Next-generation sequencing (NGS) for plant research

Presented by Daisuke Tsugama

Email: tsugama@res.agr.hokudai.ac.jp
Tel: 011-706-2471
Room: S268 (Lab of Crop Physiology)

Slides used for this class can be downloaded at
http://www.agr.hokudai.ac.jp/botagr/sakusei/materials.html

This class …

- Introduces theories and applications of NGS, which is now very popular in plant research, from an experimental biologist's viewpoint
- Aims at letting you know
  - ✓ what is NGS
  - ✓ what is usually done in NGS data analysis
  - ✓ applications of NGS
  - ✓ NGS is not something to fear
- Assesses you on the basis of a small test attached to the end of the handout

Outline

1. What is NGS like?
   - Sequencers for NGS
   - Basics of NGS data analysis
2. Applications of NGS
   - RNA-Seq
   - Genome sequencing
   - RAD-Seq
   - MutMap and QTL-Seq
   - Others

Outline

1. What is NGS like?
   - Sequencers for NGS
   - Basics of NGS data analysis
2. Applications of NGS
   - RNA-Seq
   - Genome sequencing
   - RAD-Seq
   - MutMap and QTL-Seq
   - Others

Sequencers for NGS

Sequencer	Company	Output	Read length
GS-FLX	454 Life Sciences (Roche)	~400 Mb	~500 b
Ion Proton	Life Technologies (Thermo)	~10 Gb	~200 b
HiSeq 2500	Illumina	~1 Tb	~200 b
PacBio RS II	Pacific Biosciences	~1 Gb	~40 kb

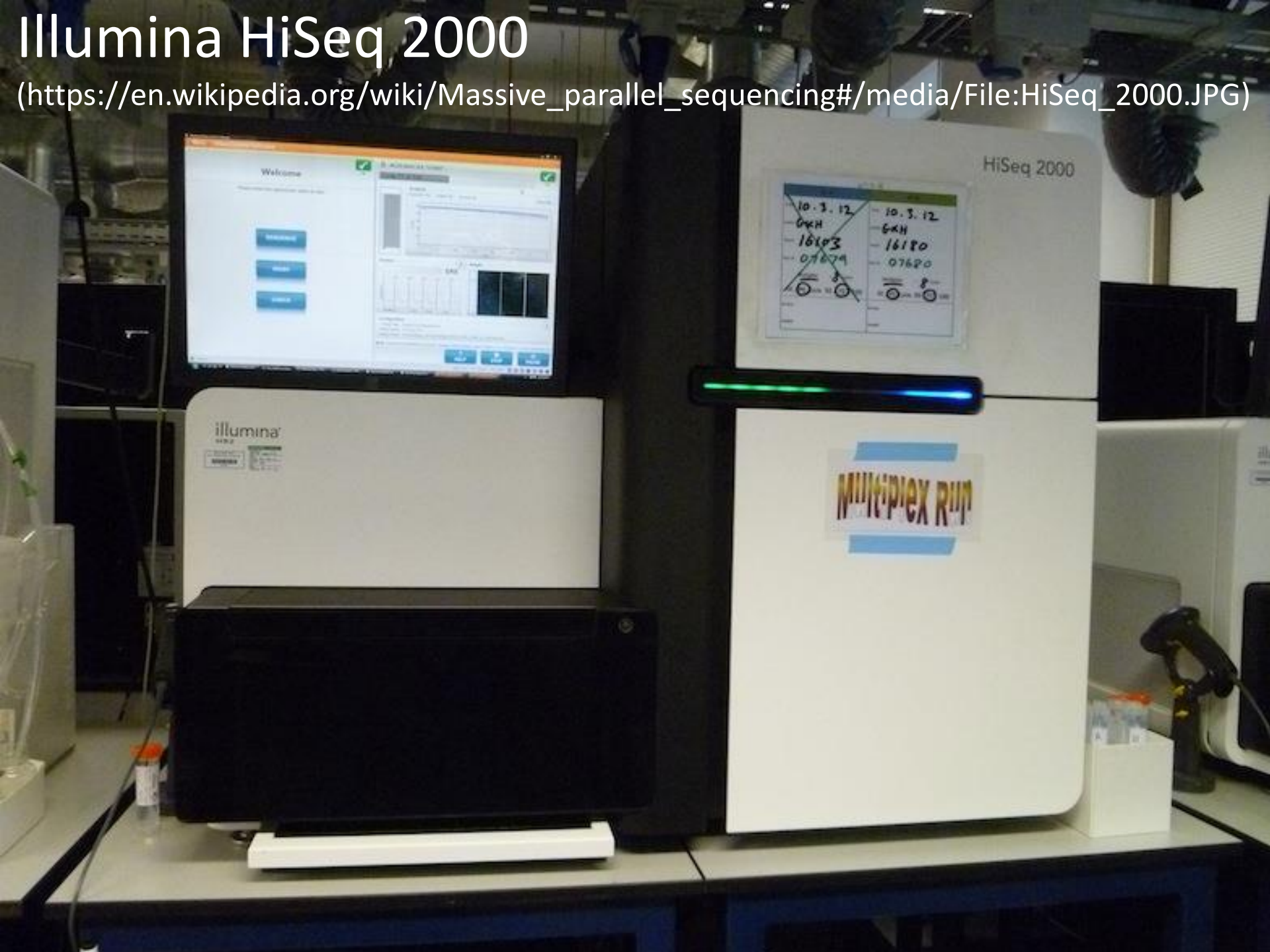
* Output (b / run) = read length (b/read) × # of reads

Ion Proton semiconductor

(https://en.wikipedia.org/wiki/Ion_semiconductor_sequencing#/media/File:Life_Technologies_-_Ion_Proton_(TM).jpg)

Illumina HiSeq 2000

(https://en.wikipedia.org/wiki/Massive_parallel_sequencing#/media/File:HiSeq_2000.JPG)



PacBio RS II

(https://en.wikipedia.org/wiki/Pacific_Biosciences#/media/File:PacBio_RSII.jpg)

Sequencers for NGS

Sequencer	Company	Output	Read length
GS-FLX	454 Life Sciences (Roche)	~400 Mb	~500 b
Ion Proton	Life Technologies (Thermo)	~10 Gb	~200 b
HiSeq 2500	**Illumina**	**~1 Tb**	**~200 b**
PacBio RS II	**Pacific Biosciences**	**~1 Gb**	**~40 kb**

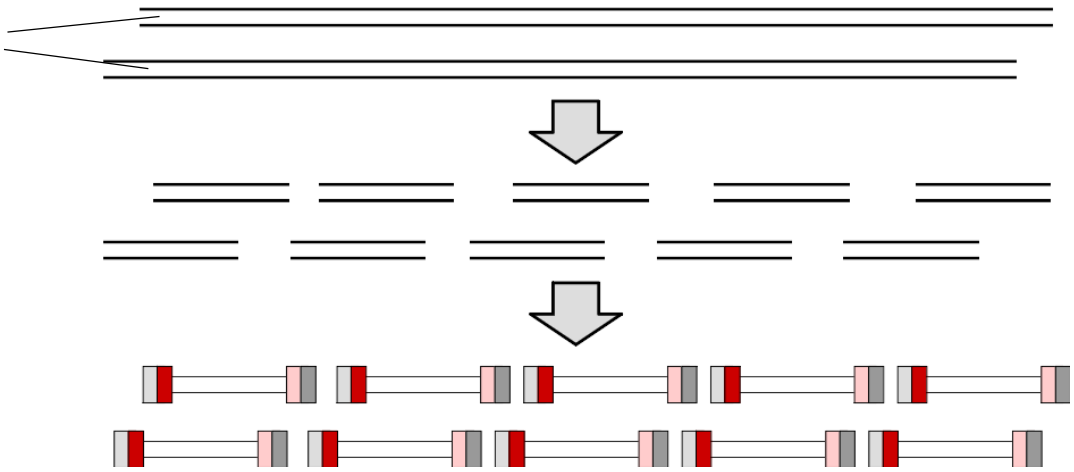
HiSeq and PacBio have been gaining popularity

