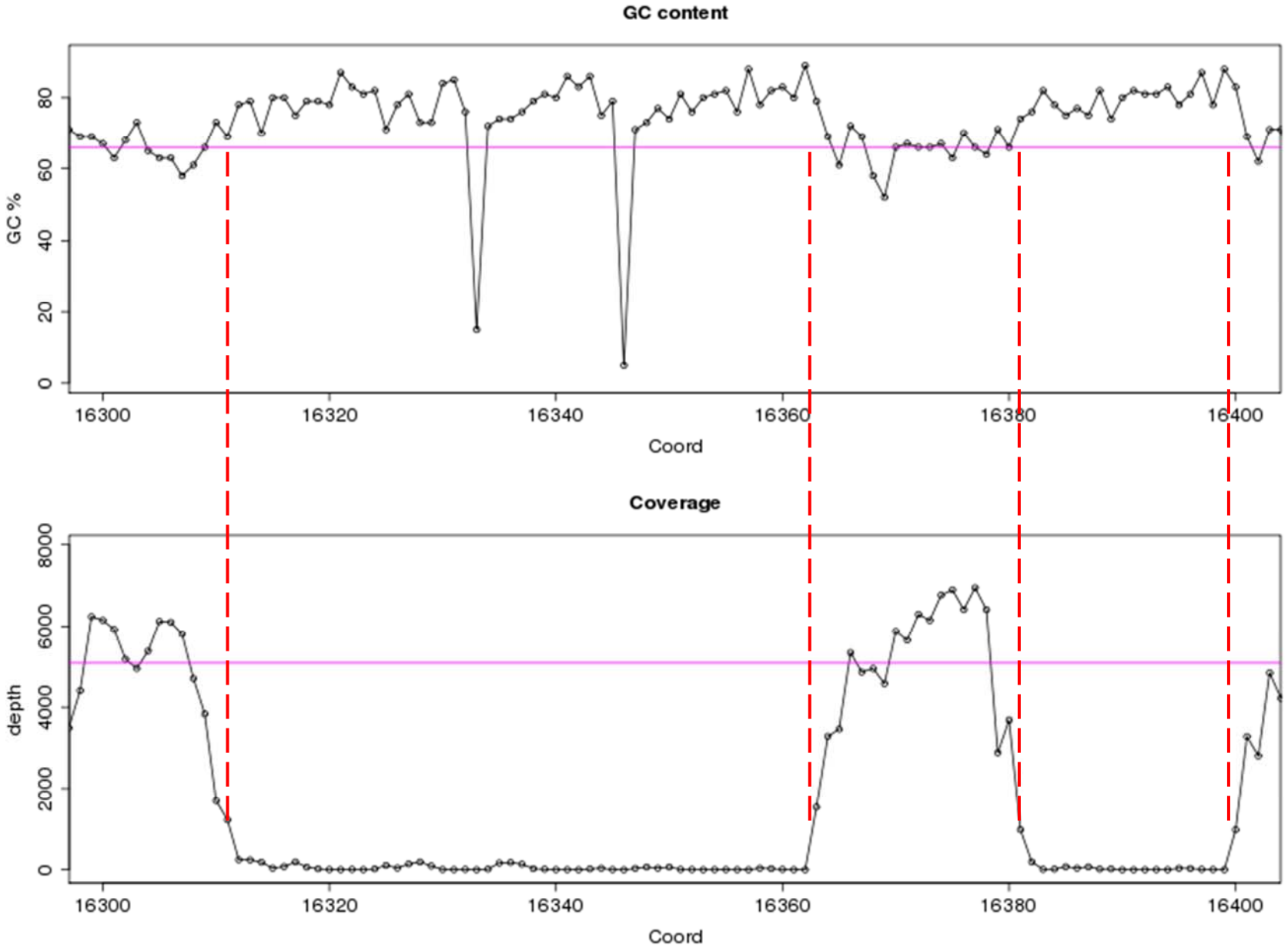


SYSTEMATIC ERROR



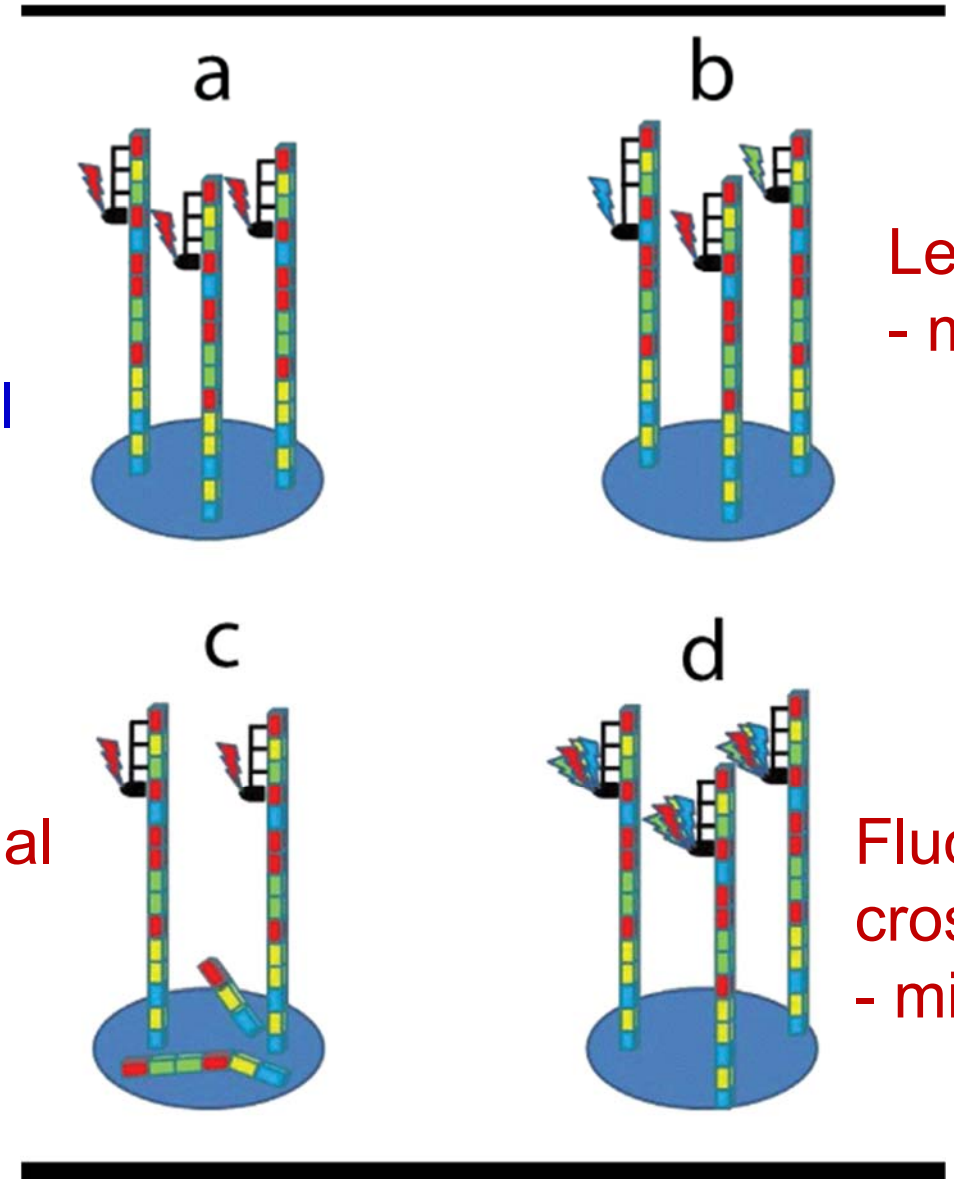
RANDOM ERROR bmcgenomics.biomedcentral.com

Coverage at high GC% regions - HiSeq data



Factors of Illumina signal noise

Normal extension
- Strong signal



Leading/lagging
- mixed signal

Loss of material
- Weak signal

Fluorophore crosstalk
- mixed signal

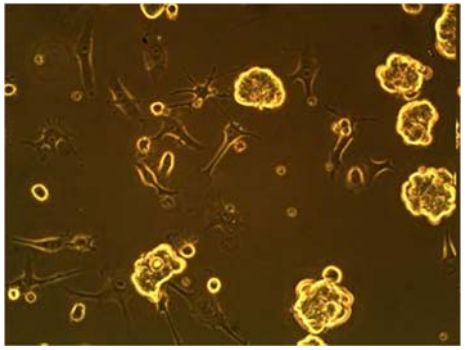
Datta et al., 2010 J ProteomicsBioinform 3: 183-1

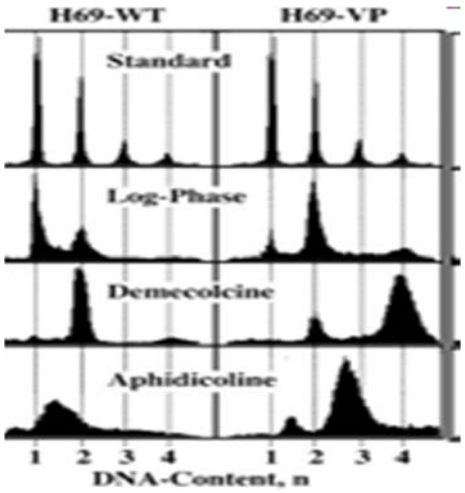
NGS project considerations (4)

- Bioinformatic algorithms:
  - Data type ~ suitable programs
  - Can handle long read length?
  - Can handle high heterozygosity?
  - Preference for SR or PE data?
- Statistical models?
  - Data normalization?
  - Significance of predictions?
- FDR vs FPR?
  - Data amount
  - Data quality

Sample considerations

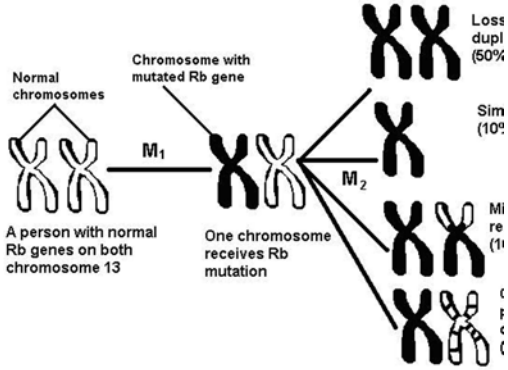
- Pure strain?

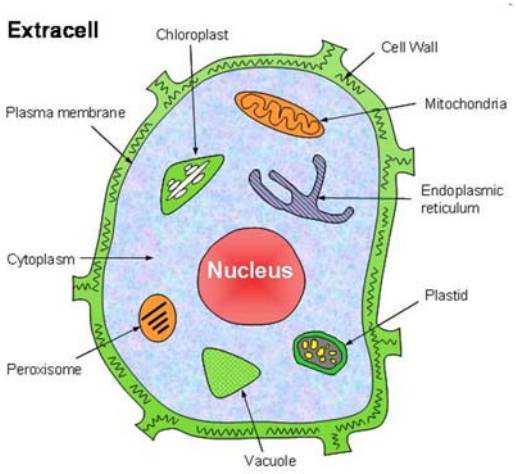




- Genome ploidity?

- Heterozygosity?

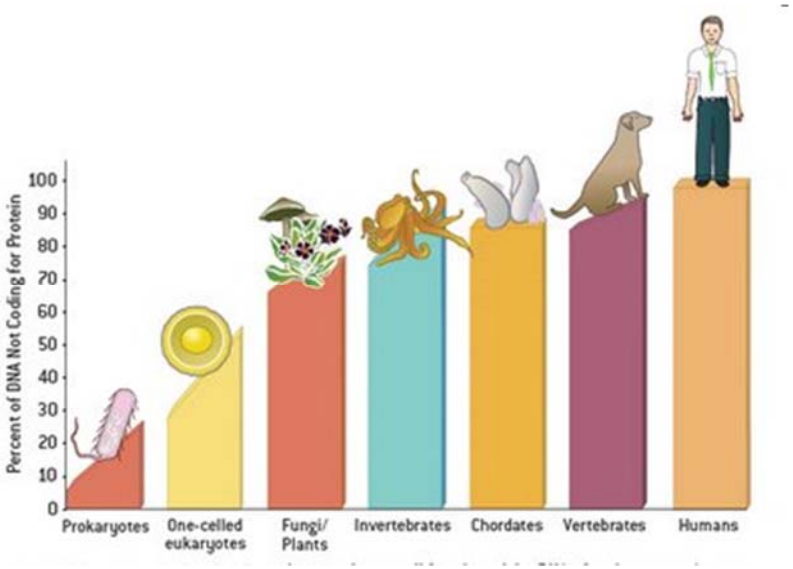




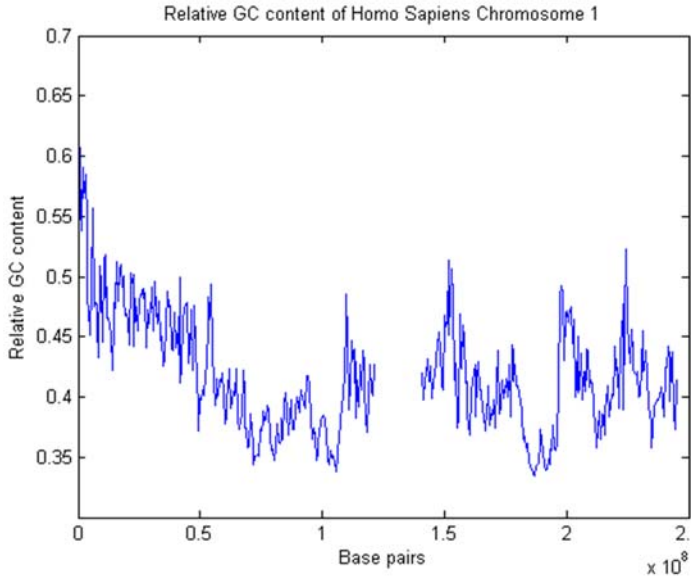
- Plastids (mitochondria, chloroplast)?

Genome consideration

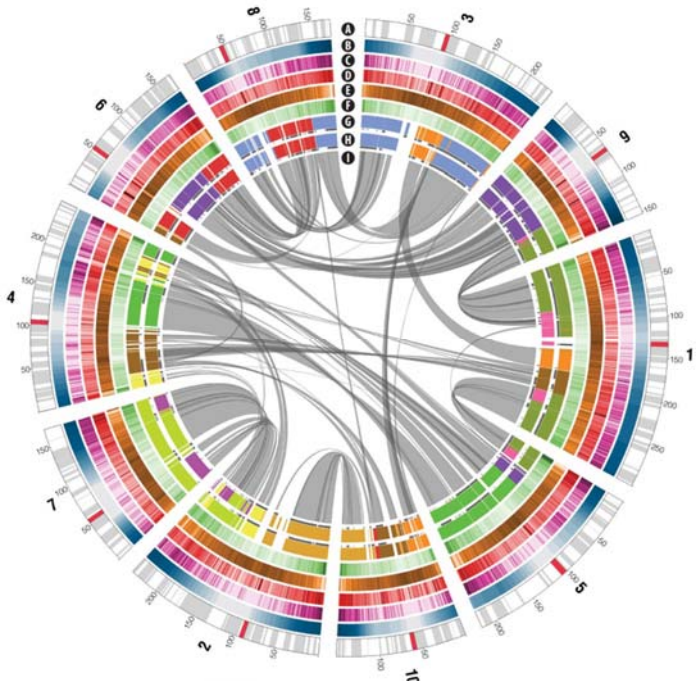
- Genome size



- GC%



- Genome complexity
  - repeats, duplications…



Data requirement
(assembly consideration)

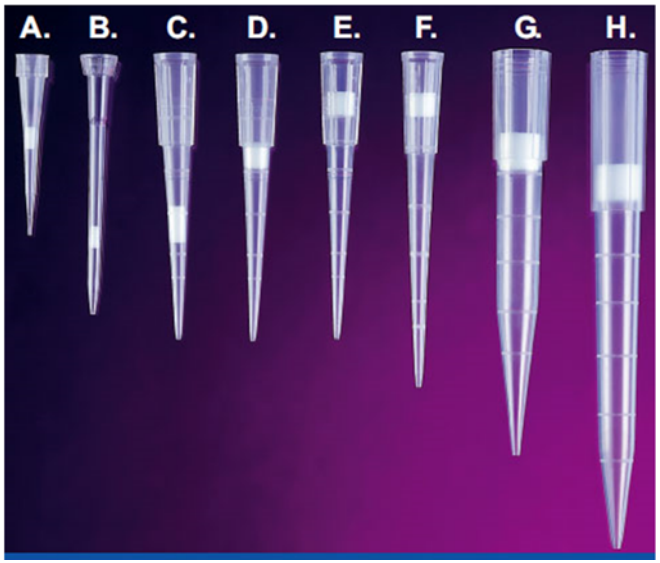
- NGS platform
- Read length
- Sequencing depth (Fold coverage)
- Single Read vs Paired-end

III. Good Lab Practices & DNA/RNA preparation

Lab wear and clean bench

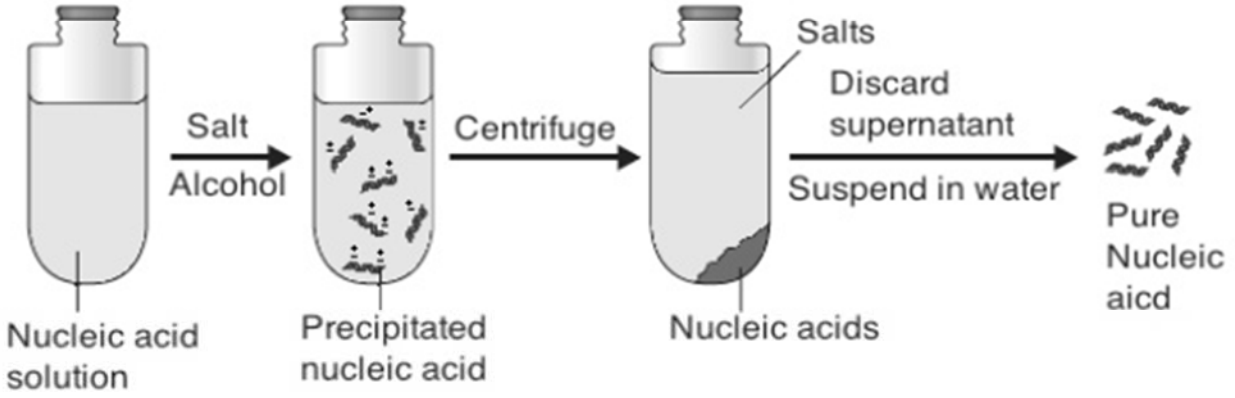
Plastic wares

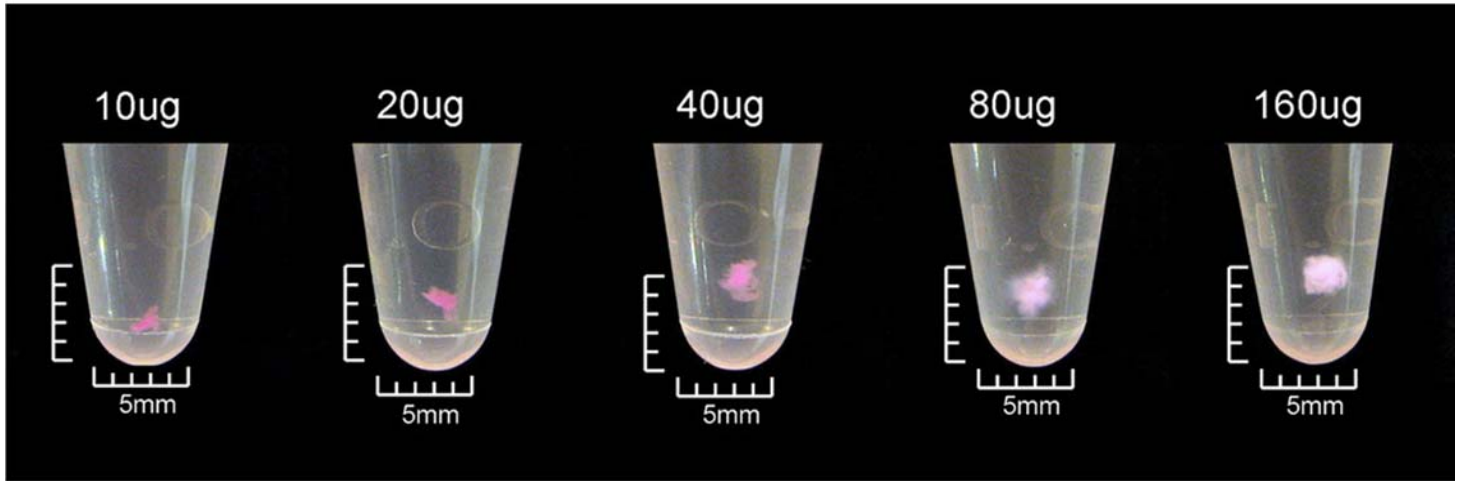
1. Dnase/Rnase free (pre-sterilized)
2. non-sticky / Low-bind



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.





