Korea Institute of Science and Technology

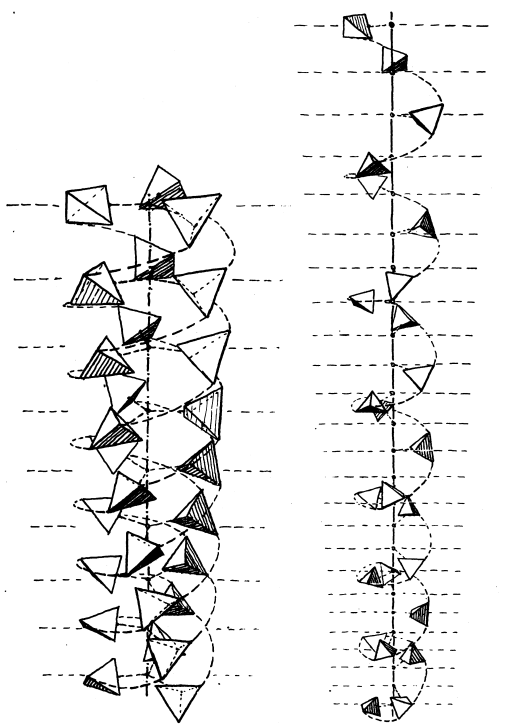
RNA-Sequencing Data Analysis

Korea Institute of Science and Technology (KIST)
Clean Energy Research Center

Sungmin Hwang

2022. 2. 11

Pioneer works to identify DNA structure

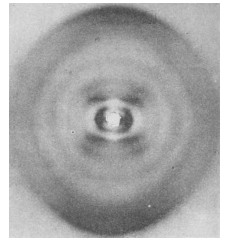


A Proposed Structure For The Nucleic Acids

Pauling and Corey, 1953 PNAS
doi.org/10.1073/pnas.39.2.84

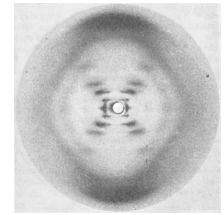
Molecular Structure of Nucleic Acids: A Structure for **Deoxyribose Nucleic Acid**

Watson and Crick, 1953 Nature
doi.org/10.1038/171737a0



Molecular Structure of Nucleic Acids: Molecular Structure of Deoxypentose Nucleic Acids

Wilkins, Stokes and Wilson, 1953 Nature
doi.org/10.1038/171738a0



Molecular Configuration in Sodium Thymonucleate

Franklin and Gosling, 1953 Nature
doi.org/10.1038/171740a0

DNA sequencing technology: 1st generation (Sanger's chain termination)

Original DNA sequence

ATGCAGCGTTACCATG . . .

Amplification

+ Pol

+ ddNTPs

A
A
ATGCA
ATGCAGCGTTA
ATGCAGCGTTACCA

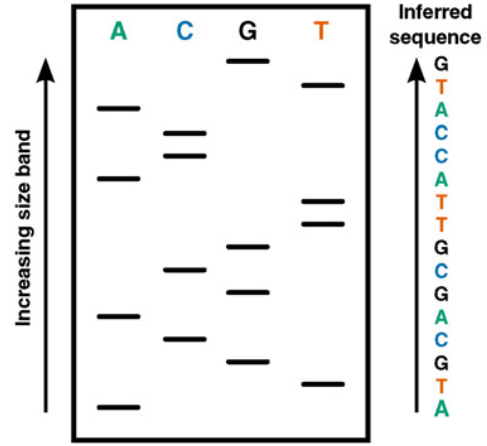
C
ATGC
ATGCAGC
ATGCAGCGTTAC
ATGCAGCGTTACC

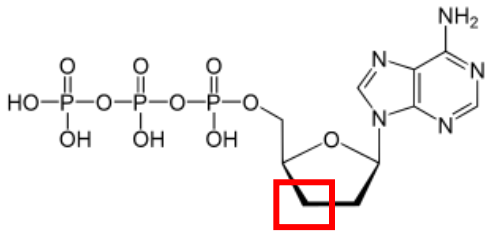
G
ATG
ATGCAG
ATGCAGCG
ATGCAGCGTTACCATG

T
AT
ATGCAGCGT
ATGCAGCGTT
ATGCAGCGTTACCAT

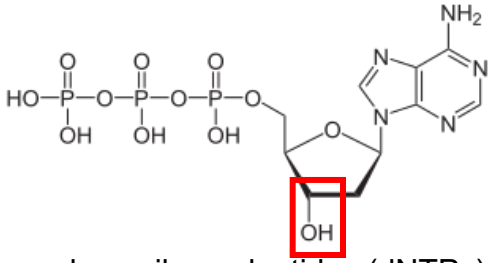
Electrophoresis

Reading signals





dideoxynucleotides (ddNTPs)
: lack the 3′ hydroxyl group that is required for extension of DNA chains



deoxyribonucleotides (dNTPs)