Illumina NGS technology

DNA to be sequenced

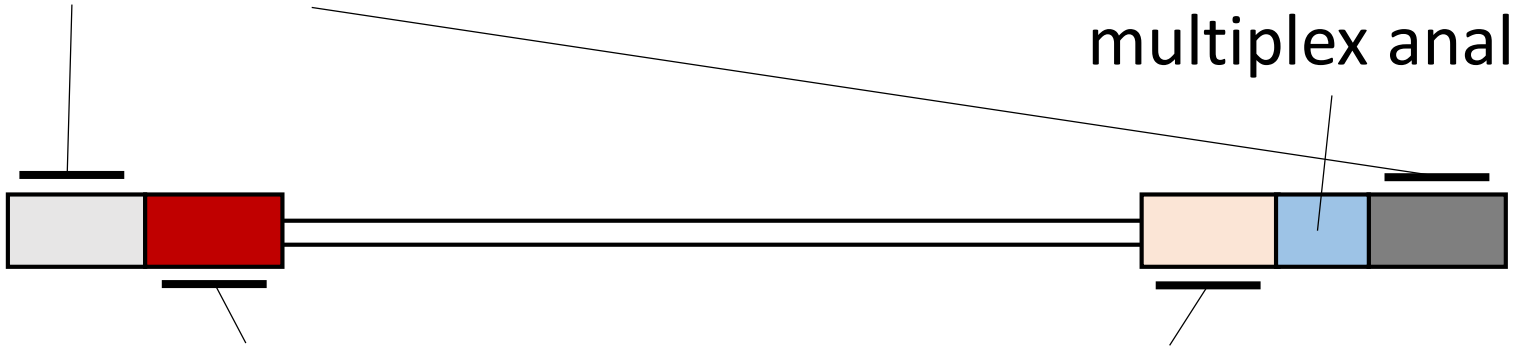
DNA fragmentation

Addition of adapters



Annealing sites for the bridge PCR

Index (barcode) for multiplex analysis



Annealing sites for sequencing primers

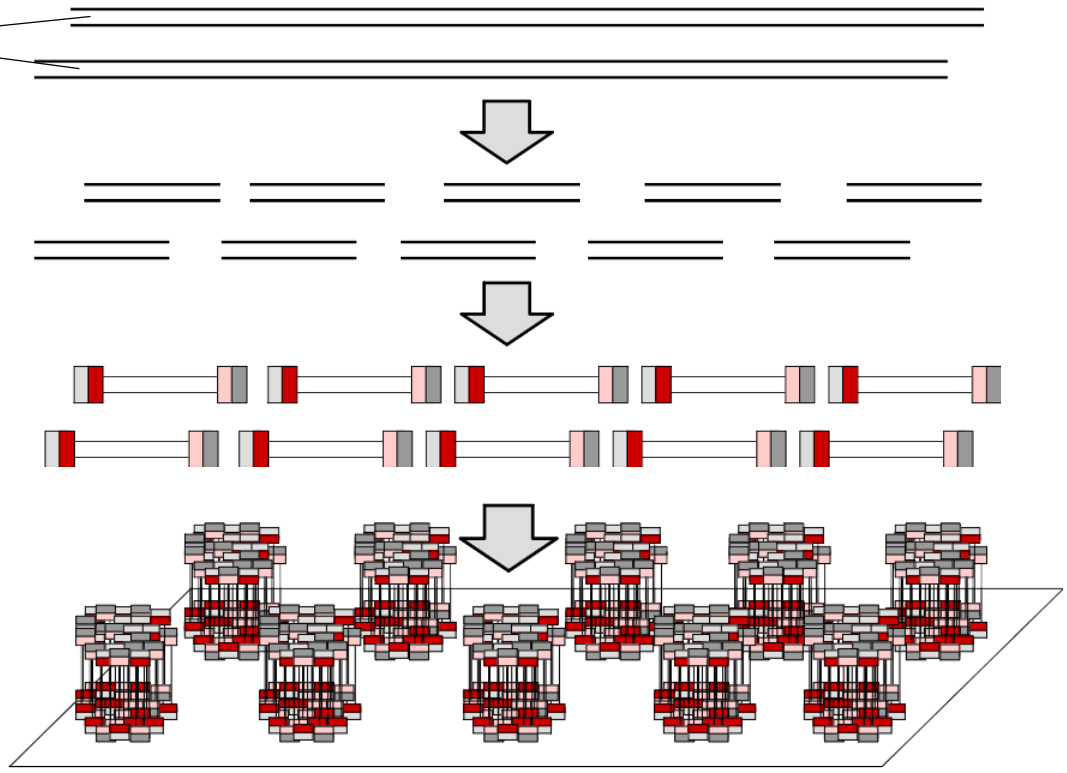
Illumina NGS technology

DNA to be sequenced

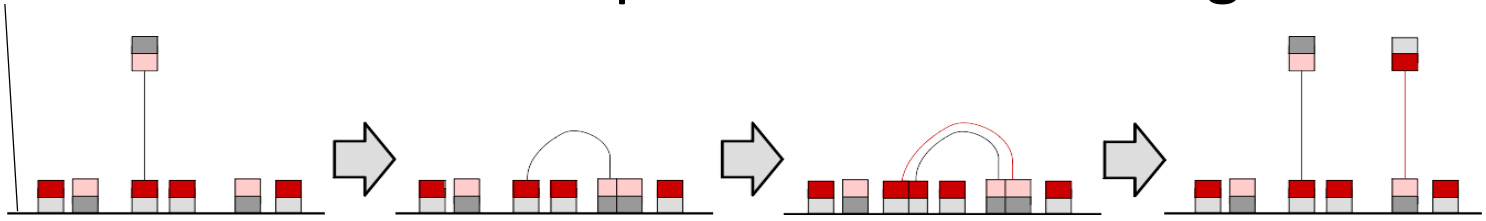
DNA fragmentation

Addition of adapters

Bridge PCR &
Cluster formation



Glass flow cell covered with primers for the bridge PCR



Illumina NGS technology

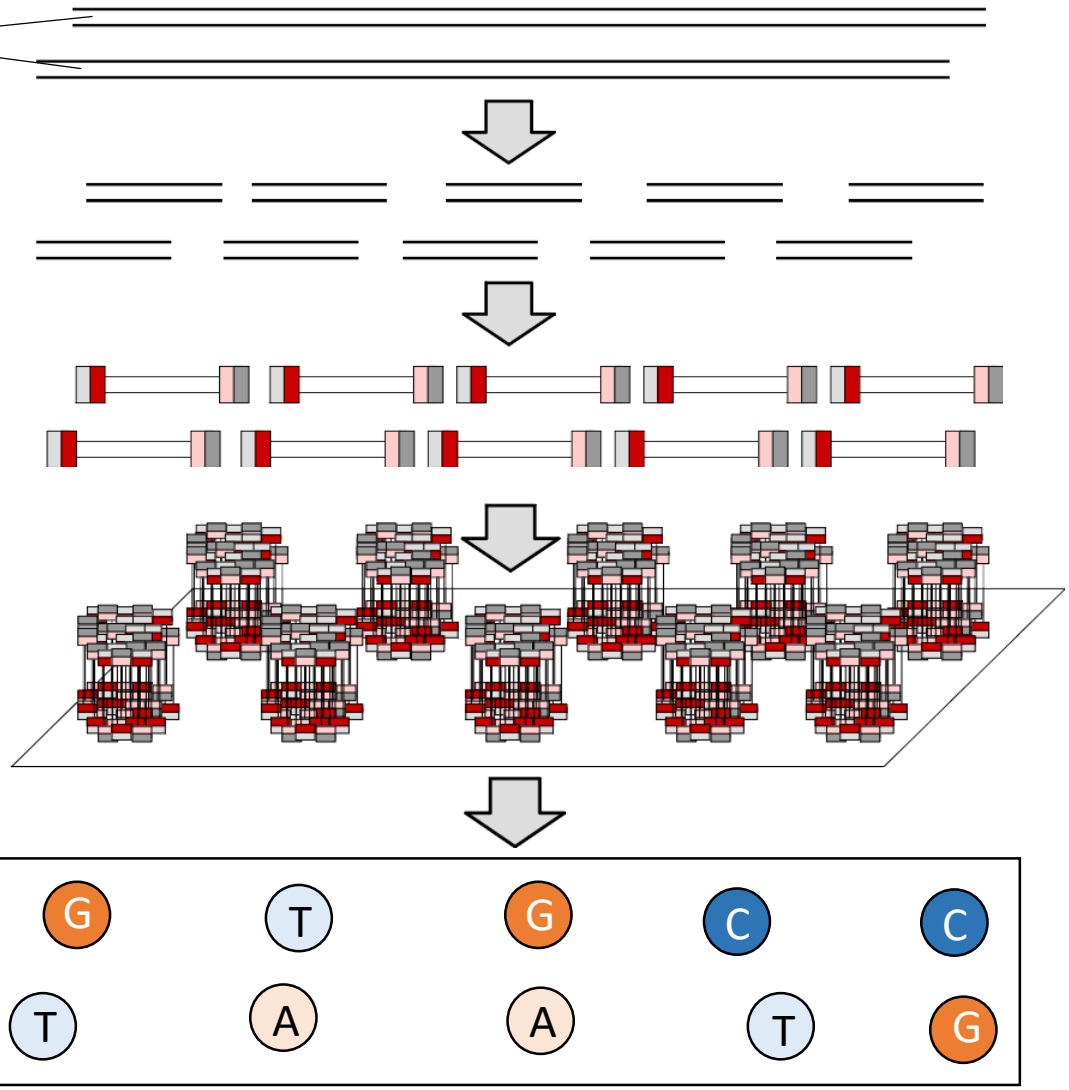
DNA to be sequenced

DNA fragmentation

Addition of adapters

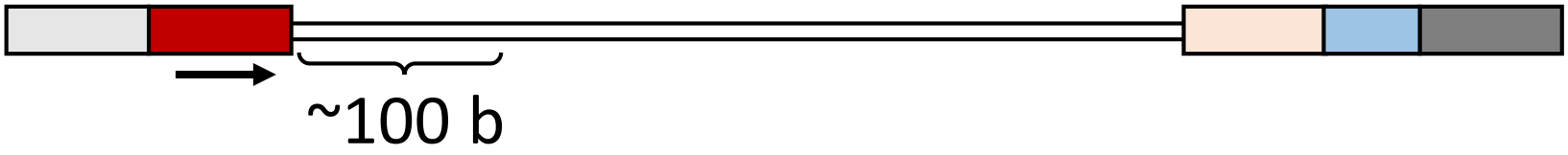
Bridge PCR &
Cluster formation

Signal detection
(~100 times)

