次世代高通量基因組測序 –

NGS平台原理和實驗設計的考量

呂美曄博士

研究副技師，生物多樣性研究中心

中央研究院 基因體定序中心

內容: 本課程將介紹常用的NGS平台、原理、和常見的應用。

內容也將包括平台選擇，實驗設計的總體考量，以及NGS應用相關的樣本問題。

2015/6/8 LSL, 9:40 – 10:20am

High Throughput Genomics Core

Outlines

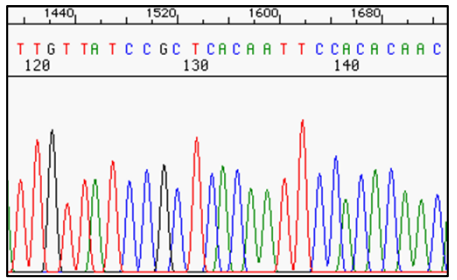
1. Revolution of sequencing technologies
2. Choice of platforms
3. Data format
4. Applications
5. Single cell
6. Project design considerations
7. Omics: genome, transcriptome, proteome, metabolome

§ Sequencing advancement

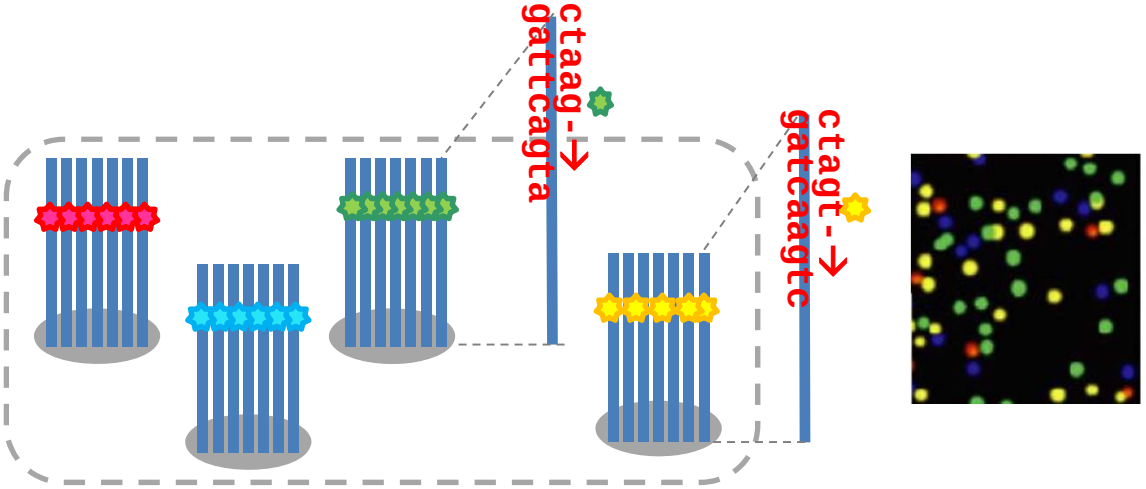
Evolution of Sequencing Technologies

Sanger
1 rxn/tube

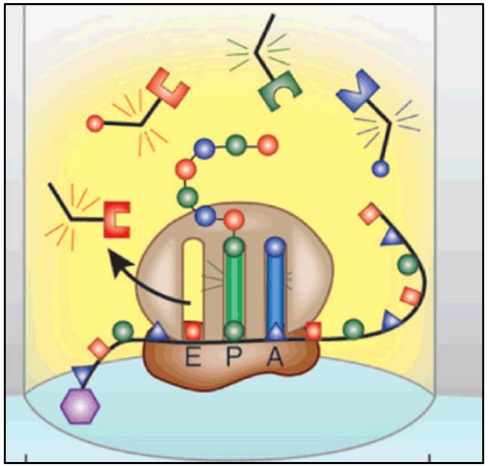
```
gctagttgaccttgaccaagcatggcgatcgat
|||||||
cgatca--->
```

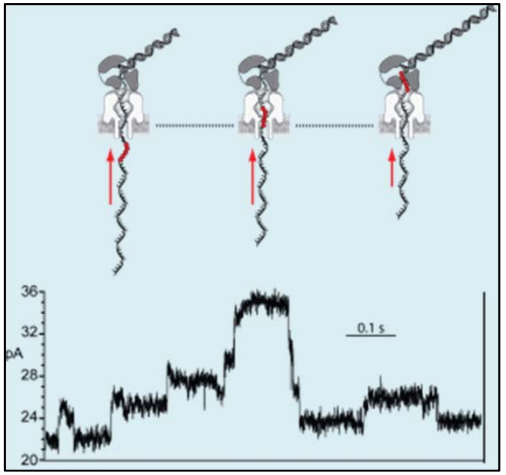


2nd-Gen
Clonal amplification



3rd-Gen
Single mol. Seq.





Choice of NGS platforms





Platform	Roche 454		Illumina			Life Technologies		PacBio
Model	GS+	454 Jr.	HSQ2500 HT	HSQ2500 Rapid	MiSeq	Ion Torrent	Ion Proton	RS II
Output /run	600-1000 Mb	40-70 Mb	1 Tb /8 lanes	90 Gb /2 lanes	15 Gb	1 Gb (318 chip)	10/100G (P1/P2)	250-375 Mb
Read length	avg. 450-950nt	450-600nt	PE 2*125	PE 2*250	PE 2*300	avg. 250nt	n.a.	Up to 30 kb
Run time	48 hrs		6-7 days	48-70 hrs	48 hrs	~3 hrs	n.a.	<2 hrs
insert range	20-30 kb			MP: 20 kb	MP: 20 kb	3-10 kb		15-20 kb
Special				SLR: 5-10kb				

As of 5/2015

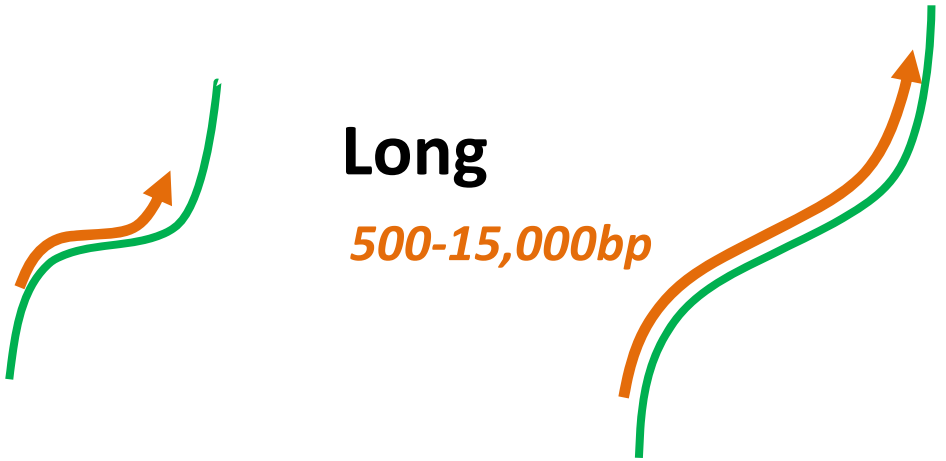
Comparison of 2nd vs 3rd-Gen sequencing

		2nd-Gen	3rd-Gen
Specs	Read length	Short-med. (200~1000nt)	Long (300nt~20k nt)
	Read count output	High	low
	Base call accuracy	High	Lower
Platforms	Manufacturer	454, Illumina, Ion Proton	PacBio, (NanoPore)
	Pros/Cons	Shorter reads, High per base Q, higher throughput	Longer reads, Lower per base Q, lower throughput
Applications	De novo genome		
	De novo transcriptome		
	Re-seq: SNP, short INDEL		
	Genome re-arrangement	Difficult	Yes
	RNA-seq profiling	Yes	Too shallow
	Isoform sequencing	Yes	Yes (read through?)
Sample requirement	Purity	Mid-high	High
	Quantity	Mid-high	High
	Integrity	Mid-high	High

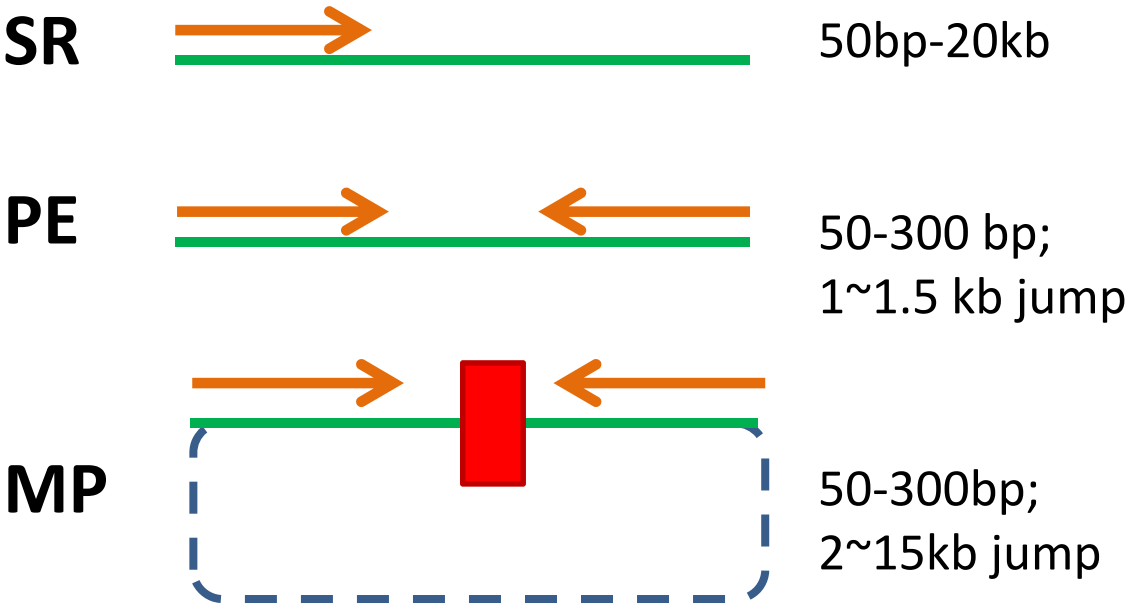
§ Data format and QC

Types and Characteristics of NGS Reads

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



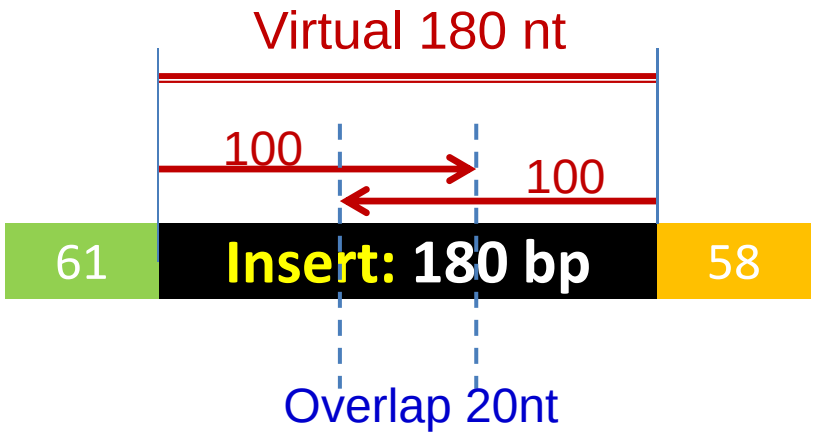
- **Read types:**



Insert size vs Library Fragment Size

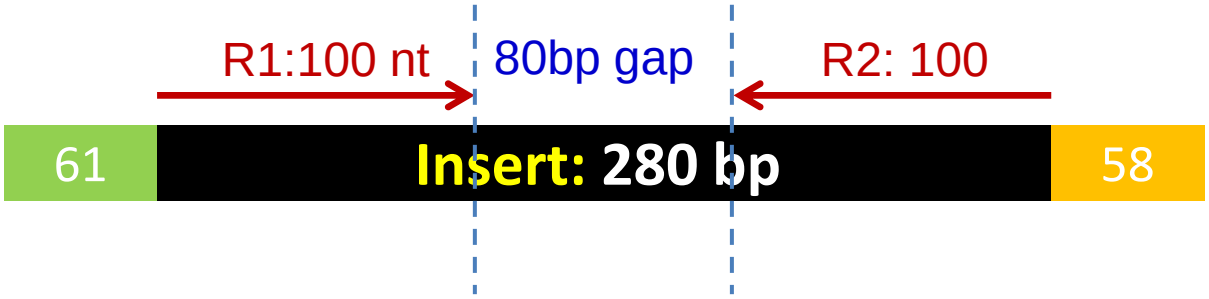
300-bp fragment:

Ends overlapped 20 bp



400-bp fragment:

Ends gapped by 80 bp



Adaptor sequence length: can vary between 60-70bp

Illumina Read – fastQ

Index sequence

no control

Y/N: failing PF or not

Sequence header *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

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

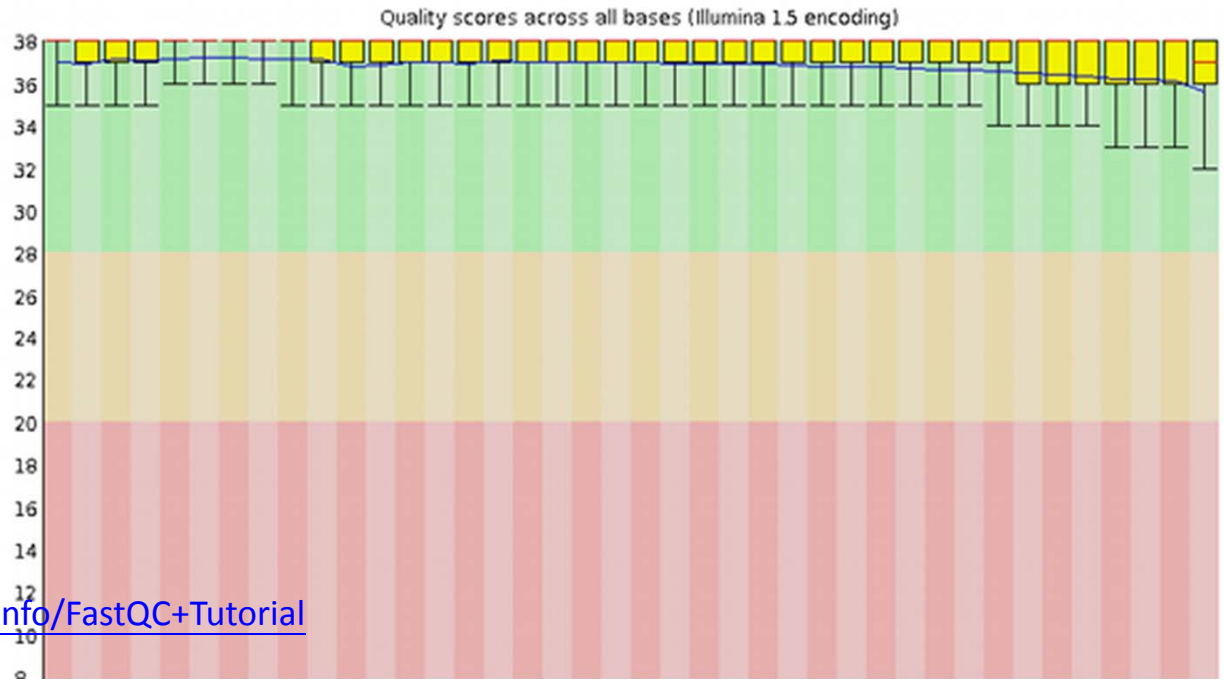
Basic Statistics

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

Read processing:

- FASTX-toolkit
- Trimmomatic
- NGS QC Toolkit

Per base sequence quality



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

§ NGS – Common applications

Genome-Seq

ChIP-Seq

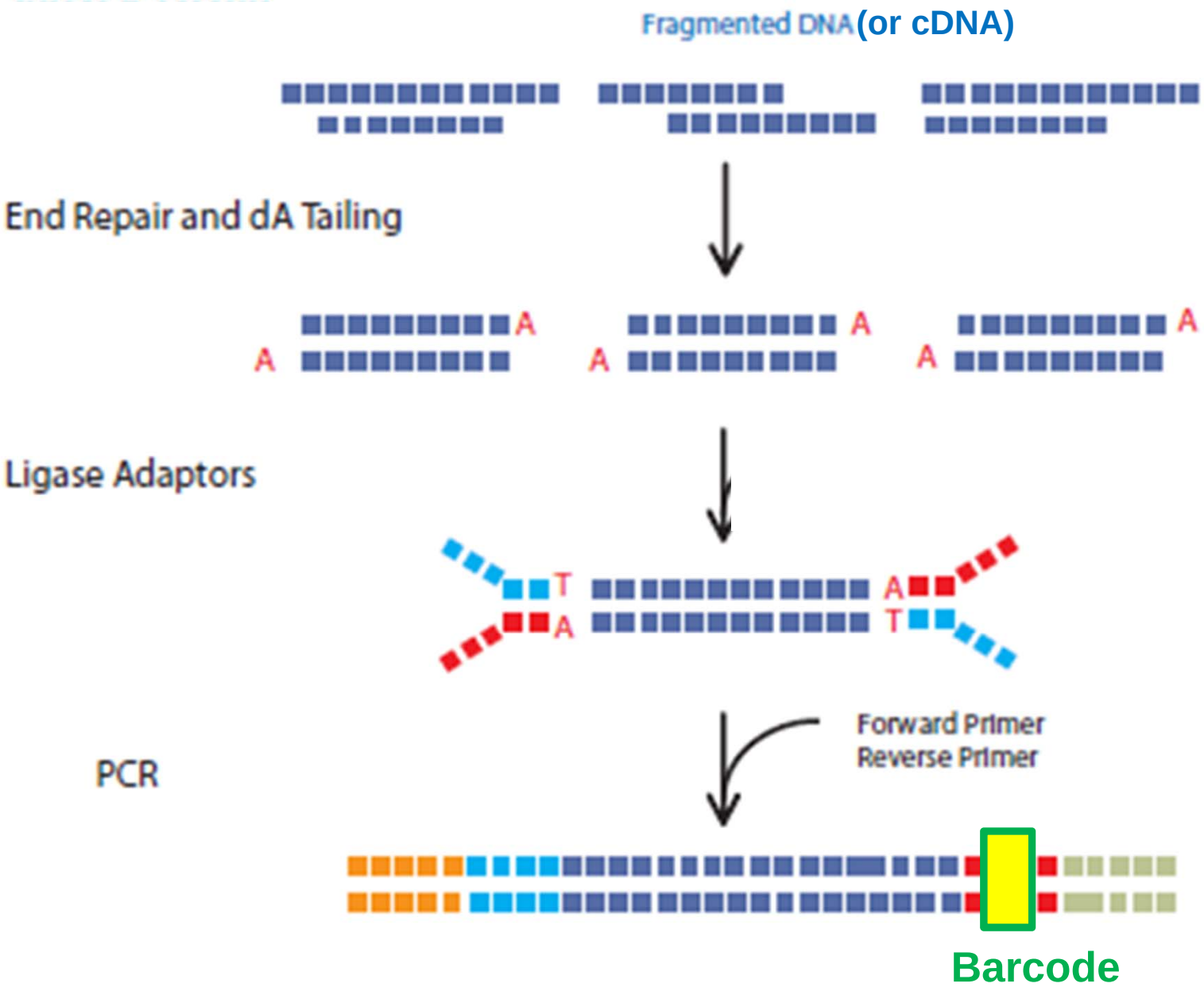
RNA-Seq

miRNA-Seq

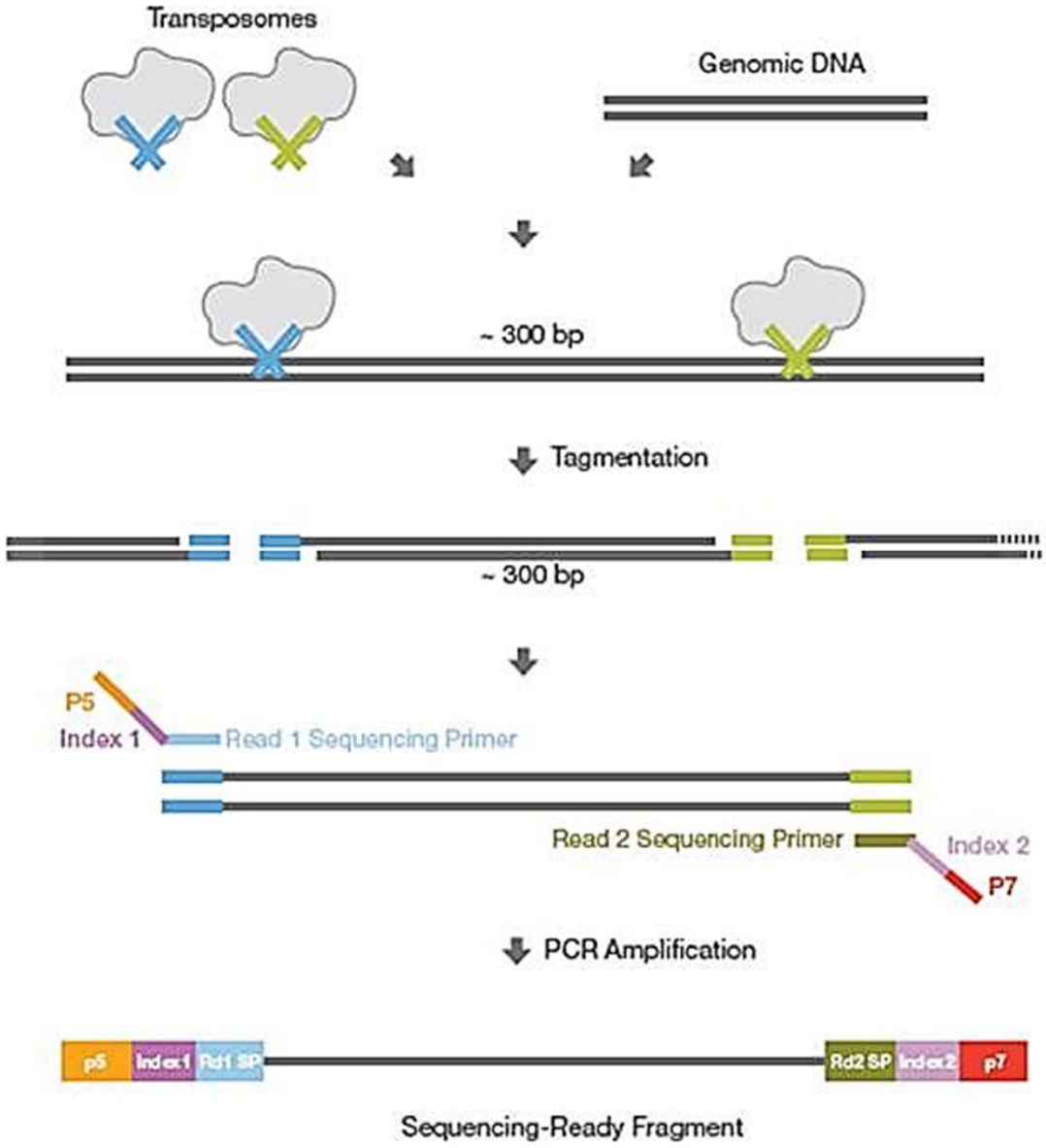
1. Genome Sequencing

- ***Re-sequencing***
- ***De novo assembly***
- ***SLR (Synthetic Long Read)***

1. Shotgun Library Preparation



2. Tn tagging - Nextera Library prep



1. **TemplateDNA + transposome complex (contain adaptor)**

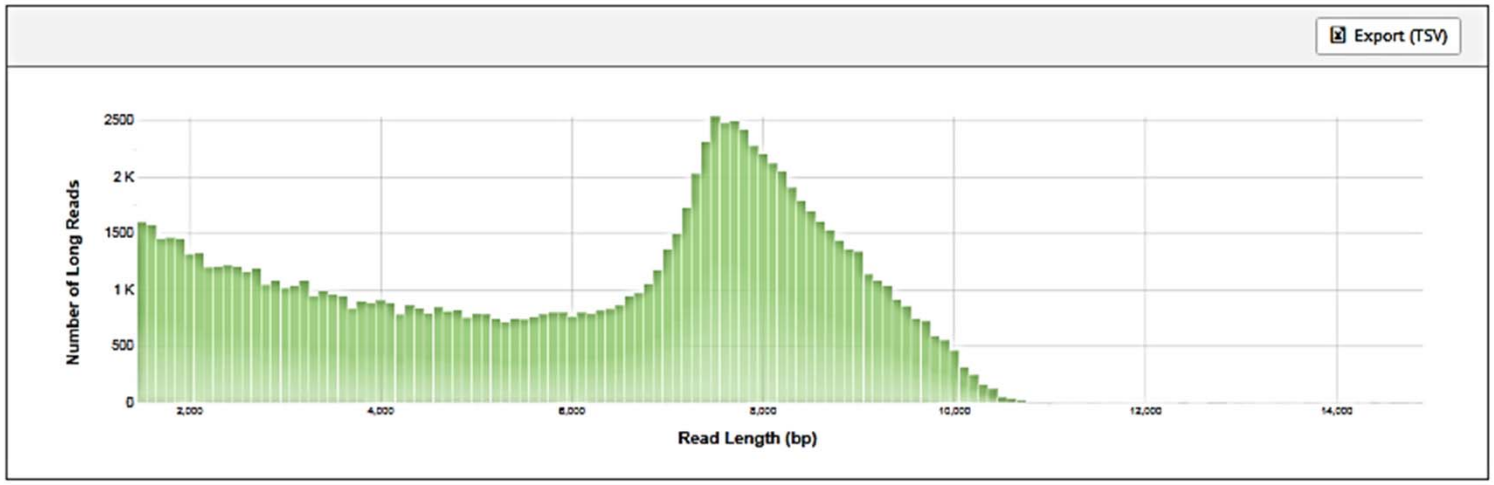
2. **Tagmentation breaks DNA and add adaptor to ends**

3. **PCR amplification to engineer barcode and sequencing primers**

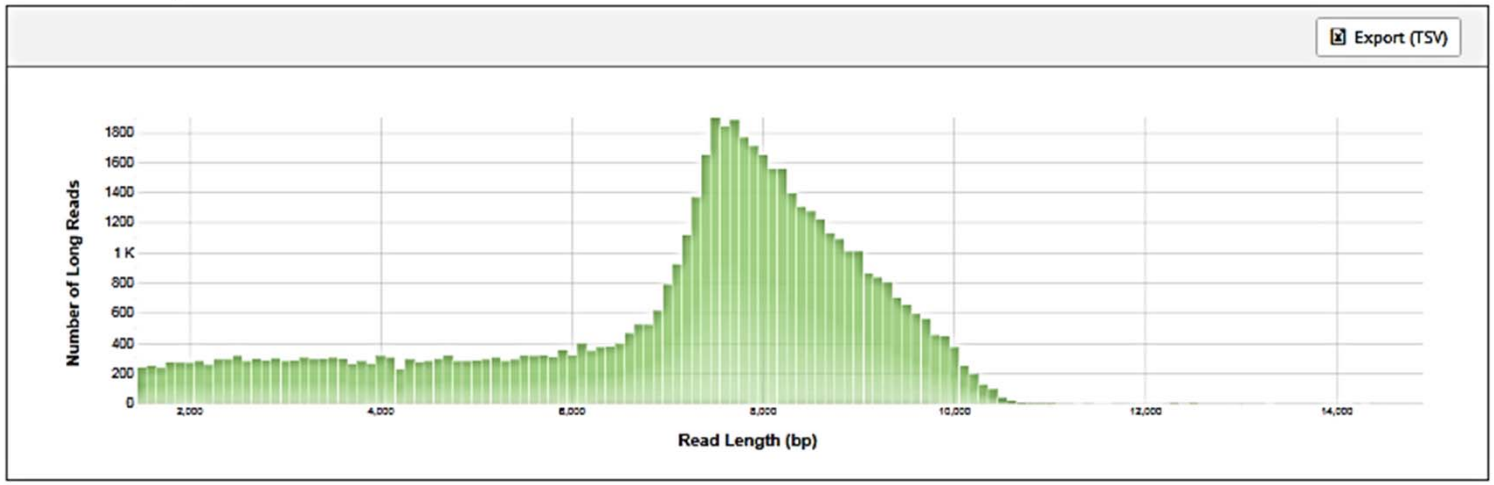
http://www.gtbiotech.com.tw/products/Nextera_XT_DNA.asp

3. Synthetic Long Read (SLR)

All Long Read Size Distribution



End-Marked Long Read Size Distribution



HiSeq reads assembled into long ones; major 5-9kb

Re-sequencing: Variant detection

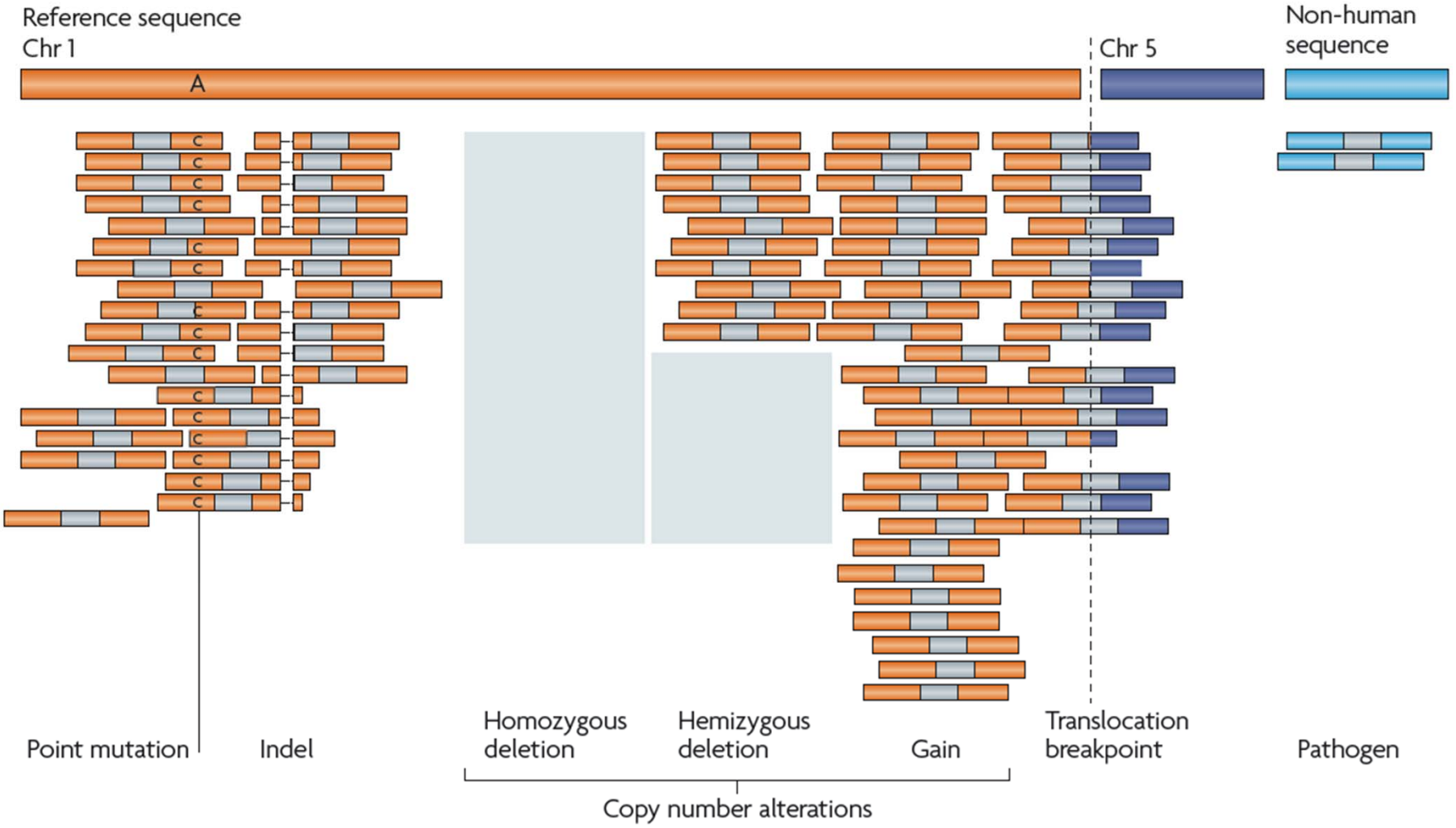
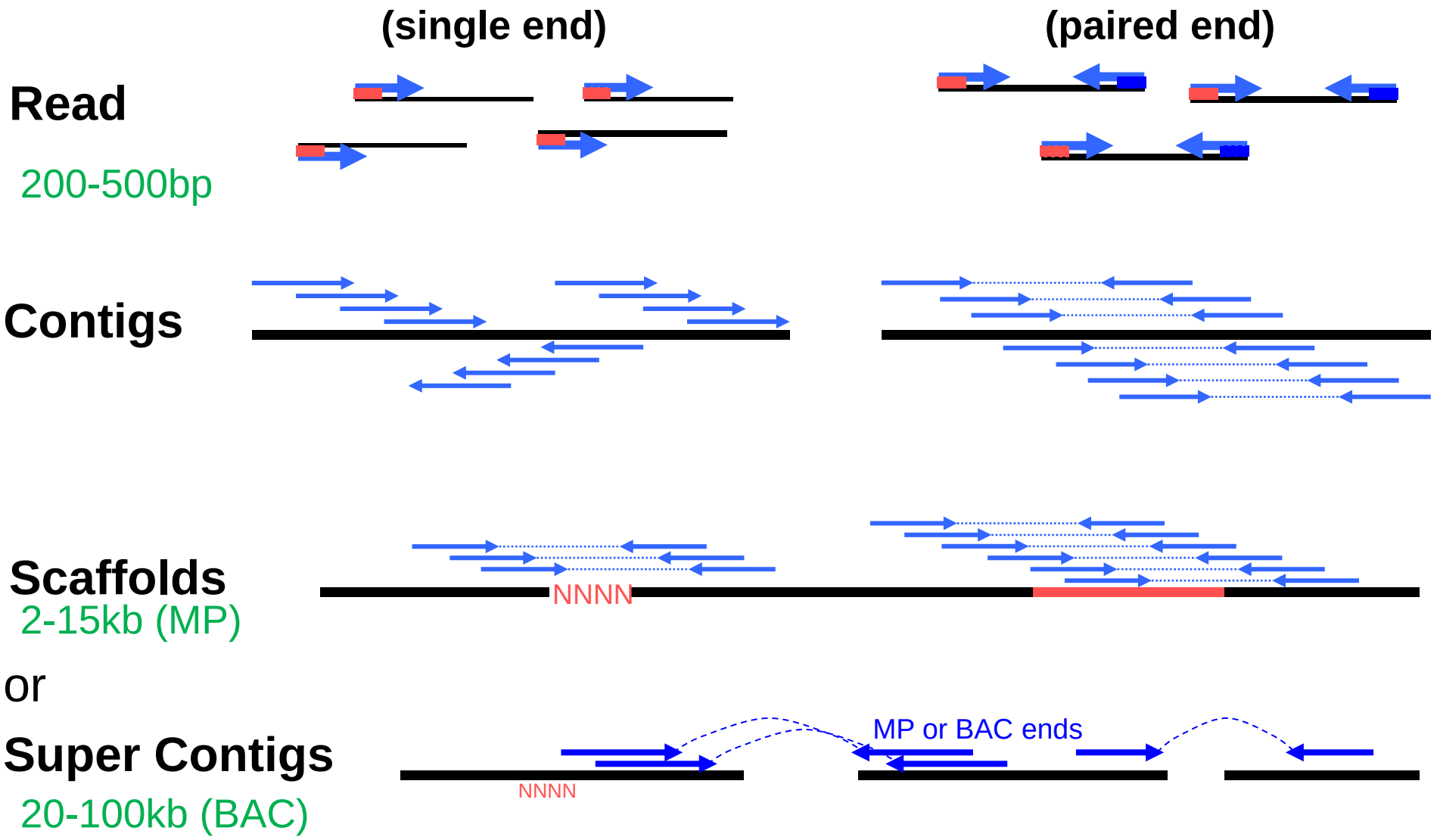