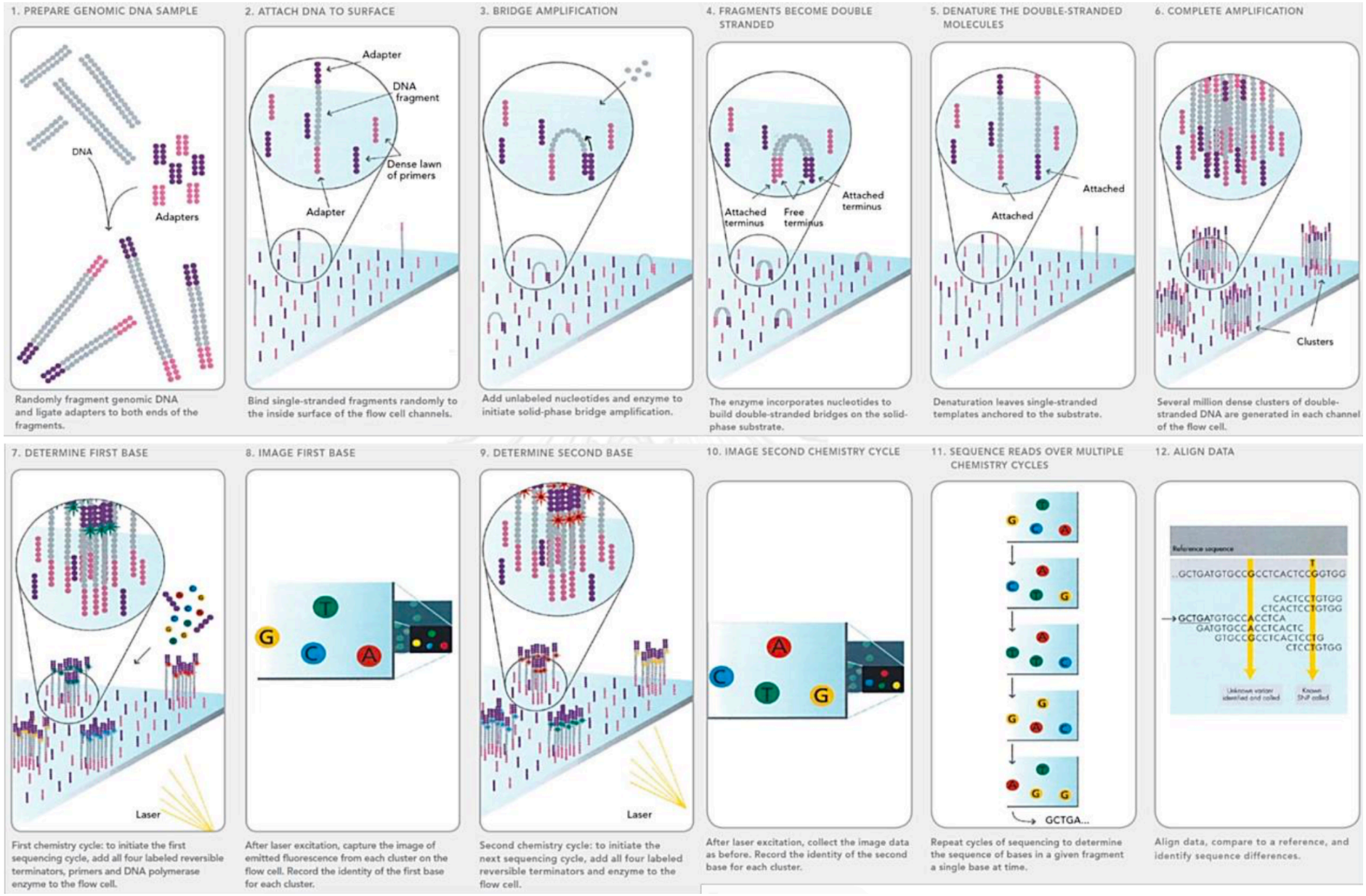
Limitation

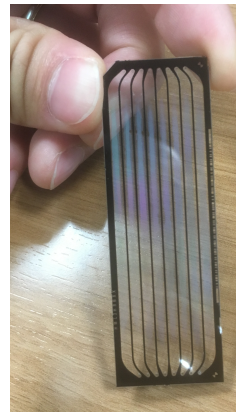
- Coverage: ~ 1 kbp
- Organisms that have a big genome size
  i.e., *E. coli* (4.6 Mbp), *Homo sapiens* (3,088 Mbp)

Heather and Chain, 2016 Genomics
https://doi.org/10.1016/j.ygeno.2015.11.003

DNA sequencing technology: Next-generation (high-throughput)