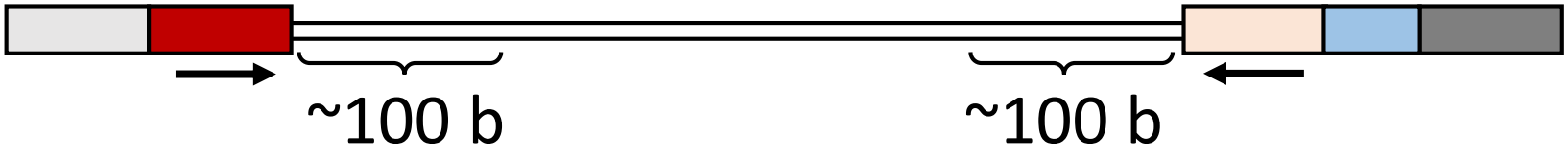


Illumina NGS technology

- Single-end read: obtained by only one primer



- Paired-end read: obtained by two primers

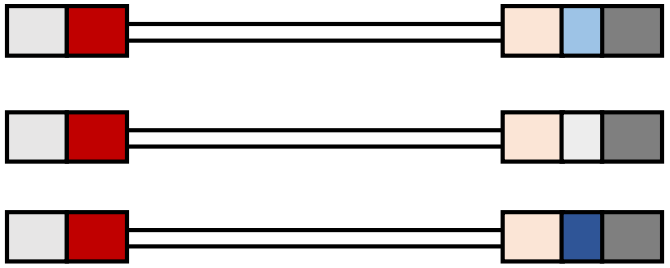


- Multiplex analysis: uses more than two indexes

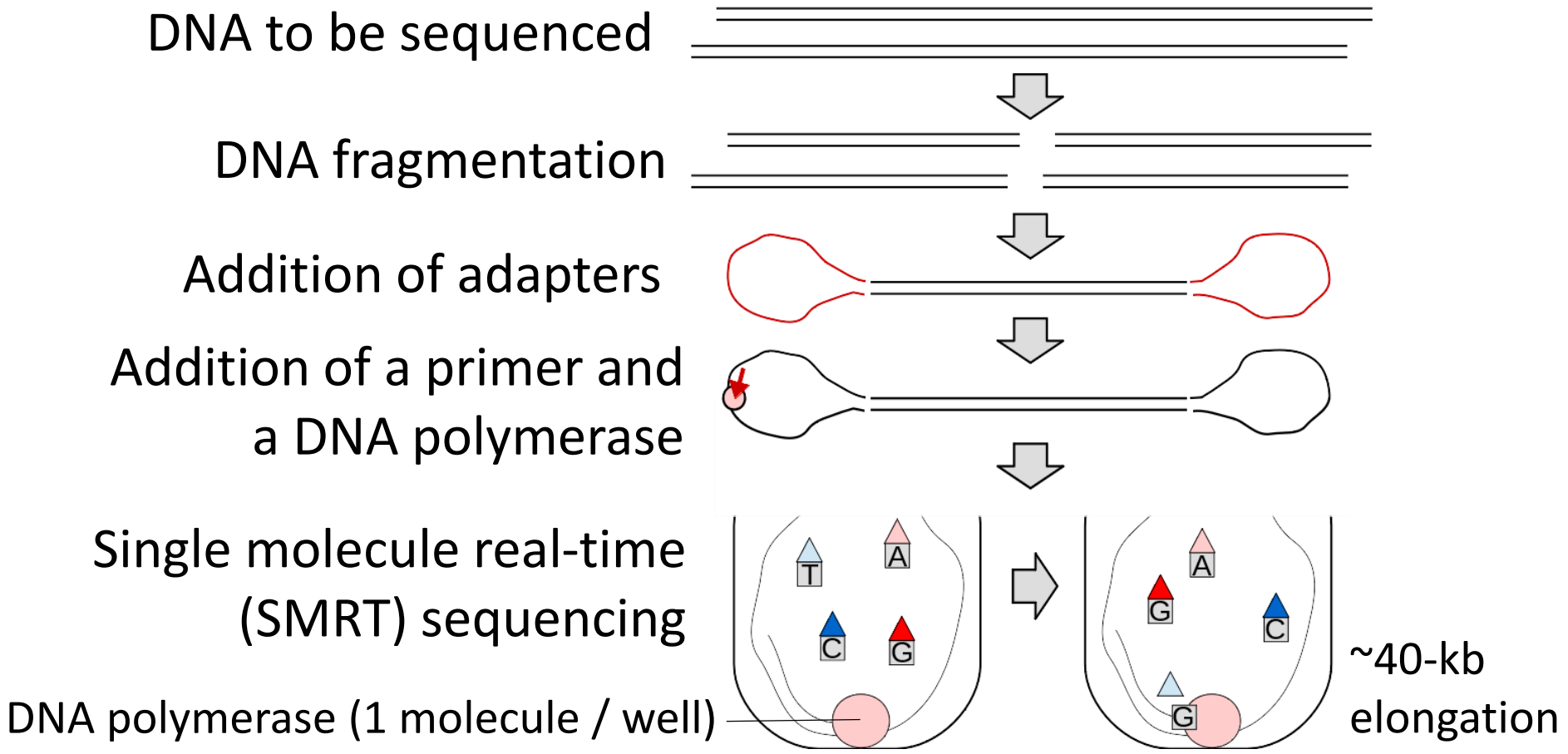
Sample A-derived read

Sample B-derived read

Sample C-derived read



PacBio NGS technology



The detector detects only fluorescent signals retained longer than 1 msec on the bottom (around the DNA pol) of the well

NGS data analysis

Run NGS to get reads

Assemble reads into contigs

Map reads to a reference

*Reference: a genome, transcripts, obtained contigs etc.

Evaluate mapping results

for further analyses

NGS data analysis – read data

Read data are often handled in the fastq format

```
@MachineX:1:1:1:1#0/1
TNAGCTTTACGTATAGGCCCCCGAT
+
#!1508<iO{TRkoI&389M|aR~y
```

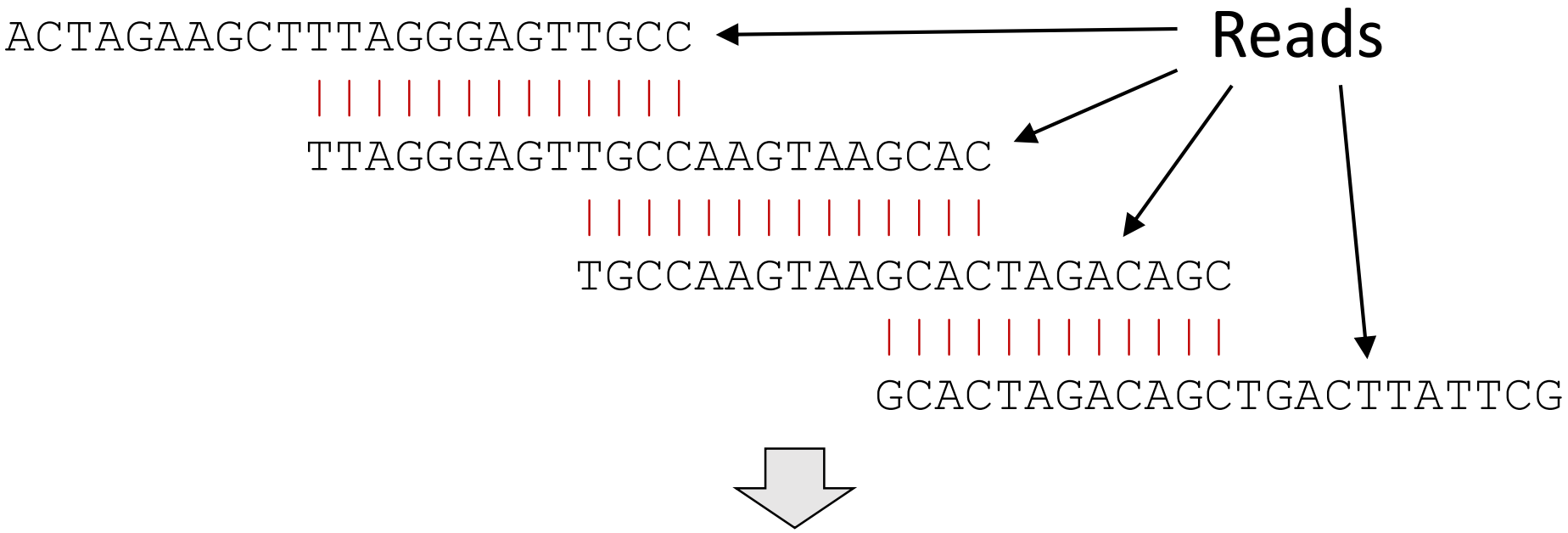
Information for the read "MachineX:1:1:1:1#0/1"

```
@MachineX:1:1:1:2#0/1
ATTGCGTTGTAAGTTGGGGCCTCTC
+
…
```

Information for the read "MachineX:1:1:1:2#0/1"

(usually a great number of reads follow)

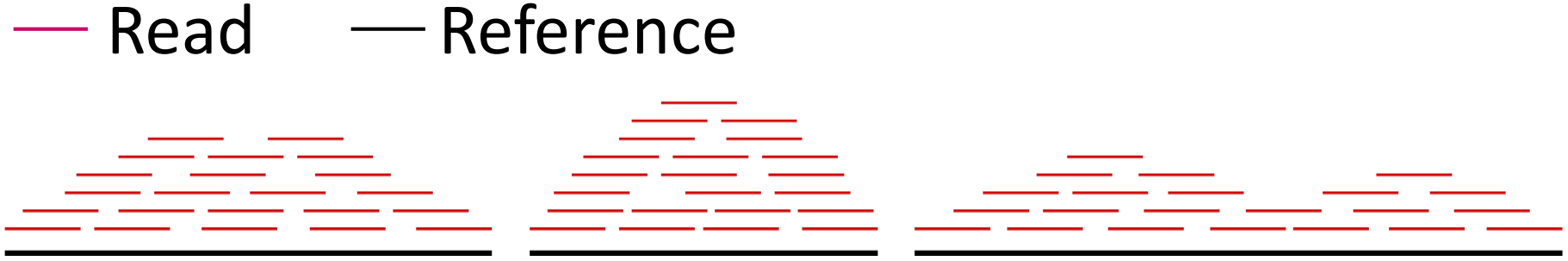
NGS data analysis – assembly



An assembly requires a lot of memory (e.g., *de novo* assembly for an ~3 Gb genome requires ~150 GB memory)