Sample QC

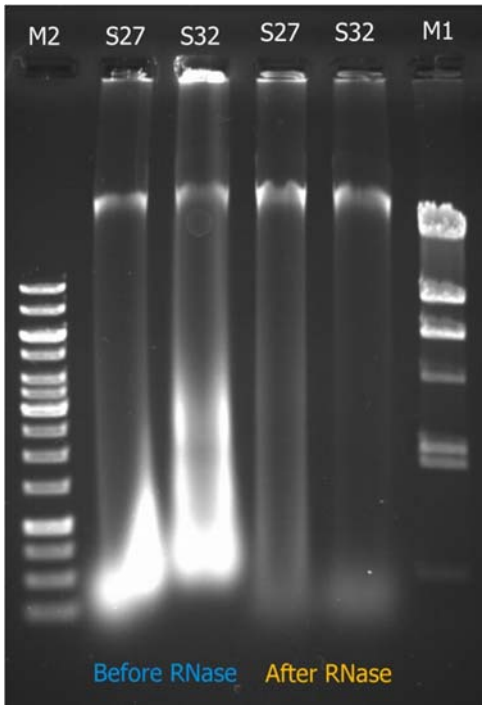
Ampure XP magnetic beads

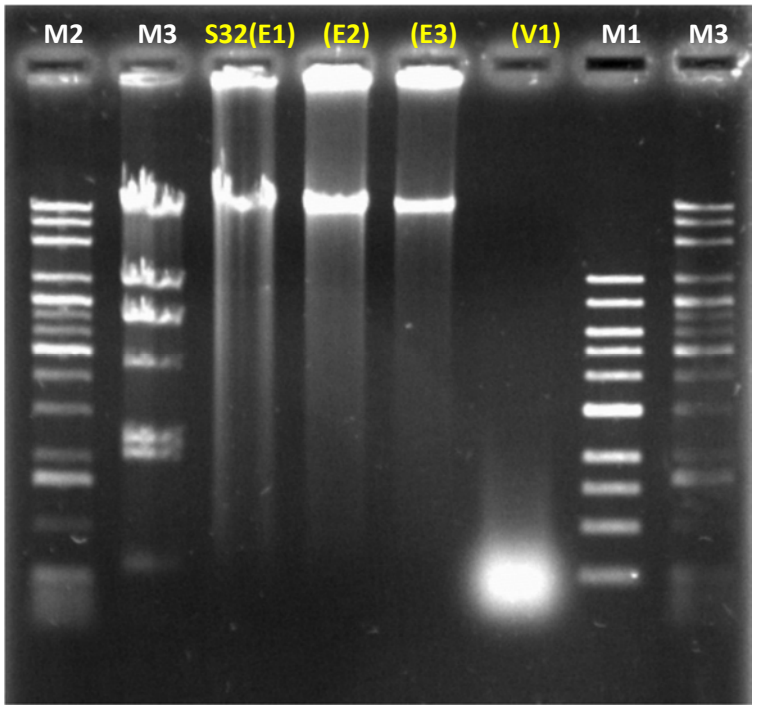
Qubit Fluorometric Assays

BioAnalyzer and Chip kits

Dark Reader & Cyber dyes

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

Genomic DNA sample info

Organism or Species	Myccna sp. (fungus)	Genome Size (Mb)/DNA Length	~44Mb.

	Sample Name[2] (tube labeling)	Sample Type[3]	Quantity & Quality[1] (measured by: ☑ NanoDrop ☑ Qubit ☐ BioAnalyzer) Conc. (ng/ul)	Vol. (ul)	Amount (ug)	OD 260/280	OD 260/230	Notes	Sample ID (core only)
1	Kd_0922_9	g	Nano 314.8 Qubit 394	~40	15.7	1.82	1.23	PE(2×300), MP	SG-HY01
2	Kd_hap_0922_12	g	427.3 / 350	~25	8.8	1.87	1.99	PE(2×300)	SG-HY02
3	CT6-001_1006	g	25.2 / 47	~50	2.4	1.94	2.04	PE(2×300)	SG-HY03
4		g							
5									
6									
7									
8									

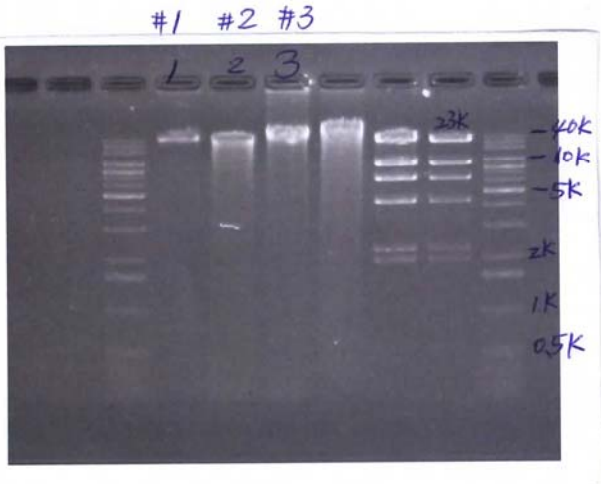
Dissolved in: ☐ H_2O ☐ EB ☑ Other: kit TE buffer extraction

Enzyme Treatment
☐ DNase
☑ RNase
☐ RNase Inhibiter

Brand & Amount of Usage

Purification Method
OmniPrep kit

Gel Images **Post-run staining only.** Please indicate **sample no. & major marker sizes** (ladder should cover at least 0.1-10kb).
If BioAnalyzer profile is available, please attach the data at the end.

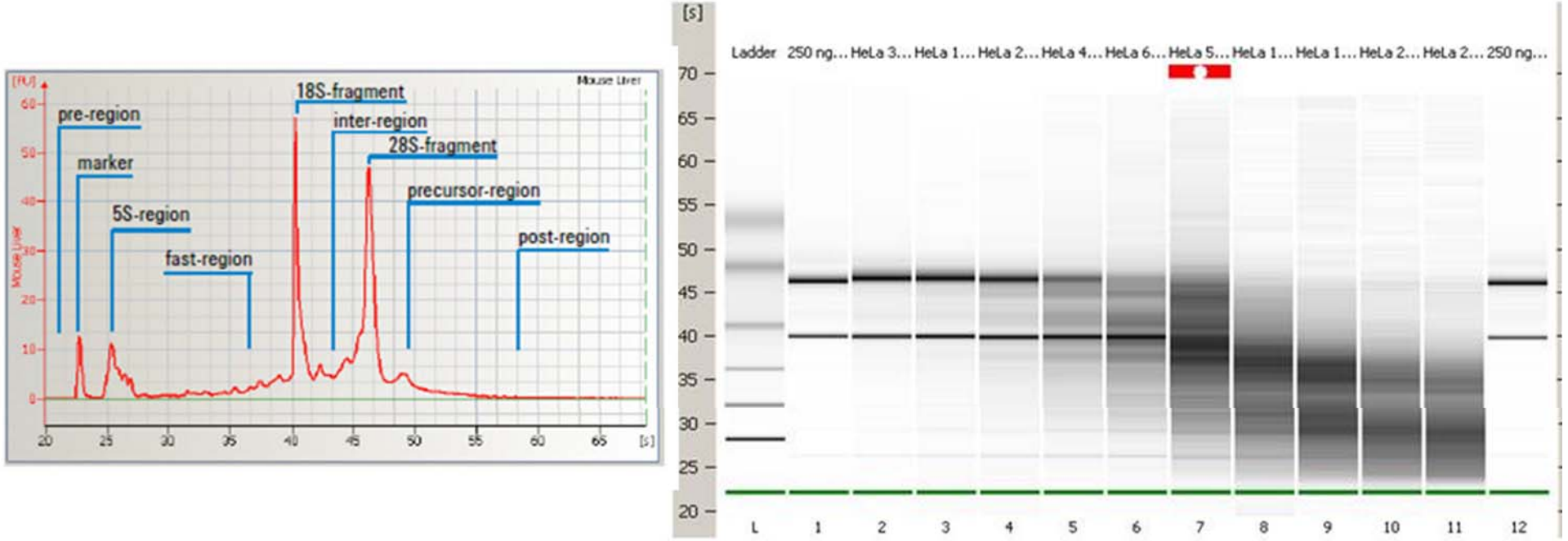


0.8% Agarose.
75V 30min
50V 80min

Loading amount:
sample 150 ng/lane
marker 100 ng/lane

外染 报仕特 clear view
20min

RNA integrity assessment by BioAnalyzer Pro



RNA integrity - BioAnalyzer

BioAnalyzer RNA ladder

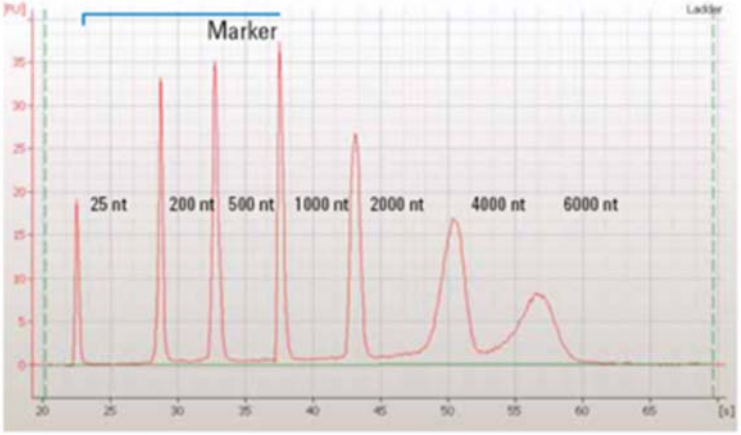
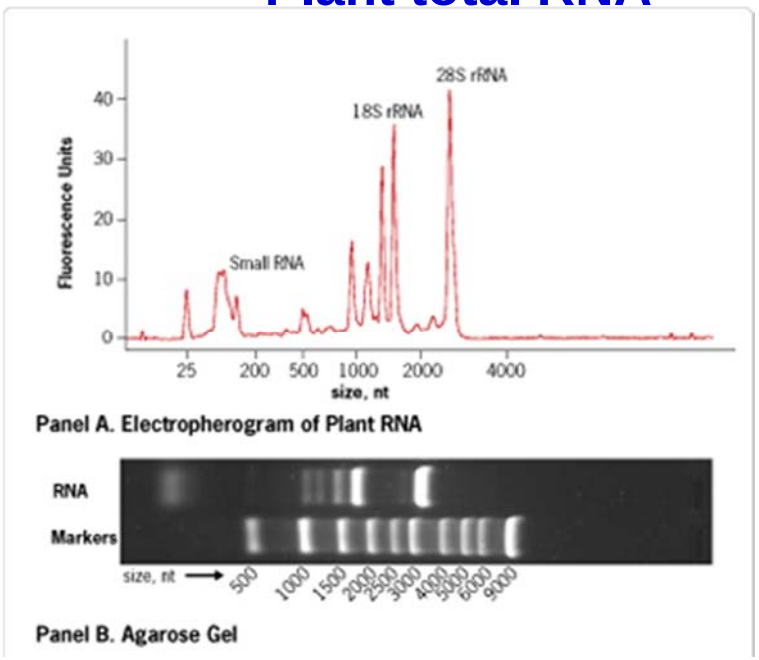
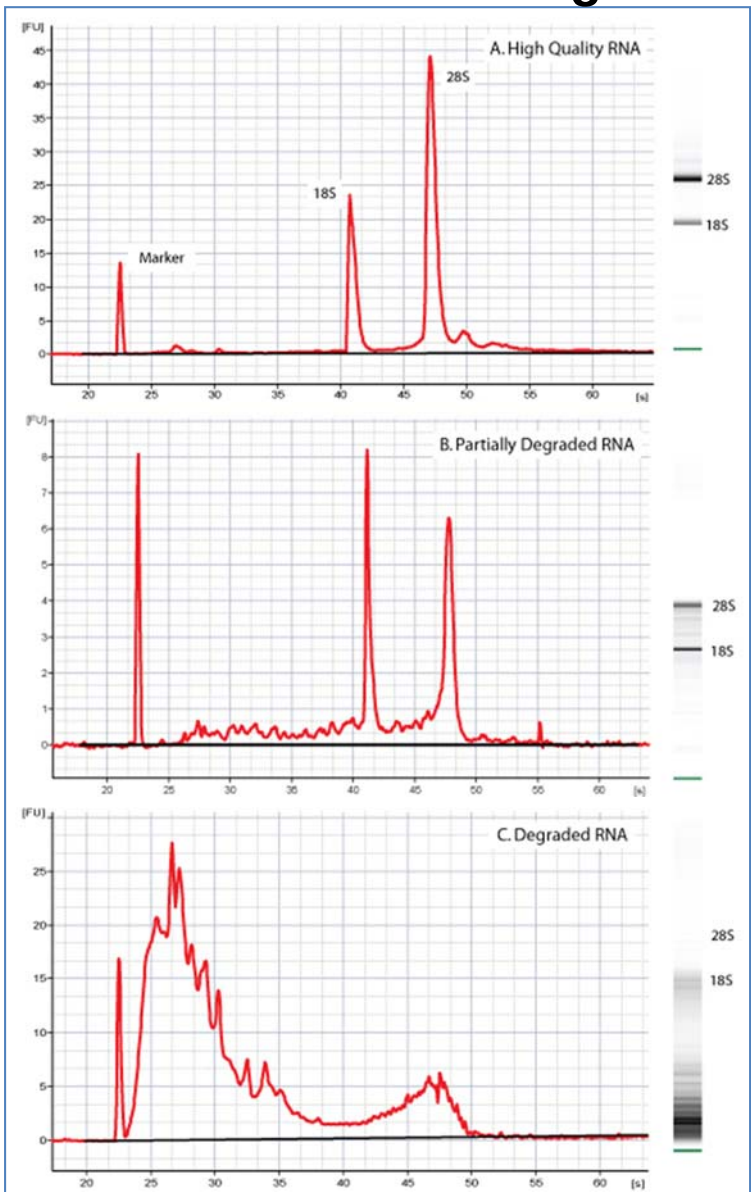


Figure 1 RNA 6000 Nano ladder

Plant total RNA

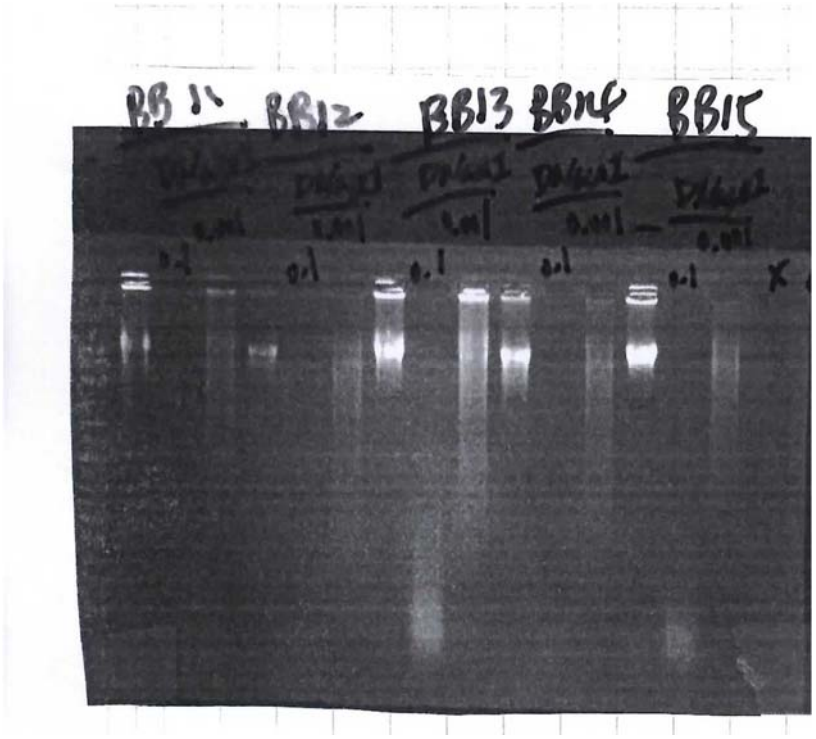


Human RNA – various degradation

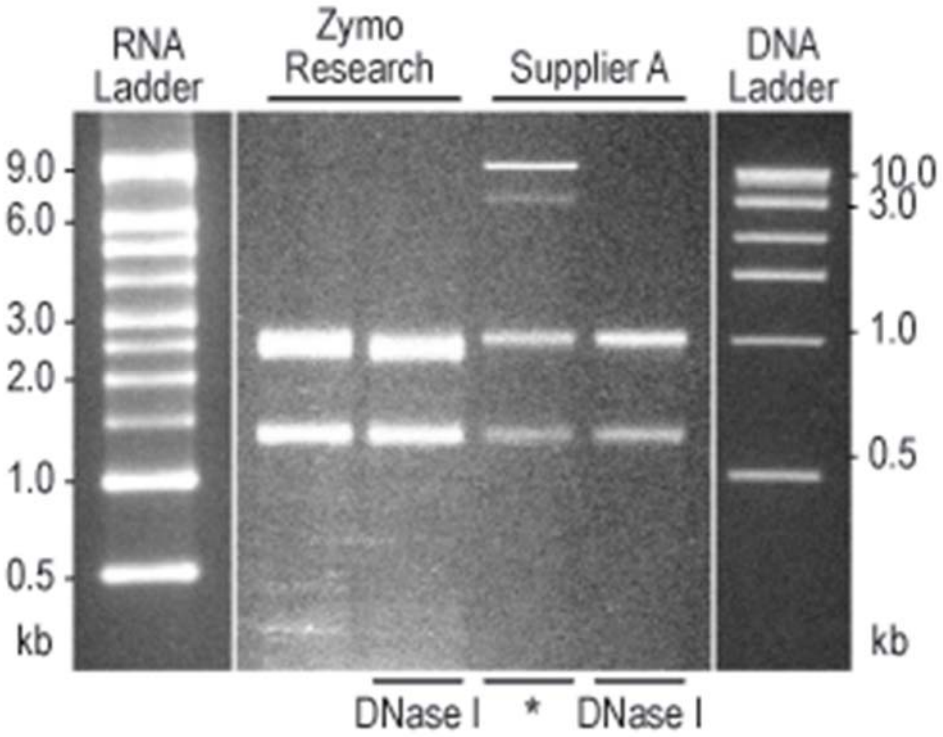


DNase I treatment

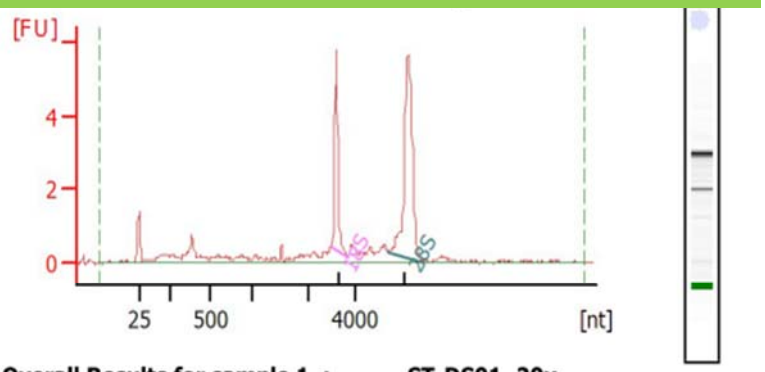
gDNA sample



RNA sample



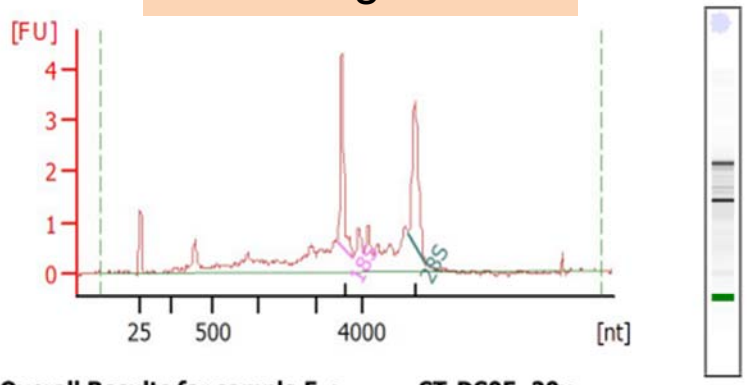
Good RNA: rRNA ratio>1.8, RIN>8



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

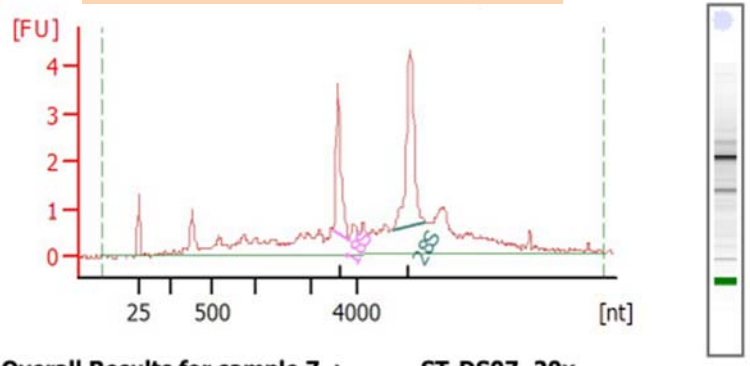
Some degradation



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

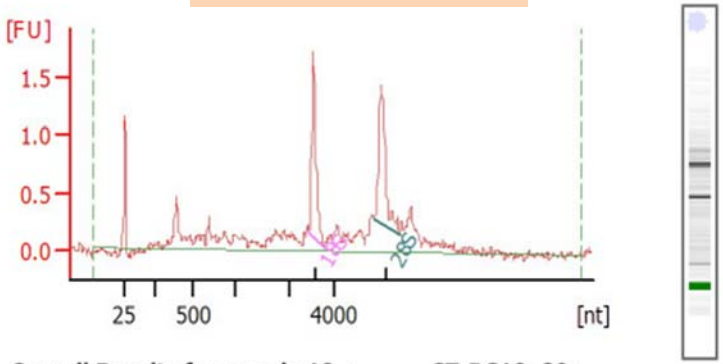
gDNA contamination



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

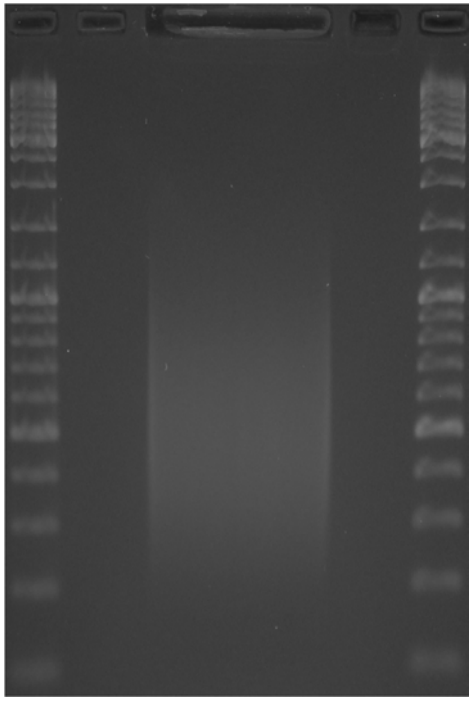
RNA Area:	14.3
RNA Concentration:	34 ng/µl
rRNA Ratio [28s / 18s]:	1.0
RNA Integrity Number (RIN):	7.1 (B.02.07)
Result Flagging Color:	
Result Flagging Label:	RIN: 7.10