Figure 3 | **Types of genome alterations that can be detected by second-generation sequencing.** Sequenced

Meyerson et al. NRG 2010

De novo Genome Assembly - hierarchical

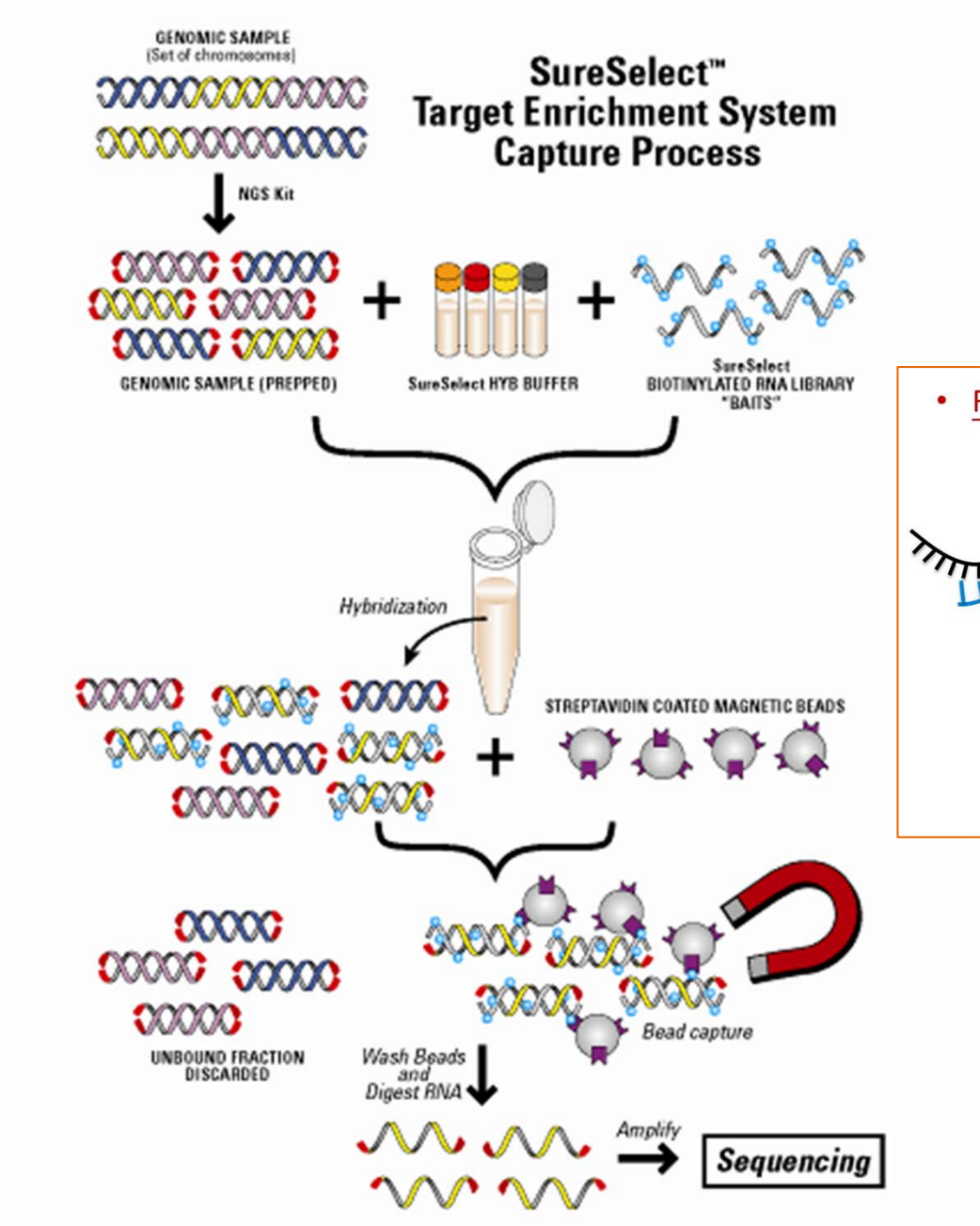


2. Capture Sequencing

- ***Design concept***
- ***Genome covergae***
- ***Selected panel***

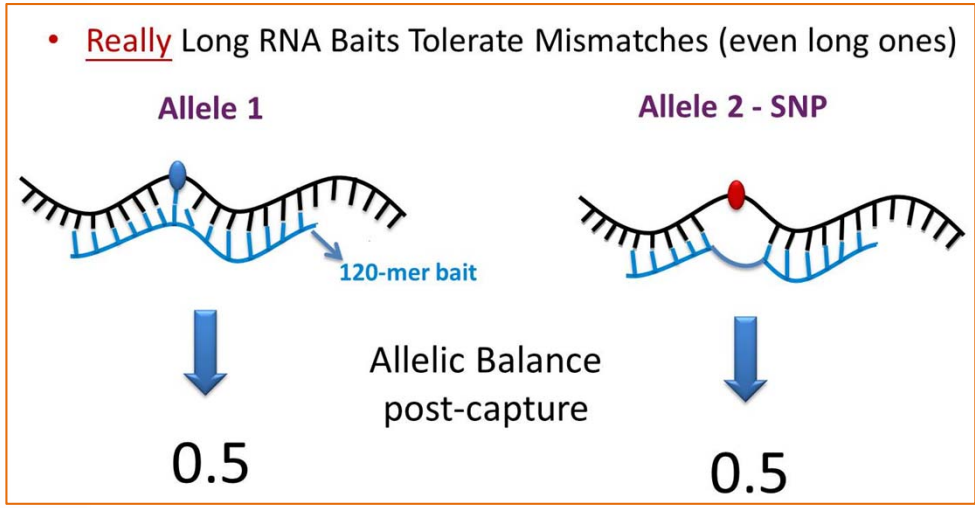
Chilamakuri et al. BMC Genomics 2014, 15:449
Performance comparison of four exome capture systems for deep sequencing.
http://www.biomedcentral.com/1471-2164/15/449

SureSelect™ Target Enrichment System: Workflow

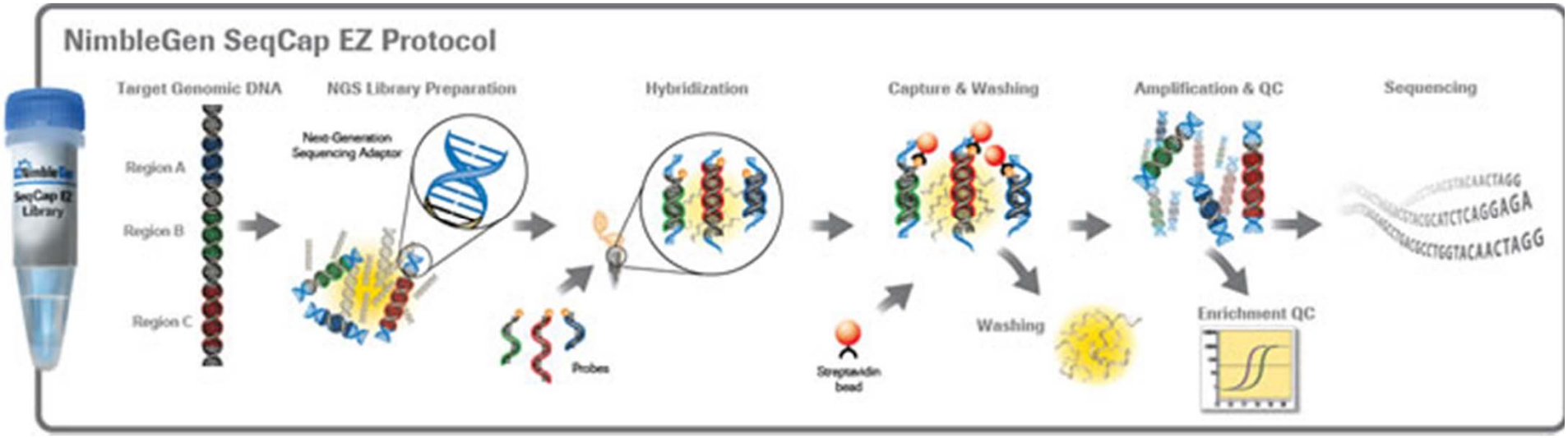


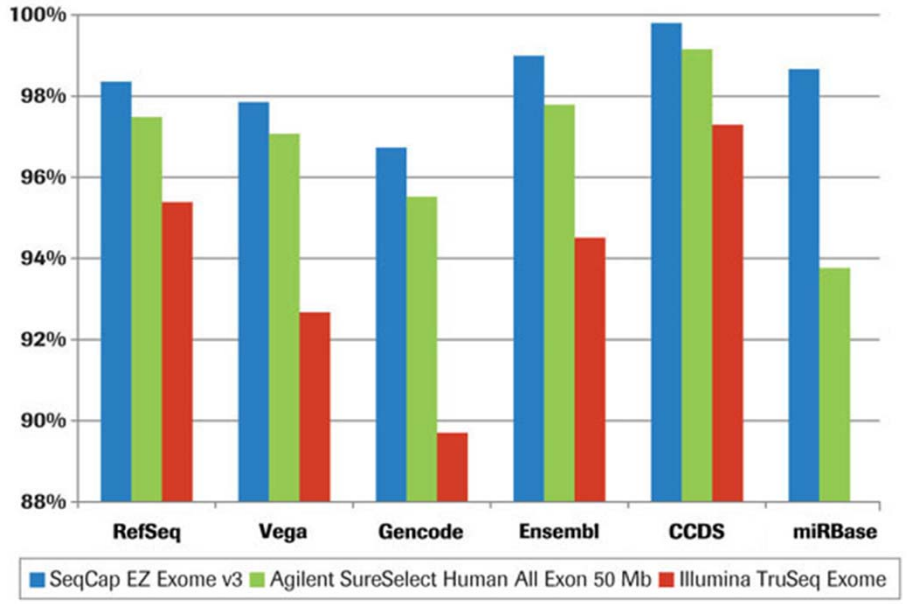
Sample requirements:

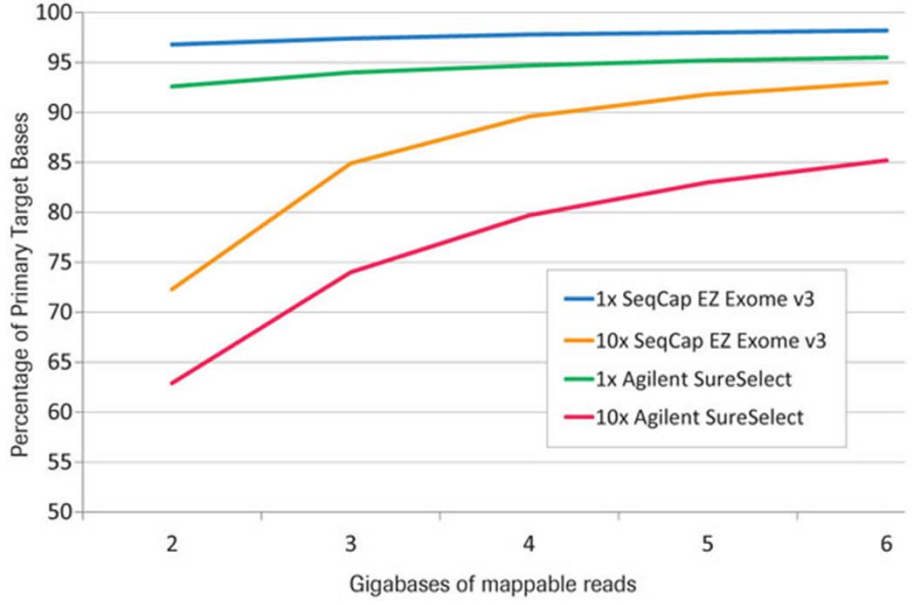
a. ds-DNA > 3 ug
b. OD260/280 >1.8



Roche SeqCap



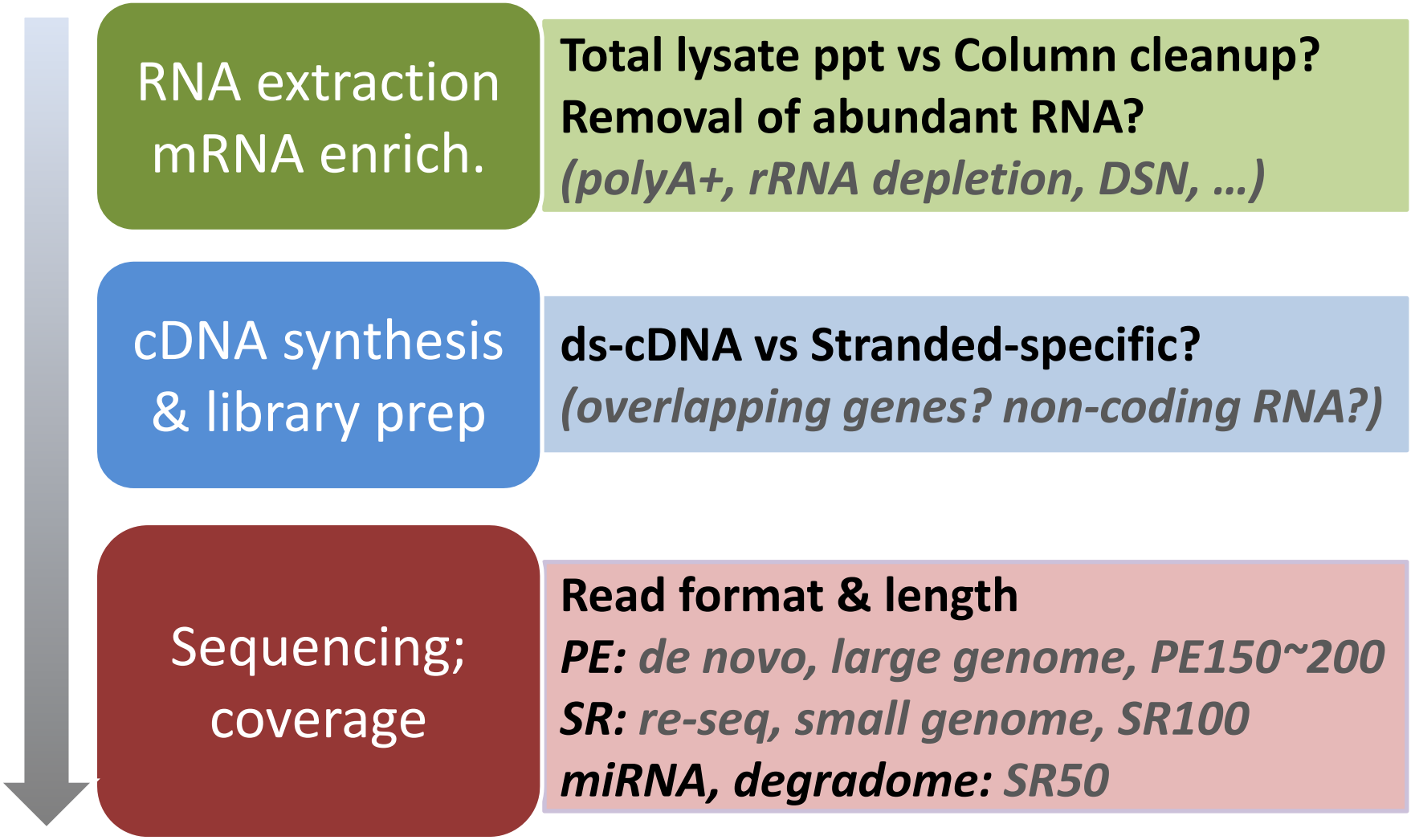




3. Transcriptome sequencing

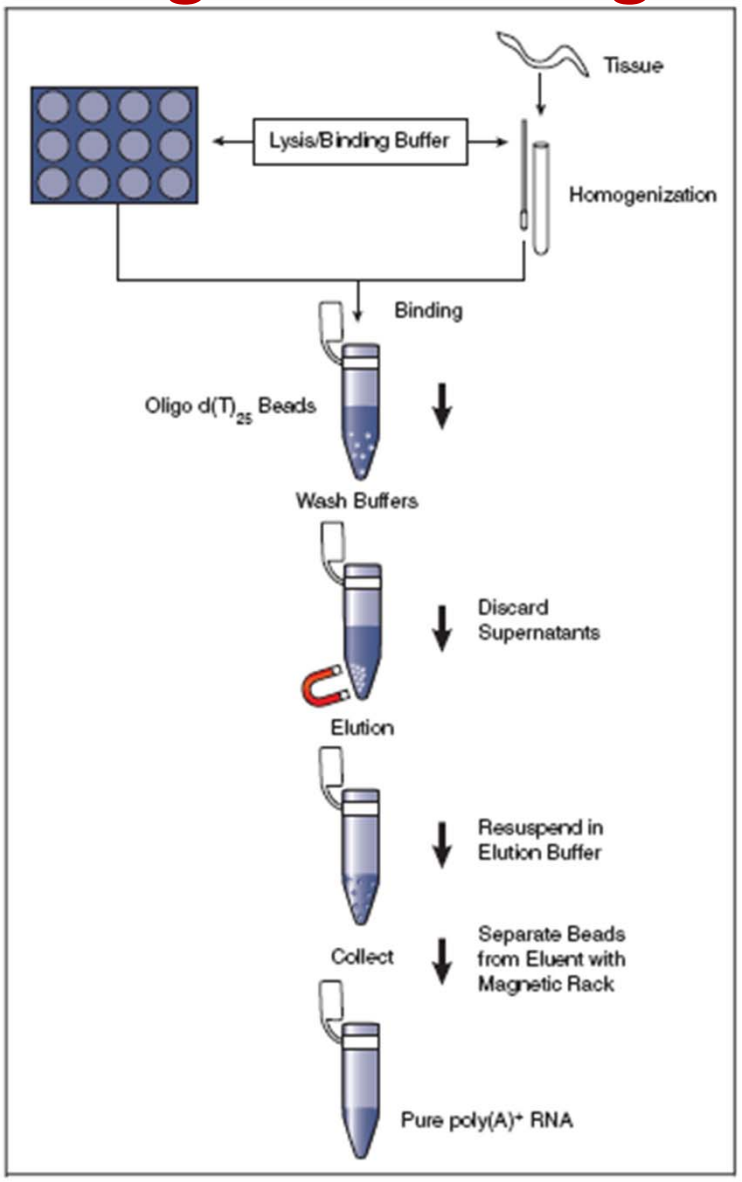
- *mRNA vs stranded*
- *smRNA, non-coding RNA*