Illumina Co.,

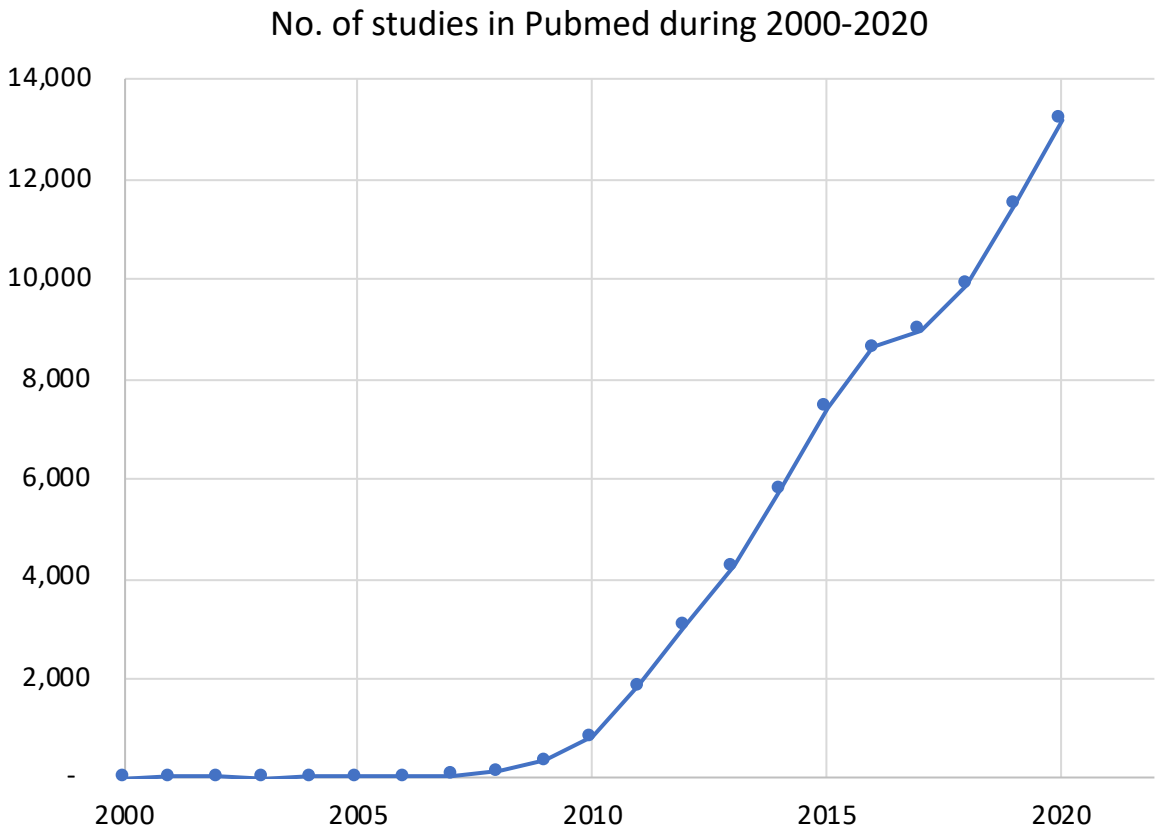




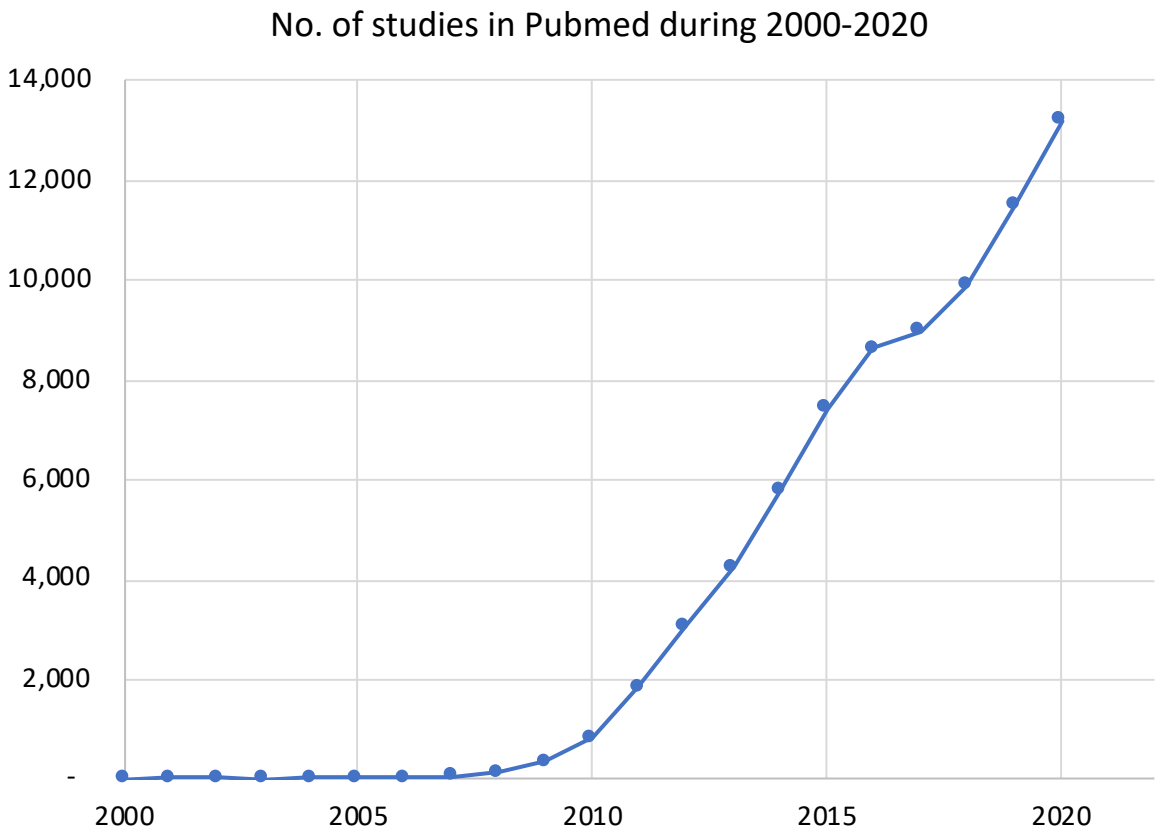
Flow-cell

https://www.youtube.com/watch?v=fCd6B5HRaZ8

DNA sequencing technology: Next-generation (high-throughput)

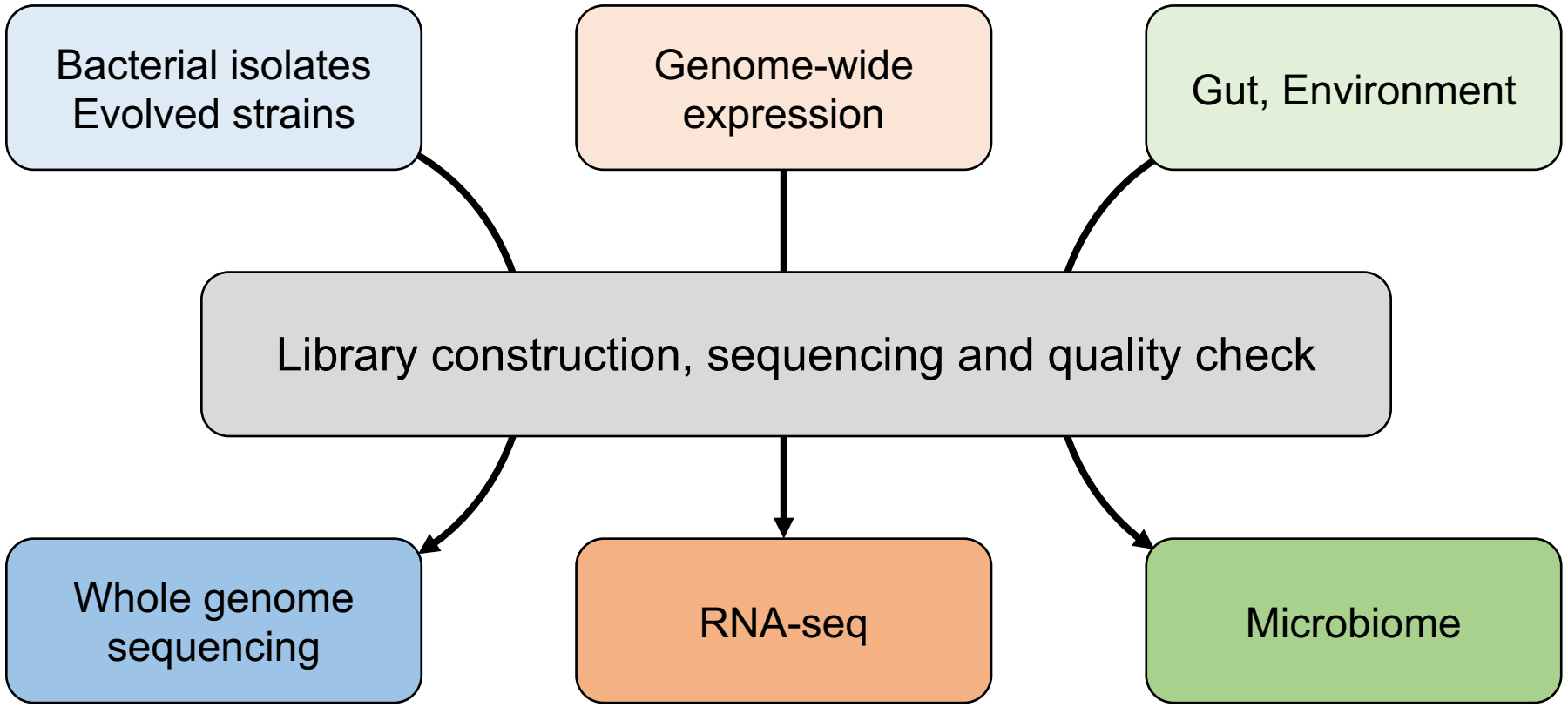


DNA sequencing technology: Next-generation (high-throughput)