Library QC

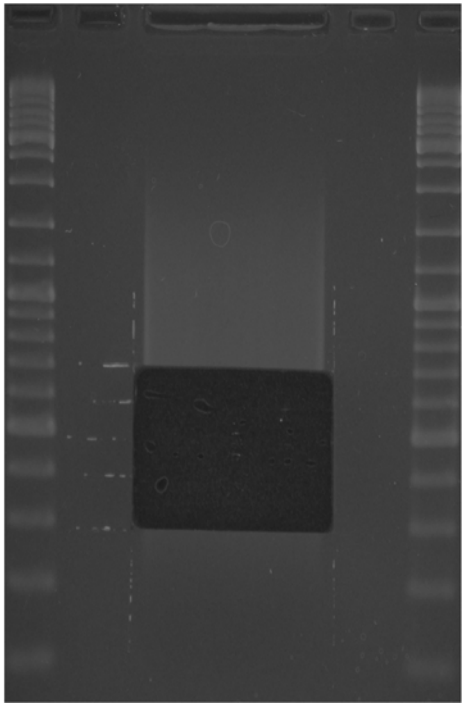
- Gel check
- BioAnalyzer
- Qubit qunaitifcation
- qPCR

Example: Shotgun gDNA library

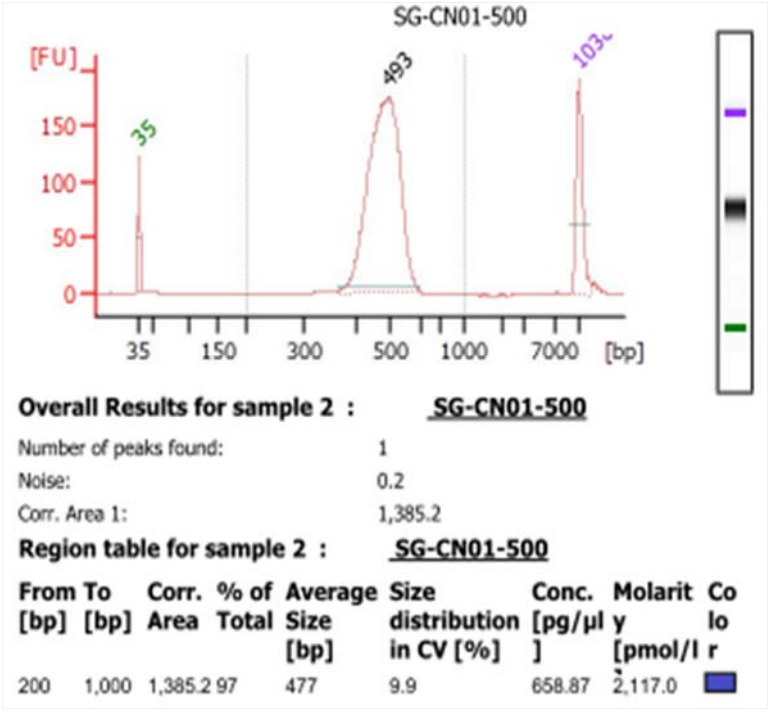
Total profile



Gel sizing



Final PCR library



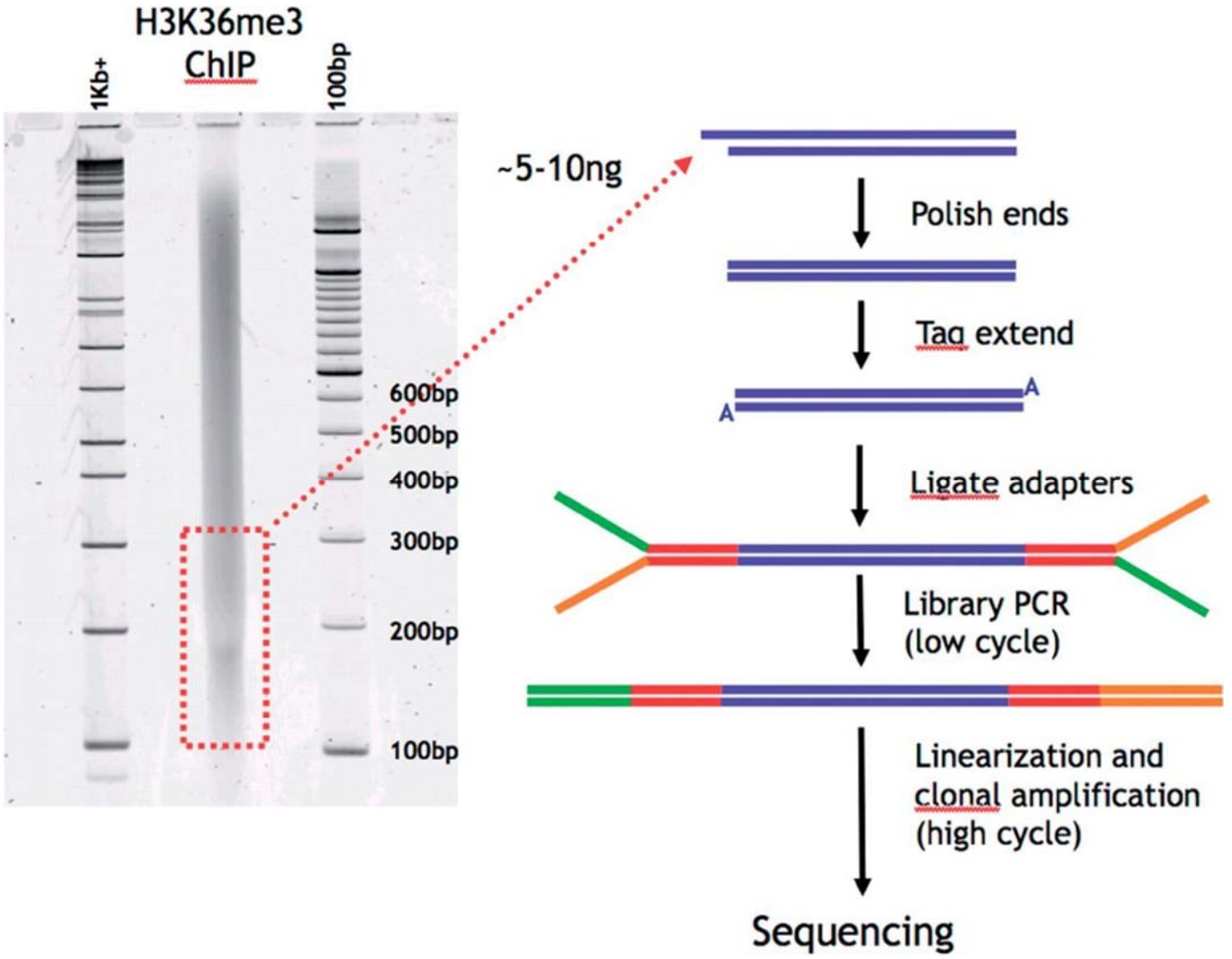
Overall Results for sample 2 : **SG-CN01-500**

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	658.87	2,117.0	

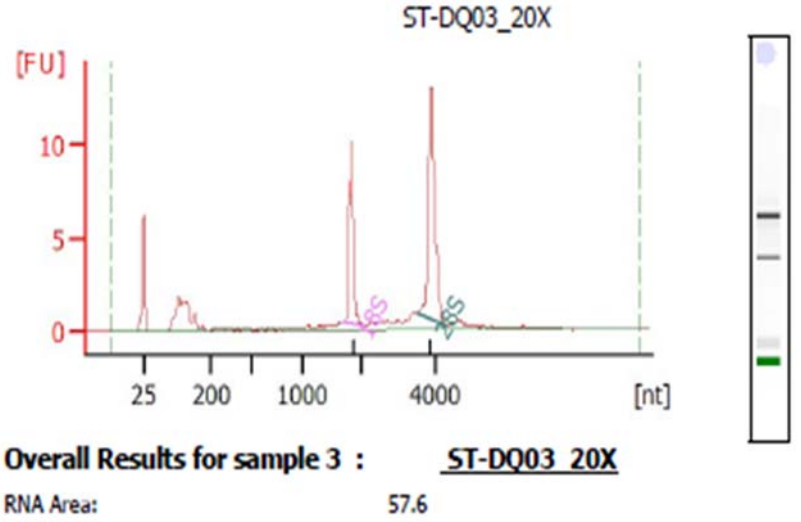
Overview of ChIP-seq construction

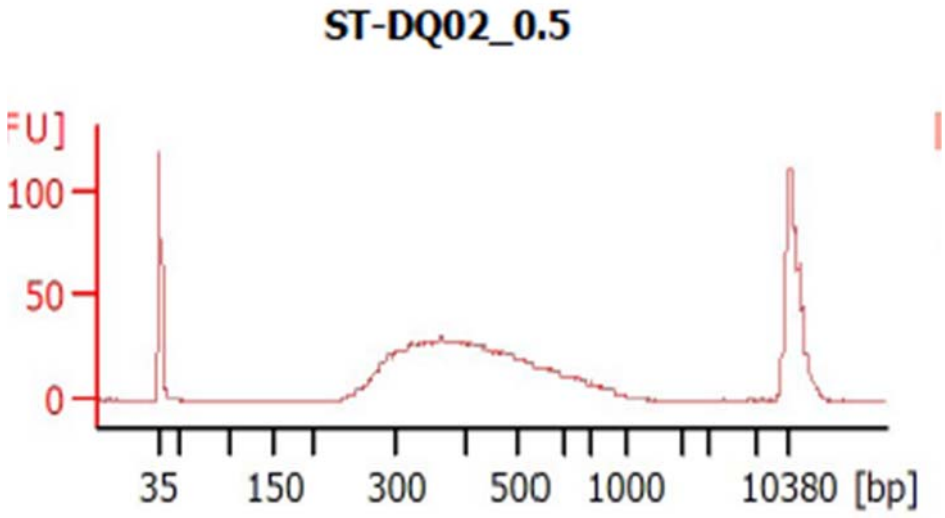


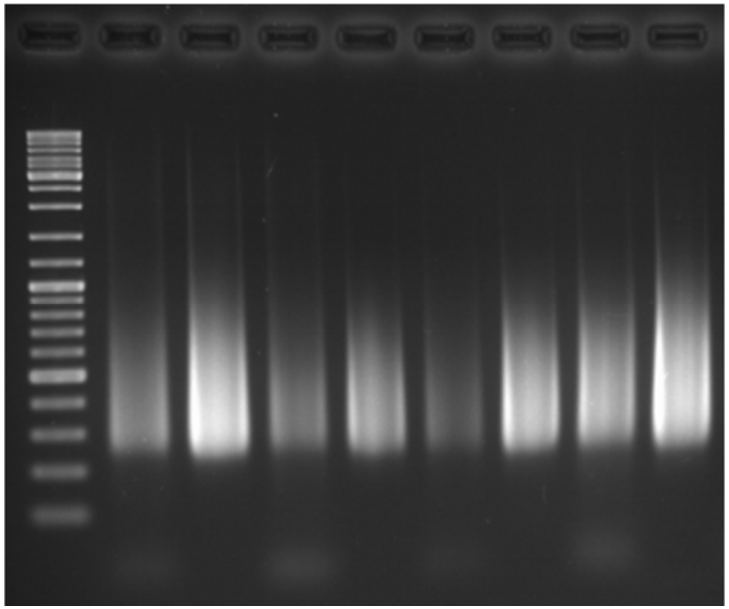
Hirst M , Marra M A Briefings in Functional Genomics 2010;9:455-465

Briefings in Functional Genomics

mRNA-seq prep
(no gel size selection)

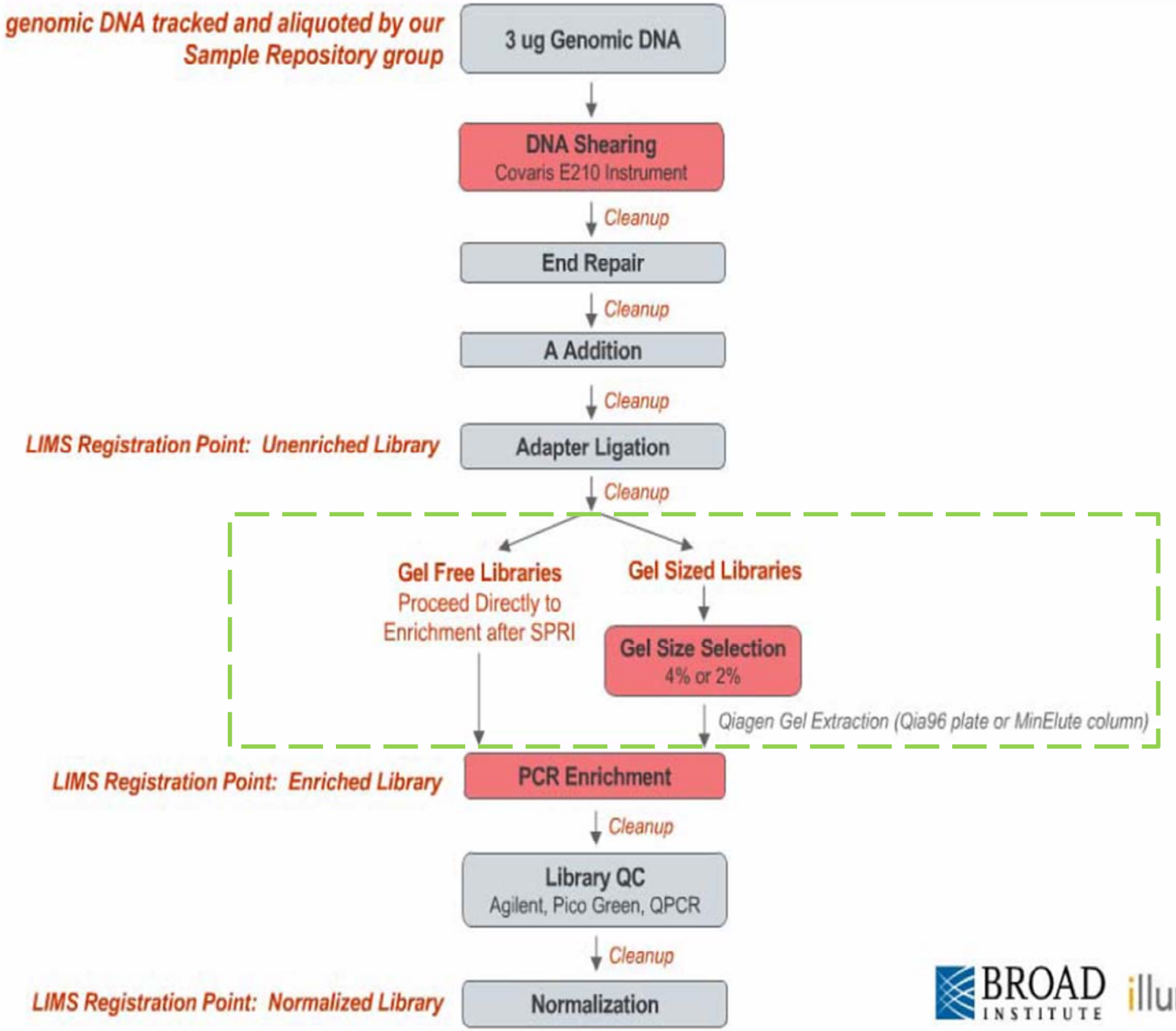






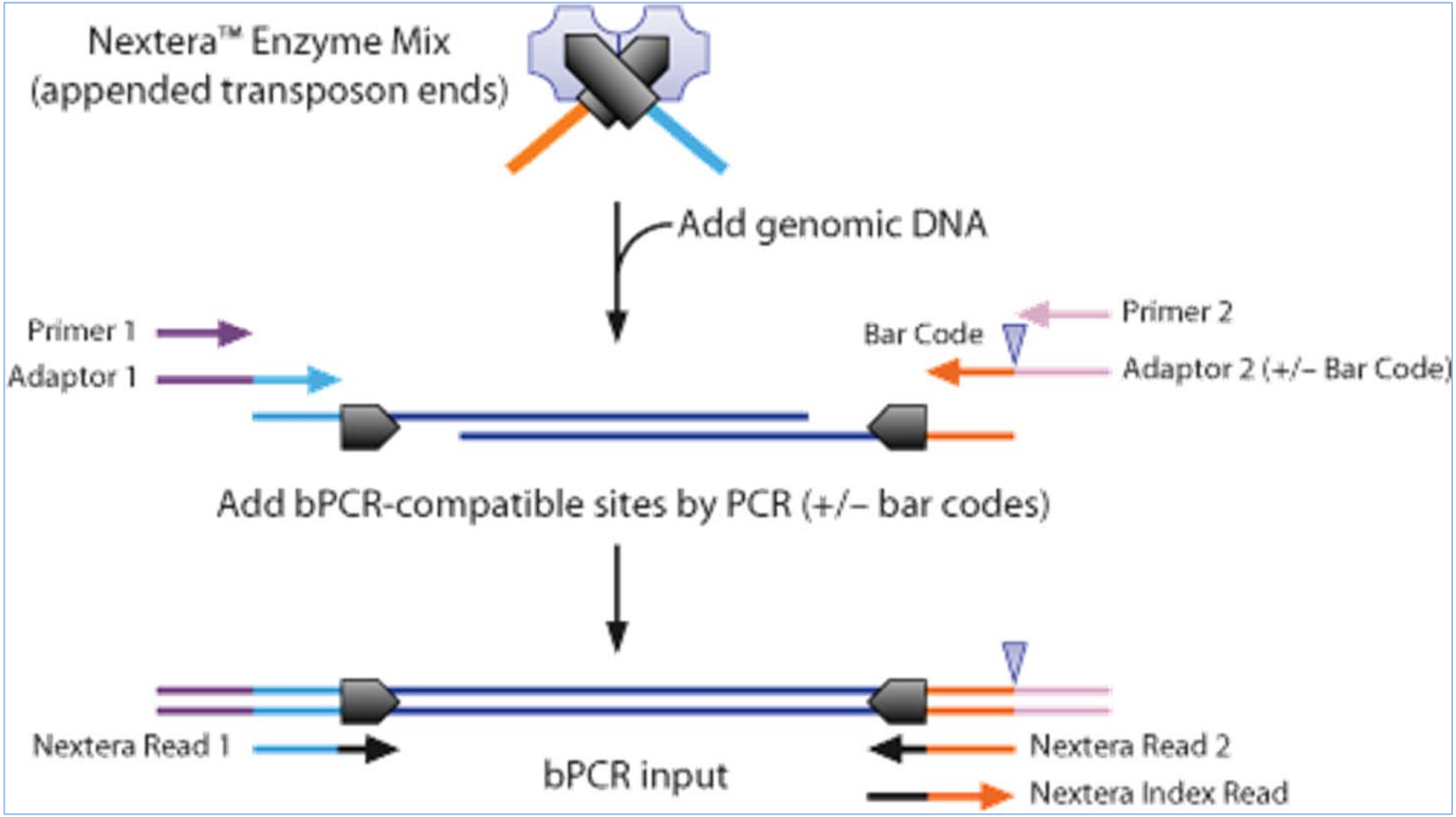
IV. Data types, preprocessing, and assessment

Sample Prep Workflow at the Broad

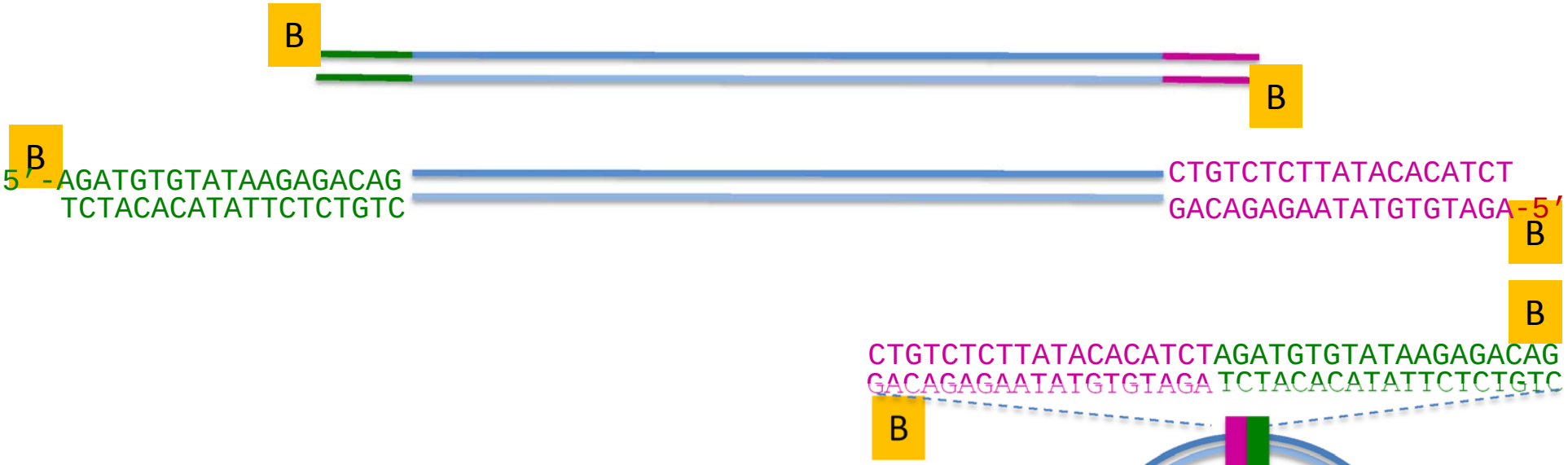


BROAD INSTITUTE illumina

gDNA library prep by Tn insertion



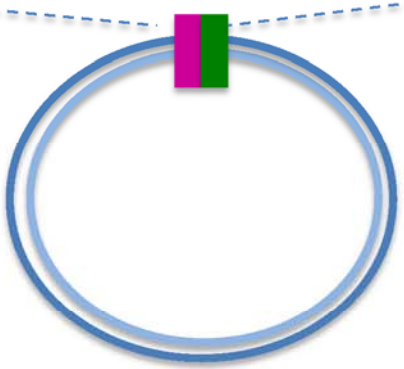
Nextera Mate Pair



New nextera MP junction adapter seq.