RNA-seq: considerations

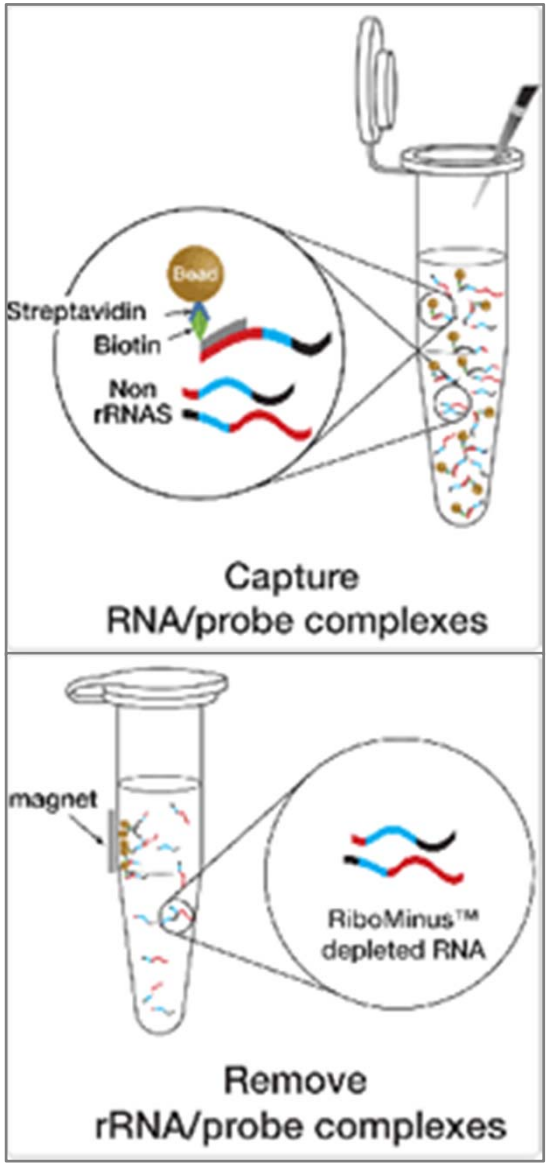


mRNA enrichment

Oligo-dT binding

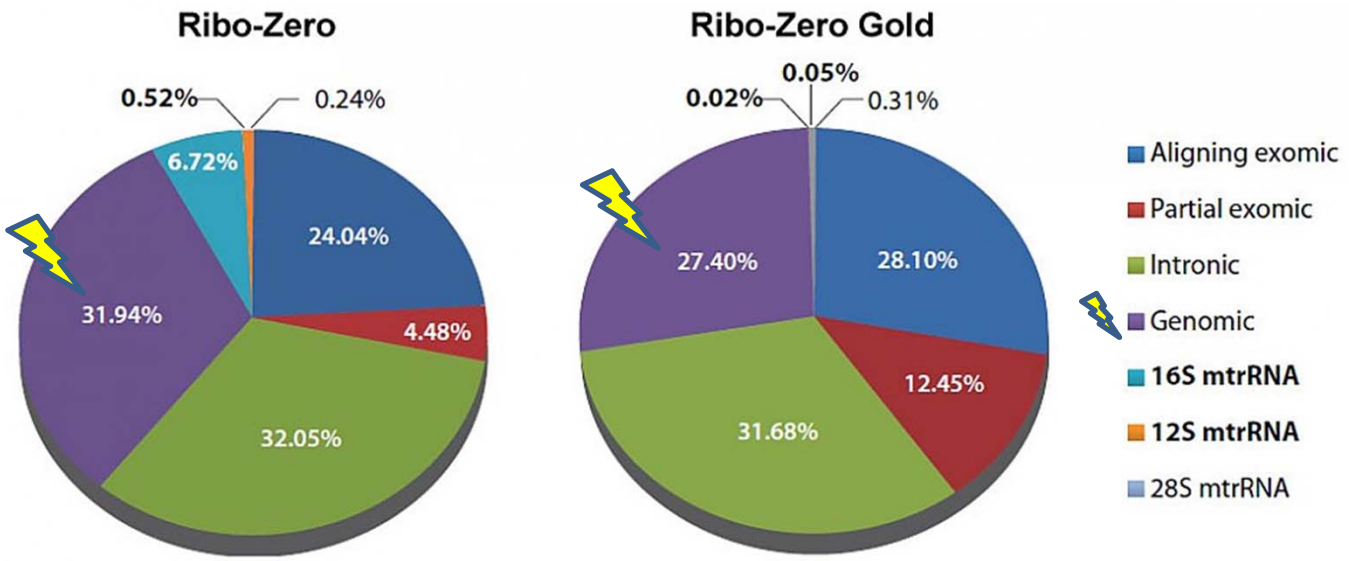


rRNA removal

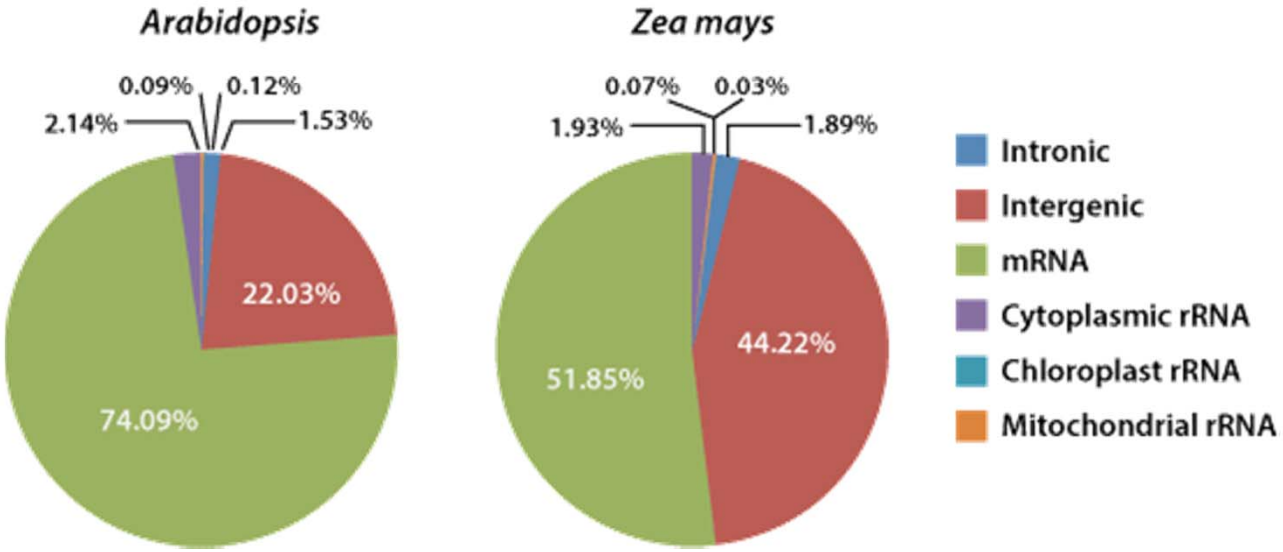


rRNA removal by RiboZero depletion

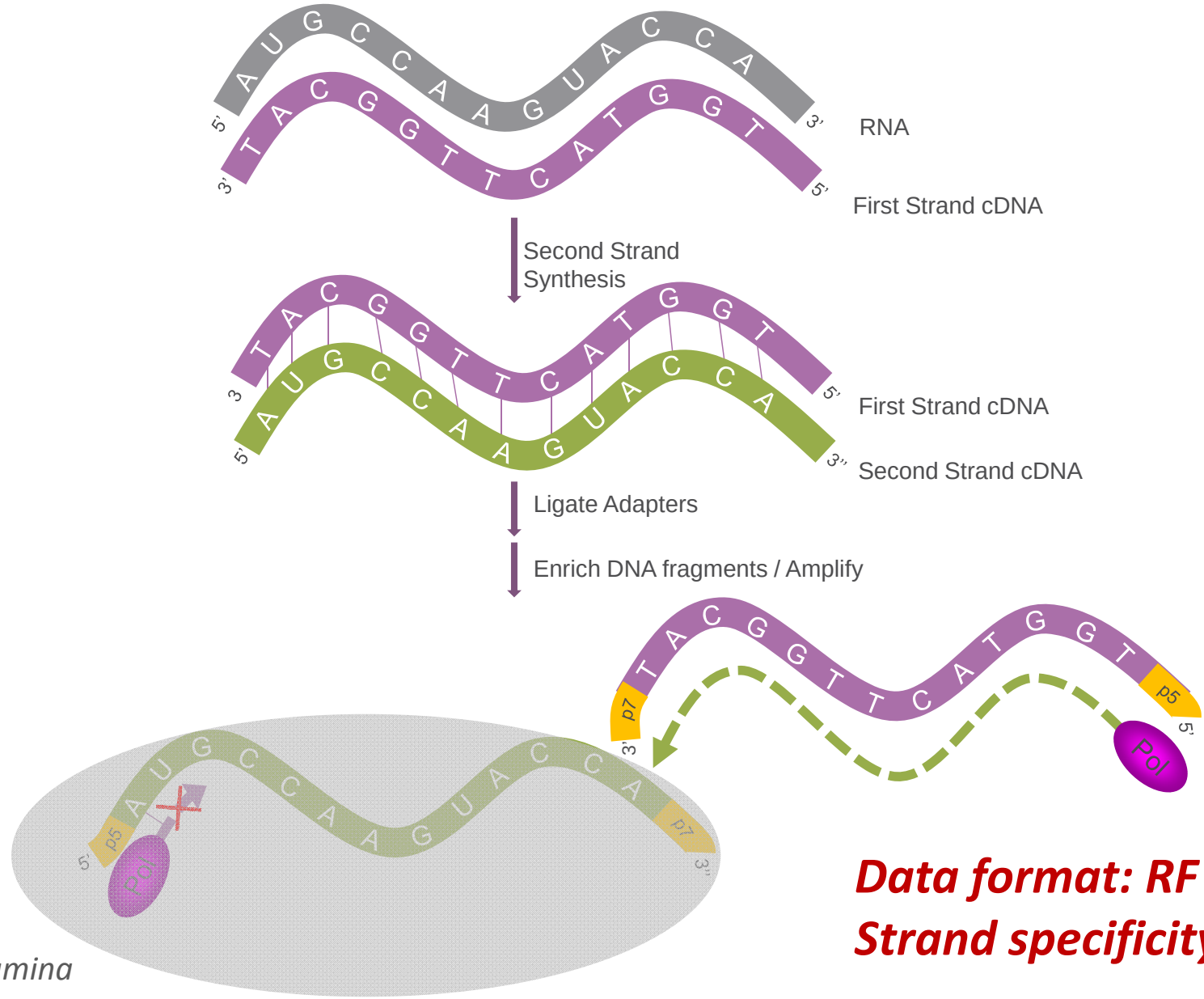
Epicentre



Illumina



Strand-specific RNA-seq prep

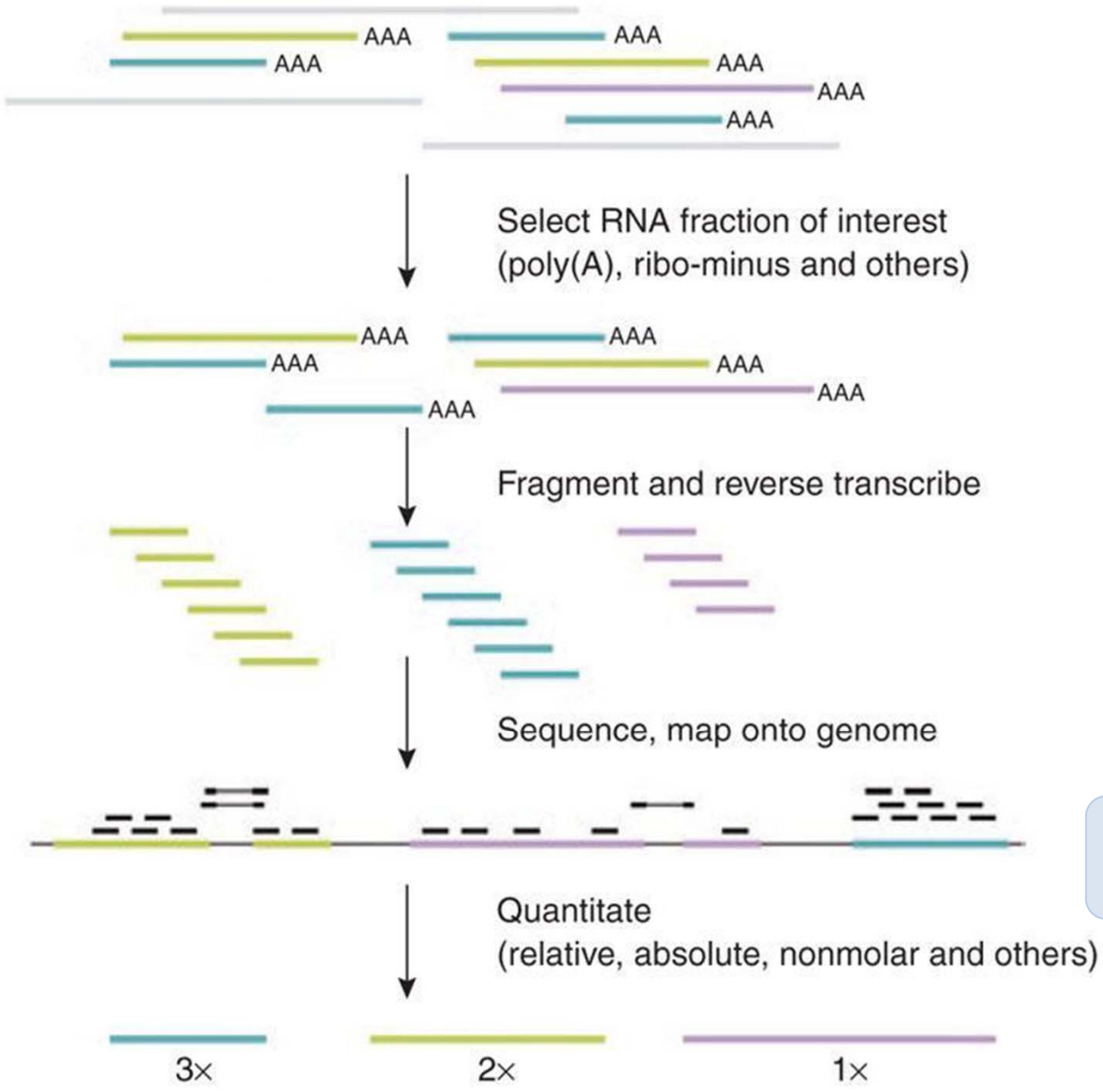


Source: Illumina

Data format: RF

Strand specificity >90%

Transcriptome profiling: RNA-seq

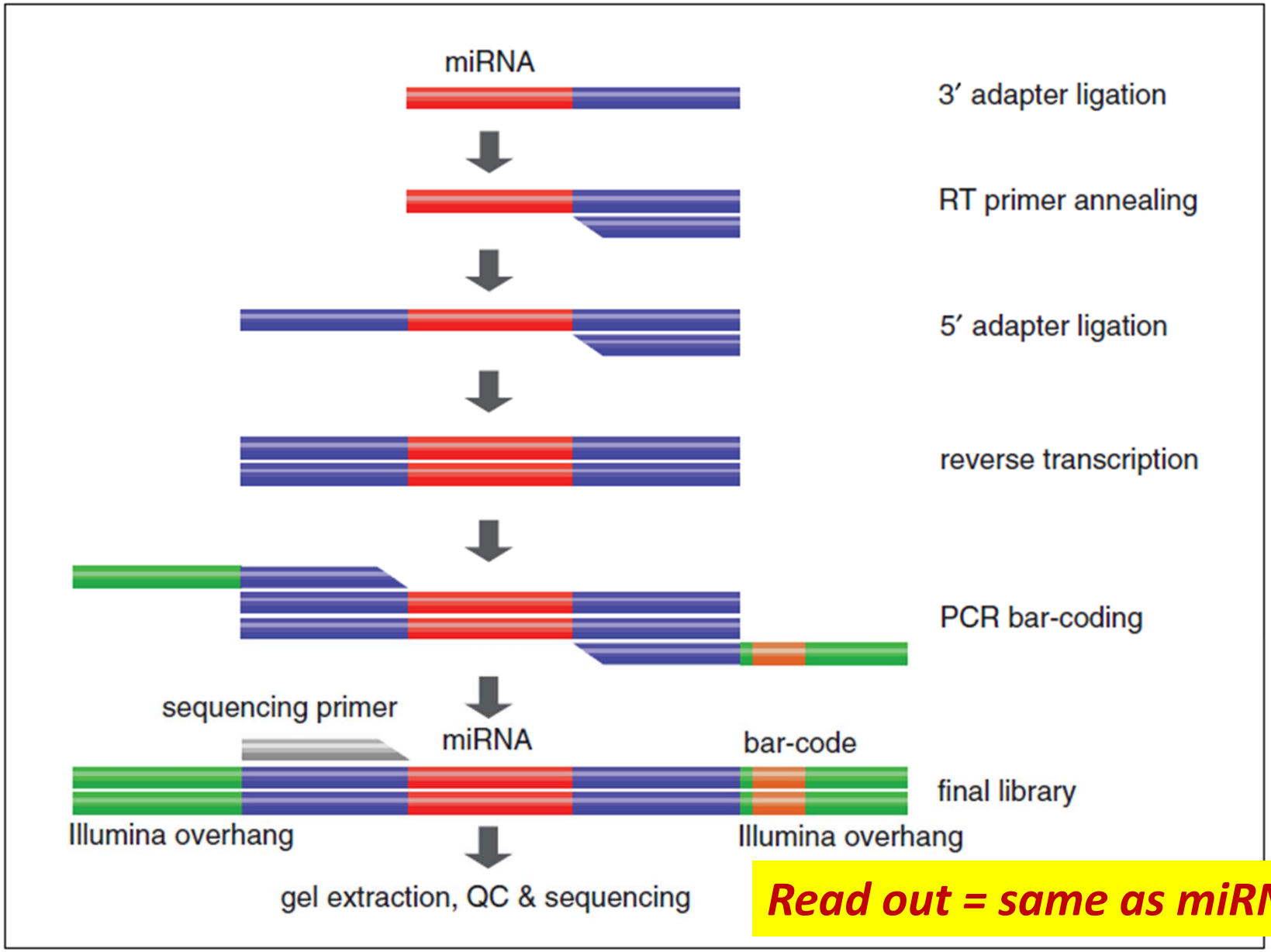


RPKM (FPKM)

$$\frac{\text{Mapped Reads}}{\text{Gene length (Kb) x total reads (M)}}$$

Nature Methods **6**, S22 - S32 (2009)