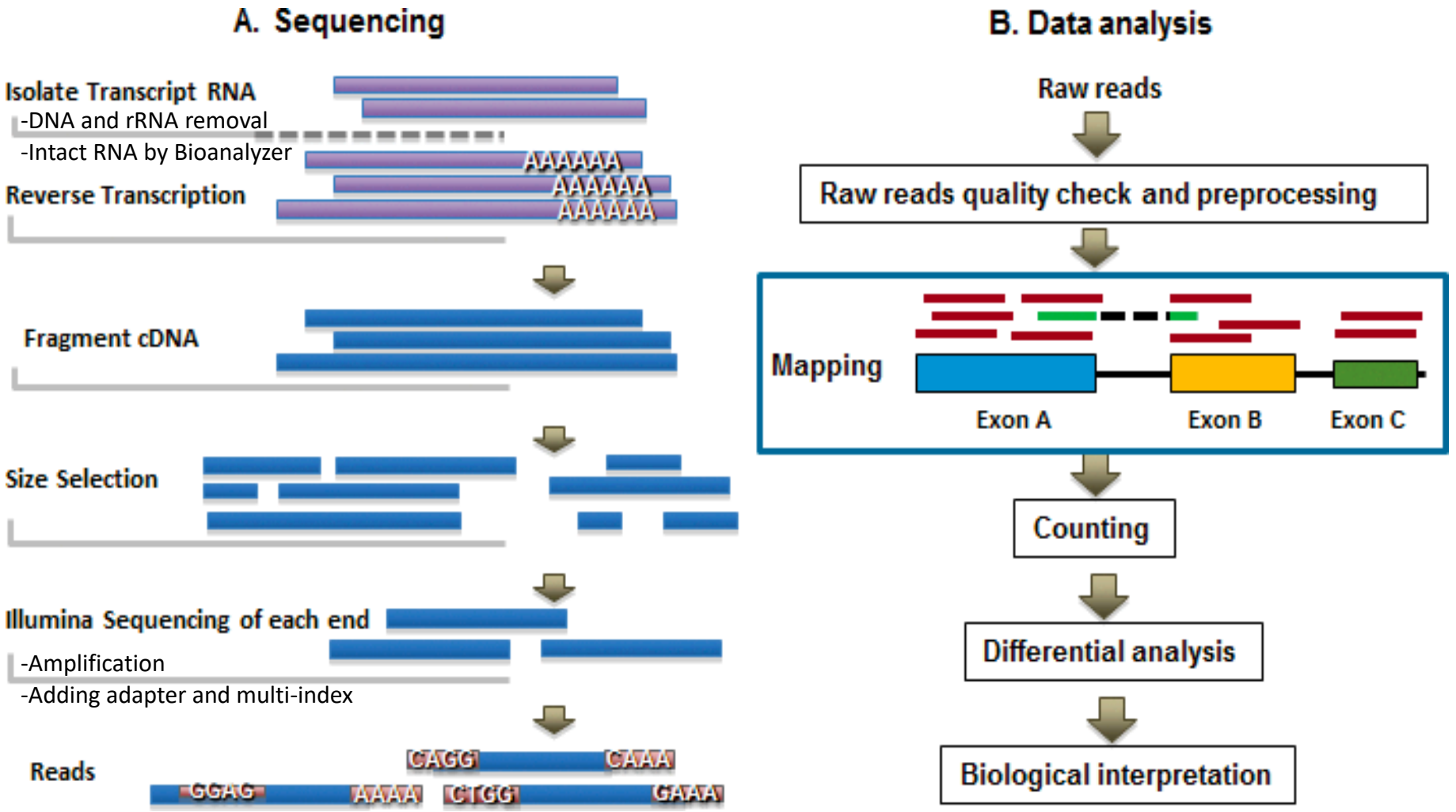


* 본 자료는 **NGS 기술을 활용한 연구 내용 전달**에 있으며, **투자 권유 및 종목 추천이 아님**을 명시합니다.