Table 1: Mate Pair Adapter Sequence Elements

Adapter Element	Sequence
Circularized Duplicate Junction Adapter	CTGTCTCTTATACACATCTAGATGTGTATAAGAGACAG
Circularized Single Junction Adapter	CTGTCTCTTATACACATCT
Circularized Single Junction Adapter Reverse Complement	AGATGTGTATAAGAGACAG (Transposase)
Read 1 External Adapter	GATCGGAAGAGCACACGTCTGAACTCCAGTCAC
Read 2 External Adapter	GATCGGAAGAGCGTCGTGTAGGGAAAGAGTGT



Old nextera PE transposase seq.

Nextera® DNA Sample Preparation Kit (Illumina) [1,2]

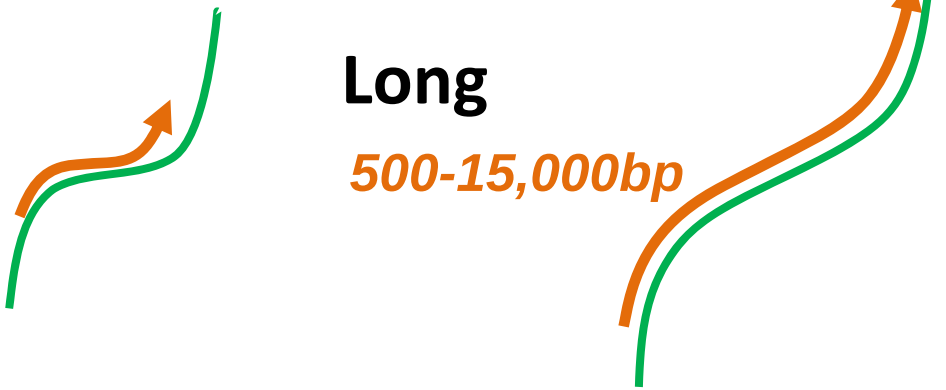
Nextera® transposase sequences (FC-121-1031, FC-121-1030)

5′ TCGTCGGCAGCGTCAGATGTGTATAAGAGACAG
(a) Read 1 -->

5′ GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAG
(d) Read 2 -->

Types and Characteristics of NGS Reads

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

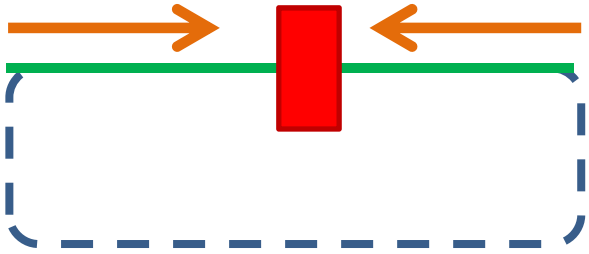


- **Read types:**

SR (single end) 50bp-20kb

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

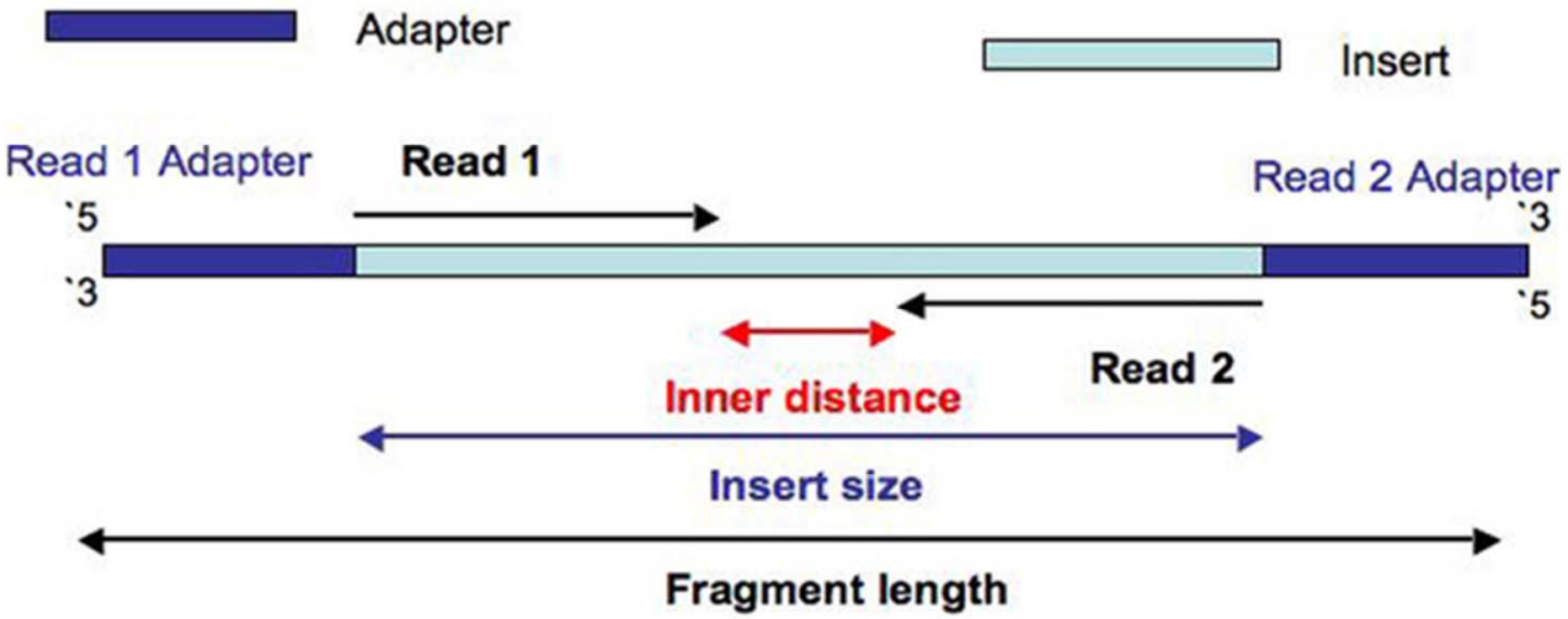
MP (mate-pair) 50-300bp; 2~15kb jump



NGS library fragment

1. architecture of library fragment
2. associated issues: read length & mappable rate

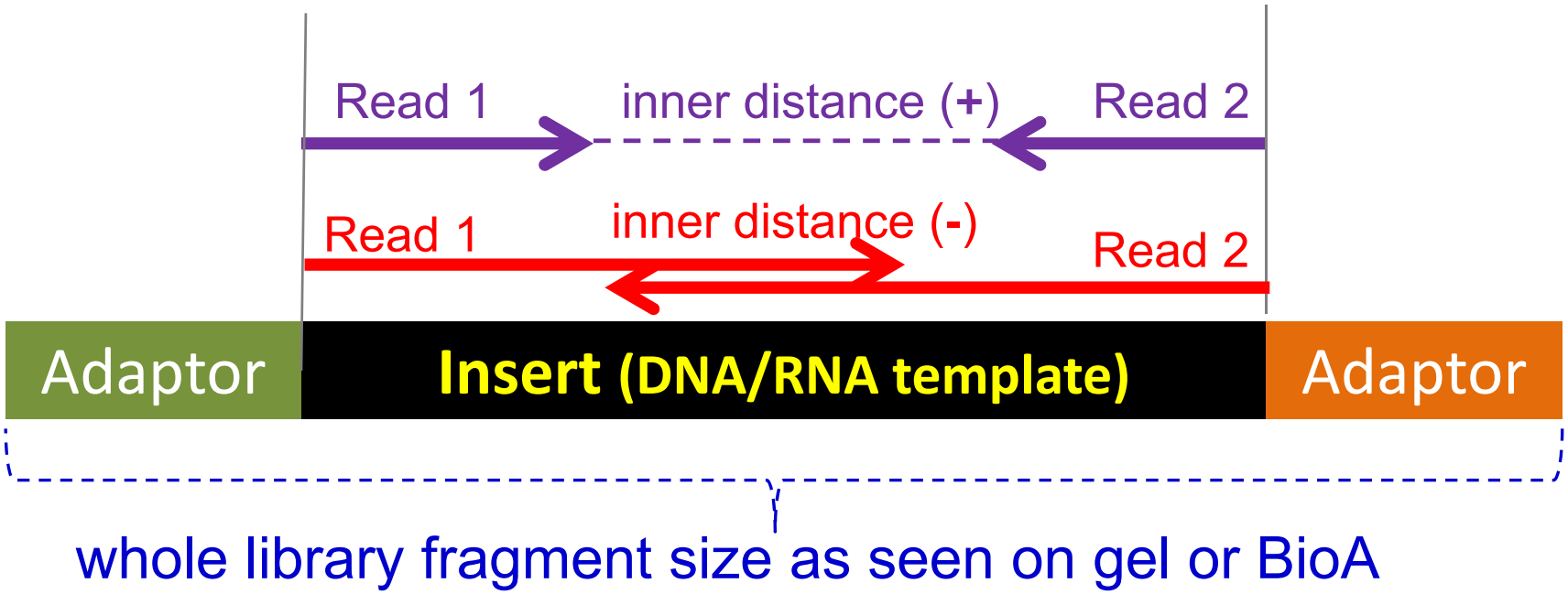
Fragment v. Insert v. Inner Distance



http://www.frontiersin.org/files/Articles/77572/fgene-05-00005-HTML/image_m/fgene-05-00005-g001.jpg

Fragment v. Insert v. Inner Distance

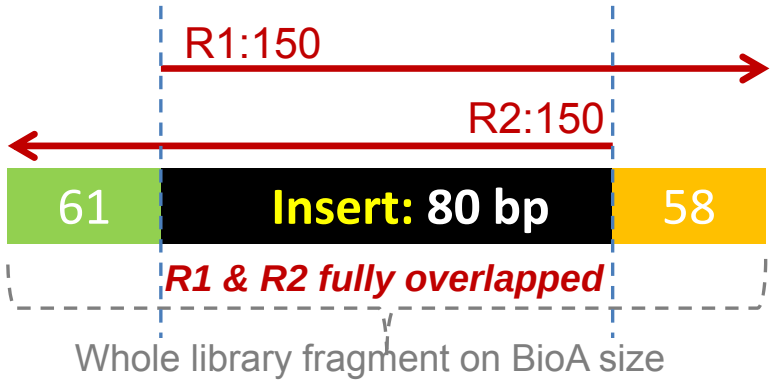
A. **Library fragment** = length detected by BioA/agarose gel
B. **Insert size** = DNA or RNA template; no adaptors included
C. **Inner distance** = distance b/w the end base of R1 and R2
    1. Positive distance = gapped ends
    2. Negative distance = overlapped ends



Insert size vs Library Fragment Size: PE2*150

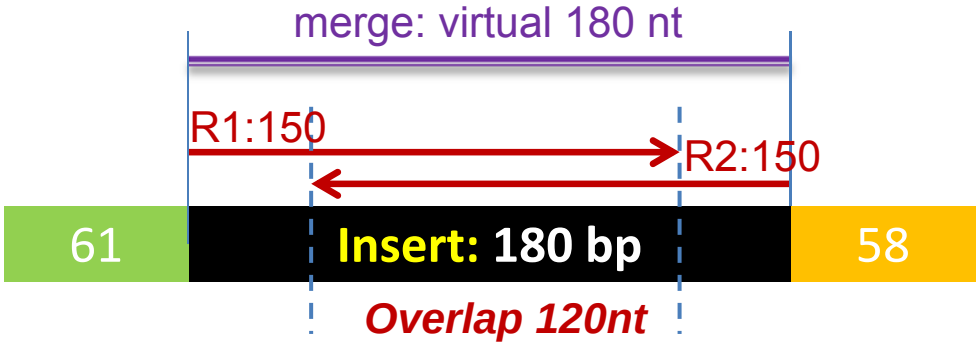
200-bp fragment (BioA):

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



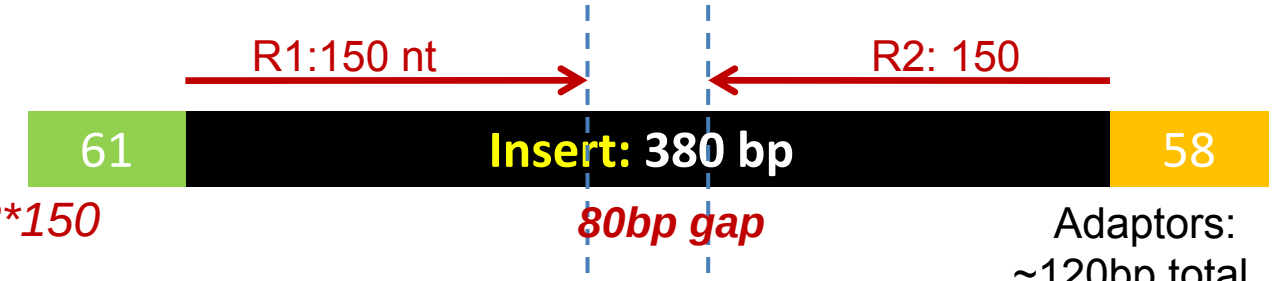
300-bp fragment (BioA):

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



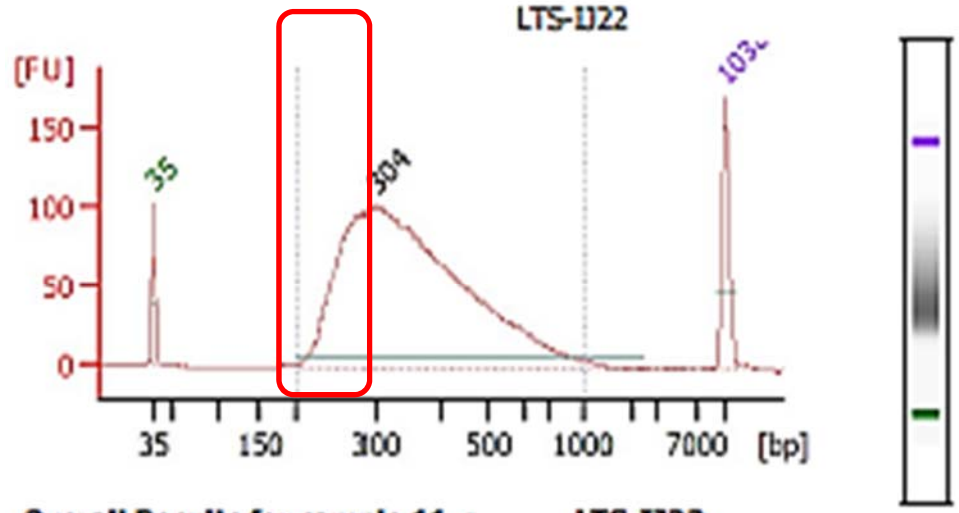
500-bp fragment:

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



Short fragment may read through to adaptors

May be un-mappable due to long adaptor in reads



Overall Results for sample 11 : **LTS-D22**

Number of peaks found:	1
Noise:	0.3
Corr. Area 1:	2,319.9

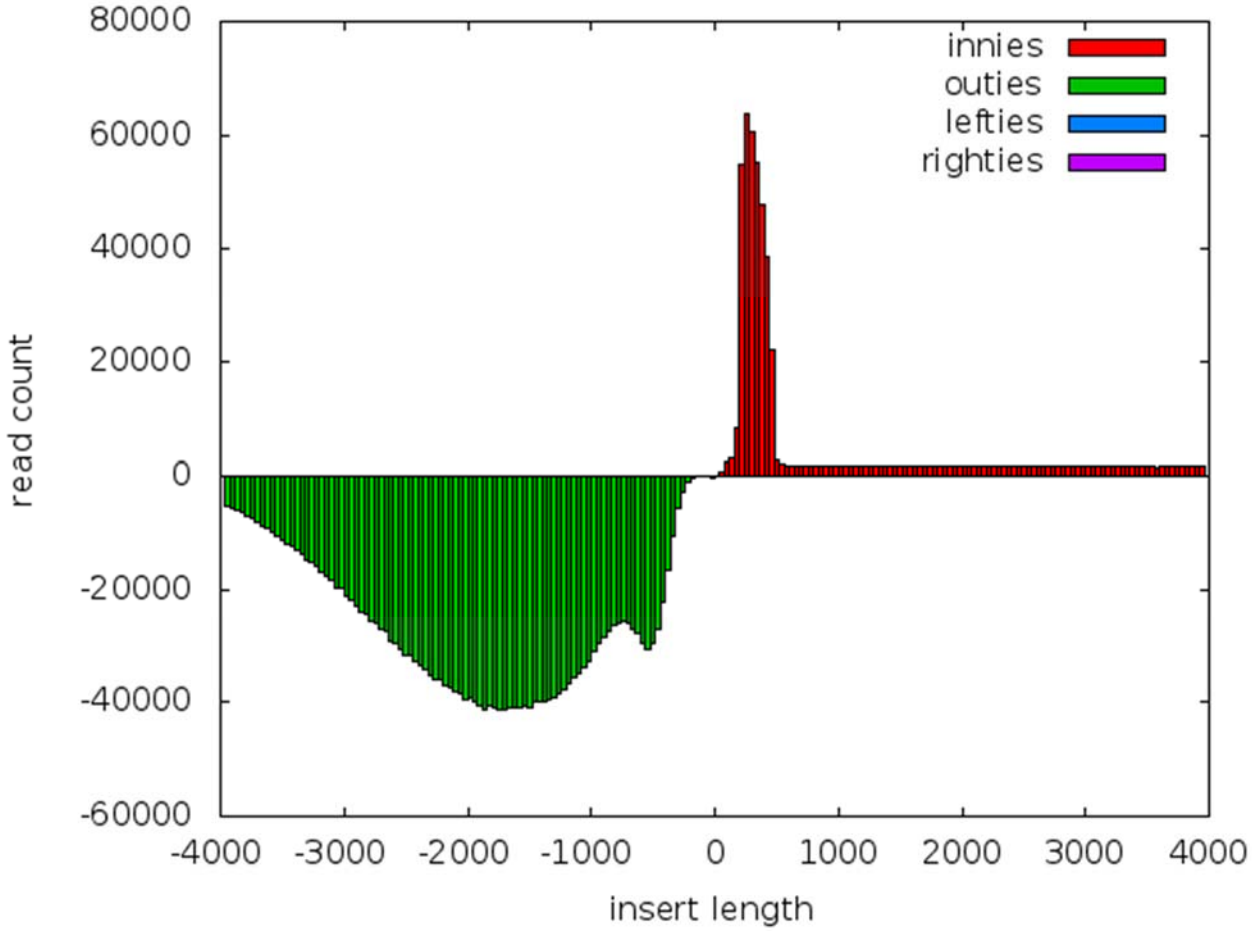
Region table for sample 11 : **LTS-D22**

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	2,319.9	98	377	31.9	1,892.19	8,511.7	

- Range: 200-1000 bp
- peak @ 304 bp
- Avg. @ 377 bp

Solution: Trim off adaptor sequences before mapping!