NGS data analysis – mapping

Mapping: associating each read with a reference

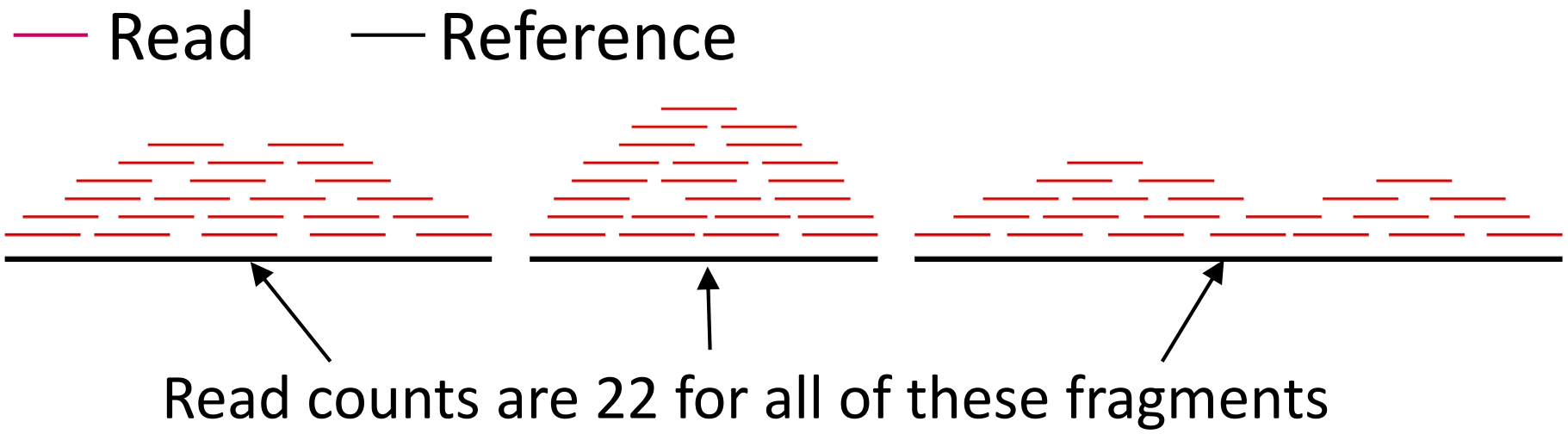


Reference:

- Known genome
- Known transcripts
- Contigs obtained by *de novo* assembly

NGS data analysis – mapping

Mapping: associating each read with a reference



Read counts for each region (or fragment) of the reference are often used to interpret the data

Outline

1. What is NGS like?
   - Sequencers for NGS
   - Basics of NGS data analysis
2. Applications of NGS
   - RNA-Seq
   - Genome sequencing
   - RAD-Seq
   - MutMap and QTL-Seq
   - Others

RNA-Seq

- Is a transcriptome analysis using NGS
- Flow:

  RNA extraction → mRNA purification →

  mRNA shearing → cDNA synthesis → NGS
- Each contig derived from a *de novo* assembly corresponds to each kind of transcripts
- Expression levels of the transcripts are evaluated with FPKM, RPKM or TPM
- They are usually used for further analyses such as clustering and a GO analysis

RNA-Seq

- FPKM:
  fragments per kb of exon per million mapped fragments
- RPKM:
  reads per kb of exon per million mapped fragments

  *FPKM = RPKM when reads are all single-end

$$\text{FPKM of the contig A} = \frac{R_A \times 10^9}{N \times L_A}$$

R_A = # of reads mapped to A

N = total # of mapped reads

L_A = size of A

RNA-Seq

- TPM: transcripts per million

$$\text{TPM of the contig A} = \left(\frac{R_A}{L_A}\right) / \Sigma \left(\frac{R_i}{L_i}\right) \times 10^6$$