Applications by using NGS



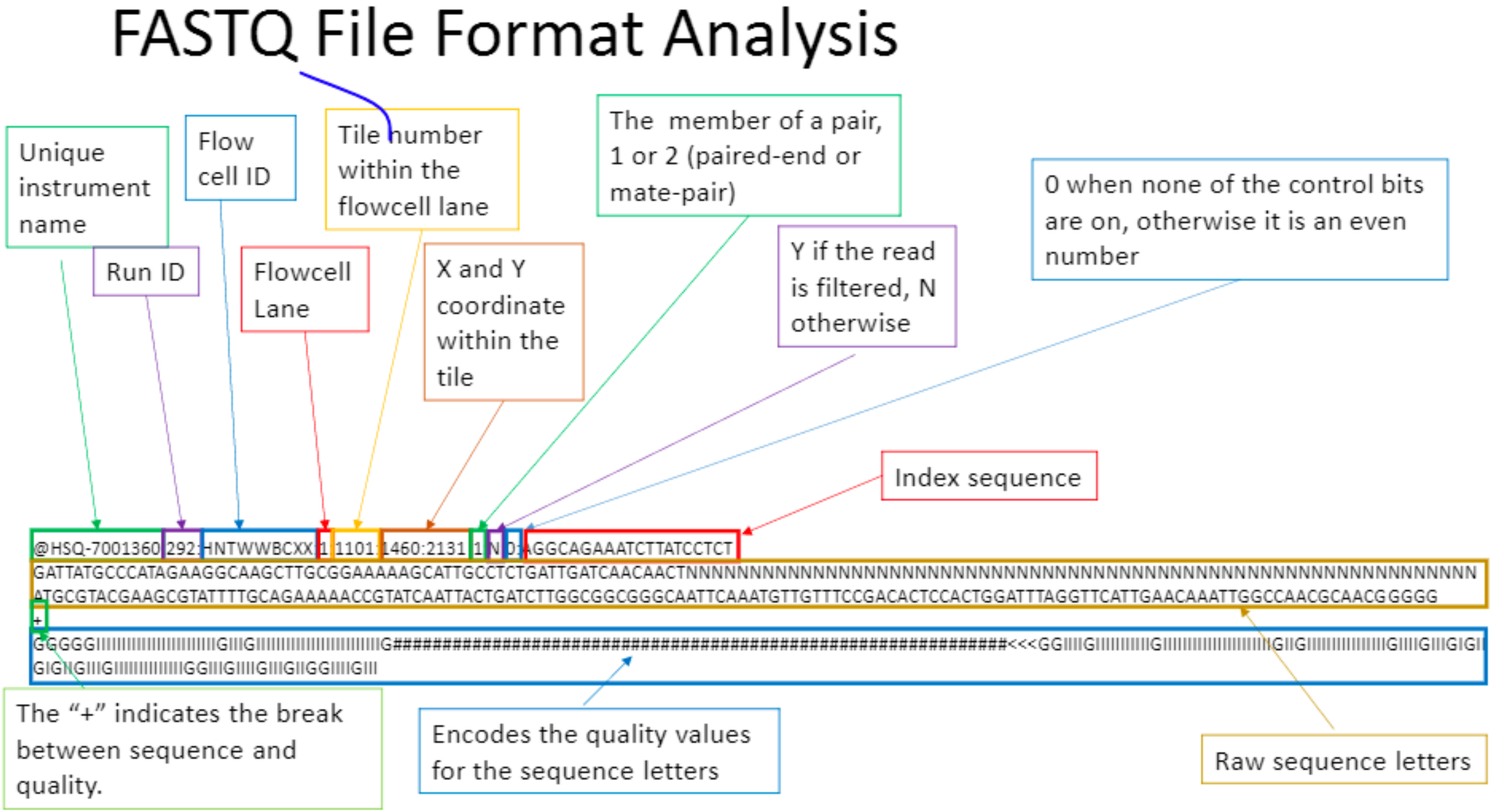
Overview of RNA-seq flowchart



Zhao et al., 2015 Bioinformatics: Updated Features and Applications
https://www.intechopen.com/chapters/50574

Raw data format of NGS data

Raw reads → Quality check → Mapping → Counting → DEG analysis → Functional enrichment → Data interpretation



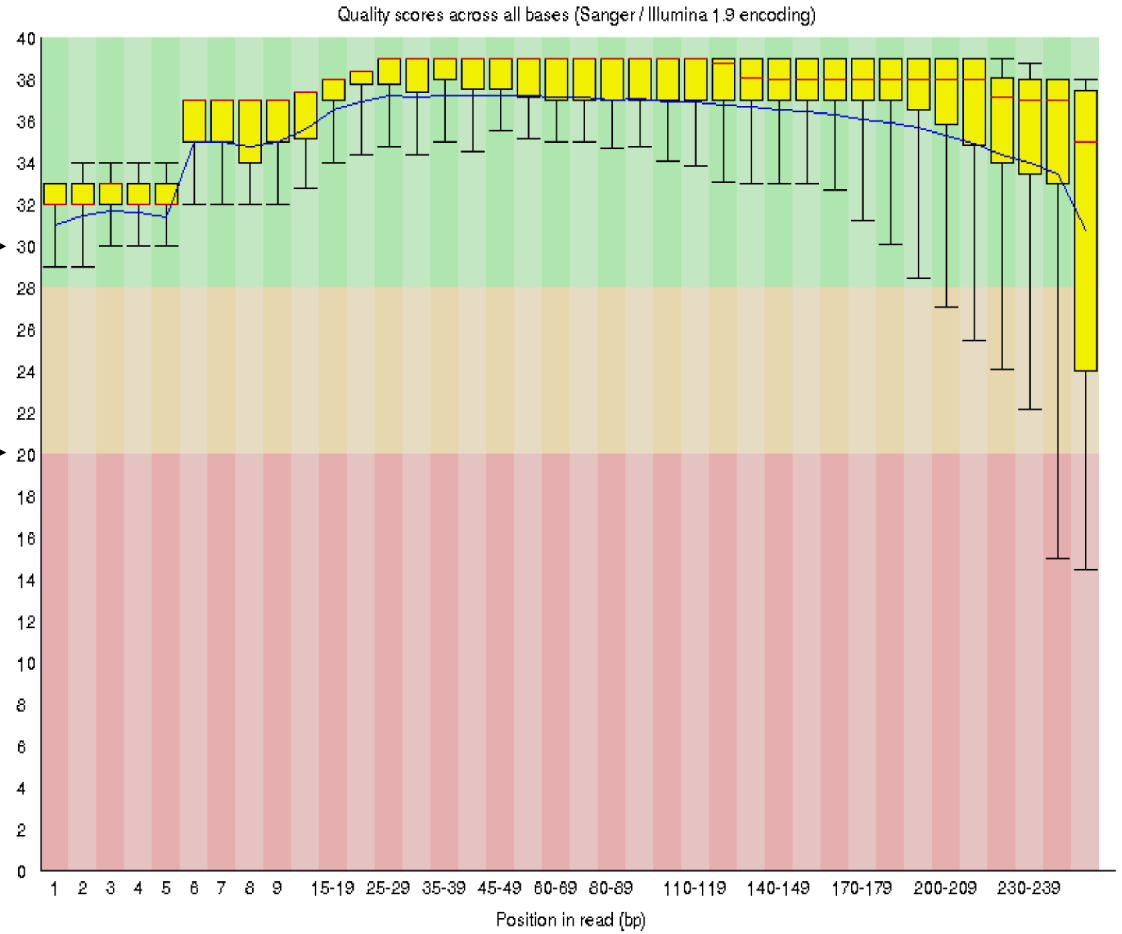
https://kscbioinformatics.wordpress.com/2017/02/03/raw-illumina-sequence-data-files-for-dummies-part-1/