Mate-pair insert- size



www.biostars.org

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 * $\log_{10}(P)$]

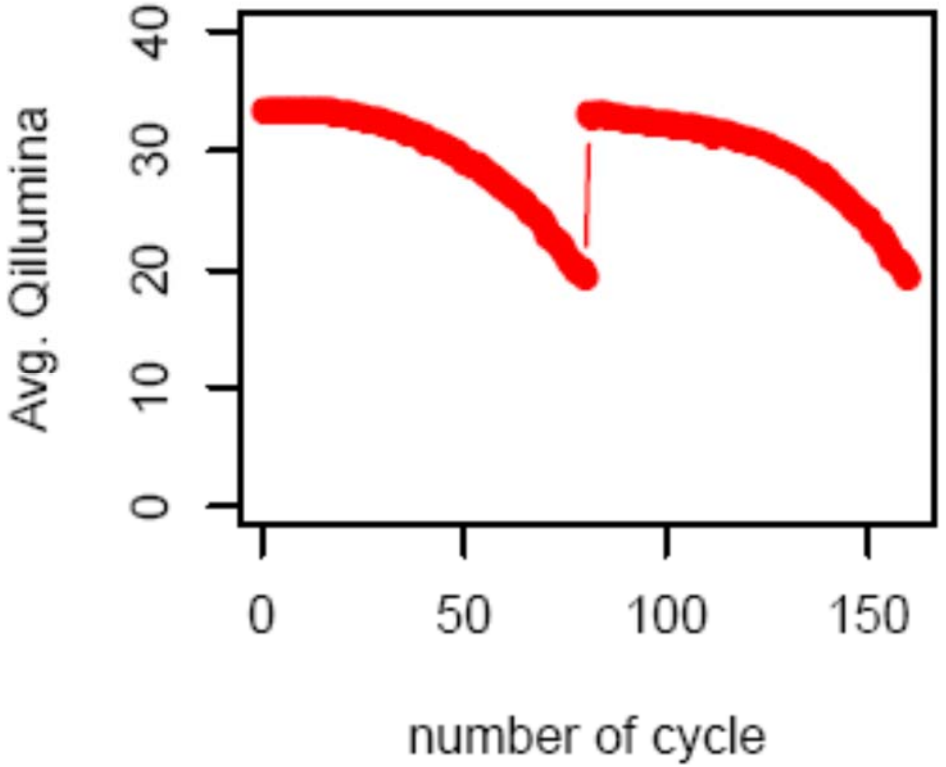
Phred Score Q	Error probability
10	1 in 10
20	1 in 100
30	1 in 1,000
40	1 in 10,000

Seq. performance assessment – Base Q

Phred quality scores Q: logarithmically related to error probabilities

P by $Q = [-10 * \log_{10}(P)]$

Phred Score Q	Error probability	Base call accuracy
10	1 in 10	90%
20	1 in 100	99%
30	1 in 1,000	99.9%
40	1 in 10,000	99.99%



FastQC Report

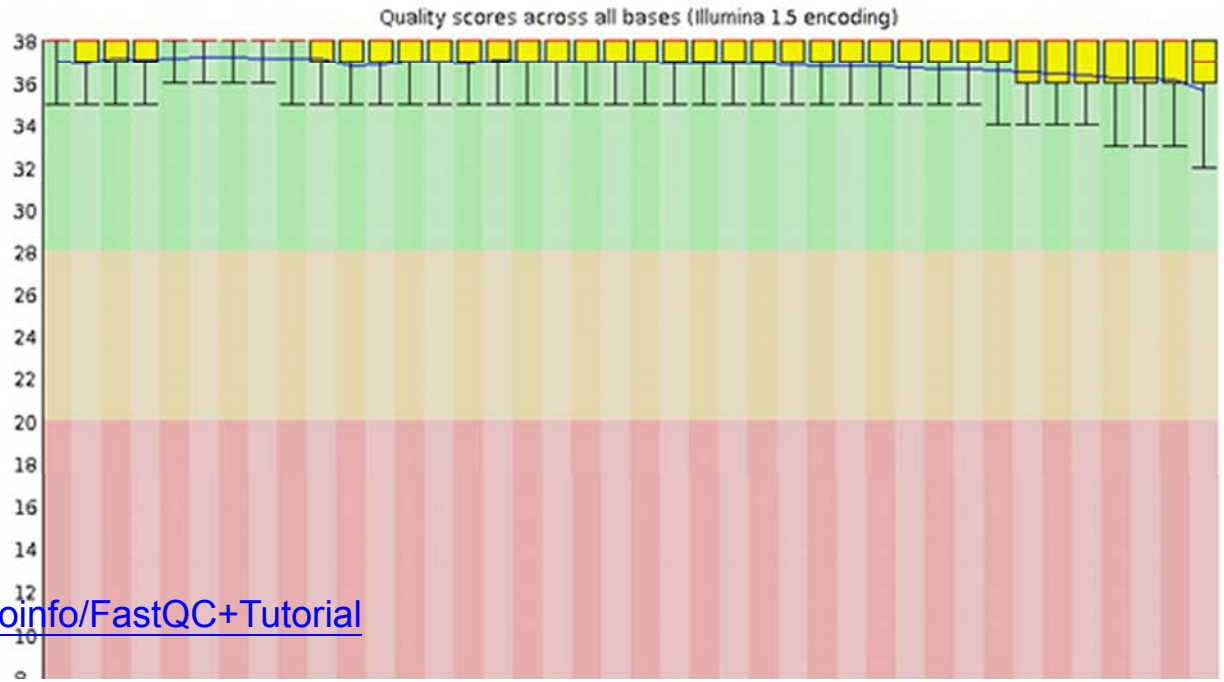
Summary

- Basic Statistics
- Per base sequence quality
- Per sequence quality scores
- Per base sequence content
- Per base GC content
- Per sequence GC content
- Per base N content
- Sequence Length Distribution
- Sequence Duplication Levels
- Overrepresented sequences
- Kmer Content

Basic Statistics

Measure	Value
Filename	good_sequence_short.fastq
File type	Conventional base calls
Encoding	Illumina 1.5
Total Sequences	250000
Filtered Sequences	0
Sequence length	40
%GC	45

Per base sequence quality



Read processing:

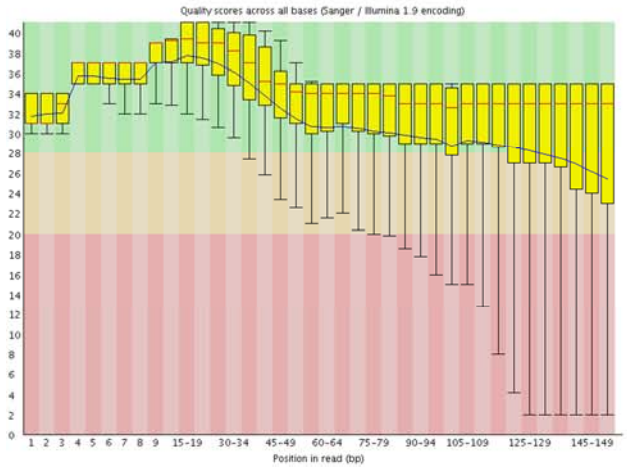
- FASTX-toolkit
- Trimmomatic
- NGS QC Toolkit

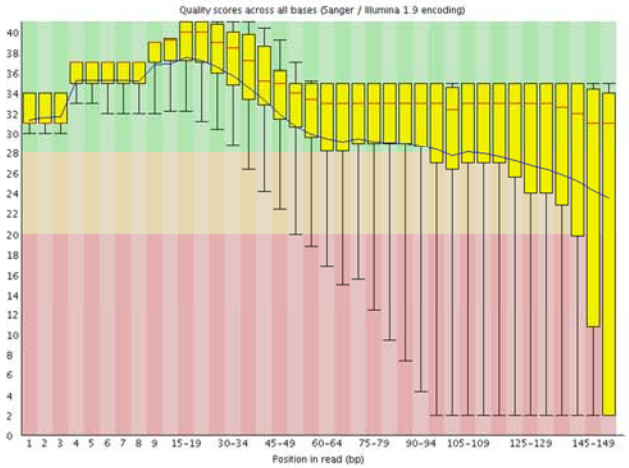
https://wiki.hpcc.msu.edu/display/Bioinfo/FastQC+Tutorial

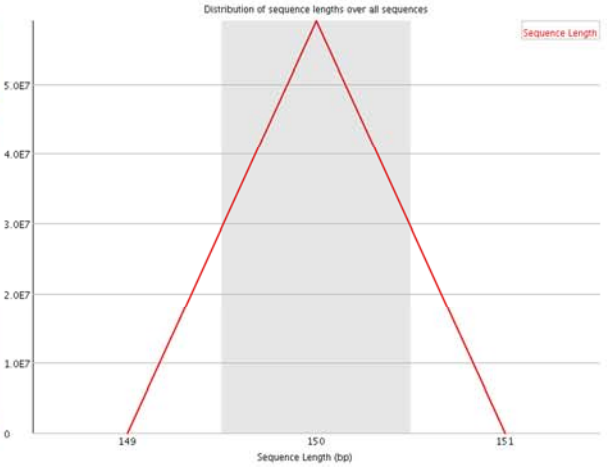
Data QC

Adapter Trimming Result of HiSeq genomic PE Reads

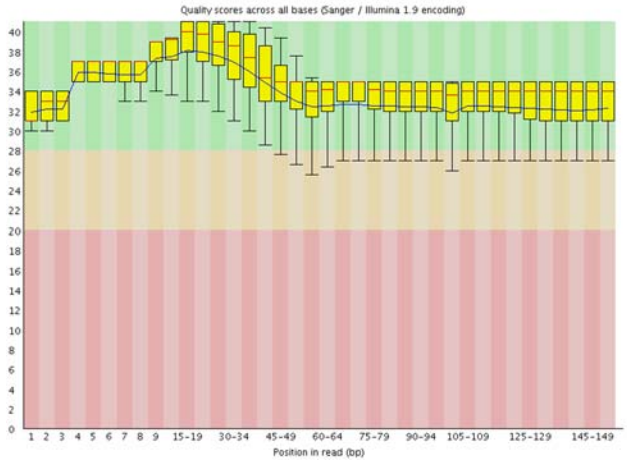
HiSeq.CDC10.raw150

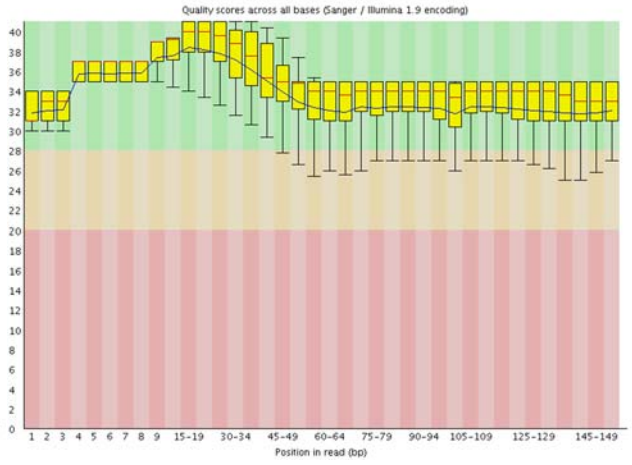


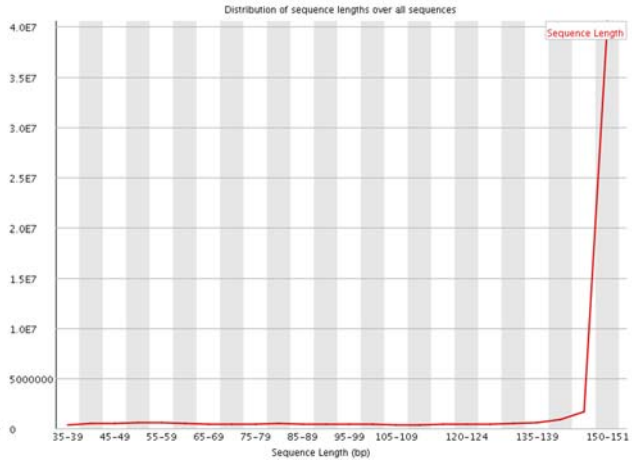




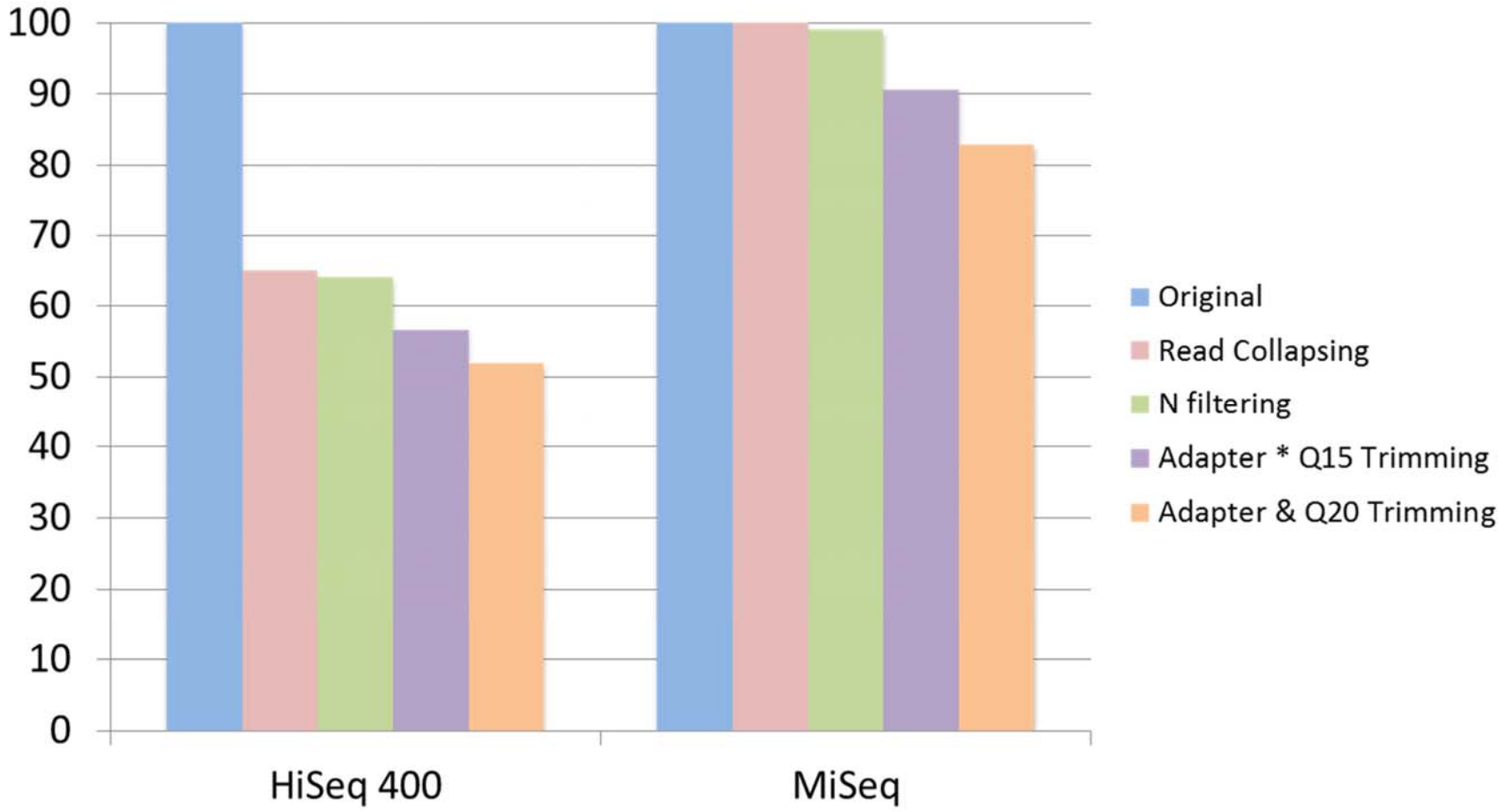
HiSeq.CDC10.raw150 (after adapter trimming by Trimmomatic)





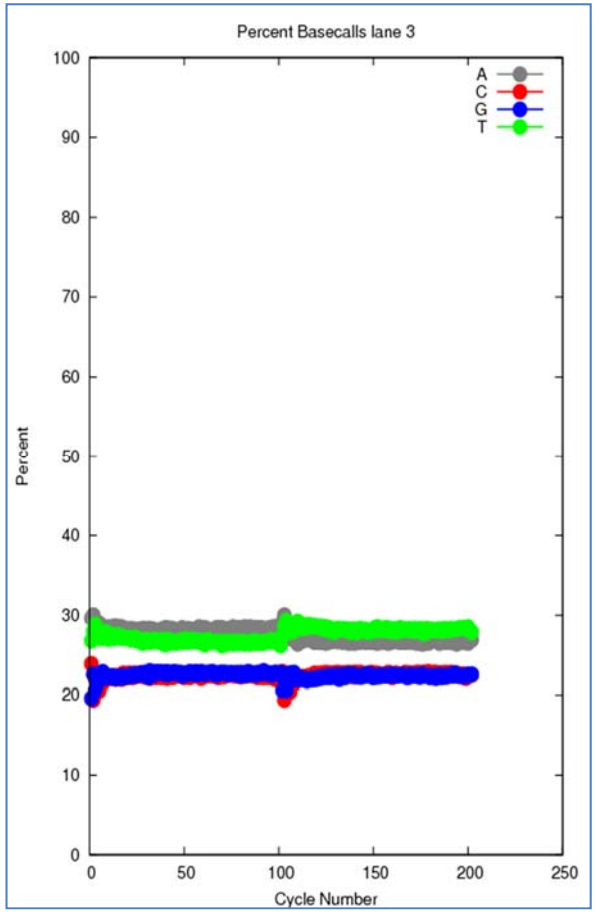


Genomic read preprocess

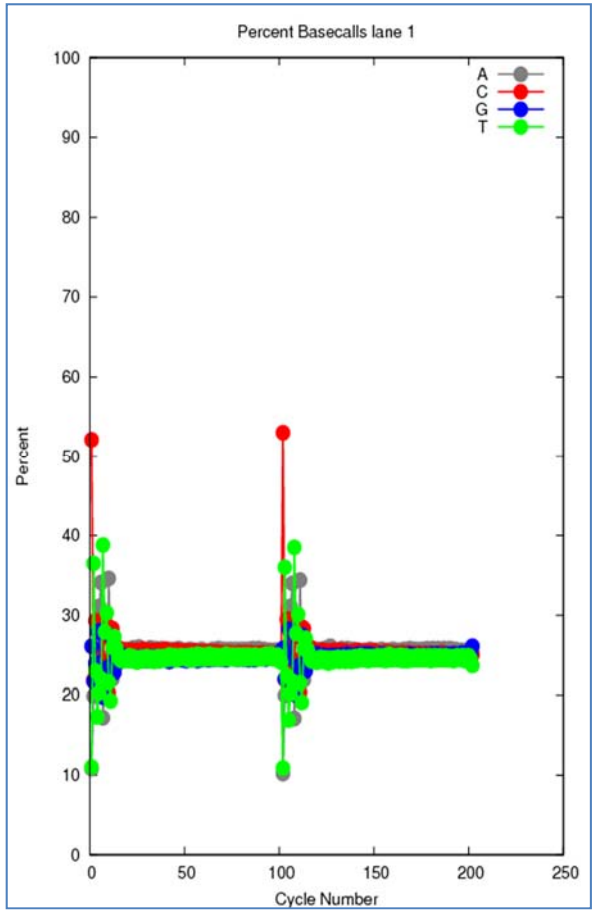


IVC plots (Intensity vs Cycle)

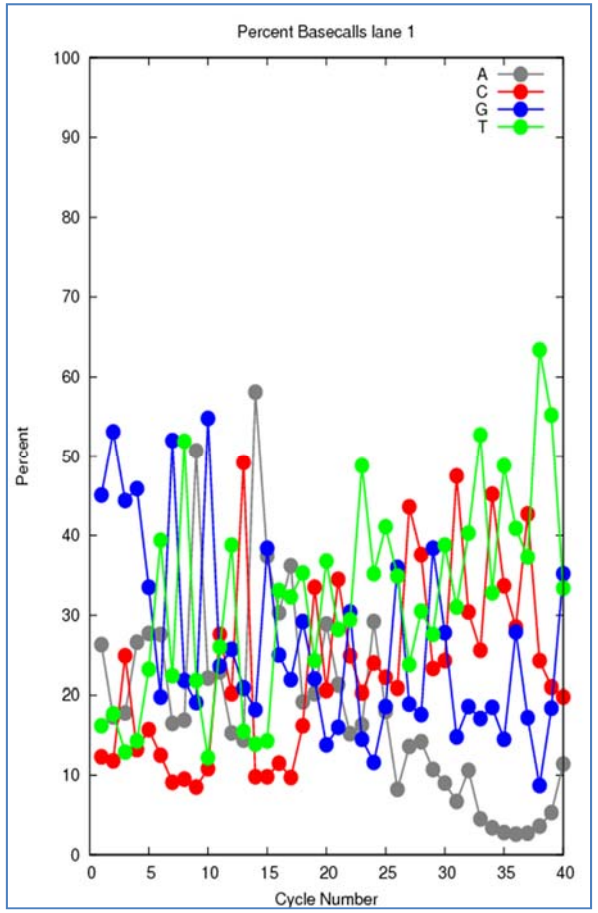
Normal GC%



mRNA-seq

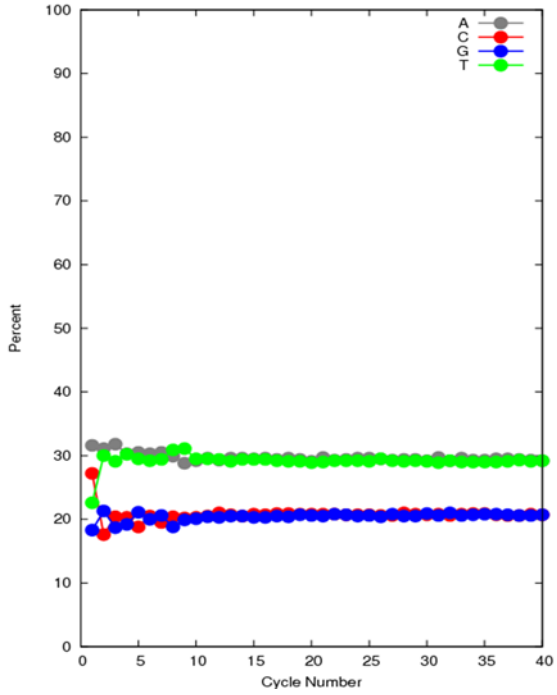


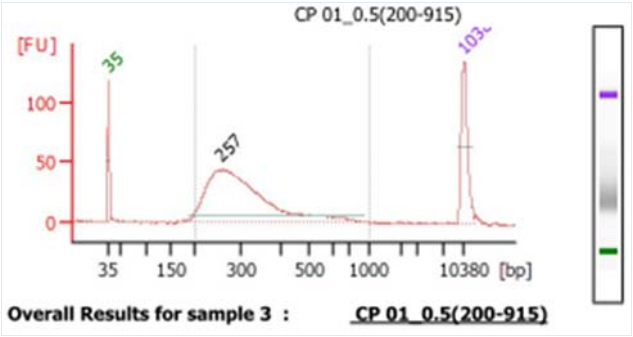
smRNA-seq



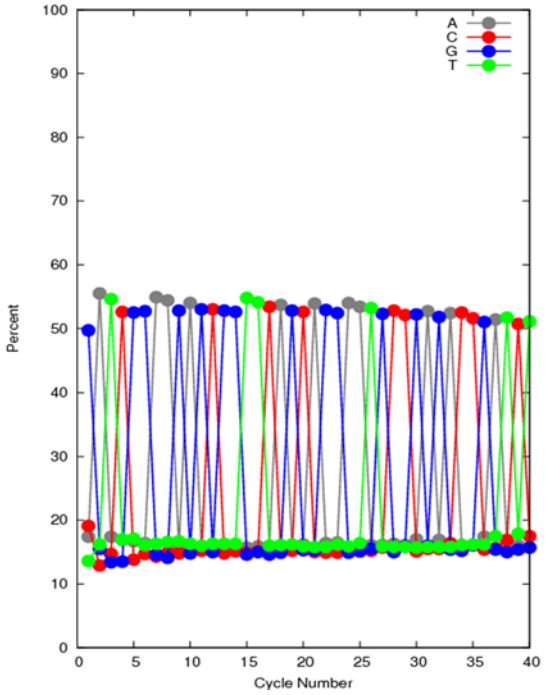
IVC plots (Intensity vs Cycle)