R_A = # of reads mapped to A

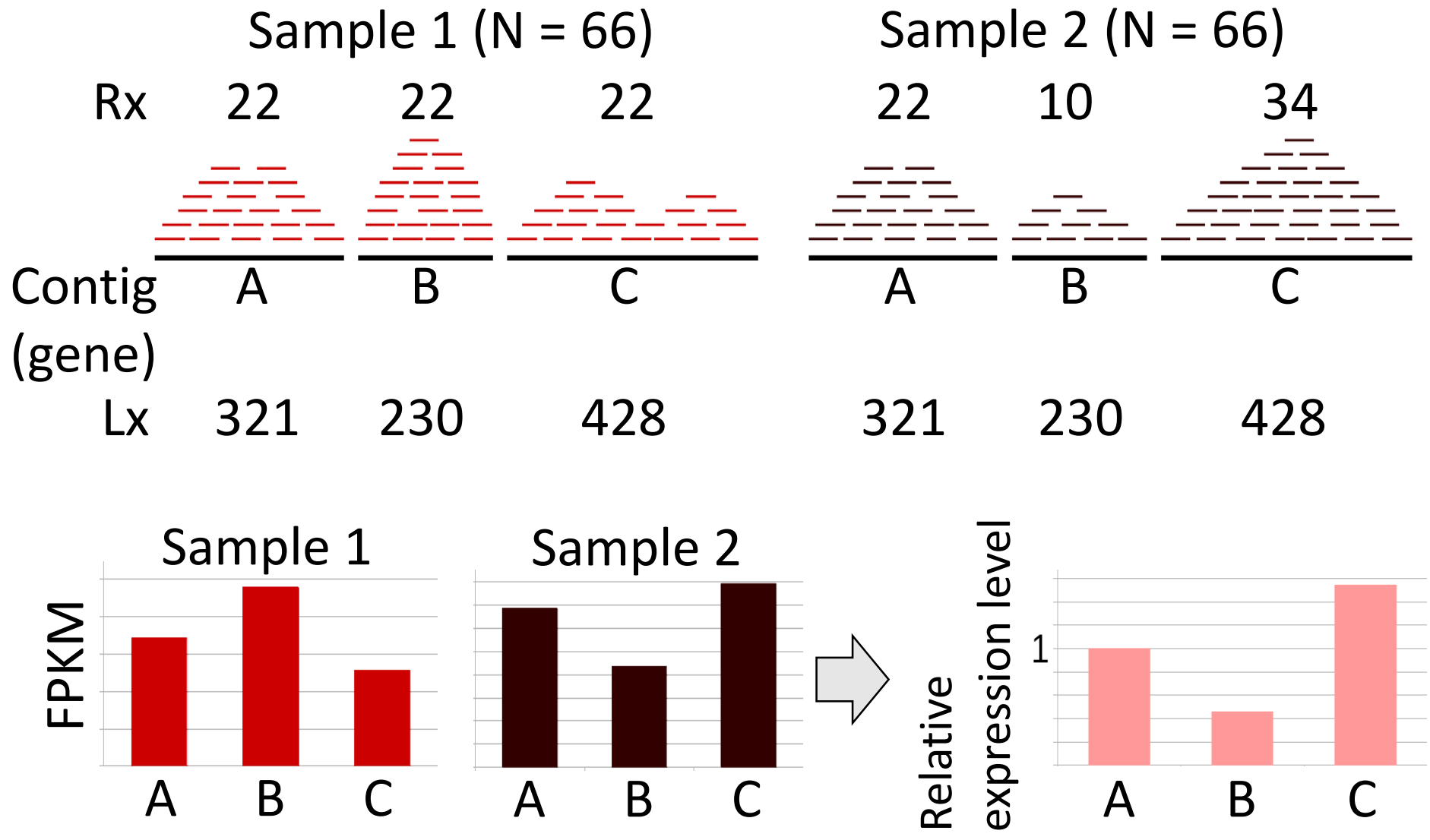
L_A = size of A

TPM is like

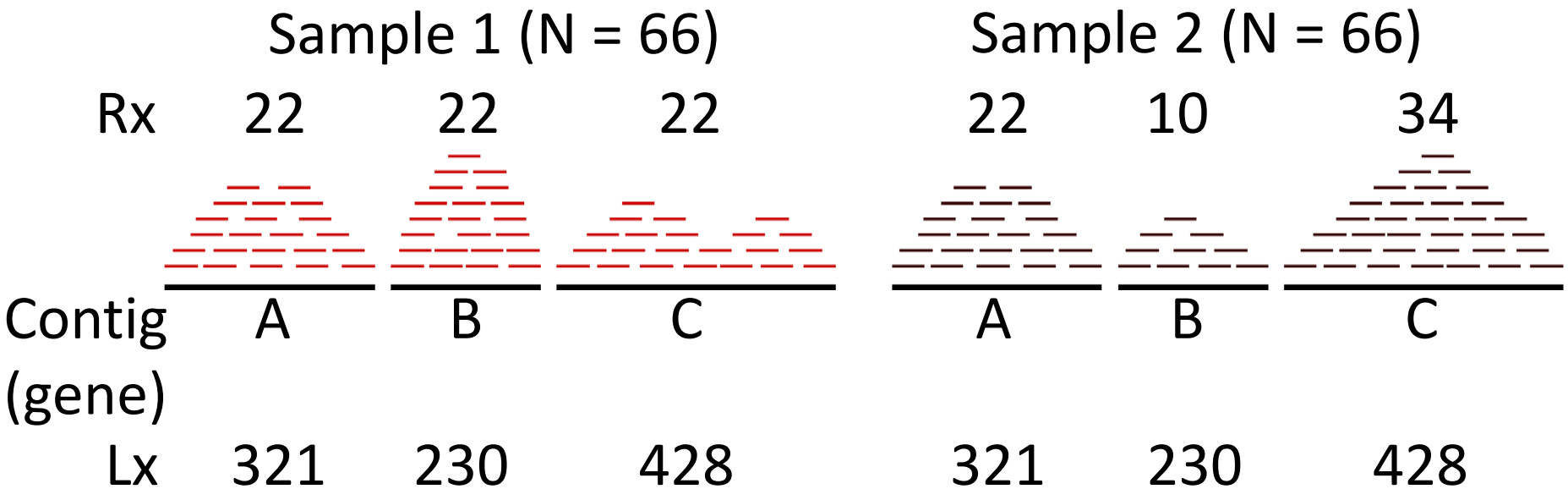
The copy number of mRNA of interest /

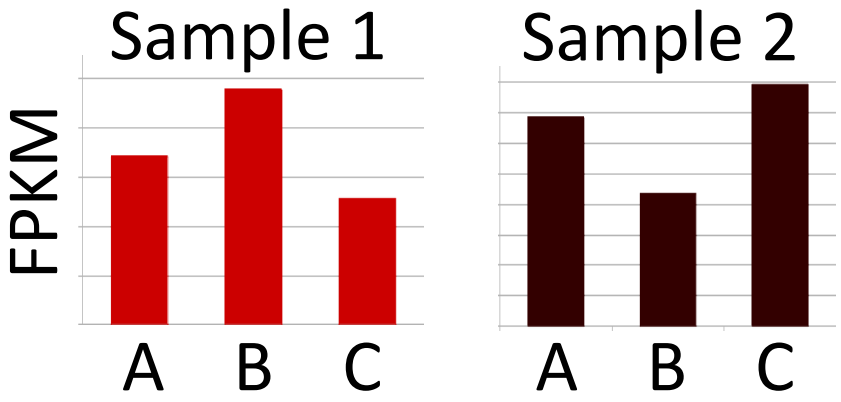
The total copy number of mRNA

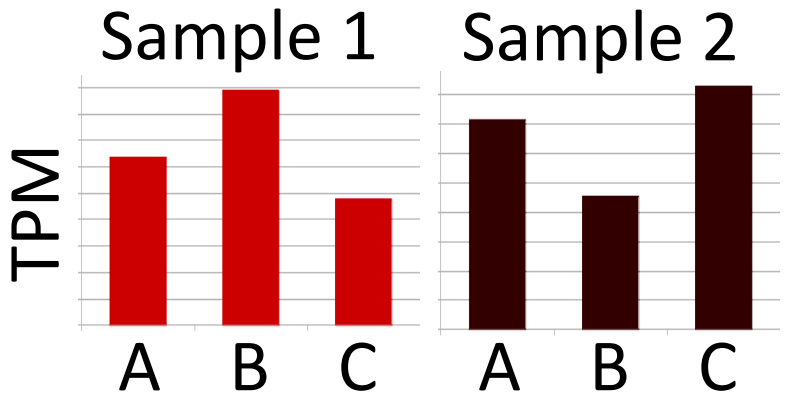
RNA-Seq


Sample 1 (N = 66)
Sample 2 (N = 66)
Rx
22
22
22
22
10
34
Contig (gene)
A
B
C
A
B
C
Lx
321
230
428
321
230
428
Sample 1
Sample 2
FPKM
A
B
C
A
B
C
Relative expression level
1
A
B
C

RNA-Seq


Sample 1 (N = 66)
Sample 2 (N = 66)
Rx
22
22
22
22
10
34
Contig
(gene)
A
B
C
A
B
C
Lx
321
230
428
321
230
428