smRNA library prep - Directional



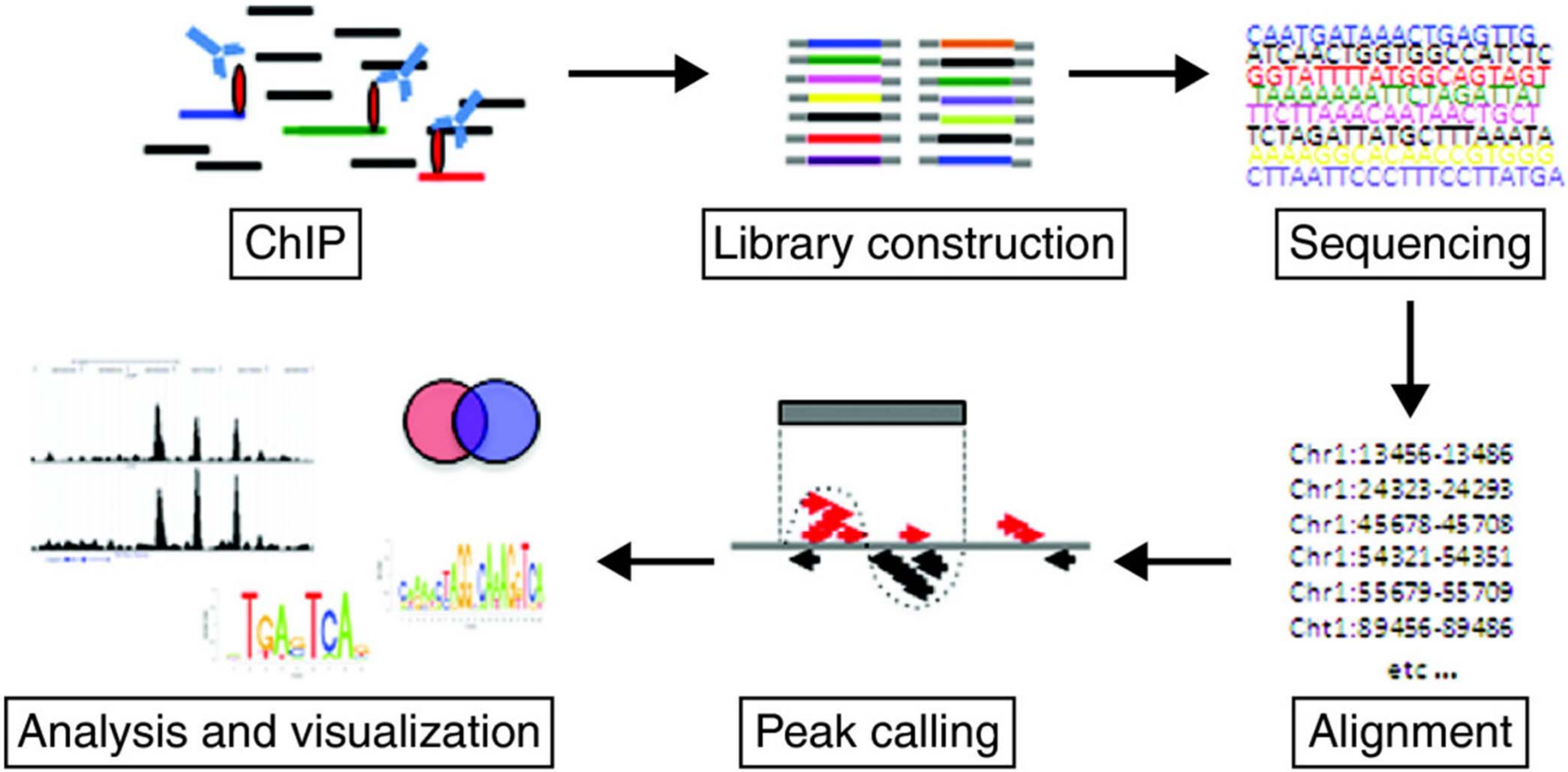
Read out = same as miRNA

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

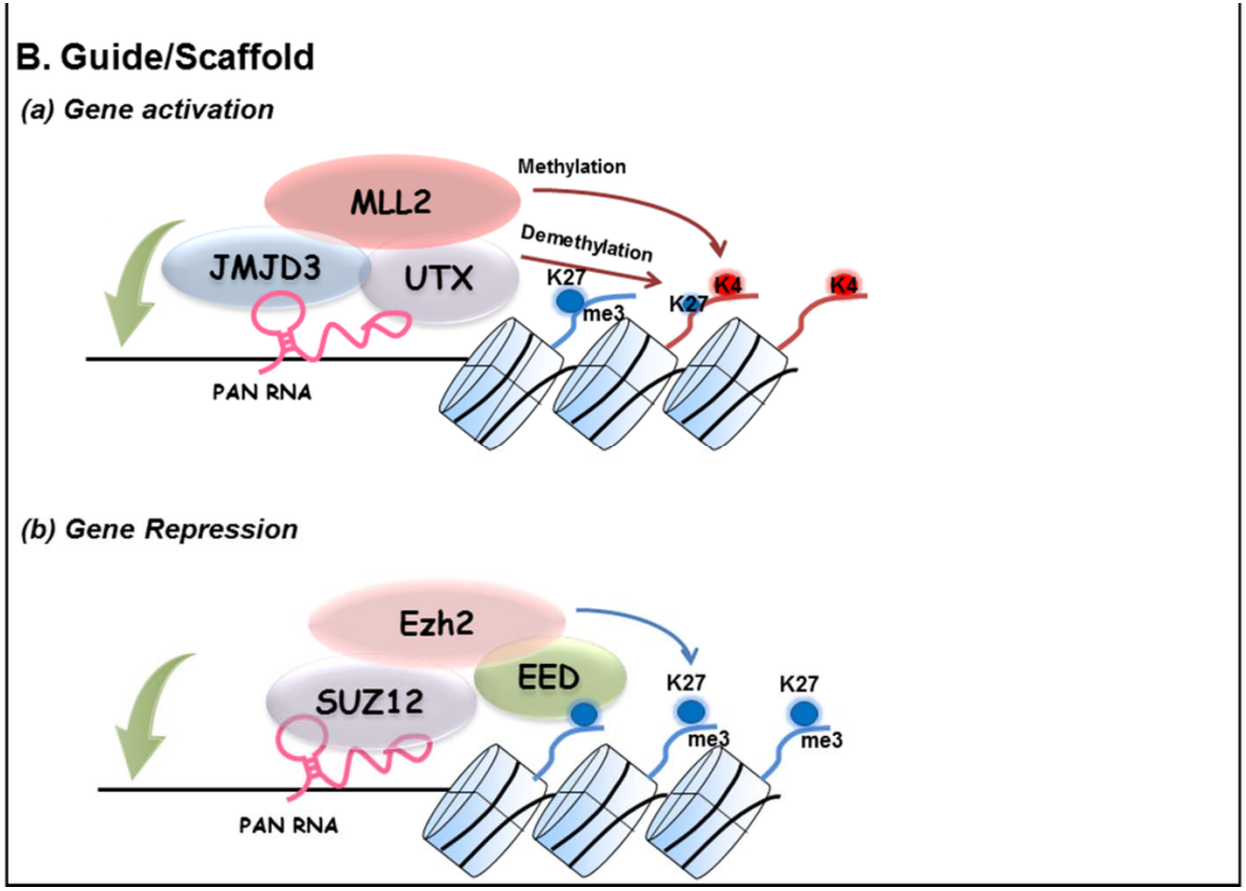
4. Epigenetic sequencing

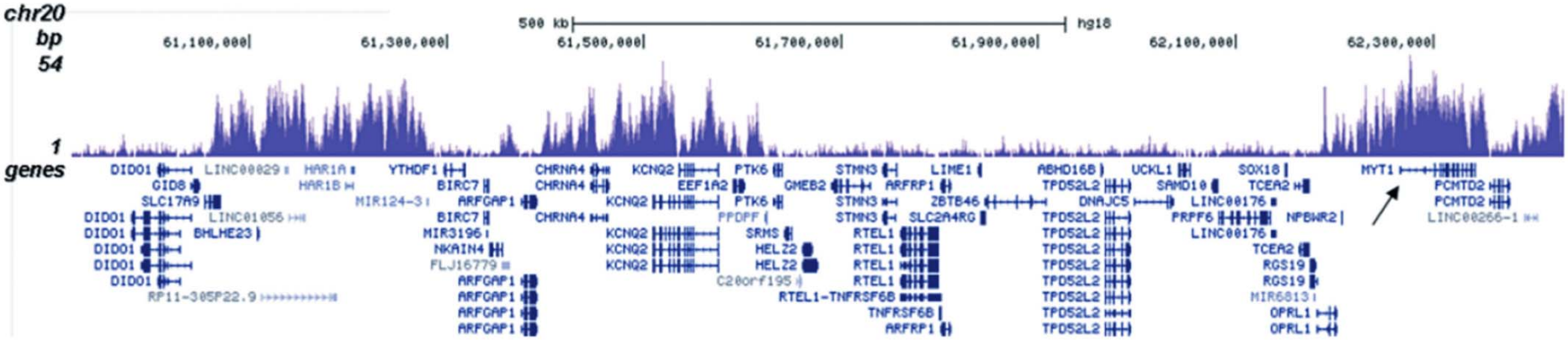
- *ChIP-seq*
- *Histone-IP*

ChIP-seq procedure



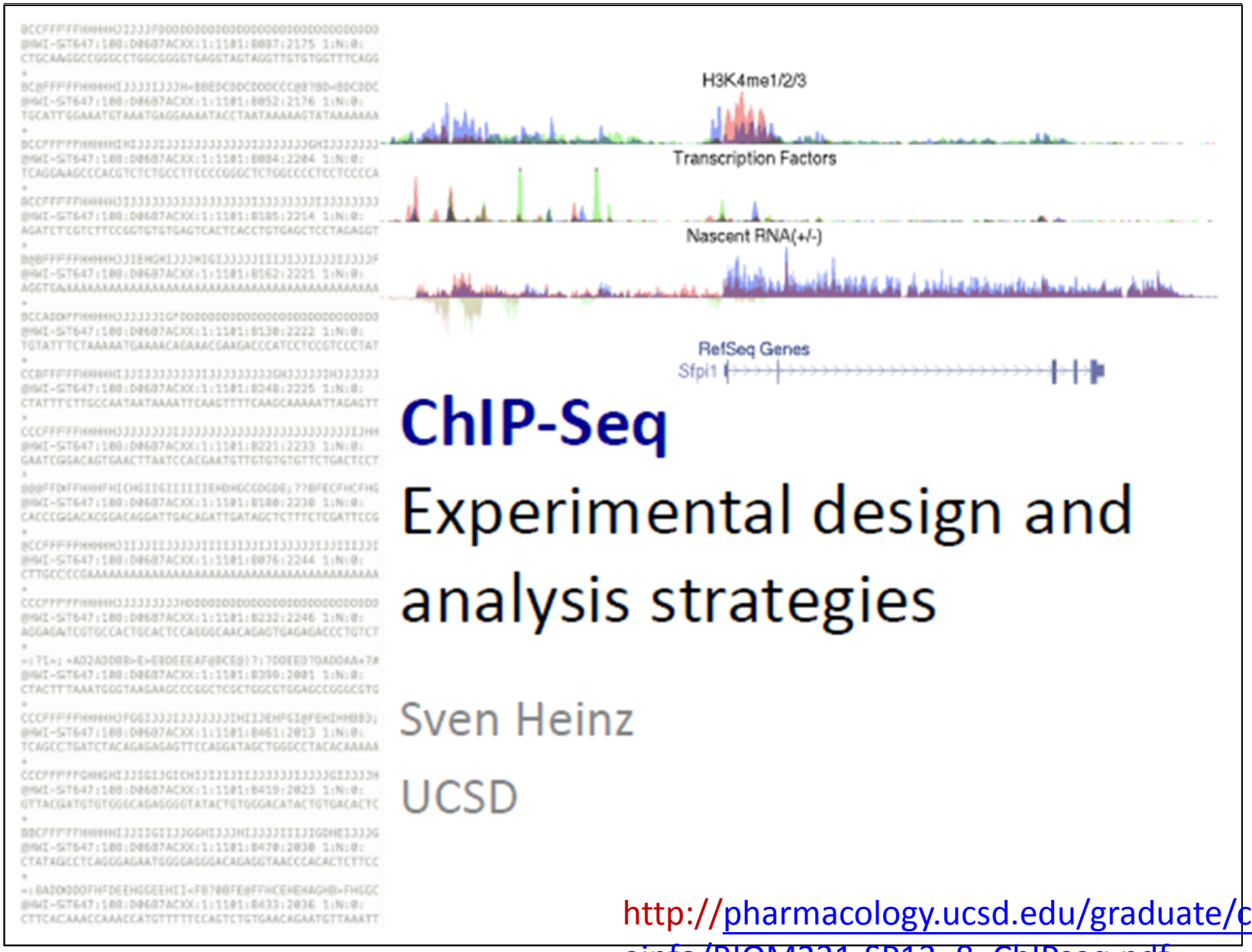
Anti-K27me3





http://www.mdpi.com/1999-4915/6/11/4165/htm

ChIP-seq Experimental Design



http://pharmacology.ucsd.edu/graduate/courseinfo/BIOM231-SP13_8_ChIPseq.pdf

Planning ChIP-Seq

- Sample: ~0.1 - 150 million cells; biological replicates; controls & INPs
- Require: ChIP-grade antibody with high specificity (must be able to recognize its epitope under cross-linked conditions). Compare targets in ChIP-PCR vs WB. Most histone modifications work well, but many are highly correlated…
- Use well-working and -accepted marks: H3K4me1/2/3, H3K27ac/me3, H3K36me3
- Carrier DNA blocking not preferred

5. Amplicon sequencing

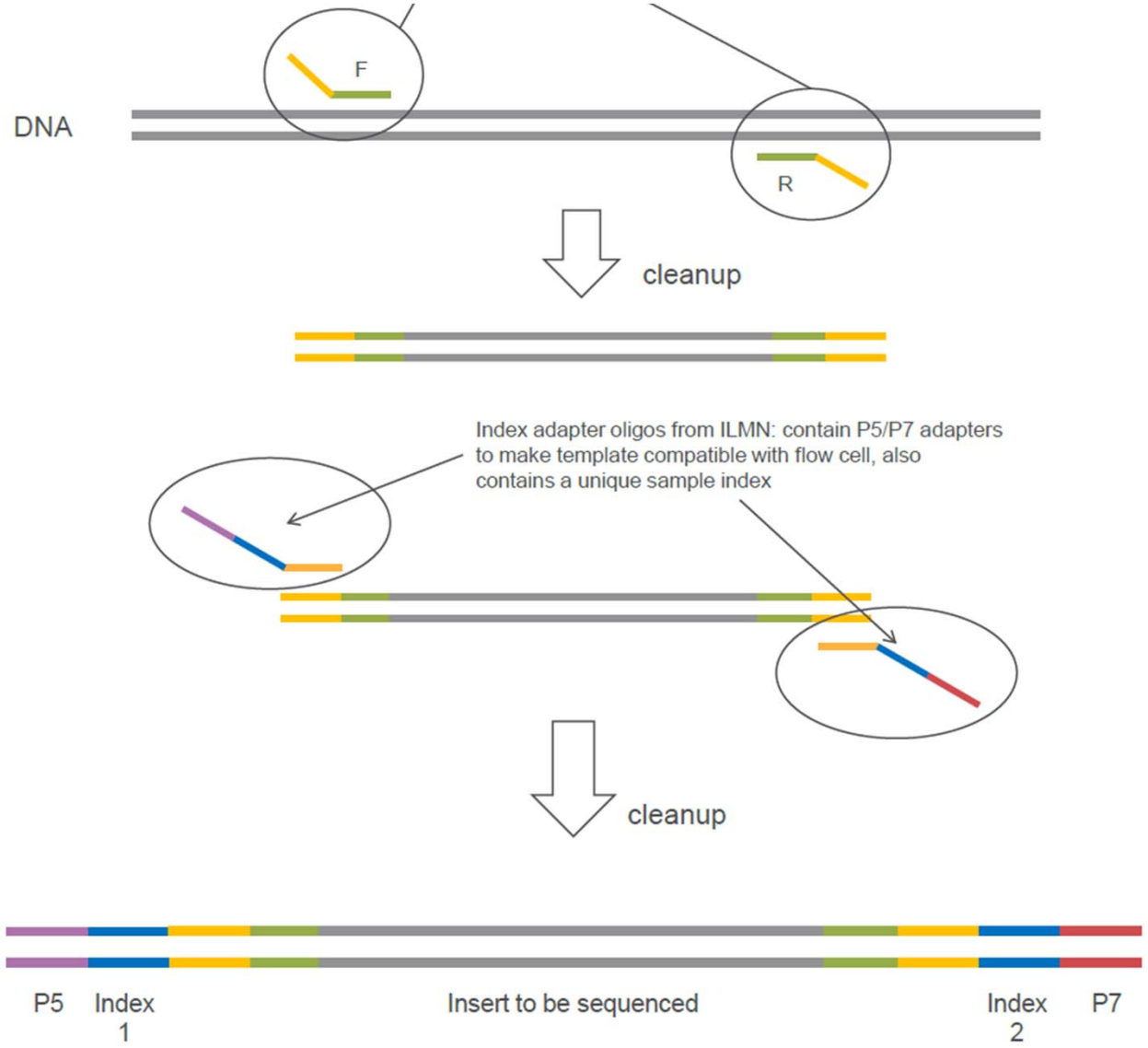
- ***Two-step PCR***
- ***Nextera XT tagmentation***

Amplicon library: 2-step PCR

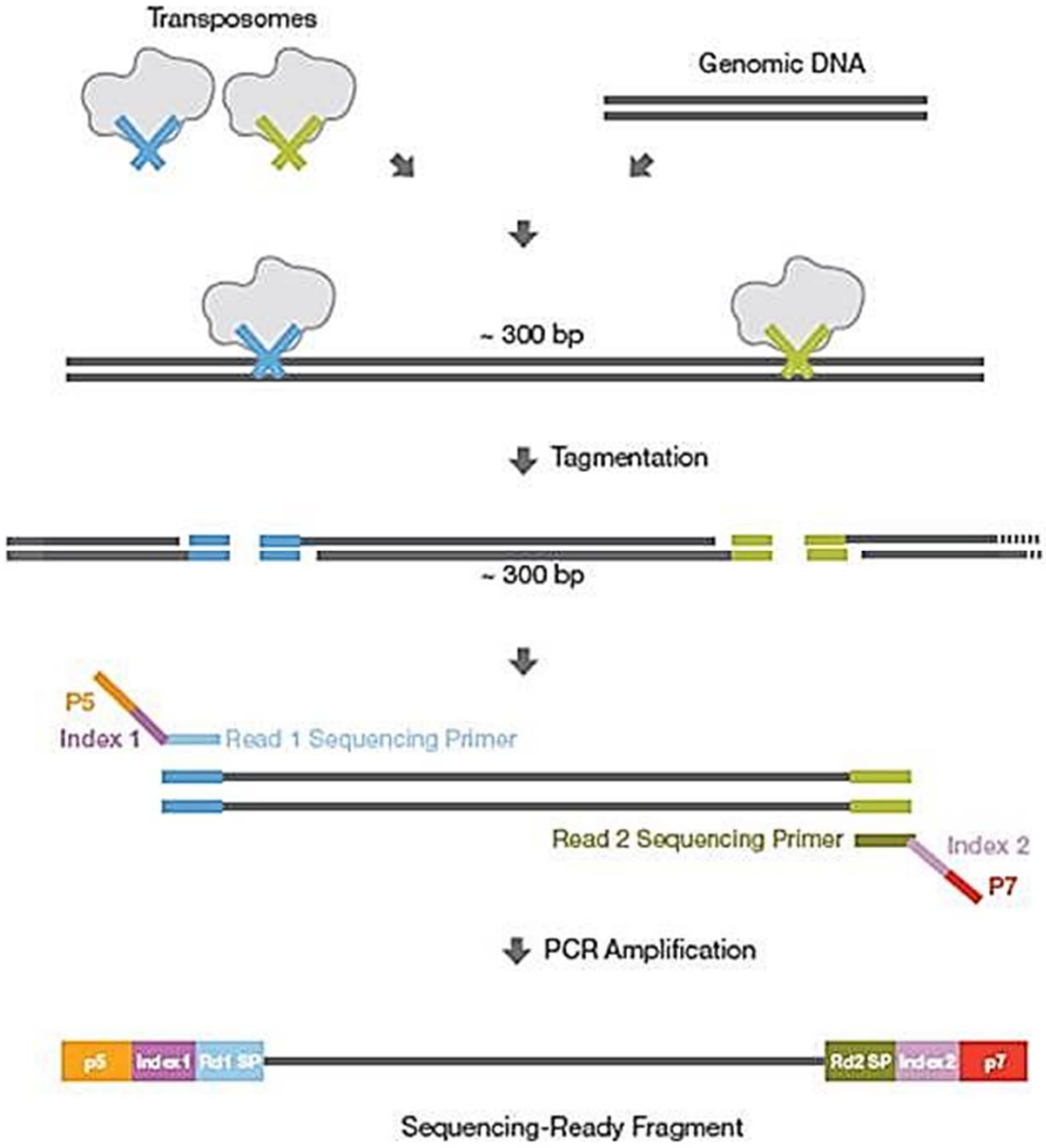
1st round:
Amplify ROI

2nd round:
add indices + adapters

Best: 400-600bp



Nextera Library prep - Tn tagging



1. **TemplateDNA + transposome complex (contain adaptor)**

2. **Tagmentation breaks DNA and add adaptor to ends**

3. **PCR amplification to engineer barcode and sequencing primers**

http://www.gtbiotech.com.tw/products/Nextera_XT_DNA.asp

6. Single cell/ Low input sequencing

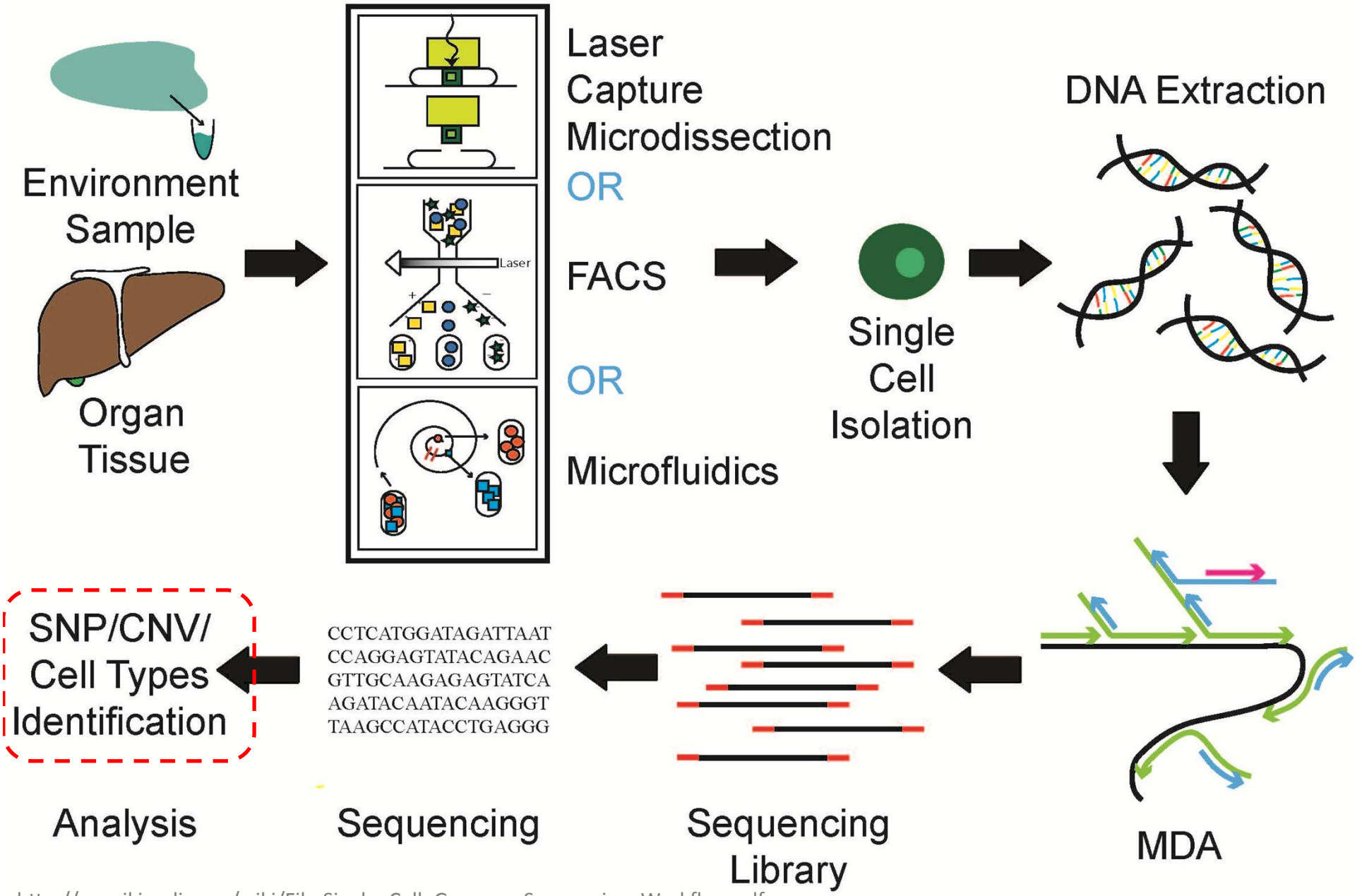
Methods:

- Smarter (Clonetech)
- Transplex (Sigma-Aldrich)
- C1 + Nextera
- ThruPLEX (Rubcon Genomics)

Issues:

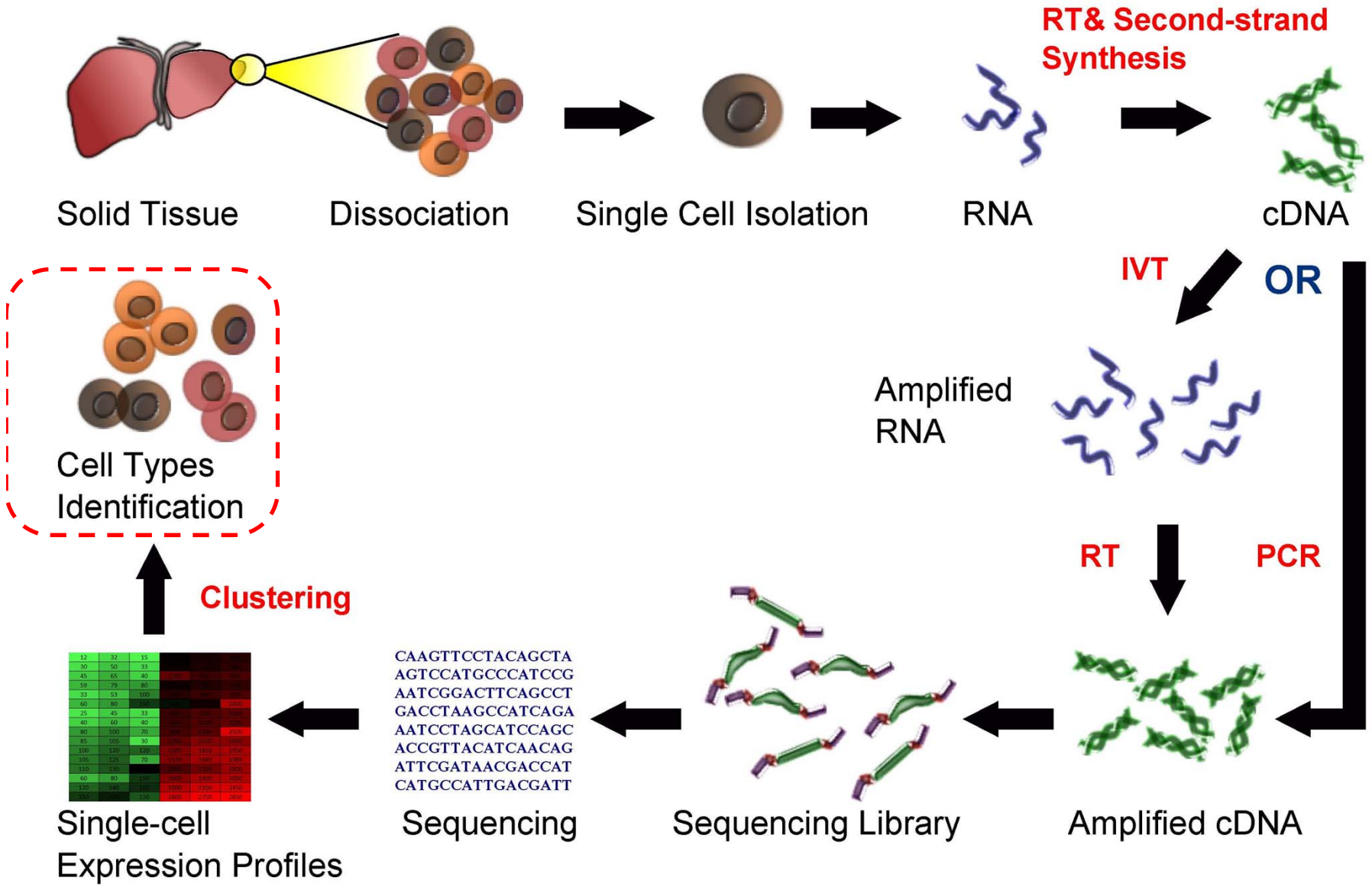
- Contamination
- Reproducibility
- Sensitivity
- Correlation to bulk

Single Cell Genome Sequencing Workflow



http://en.wikipedia.org/wiki/File:Single_Cell_Genome_Sequencing_Workflow.pdf

Single Cell RNA Sequencing Workflow



http://en.wikipedia.org/wiki/File:RNA-Seq_workflow-5.pdf

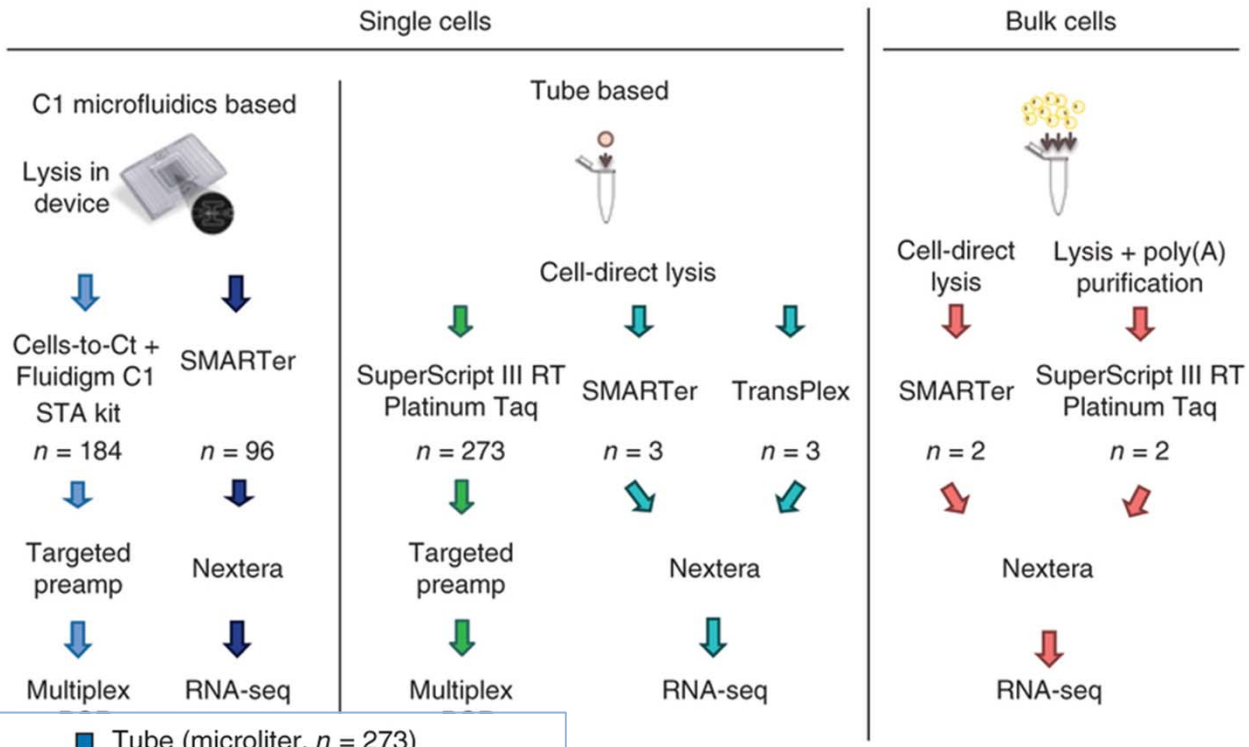
nature methods
Techniques for life scientists and chemists
Home | Current issue | Comment | Research | Arch
home ▸ archive ▸ issue ▸ analysis ▸ full text
Take part in Nature Publishing Group's annual reader
NATURE METHODS | ANALYSIS

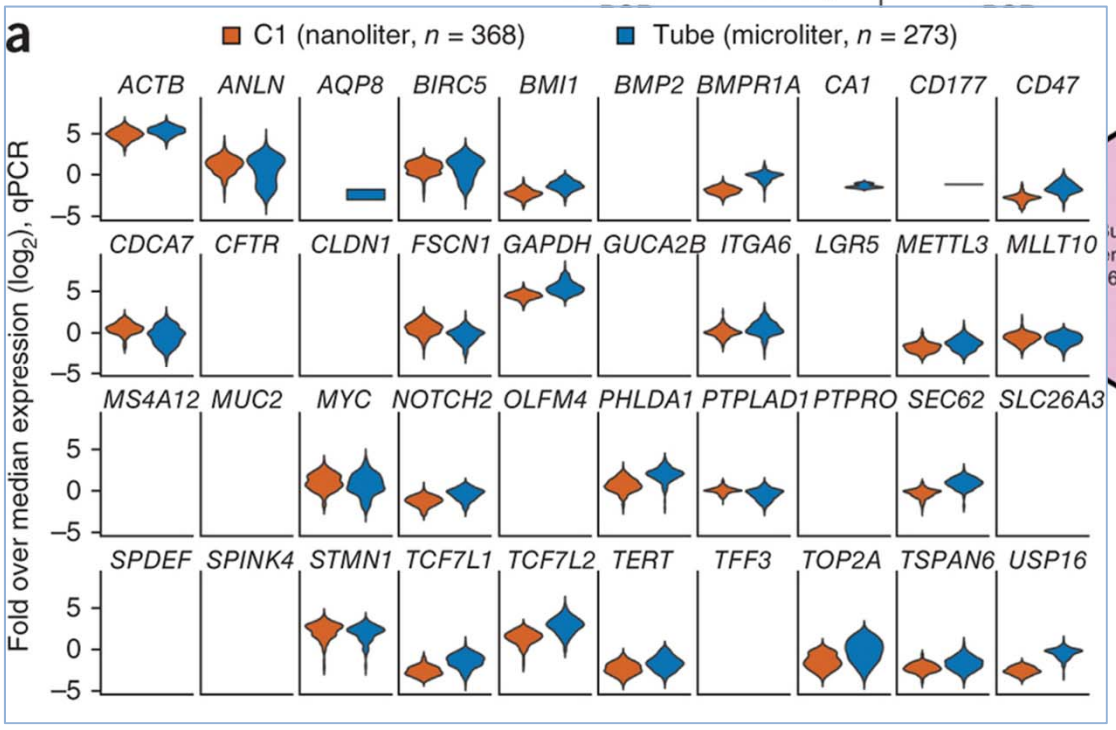
Quantitative assessment c sequencing methods

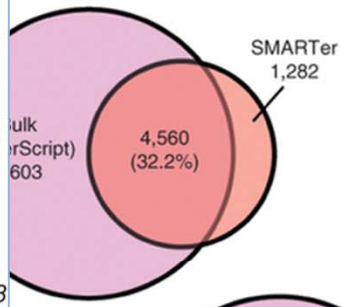
Angela R Wu, Norma F Neff, Tomer Kalisky, Pi Rothenberg, Francis M Mburu, Gary L Mantala R Quake

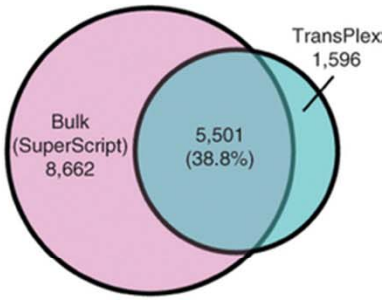
Affiliations | Contributions | Corresponding a

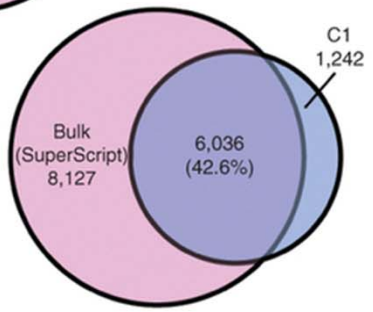
a











Summary of NGS Expt. Design

1. Application type
2. Sample nature
3. Biological replicate (n=3?)
4. Control set?
5. Seq. format and coverage depth
6. barcode multiplexity
7. Sequence bias
8. NGS bioinformatic algorithms

General NGS project design

	gDNA	RNA-seq	miRNA	ChIP	Single cell
Sample purification	genome, amplicon	poly-A, rRNA-depletion	total RNA, spin column	PCI/EtOH ppt, spin column	
Library prep	Sonication, Tn tagging	ds-cDNA, Strand-specific	Strand-specific, PAGE sizing		
Capture panel	Design concept, total length, genes in panel, genome divergence				
Seq. format	SR or PE	SR or PE	SR	SR or PE	
Read length	50~300nt	10~200nt	50nt	100~150nt	
Data scale (per sample)	Assembly: Cov: 50~100X	Assembly: 50~100X	2~20 M	10~30M reads	
	Re-seq Cov: >30X	Re-seq: >30X			
Issues	genome size, complexity, Repeats, GC%	Splicing, fusion? # bio. replicates, time course, mostly RNA>120nt	only mature form	Ab (sen. & spec.), Control, replicates(?)	Amp. bias, Contamination, Cell lysis effic., n size =?

§ Omics combination

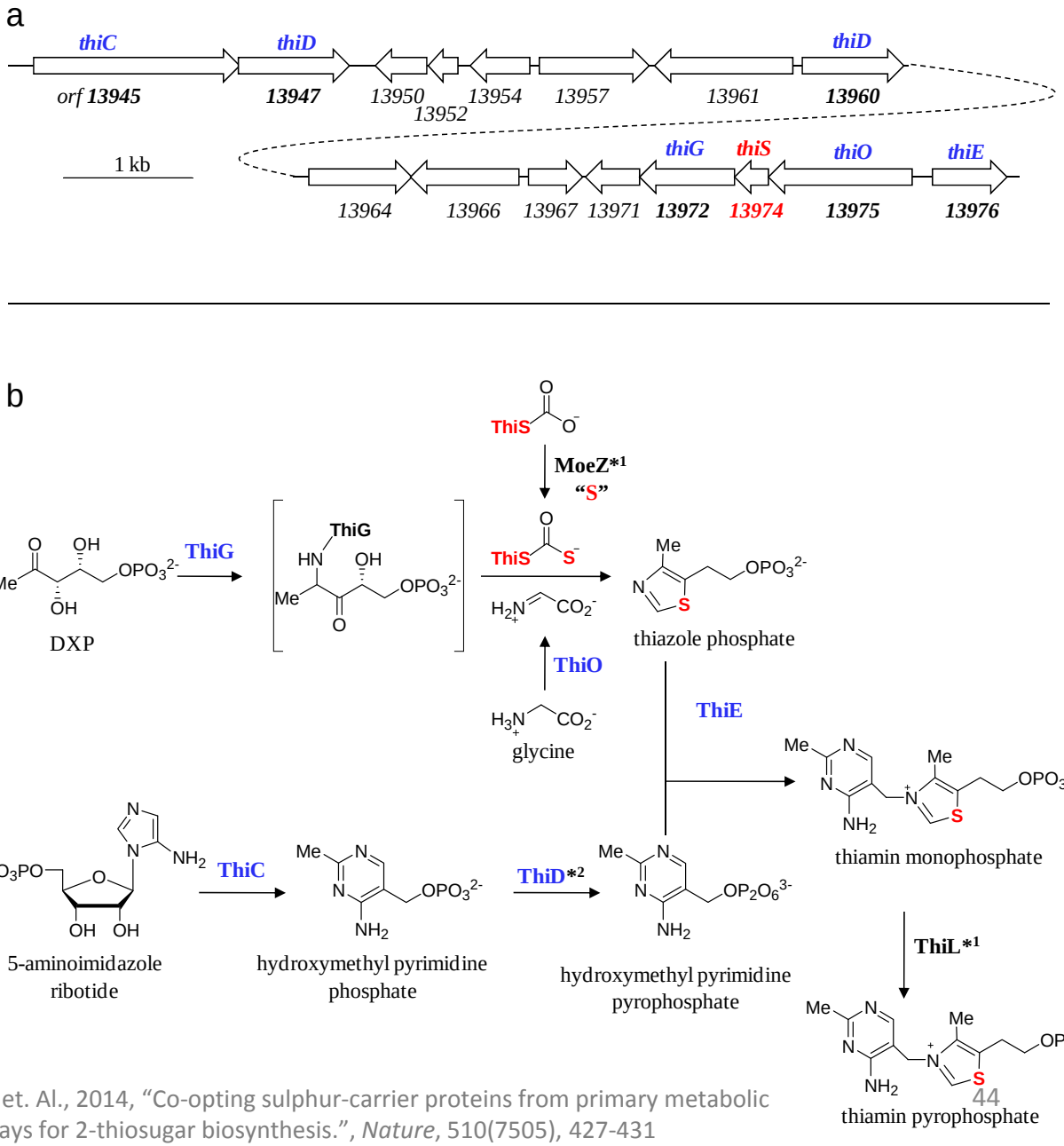
- *examples*

Identify operons & gene clusters in *Amycolatopsis orientalis*

Genome + Metabolome

AOV de novo	454 only
Total# Reads	992,862
Tatoal # fBases	423,153,549
# Contigs	**119**
Assembled Bases	**9,786,980**
N50ContigSize	185,605
avgContigSize	82,243
largestSize	663,909
a/fpipeline process	**454+GA**
numberOfContigs	**106**
numberOfBases	**9,787,625**
N50ContigSize	223,668
avgContigSize	92,336
largestSize	663,912

(avg. 77% GC)

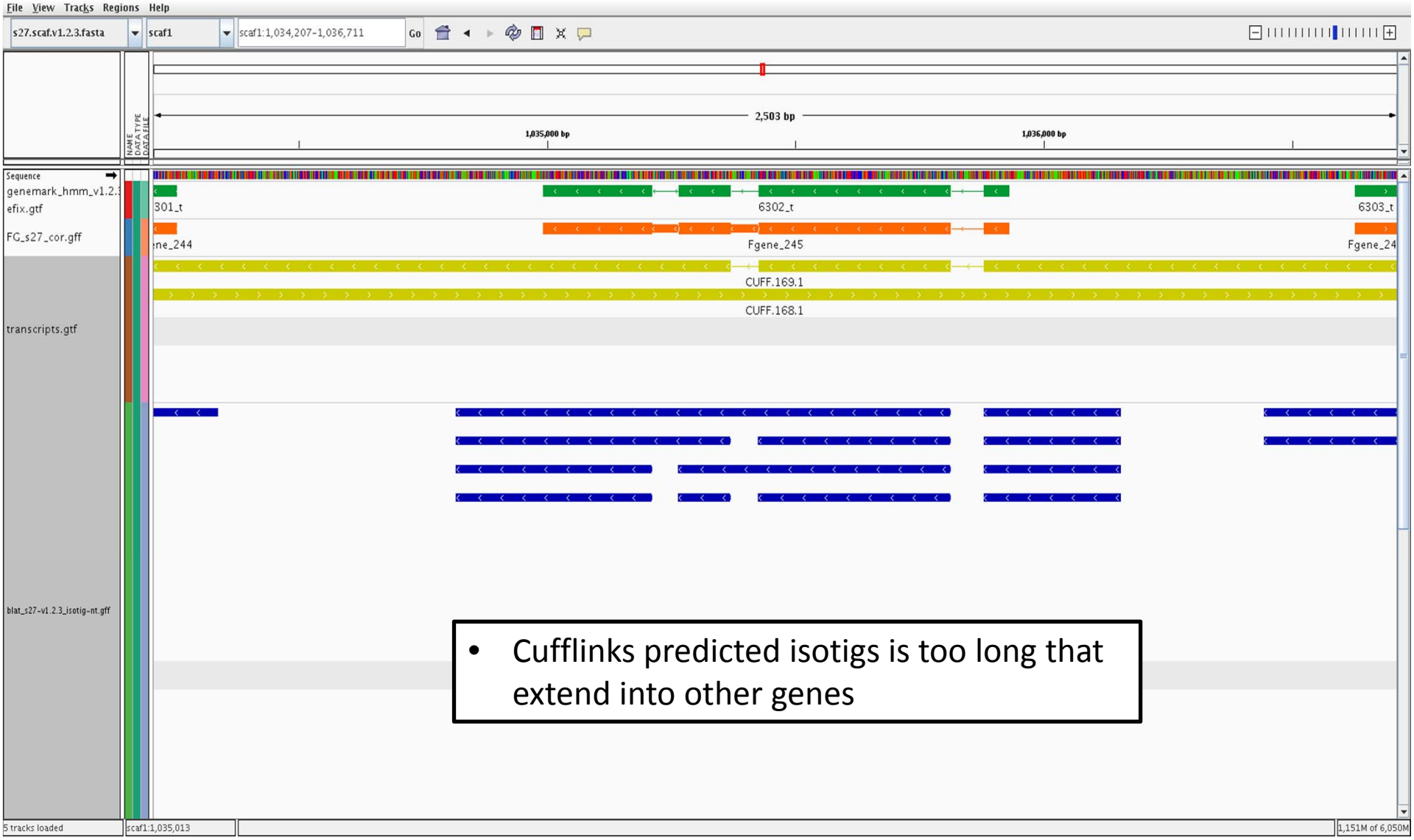


Sasaki et. Al., 2014, "Co-opting sulphur-carrier proteins from primary metabolic pathways for 2-thiosugar biosynthesis.", *Nature*, 510(7505), 427-431

Genome + Transcriptome

Gene model curation

Align transcripts to genome: help select ORF



- Cufflinks predicted isotigs is too long that extend into other genes

By: Ting-Chun Chen

Transcriptome + proteome

Omics combination

de novo transcriptome assembly and gene hunting for a cellulolytic fungus.