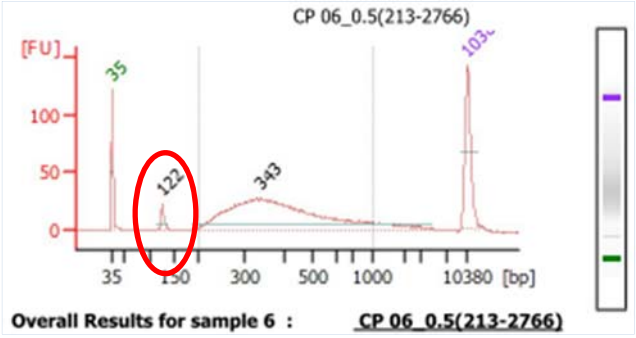
Normal input- Little bias





Low input - Strong bias





IV. Applications

- Whole genome
- Transcriptome
- Targeted sequencing

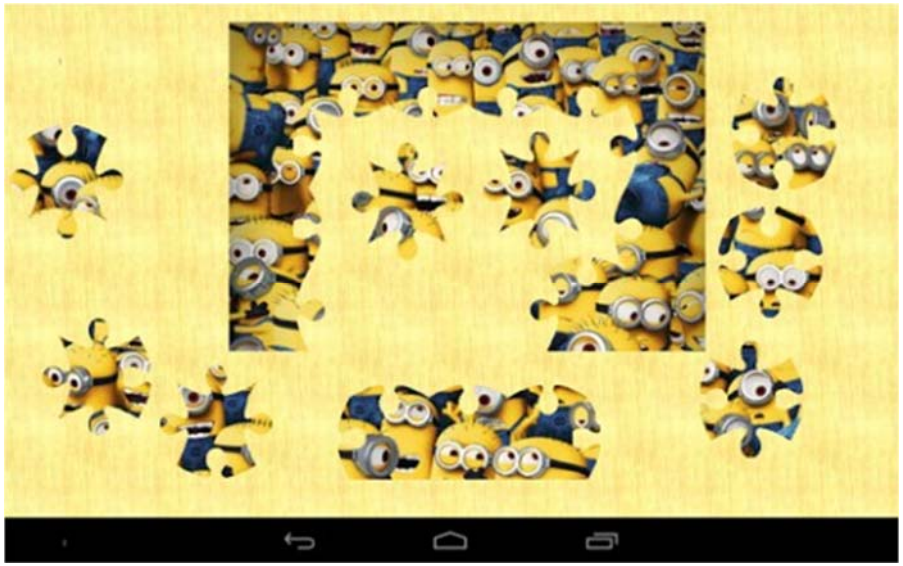
Genome Sequencing

- *De novo (assembly)*
- *Re-sequencing (mapping)*

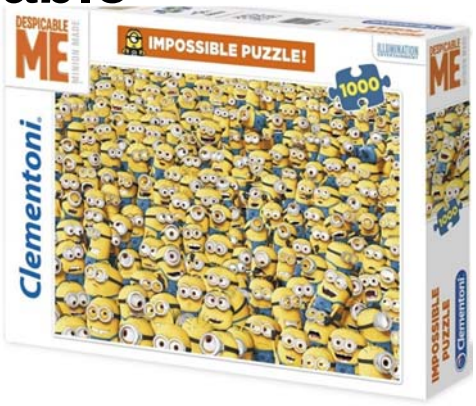
De novo seq. vs re-sequencing

***"de novo" sequencing*:**
no reference geneome available

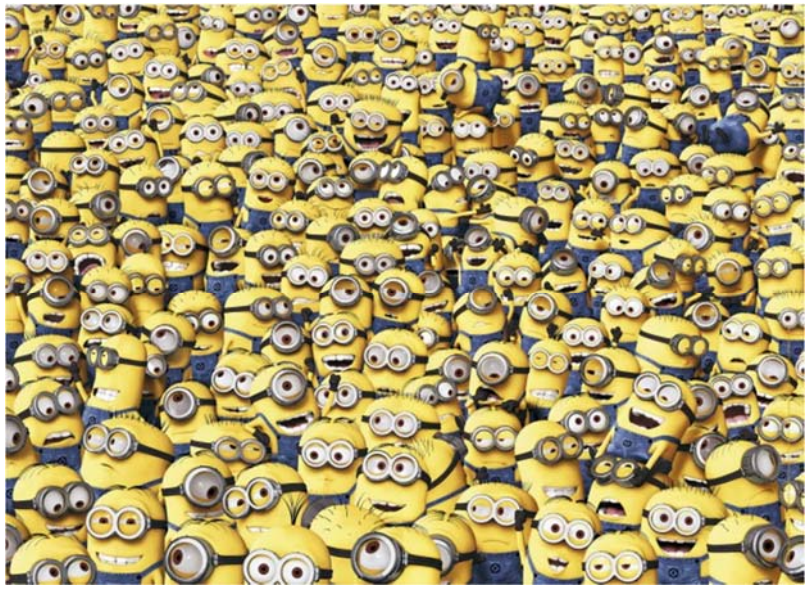
assemble



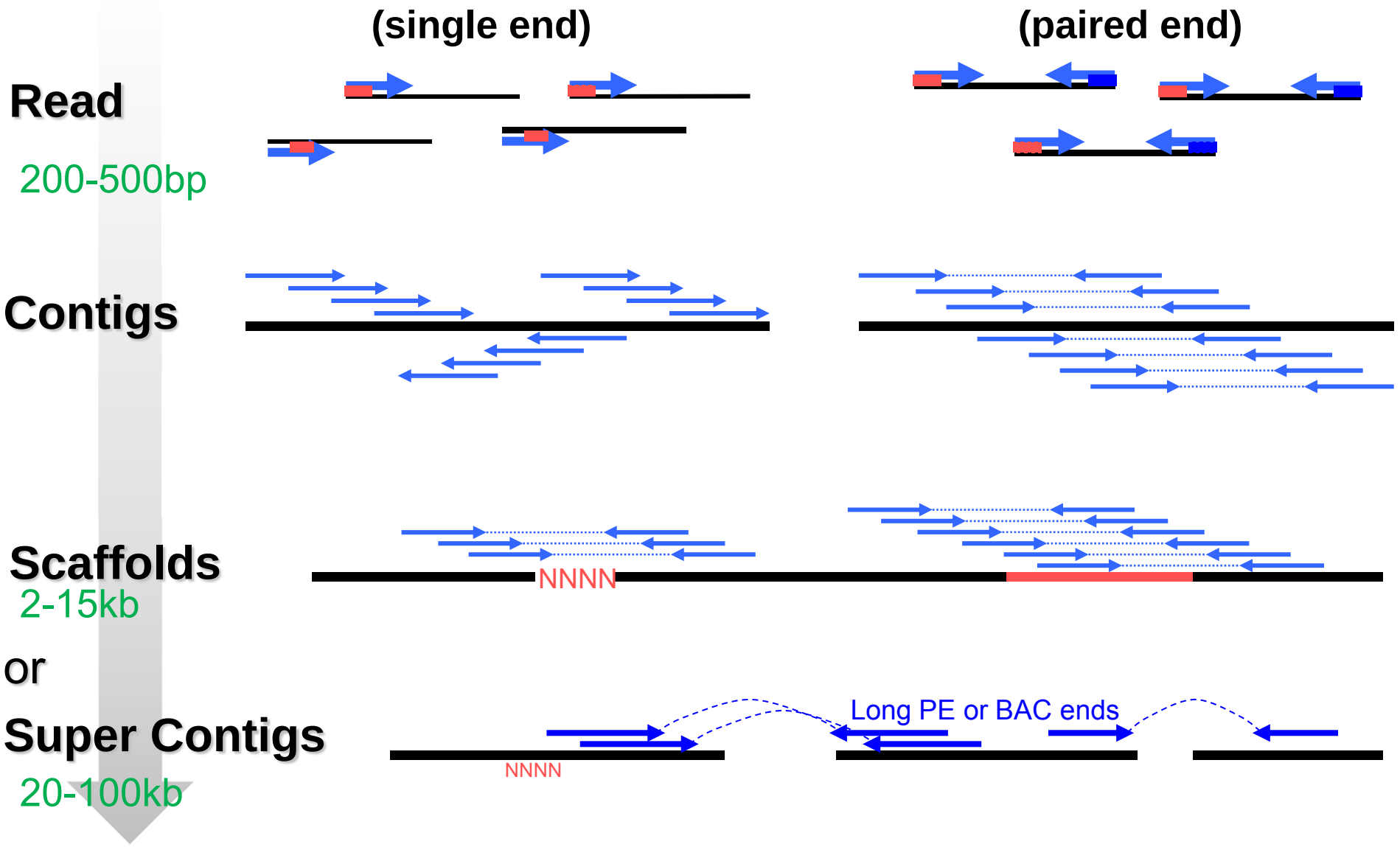
***Re-sequencing*:**
reference genome available



align

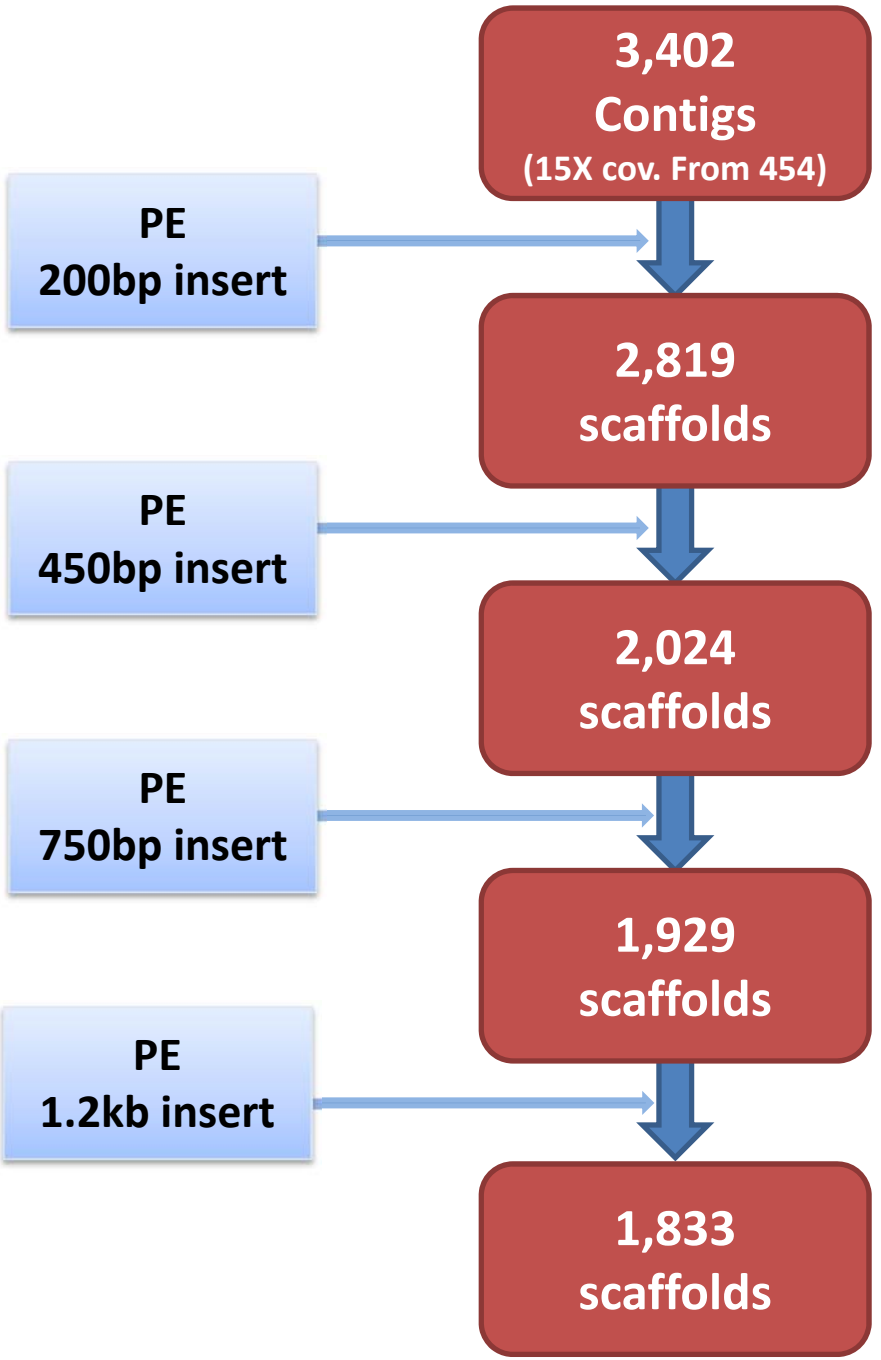


Hierarchical Genome Assembly

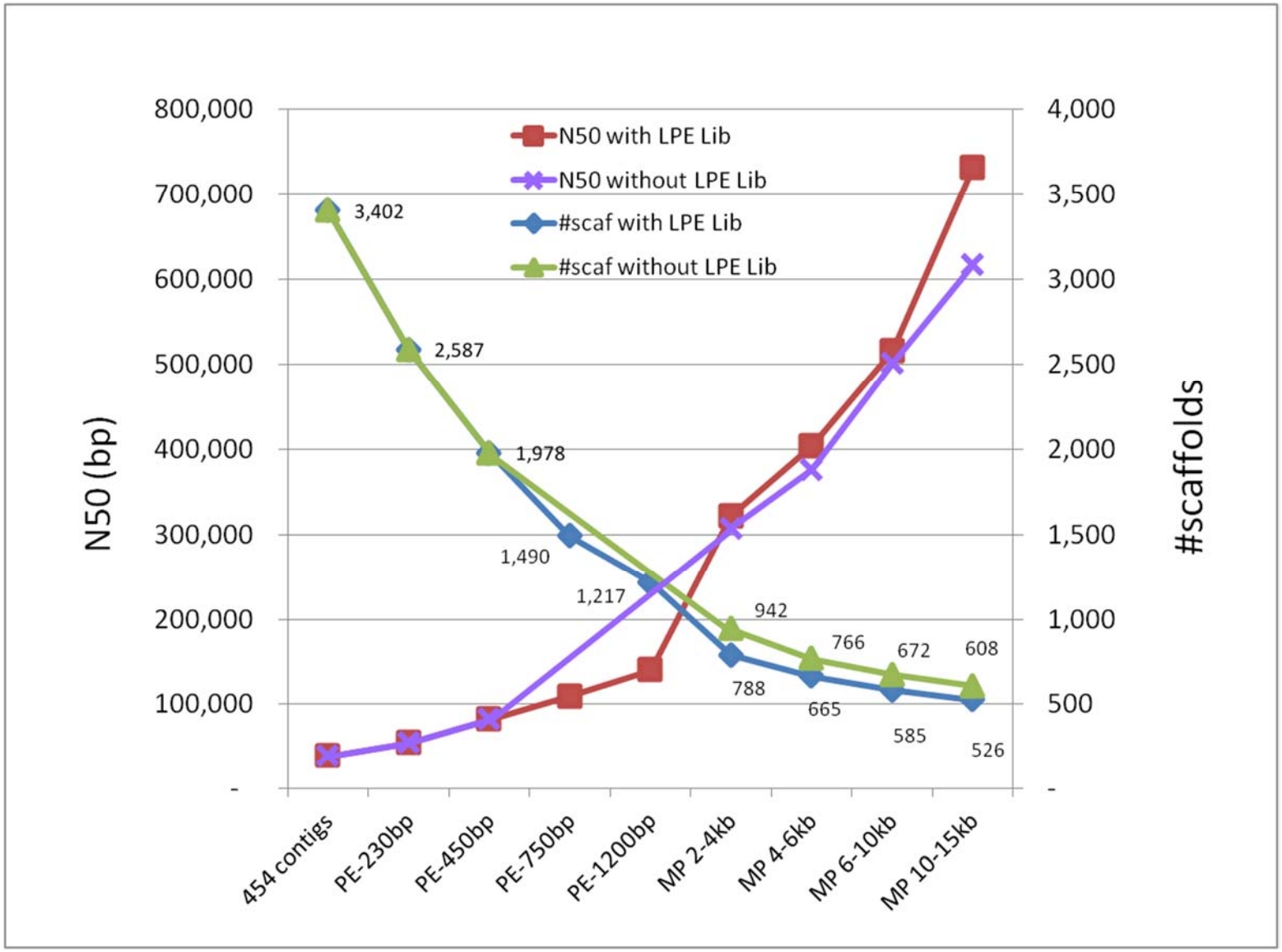


Hierarchical bridging by increasing PE insert sizes

S32 assembly	N50 (kb)	scaffolds
454 contigs	5	3,402
PE-200	10	2,819
PE-450	15	2,024
PE-750	30	1,929
PE-1200	100?	1,833

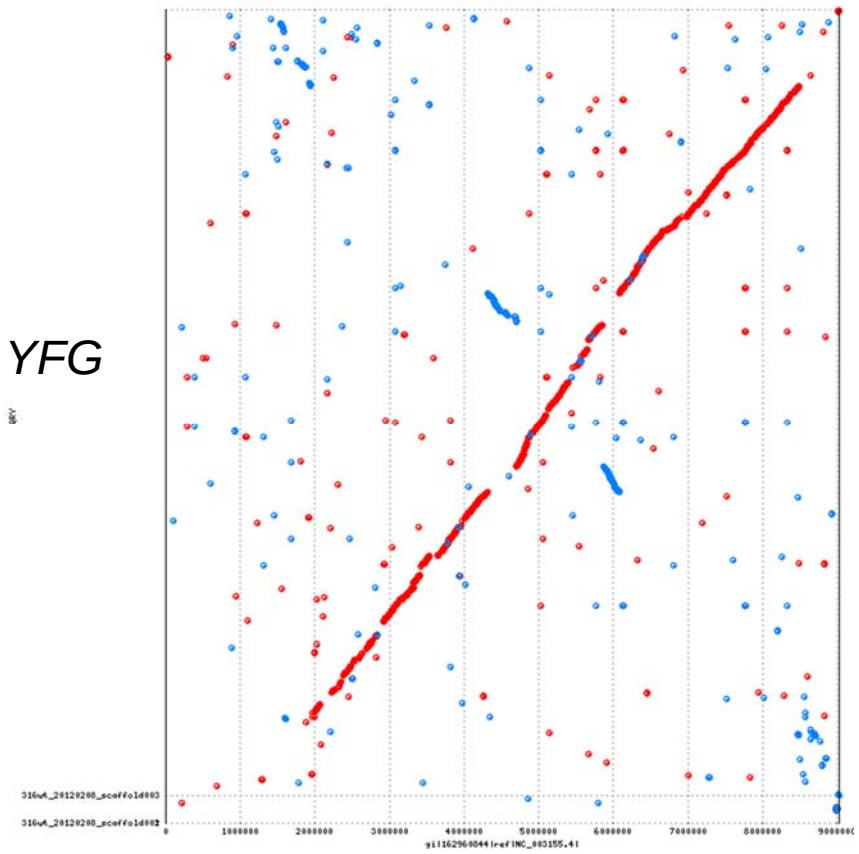


Genome assembly progress - Fungus

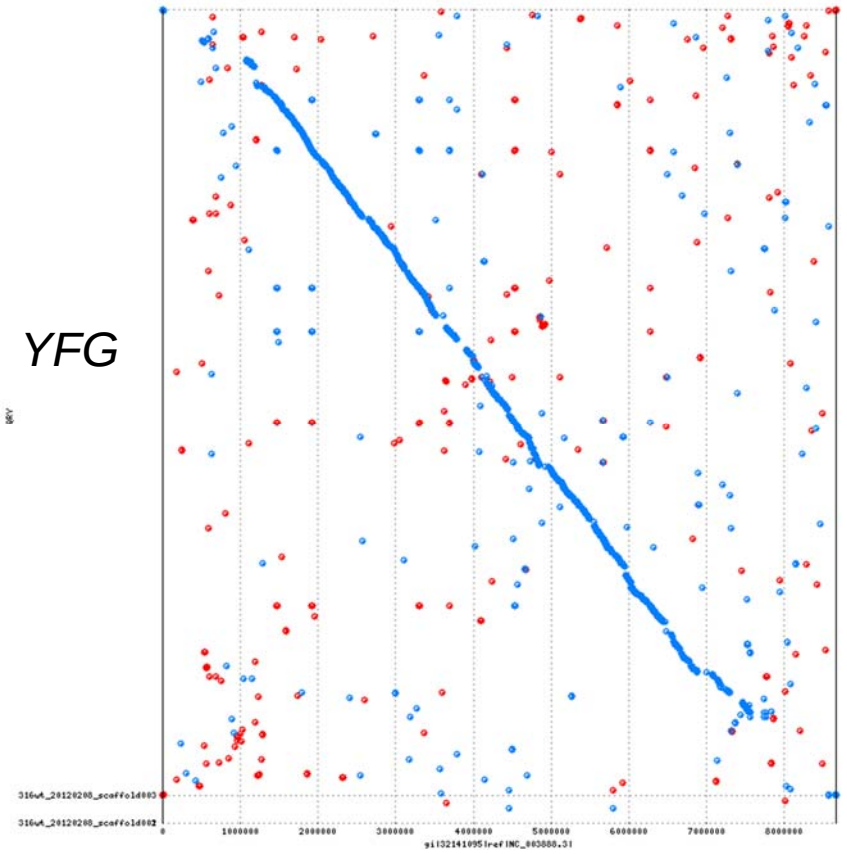


Ortholog plots – Mummer plot

• The core region of *S. avermitilis* and *S. coelicolor* have very good alignment with YFG