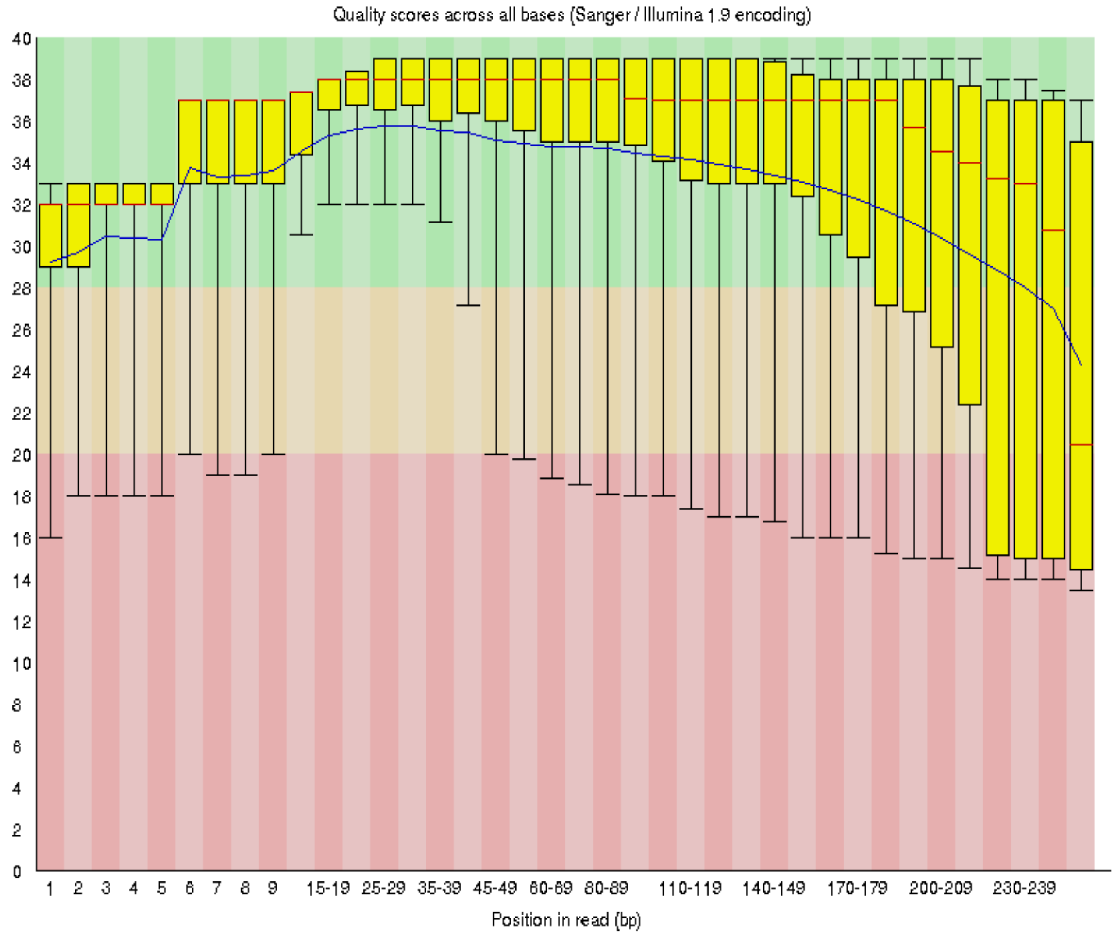
Quality check by FastQC

Accuracy

99.9% →

99% →





- 4.6 Mbp of *E. coli*
  : 1% → 46,000
  : 0.1% → 4,600

- 3,088 Mbp of Human
  : 1% → 30,880,000
  : 0.1% → 3,088,000

https://microsizedmind.wordpress.com/2014/02/21/is-improvement-possible-quality-control-of-a-illumina-nextera-dataset-for-the-novo-genome-assembly/

Mapping by bowtie2

Raw reads → Quality check → **Mapping** → Counting → DEG analysis → Functional enrichment → Data interpretation

```
10414362 reads; of these:
  10414362 (100.00%) were paired; of these:
    374434 (3.60%) aligned concordantly 0 times
    9833851 (94.43%) aligned concordantly exactly 1 time
    206077 (1.98%) aligned concordantly >1 times
    ----
    374434 pairs aligned concordantly 0 times; of these:
      234742 (62.69%) aligned discordantly 1 time
    ----
    139692 pairs aligned 0 times concordantly or discordantly; of these:
      279384 mates make up the pairs; of these:
        148566 (53.18%) aligned 0 times
        117701 (42.13%) aligned exactly 1 time
        13117 (4.69%) aligned >1 times
99.29% overall alignment rate
```

Table 1. The adopted mapping methods.

Name	Version	Mapping
Bowtie	2.2.6	Unspliced read aligner
BWA } eukaryotes	0.7.12-r1039	Unspliced read aligner
TopHat } eukaryotes	2.10	Spliced read aligner
STAR	2.5.3	Spliced read aligner
kallisto	0.43.1	pseudo-alignment
Salmon	0.8.2	pseudo-alignment

An example alignment result

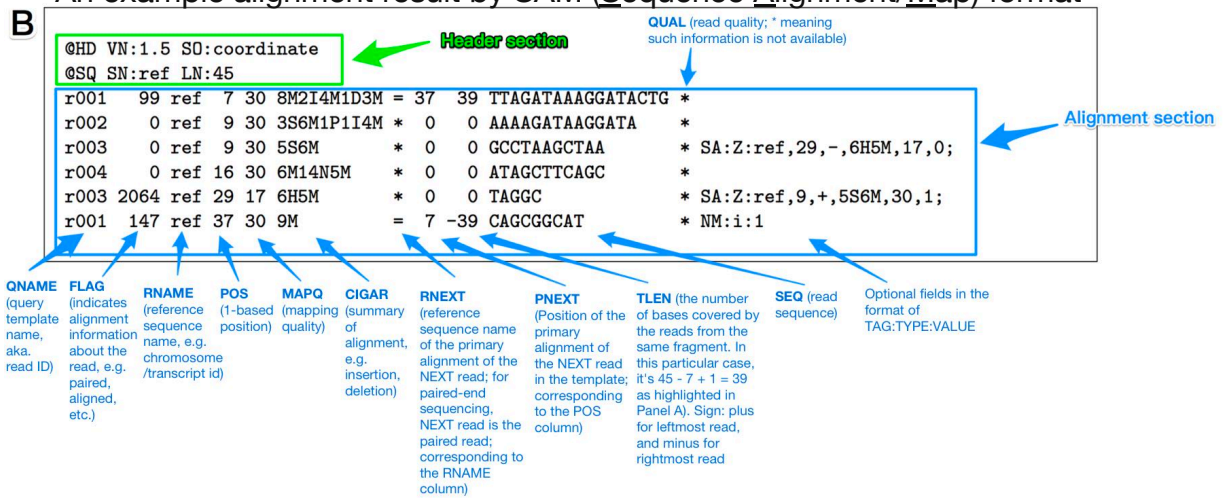
A

```
                    10        20        30        40
Coor     12345678901234  5678901234567890123456789012345
ref      AGCATGTTAGATAA**GATAGCTGTGCTAGTAGGCAGTCAGCGCCAT

+r001/1        TTAGATAAAGGATA*CTG
+r002         aaaAGATAA*GGATA
+r003       gcctaAGCTAA
+r004                   ATAGCT..............TCAGC
-r003                         ttagctTAGGC
-r001/2                                    CAGCGGCAT
```