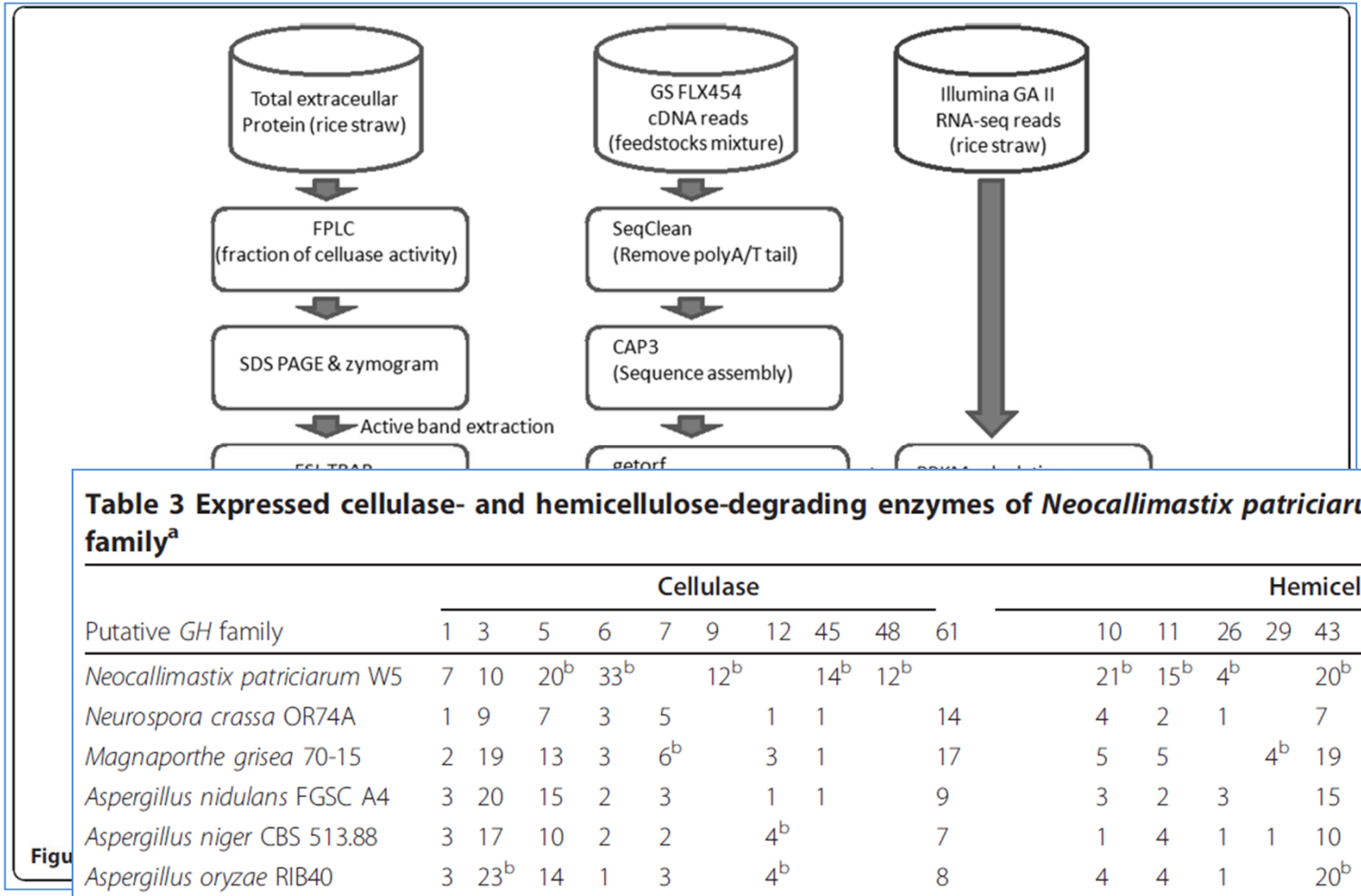


Table 3 Expressed cellulase- and hemicellulose-degrading enzymes of *Neocallimastix patriciarum* arranged by *GH* family[a]

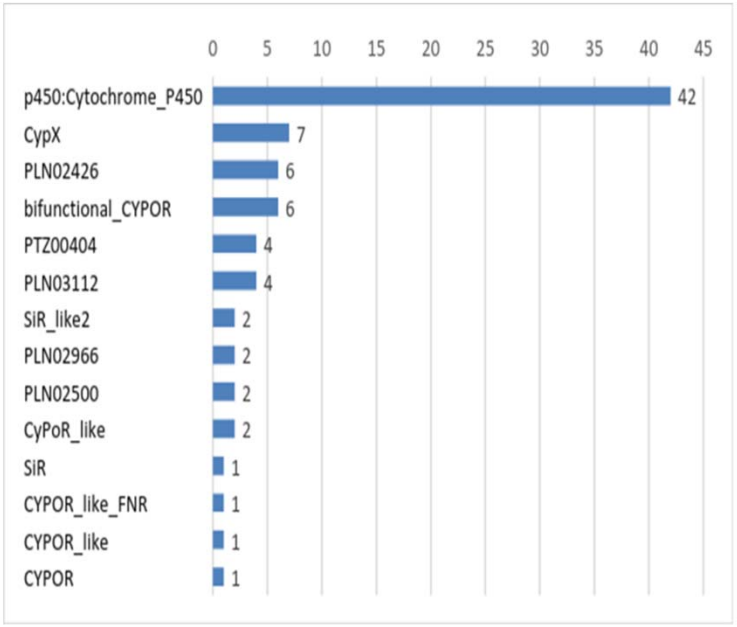
	Cellulase										Hemicellulase											
Putative *GH* family	1	3	5	6	7	9	12	45	48	61	10	11	26	29	43	51	53	54	62	67	74	95
Neocallimastix patriciarum W5	7	10	20[b]	33[b]		12[b]		14[b]	12[b]		21[b]	15[b]	4[b]		20[b]		1					
Neurospora crassa OR74A	1	9	7	3	5		1	1		14	4	2	1		7	1	1	1[b]		1[b]	1	
Magnaporthe grisea 70-15	2	19	13	3	6[b]		3	1		17	5	5		4[b]	19	3	1	1[b]	3[b]	1[b]	1	1
Aspergillus nidulans FGSC A4	3	20	15	2	3		1	1		9	3	2	3		15	2	1	1[b]	2	1[b]	2[b]	3[b]
Aspergillus niger CBS 513.88	3	17	10	2	2		4[b]			7	1	4	1	1	10	4[b]	2[b]	1[b]	1	1[b]	1	2
Aspergillus oryzae RIB40	3	23[b]	14	1	3		4[b]			8	4	4	1		20[b]	3	1	1[b]	2	1[b]		3[b]
Leptosphaeria maculans v23.1.3	3	13	15	3	3		3	2		20[b]	3	2	1	1	11	3	1		1	1[b]		1
Trichoderma reesei[c]	2	13	11	1	2		2	1		2	1	3			2				1	1[b]		

[a]*GH*: glycosyl hydrolase; CAZy: Carbohydrate-Active enZYmes database. [b]Highest gene number among fungi from JGI database and CAZy annotation. [c]Adapted from JGI database.

Wang et al. Biotechnology for Biofuels 2011, 4:24

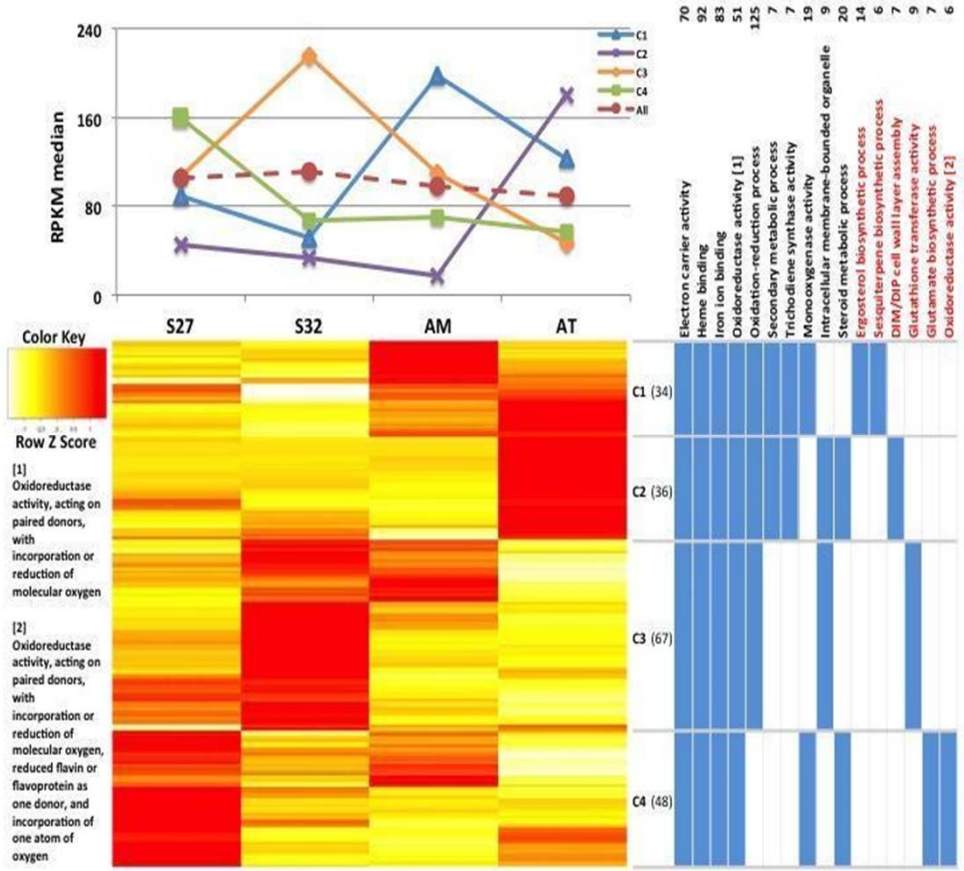
Genome + Transcriptome

Gene ontology

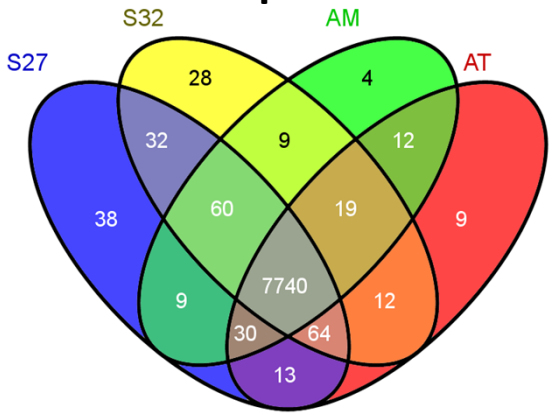


Distribution of P450 genes in different cytochrome P450 domain categories.

Clustering of DEGs

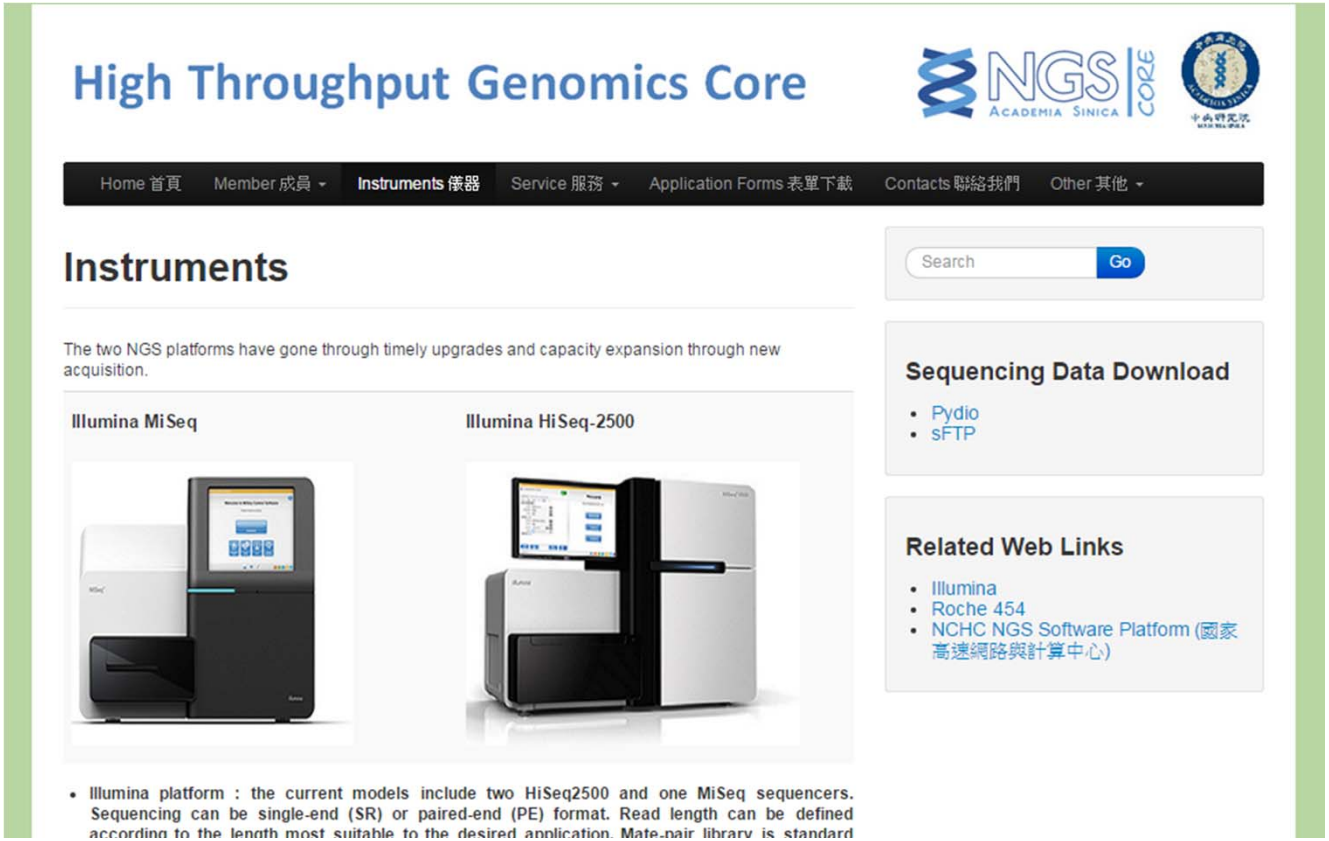


Tissue-specific DEGs



Lu et al., 2014. PNAS.


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


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
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)
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

Thank you!