S. avermitilis



S. coelicolor

Re-sequencing: Variant detection

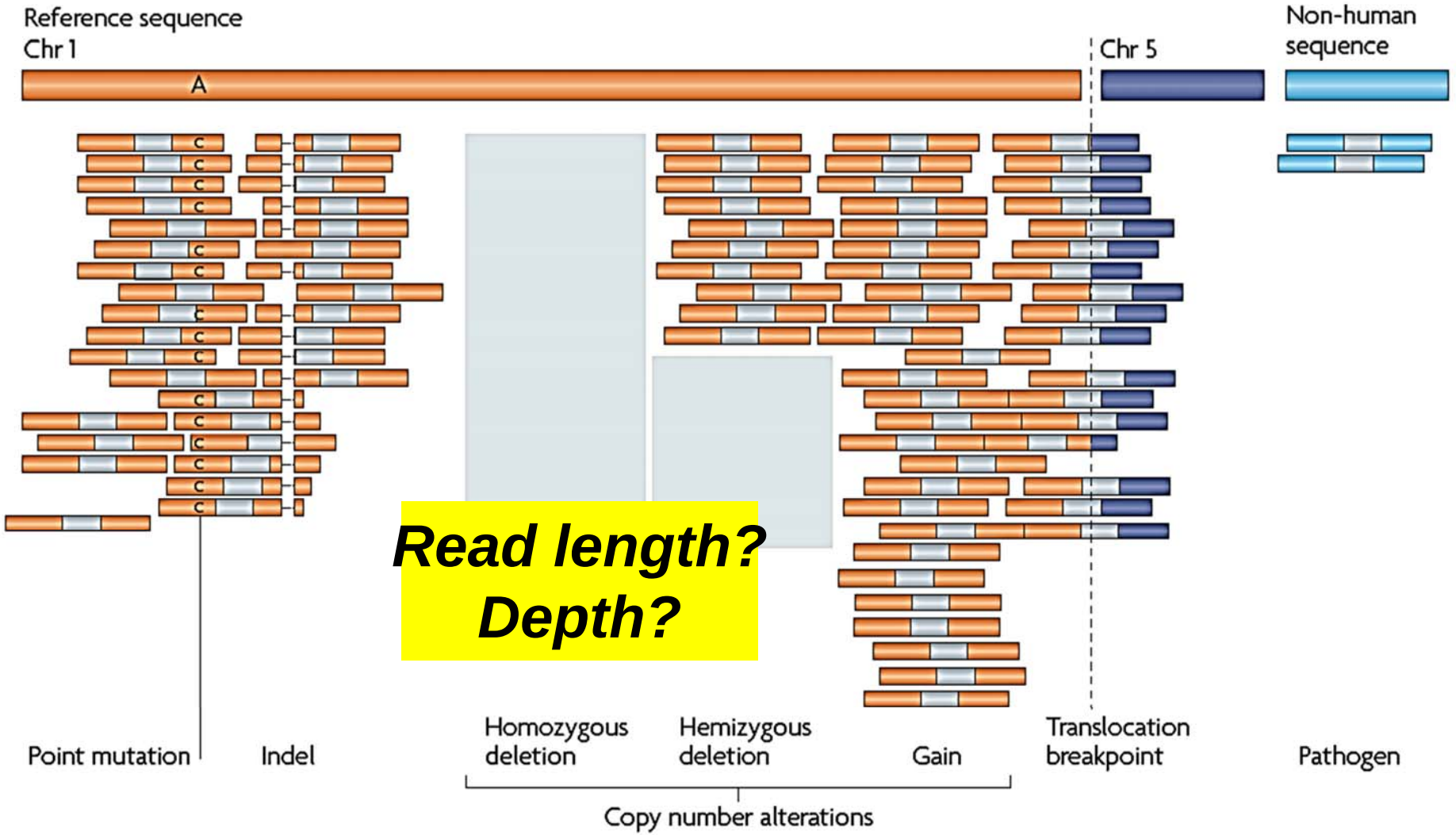
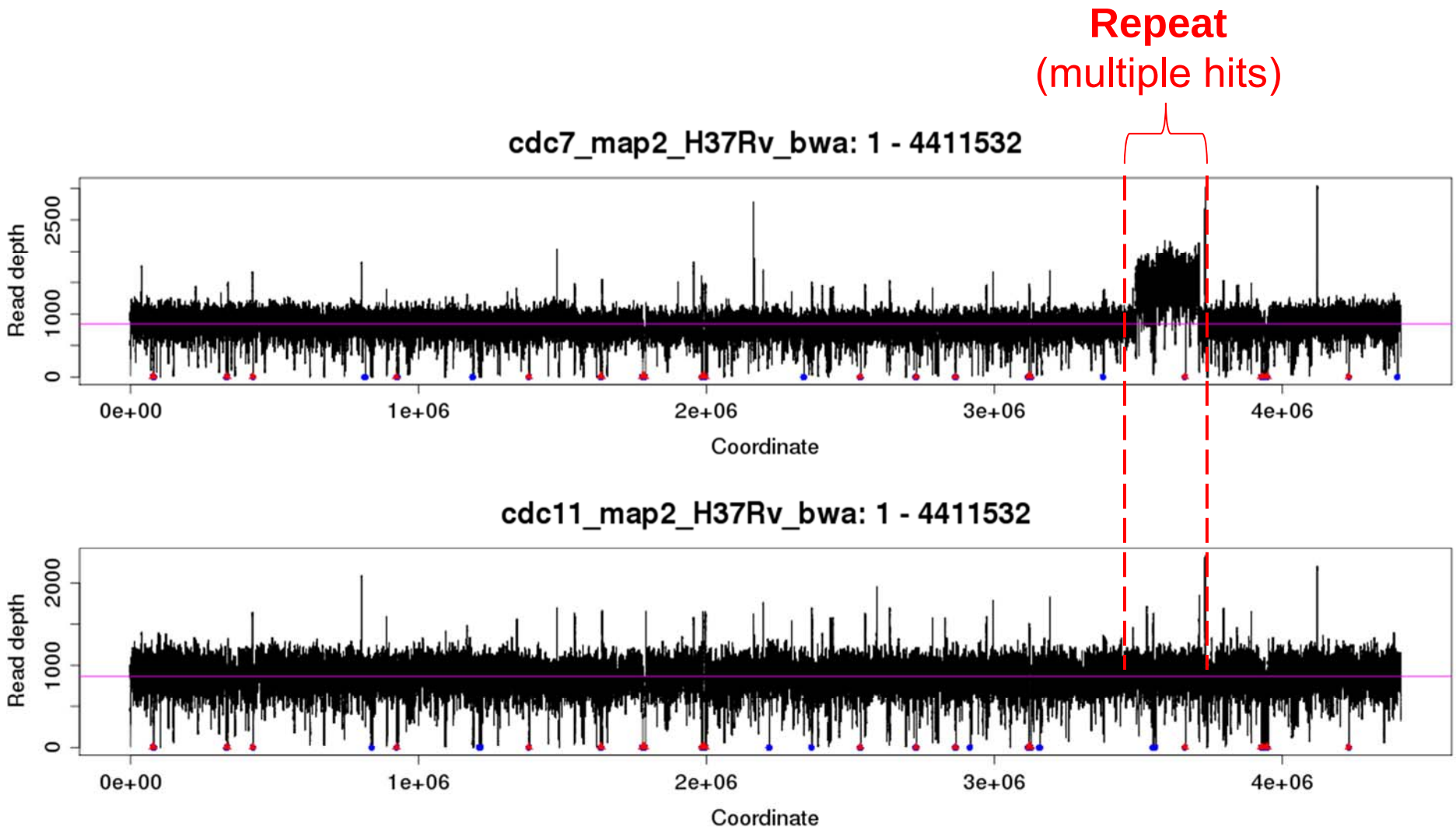


Figure 3 | **Types of genome alterations that can be detected by second-generation sequencing.** Sequenced

Meyerson et al. NRG 2010

Coverage profile



Considerations: Sequencing depth for SNP discovery

Type of Experiment	Coverage Required
Haploid SNPs/divergence	≥ 10 x
Diploid SNPs/divergence	≥ 30 x
Aneuploid/somatic mutations	≥ 50 x
Population sequencing	≥ 200 x

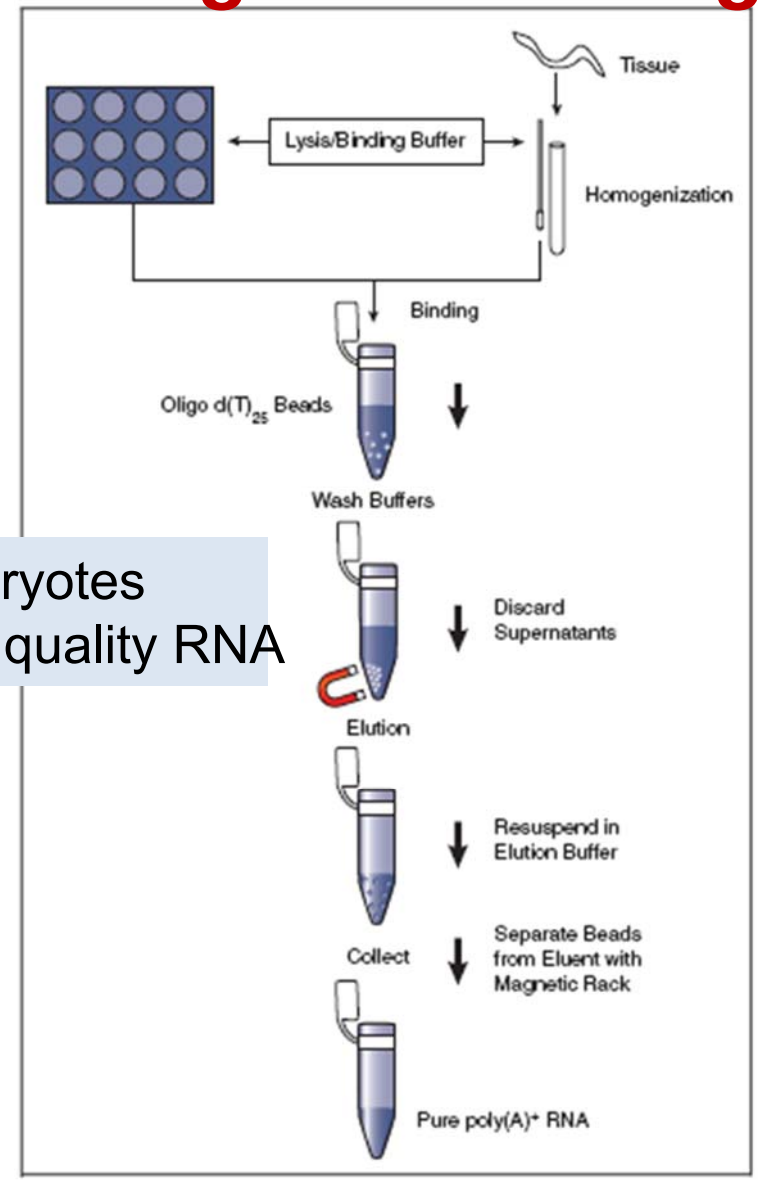
Source: Dr. Michael C. Zody

Transcriptome sequencing

- *mRNA, stranded*
- *Small RNA*

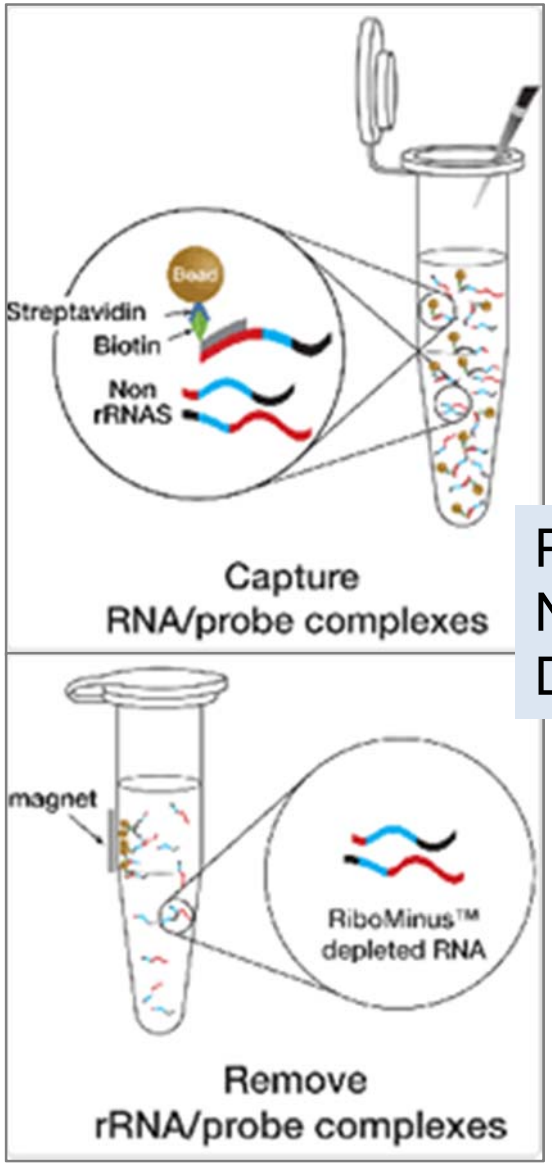
Q: mRNA enrichment?