An example alignment result by SAM (Sequence Alignment/Map) format

B

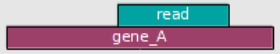
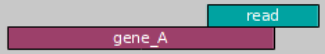
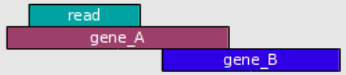
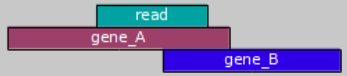
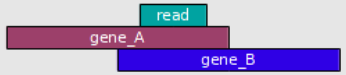
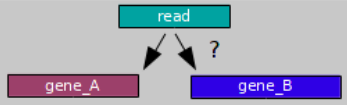


*BAM: a compressed binary version of SAM

http://zyxue.github.io/2017/09/26/sam-format-example.html
https://samtools.github.io/hts-specs/SAMv1.pdf
Costa-Silva et al., 2017 PlosOne
https://doi.org/10.1371/journal.pone.0190152

Read counting by HTseq

PSPTO_RS00025	0
PSPTO_RS00030	0
PSPTO_RS00035	0
PSPTO_RS00040	0
PSPTO_RS00045	0
PSPTO_RS00050	0
PSPTO_RS00060	0
…	
__no_feature	108
__ambiguous	162
__too_low_aQual	181
__not_aligned	126
__alignment_not_unique	0

	union	intersection_strict	intersection_nonempty
	gene_A	gene_A	gene_A
	gene_A	no_feature	gene_A
	gene_A	no_feature	gene_A
	gene_A	gene_A	gene_A
	gene_A	gene_A	gene_A
	ambiguous (both genes with --nonunique all)	gene_A	gene_A
	ambiguous (both genes with --nonunique all)		
	alignment_not_unique (both genes with --nonunique all)		

https://htseq.readthedocs.io/en/release_0.11.1/count.html

Differentially expressed genes by DESeq2

Raw reads → Quality check → Mapping → Counting → **DEG analysis** → Functional enrichment → Data interpretation

Table 2. Adopted methods for DEGs identification.

Name	Version	Normalization
baySeq	2.4.1	Scaling factors (quantile/ TMM/ total)
DESeq	1.22.1	DESeq size factors
EBSeq	1.12.0	DESeq median normalization
edgeR	3.12.1	TMM/ Upper quartile / RLE / None (all scaling factors are set to be one)
limma+voom	3.26.9	TMM
NOIseq	2.14.1	RPKM / TMM / Upper quartile
SAMseq (samr)	2.0	Based on the read count mean over the null features of data set.
DESeq2	1.10.1	DESeq size factors
sleuth	0.29.0	DESeq size factors (with slight modifications)

https://hbctraining.github.io/DGE_workshop/lessons/02_DGE_count_normalization.html
Costa-Silva et al., 2017 PlosOne
https://doi.org/10.1371/journal.pone.0190152

Differentially expressed genes by DESeq2

Raw reads → Quality check → Mapping → Counting → **DEG analysis** → Functional enrichment → Data interpretation

Normalization method	Description	Accounted factors	Recommendations for use
CPM (counts per million)	counts scaled by total number of reads	sequencing depth	gene count comparisons between replicates of the same samplegroup; **NOT for within sample comparisons or DE analysis**
TPM (transcripts per kilobase million)	counts per length of transcript (kb) per million reads mapped	sequencing depth and gene length	gene count comparisons within a sample or between samples of the same sample group; **NOT for DE analysis**
RPKM/FPKM (reads/fragments per kilobase of exon per million reads/fragments mapped)	similar to TPM	sequencing depth and gene length	gene count comparisons between genes within a sample; **NOT for between sample comparisons or DE analysis**
DESeq2's **median of ratios**	counts divided by sample-specific size factors determined by median ratio of gene counts relative to geometric mean per gene	sequencing depth and RNA composition	gene count comparisons between samples and for **DE analysis**; **NOT for within sample comparisons**
EdgeR's **trimmed mean of M values (TMM)**	uses a weighted trimmed mean of the log expression ratios between samples	sequencing depth, RNA composition, and gene length	gene count comparisons between and within samples and for **DE analysis**

https://hbctraining.github.io/DGE_workshop/lessons/02_DGE_count_normalization.html
Costa-Silva et al., 2017 PlosOne
https://doi.org/10.1371/journal.pone.0190152

Table 2. Adopted methods for DEGs identification.

Name	Version	Normalization
baySeq	2.4.1	Scaling factors (quantile/ TMM/ total)
DESeq	1.22.1	DESeq size factors
EBSeq	1.12.0	DESeq median normalization
edgeR	3.12.1	TMM/ Upper quartile / RLE / None (all scaling factors are set to be one)
limma+voom	3.26.9	TMM
NOIseq	2.14.1	RPKM / TMM / Upper quartile
SAMseq (samr)	2.0	Based on the read count mean over the null features of data set.
DESeq2	1.10.1	DESeq size factors
sleuth	0.29.0	DESeq size factors (with slight modifications)