Sample 1
Sample 2
FPKM
A
B
C
A
B
C

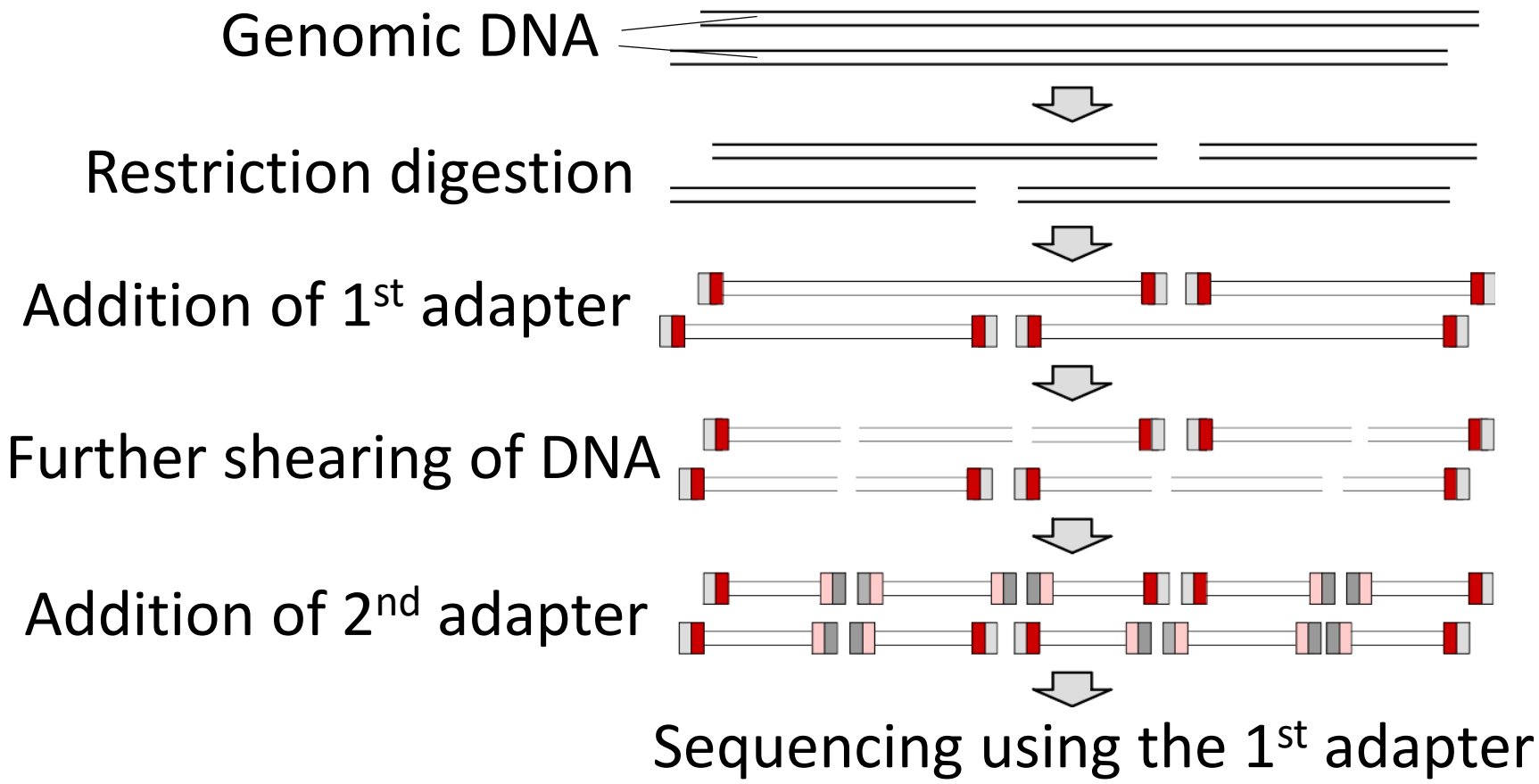

Sample 1
Sample 2
TPM
A
B
C
A
B
C

Genome sequencing

- Is sequencing a genome with NGS
- >30 × coverage is usually recommended
  E.g., for the human genome (~3 Gb), getting >90 Gb reads is preferable
- $2000 / 90 Gb if HiSeq X Ten is used
- $1000 / 1 Gb if PacBio RS II is used
- Plant genomes in general have large intergenic regions with many repetitive sequences
  → PacBio RS II has advantages over HiSeq X Ten if budget is sufficient

RAD-Seq

RAD-Seq: restriction site-associated DNA sequencing



RAD-Seq

Benefits

- Regions in the vicinity of the restriction sites can be deeply (again and again) sequenced (thus accuracy is good)