Oligo-dT binding



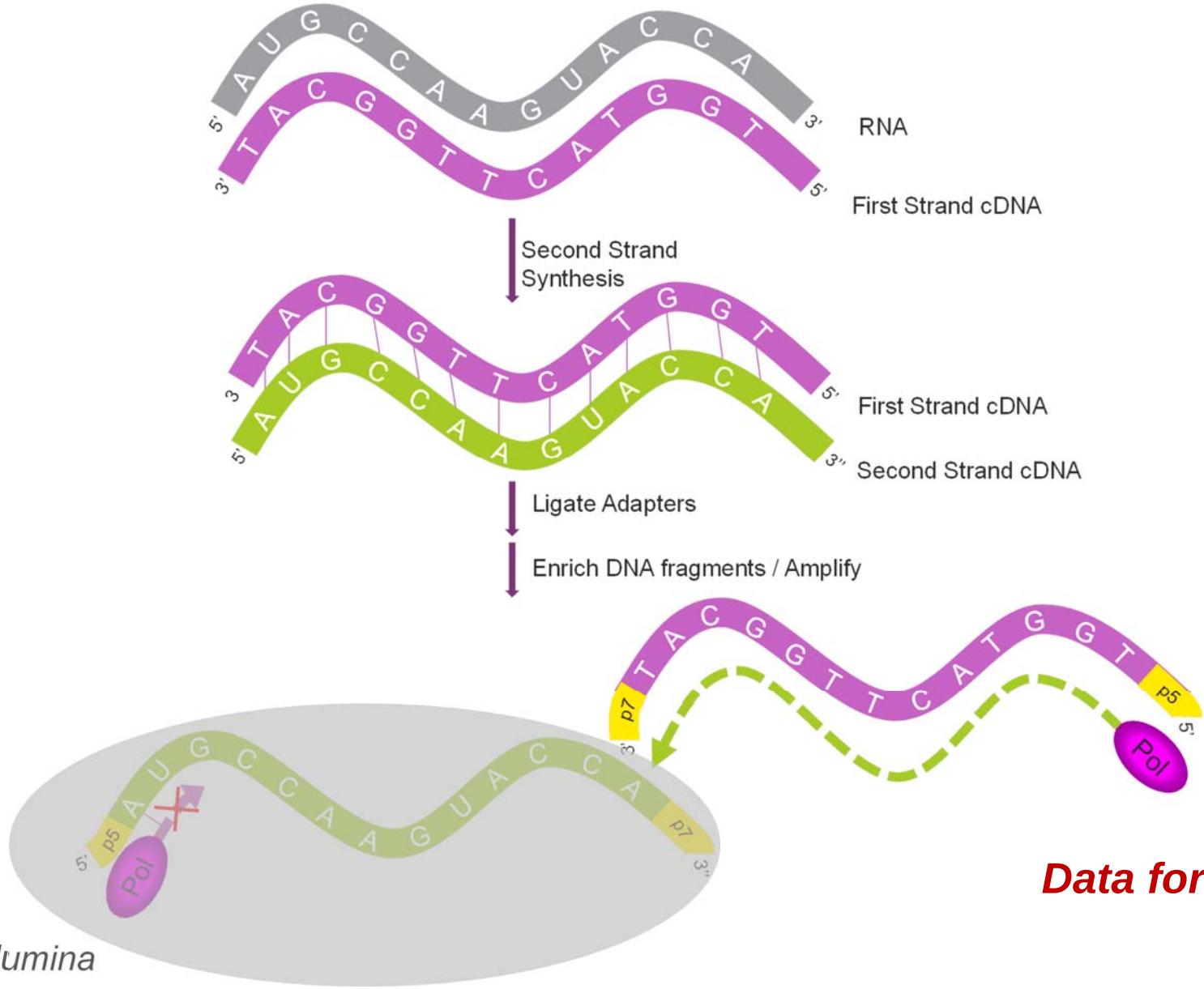
Eukaryotes
High quality RNA

rRNA removal



Prokaryotes
Non-A-tailed RN
Degraded RNA

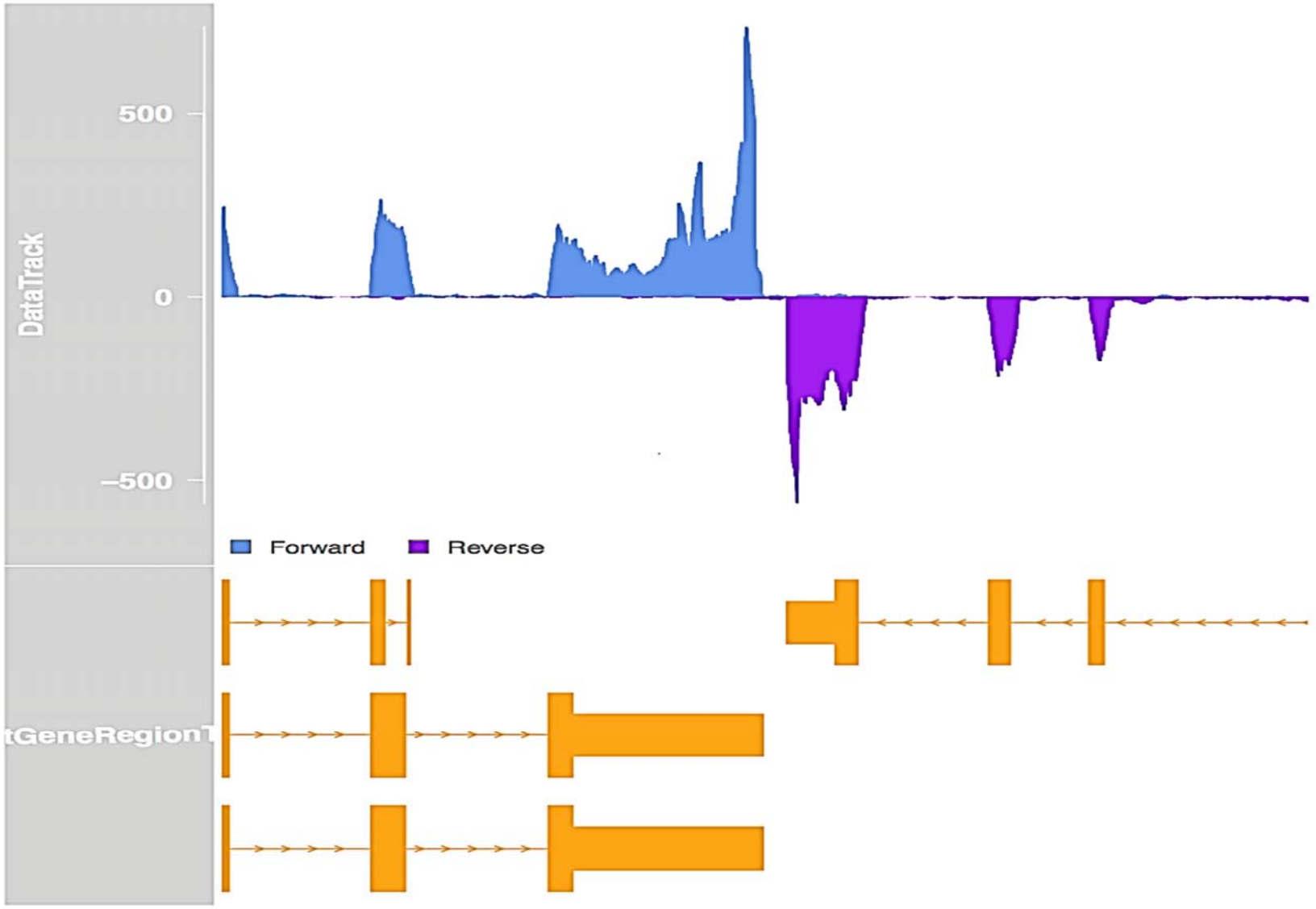
Strand-specific RNA-seq prep



Data format: RF

Source: Illumina

Visualizing stranded RNA-seq data with Gviz/Bioconductor



https://sidderb.files.wordpress.com/2014/11/blog_stranded_rnaseq_img1.png

Example of RiboZero purificatic

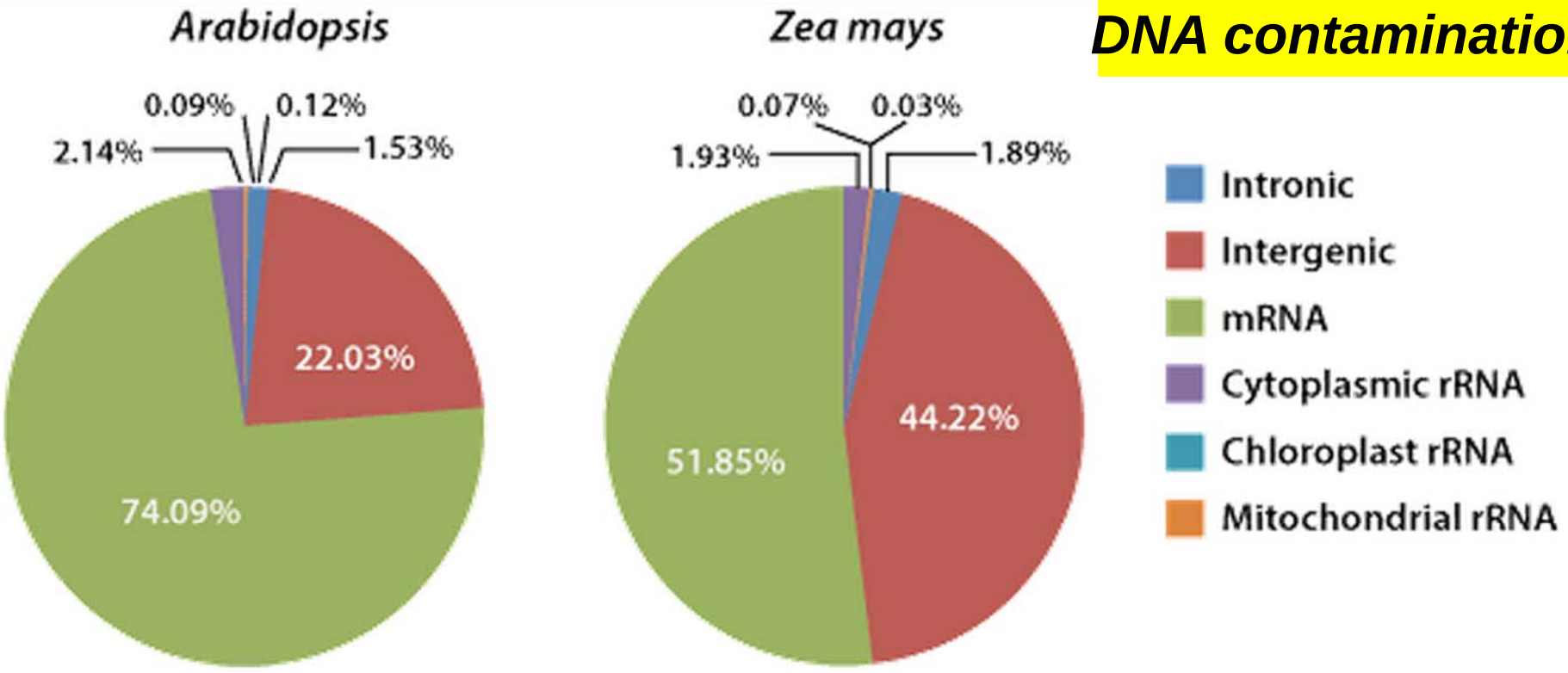
LARGE-SCALE BIOLOGY ARTICLE

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



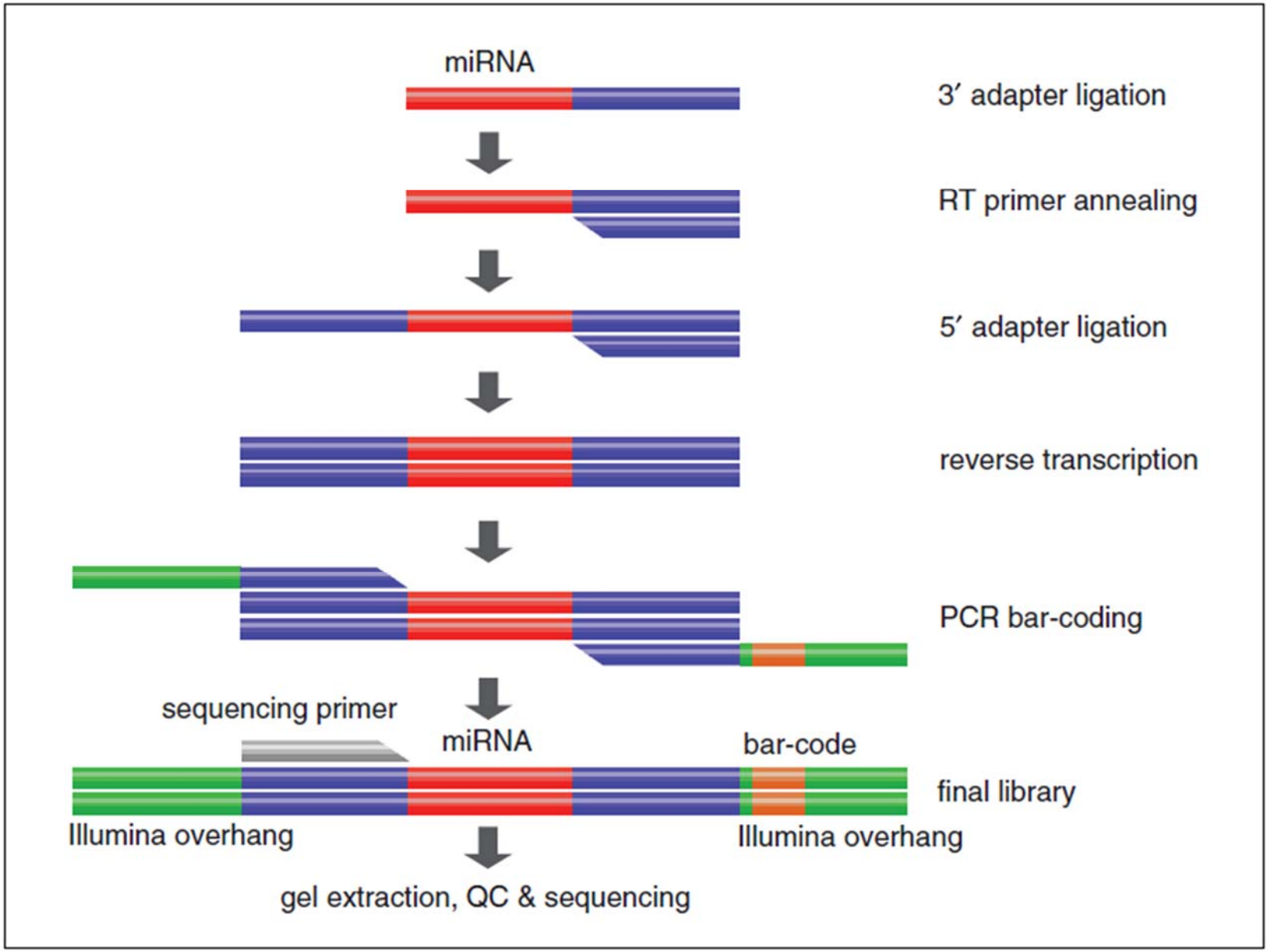
Example RNA-Seq Runs

- Human expression (per condition):
  ¼ lane HiSeq, 76bp paired
- Vertebrate annotation (per tissue):
  ¼ lane HiSeq, 101 bp paired, strand-specific
- Bacterial and fungal annotation:
  1/12 lane HiSeq, 101 bp paired, strand-specific

- Always generate paired reads if possible
- Read pairing is used to assemble transcripts
- Exception: Aligning to known transcripts for expression

Source: Dr. Michael C. Zody

smRNA library prep - Directional



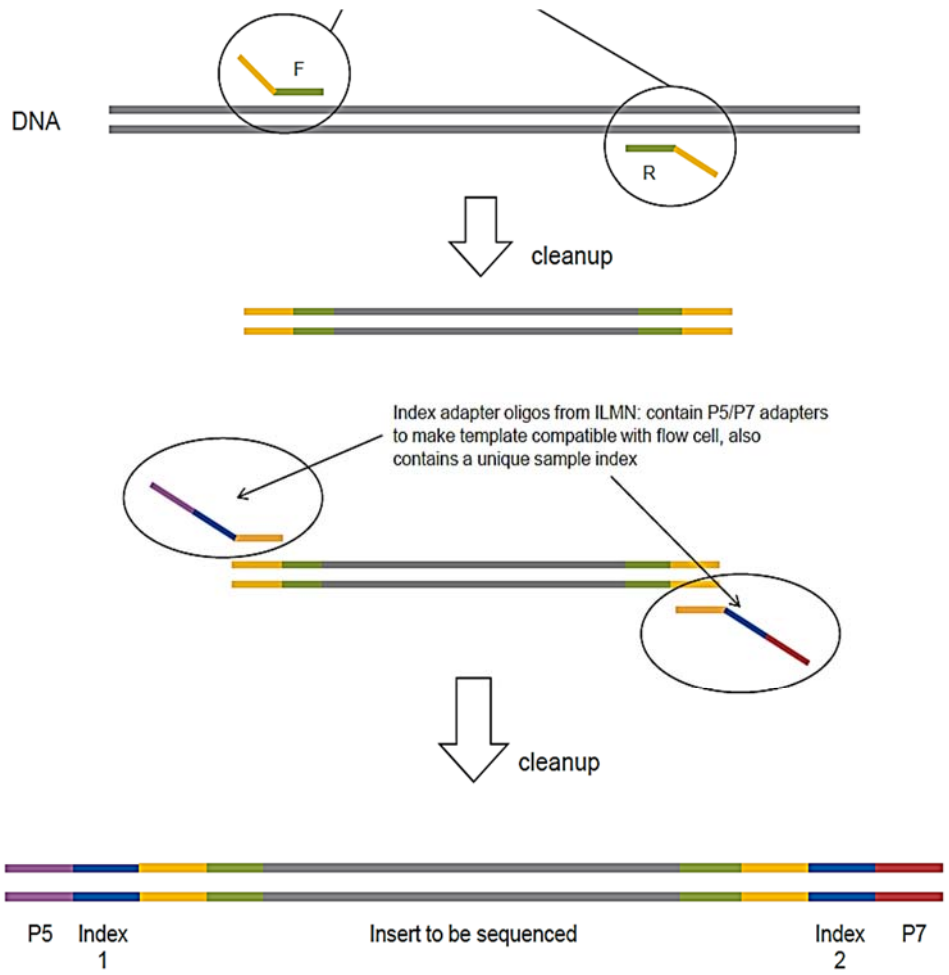
Curr Protoc Hum Genet. 2012 Apr;Chapter 11:Unit 11.12.1-10.

Targeted sequencing

- ***Amplicon***
- ***Exome Capture***

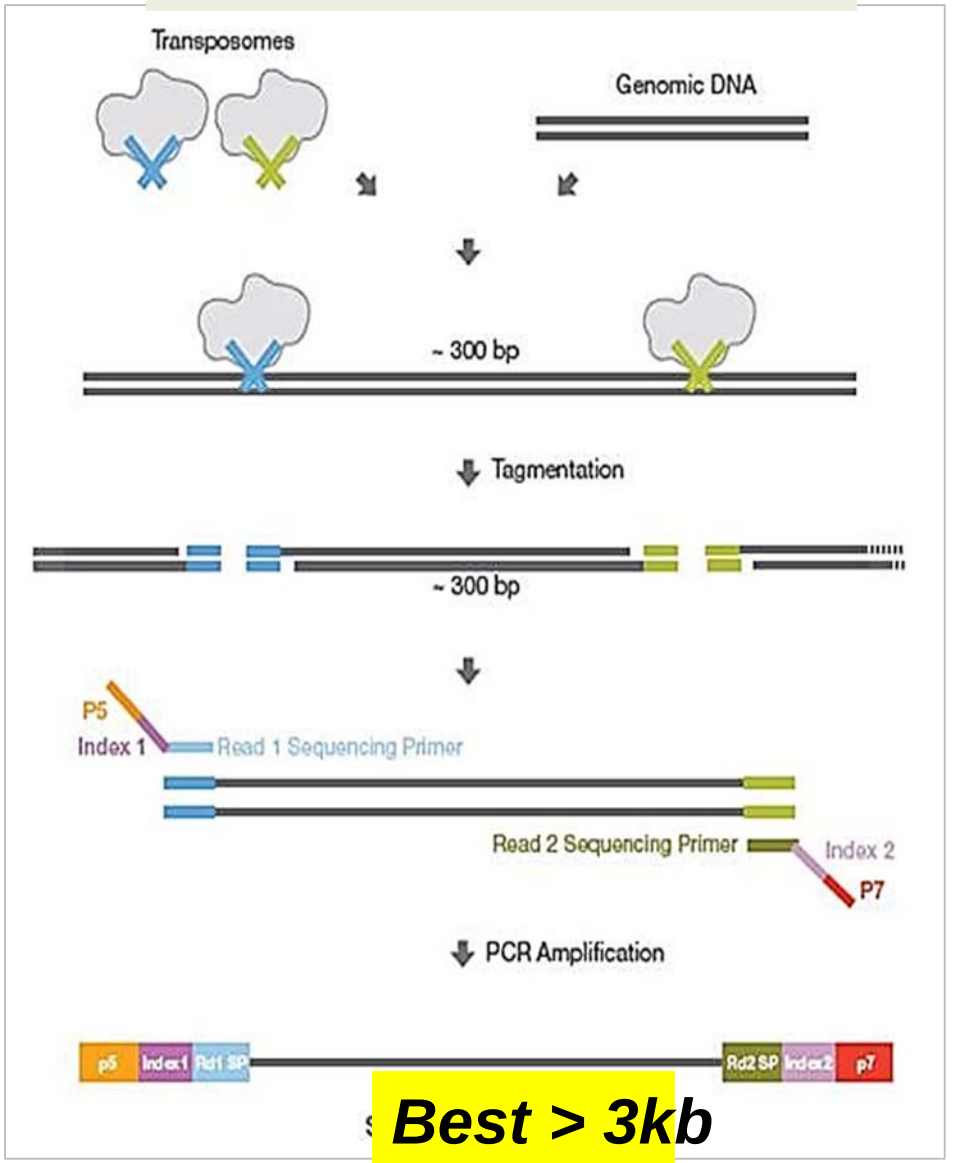
Amplicon sequencing

Indexing PCR



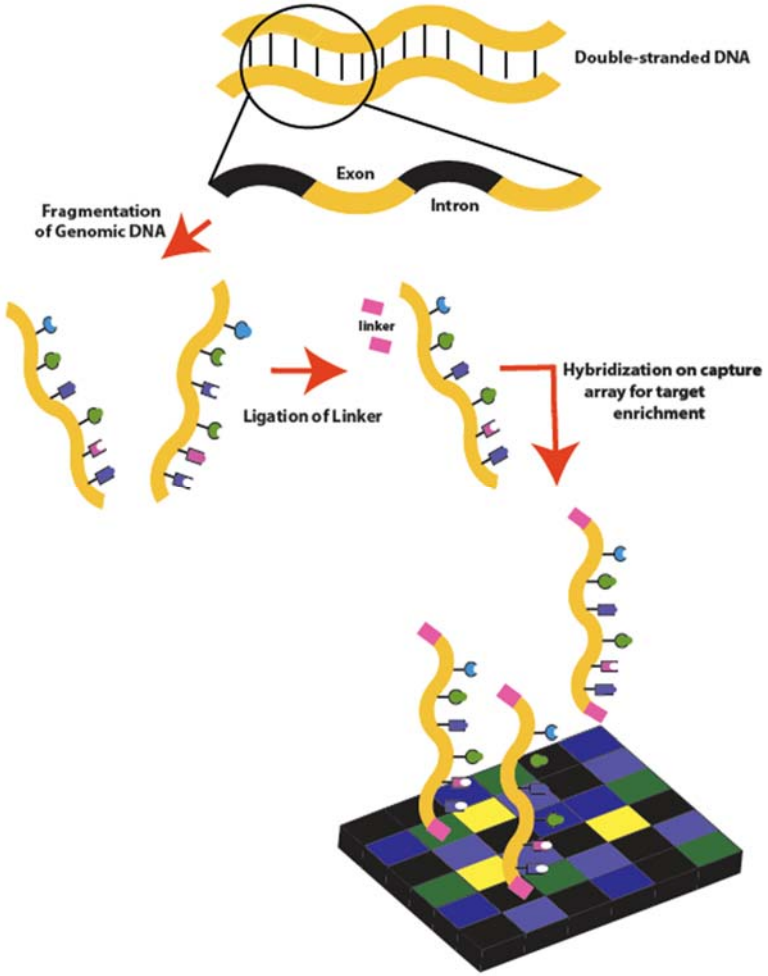
400-600bp

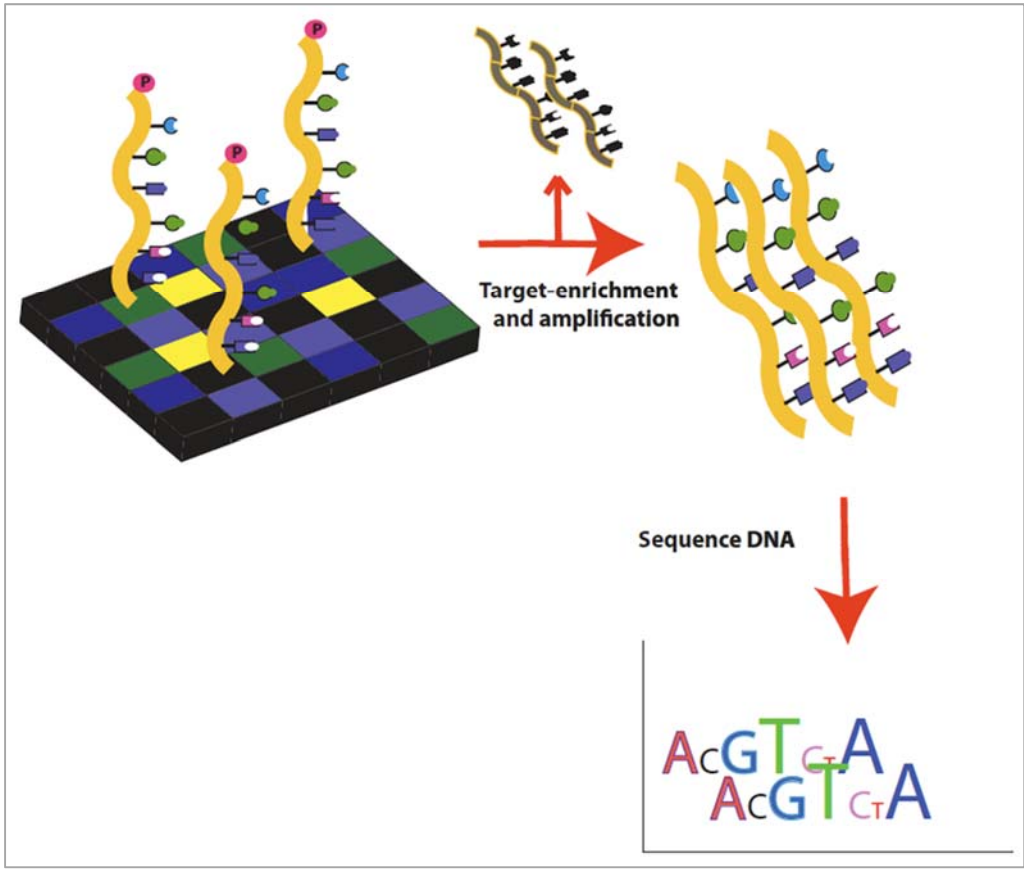
Nextera Tagmentation



Best > 3kb

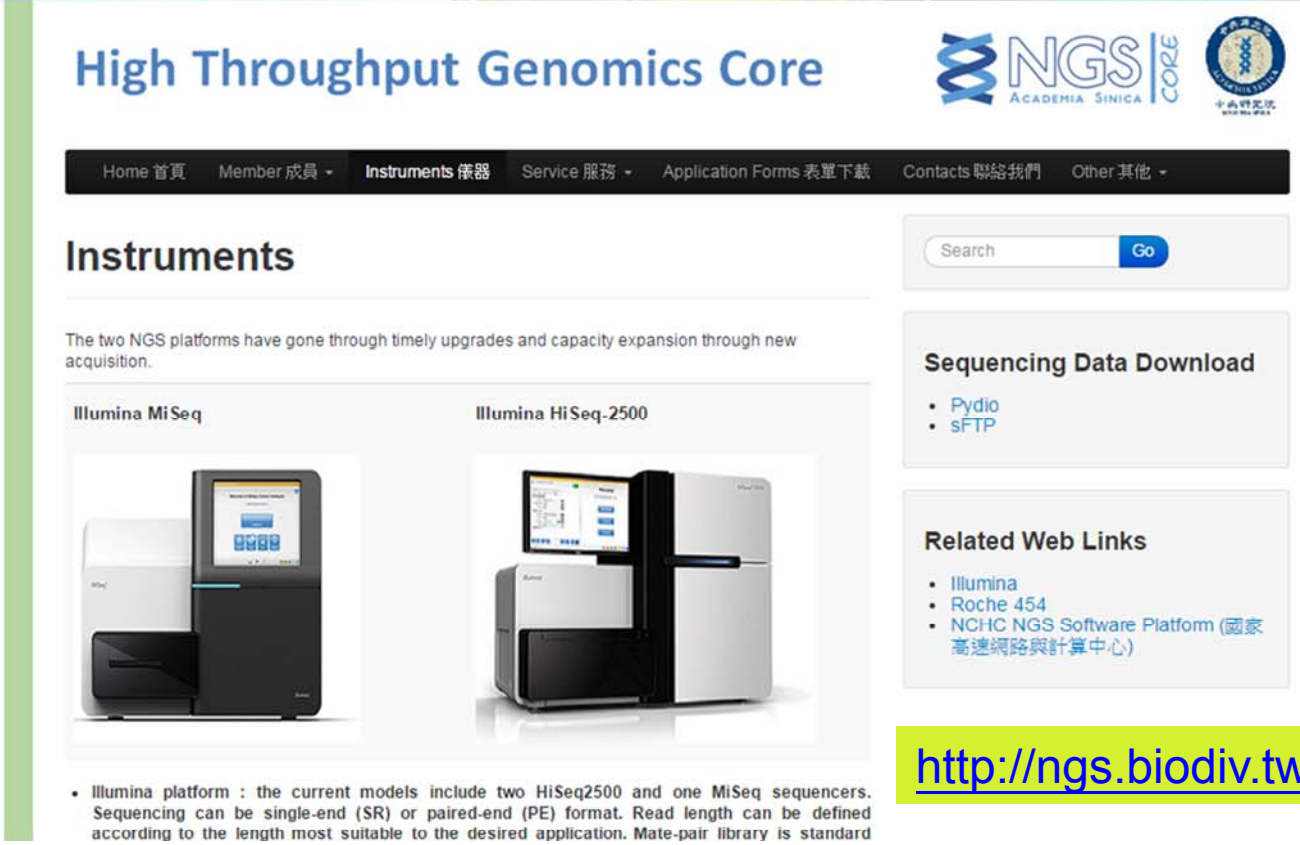
Exome/Capture Sequencing Workflow





Panel: commercial or custom

https://commons.wikimedia.org/wiki/File:Exome_Sequencing_Workflow_1a.png#/media/File:Exome_Sequencing_Workflo



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

Thank you!