- SNPs (single nucleotide polymorphisms) can be detected on a genome-wide scale

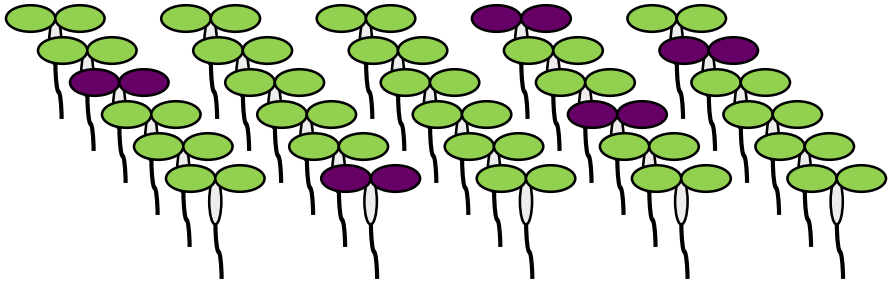
  *Regions sequenced by RAD-Seq is said to be 0.1-1% of the whole genome

  If an 8 b-recognizing restriction enzyme and single-end sequencing are used, the expected coverage would be:

  $100 \times 100 / 4^8 = 10000 / 65536 = 0.152\ldots$ (%)
- Many samples can be handled in each run using indexes
- RAD-Seq was used for developing GWAS with sorghum etc.

GWAS: genome-wide association study

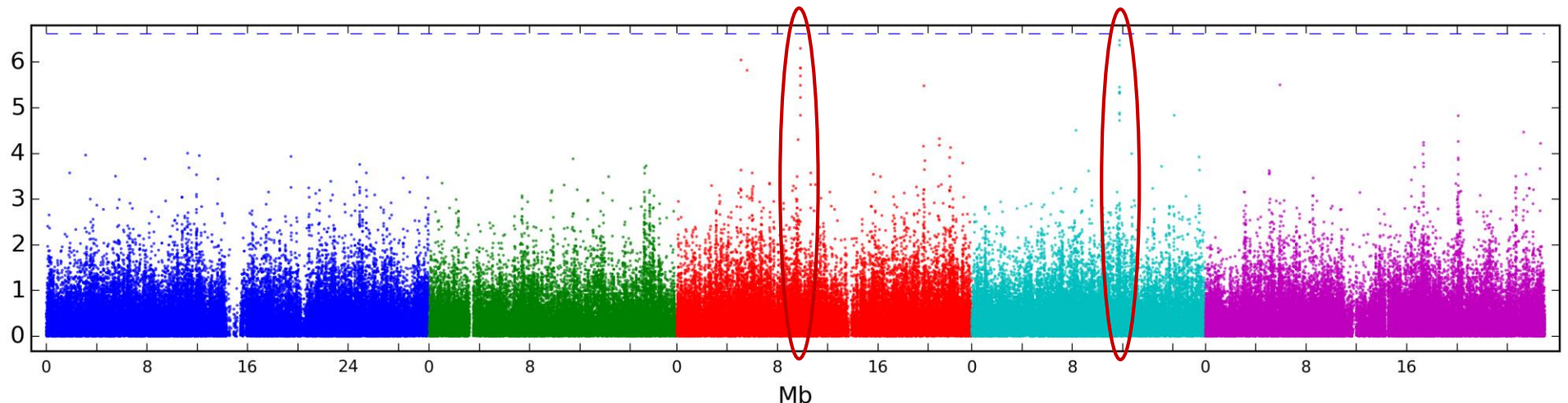
Assessment of phenotypes of various cultivars



Assessment of their SNPs

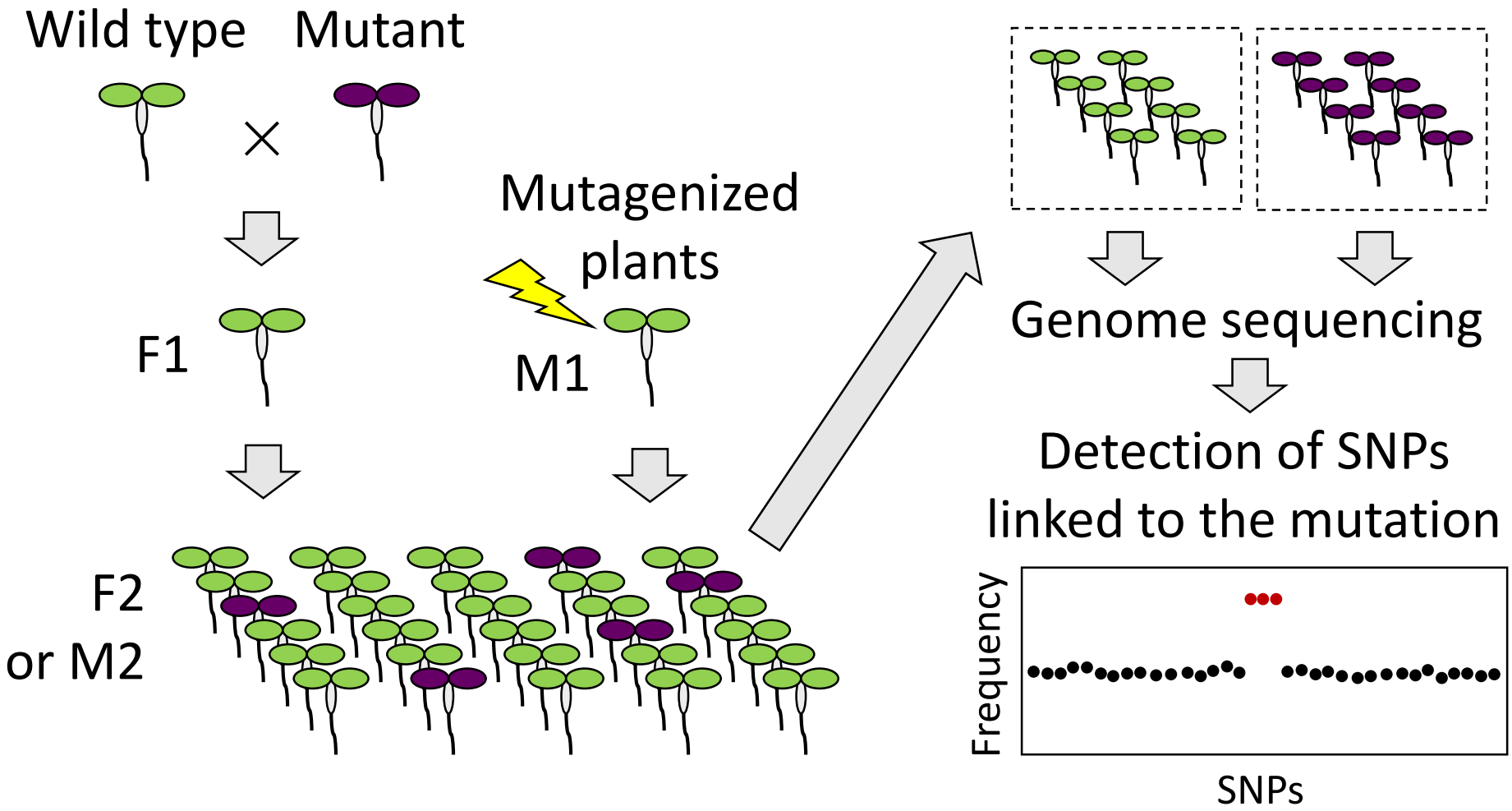
	CV1	CV2	CV3	CV4	CV5	⋯
SNP1	A	A	A	A	A	
SNP2	T	T	C	T	T	
SNP3	G	G	G	G	G	
SNP4	C	C	A	A	C	
SNP5	C	C	C	C	C	
⋮						

Detection of the SNPs associated with the phenotype of interest



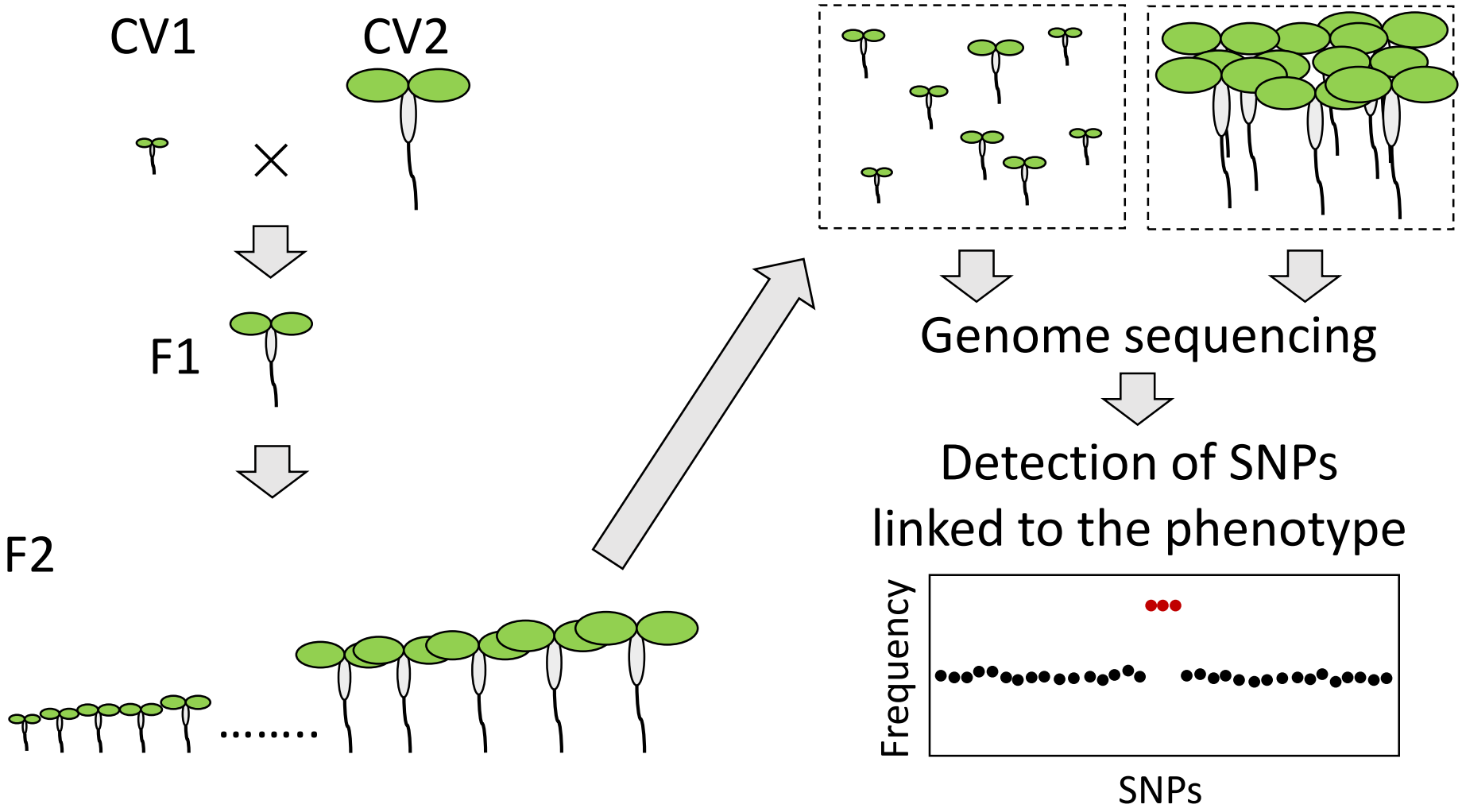
MutMap

Was developed to accelerate gene mapping



QTL-Seq

Was developed to accelerate QTL analysis



Others (not really for plant research)

- Exome sequencing:
  targets genomic regions corresponding to exons
- Amplicon-Seq:
  targets PCR products to find rare SNPs in genetic disease-causing genes or to analyze microbiota (communities of microorganisms)
- Whole genome bisulfite sequencing:
  targets genomic DNA treated with bisulfite ion, which converts unmethylated cytosine to uracil

How target DNA is prepared is important!

Summary

- Sequencers of Illumina and PacBio are often used for NGS
- Illumina sequencers output numerous short reads
- PacBio sequencers output very long reads
- It is necessary to generate contigs by *de novo* assembly if an appropriate reference is unavailable
- Mapping is often performed in NGS data analysis
- RNA-Seq and genome sequencing are the simplest yet the most useful applications of NGS
- It matters how to prepare or enrich target DNA

References

- Illumina sequencing technology:
  http://www.illumina.com/content/dam/illumina-marketing/documents/products/illumina_sequencing_introduction.pdf
- PacBio sequencing technology:
  Rhoads A, Au KF (2015) PacBio Sequencing and Its Applications. Genomics Proteomics Bioinformatics. 13(5):278-289
- MutMap:
  Abe A et al. (2012) Genome sequencing reveals agronomically important loci in rice using MutMap. Nat Biotechnol. 30(2):174-178
- QTL-Seq:
  Takagi et al. (2013) QTL-seq: rapid mapping of quantitative trait loci in rice by whole genome resequencing of DNA from two bulked populations. Plant J. 74(1):174-183.

Questions

1. It may be difficult to get a whole-genome sequence of a plant without any reference using an Illumina sequencer. Why?
2. In what situation(s), is RAD-Seq better than whole genome sequencing?
3. In RNA-Seq using model species, genome sequences are more often used as a reference for mapping than mRNA sequences. Why?
4. What would you like to do with NGS?
5. Any suggestions and/or comments?