RPKM-normalized counts table

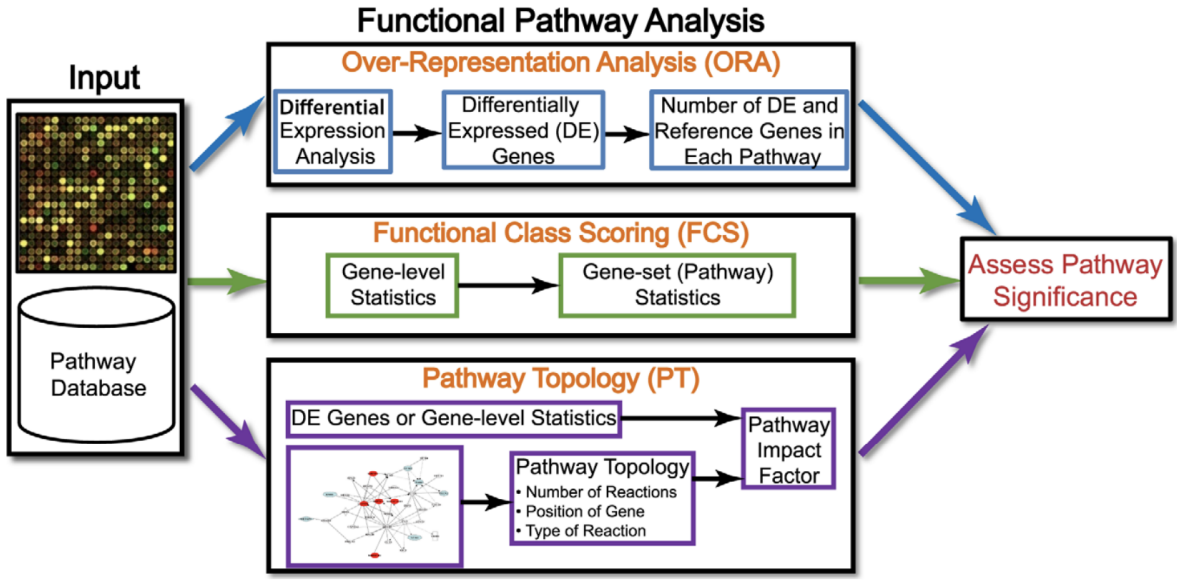
gene	sampleA	sampleB
XCR1	5.5	5.5
WASHC1	73.4	21.8
…	…	…
Total RPKM-normalized counts	1,000,000	1,500,000

DESeq2-normalized counts: Median of ratios method

gene	sample A	sample B	pseudo-reference sample	ratio of sampleA/ref	ratio of sampleB/ref	Normalized A	Normalized B
EF2A	1489	906	sqrt(1489*906) = 1161.5	1489/1161.5 = **1.28**	906/1161.5 = **0.78**	1489 / 1.3 = 1145.39	906 / 0.77 = 1176.62
ABCD1	22	13	sqrt(22*13) = 16.9	22/16.9 = **1.30**	13/16.9 = **0.77**	22 / 1.3 = 16.92	13 / 0.77 = 16.88
MEFV	793	410	sqrt(793*410) = 570.2	793/570.2 = **1.39**	410/570.2 = **0.72**	793 / 1.3 = 610	410 / 0.77 = 532.46
				Mean (1.28, 1.30, 1.39 …) = 1.3	**Mean** (0.78, 0.77, 0.72 …) = 0.77		

Functional enrichment

Raw reads → Quality check → Mapping → Counting → DEG analysis → **Functional enrichment** → Data interpretation



Analysis	What is required for input	What output looks like	✅ Pros	⚠️ Cons
[ORA (Over-representation Analysis)](#) **e.g., Gene ontology (GO), eggNOG-mapper, Pathview**	A list of gene IDs (no stats needed)	A per-pathway hypergeometric test result	- Simple - Inexpensive computationally to calculate p-values	- Requires arbitrary thresholds and ignores any statistics associated with a gene - Assumes independence of genes and pathways
[GSEA (Gene Set Enrichment Analysis)](#)	A list of genes IDs with gene-level summary statistics	A per-pathway enrichment score	- Includes all genes (no arbitrary threshold!) - Attempts to measure coordination of genes	- Permutations can be expensive - Does not account for pathway overlap - Two-group comparisons not always appropriate/feasible
[GSVA (Gene Set Variation Analysis)](#)	A gene expression matrix (like what you get from refine.bio directly)	Pathway-level scores on a per-sample basis	- Does not require two groups to compare upfront - Normally distributed scores	- Scores are not a good fit for gene sets that contain genes that go up AND down - Method doesn't assign statistical significance itself - Recommended sample size n > 10

https://alexslemonade.github.io/refinebio-examples/03-rnaseq/pathway-analysis_rnaseq_01_ora.html

Costa-Silva et al., 2017 PlosOne
https://doi.org/10.1371/journal.pcbi.1002375

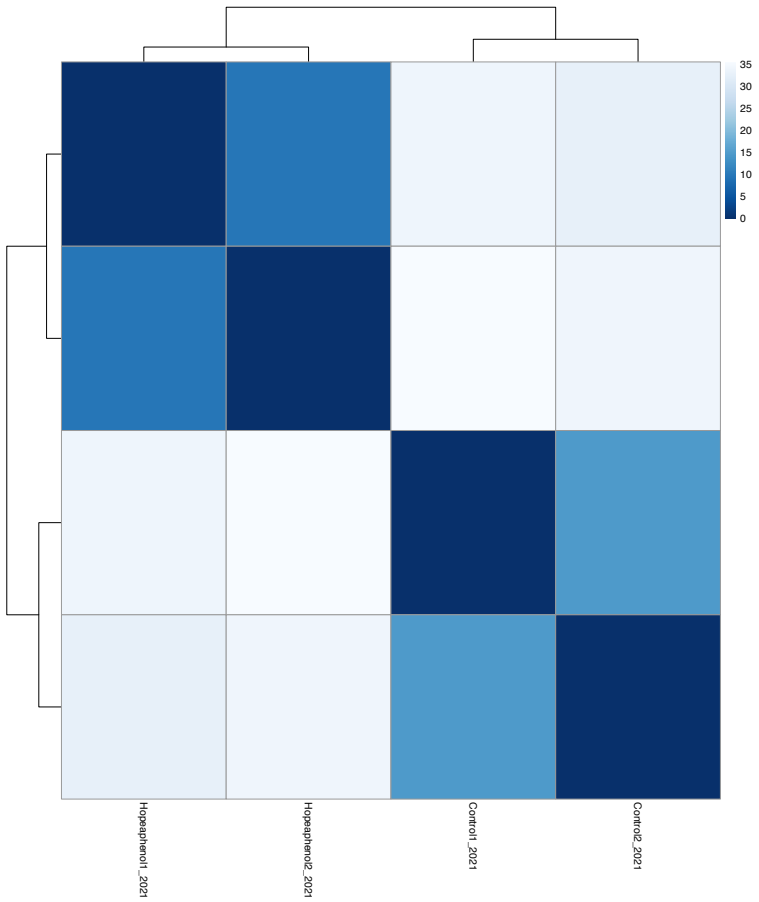
Transcriptomics: Effect of hopeaphenol to *Pectobacterium atrosepticum* SCRI1043

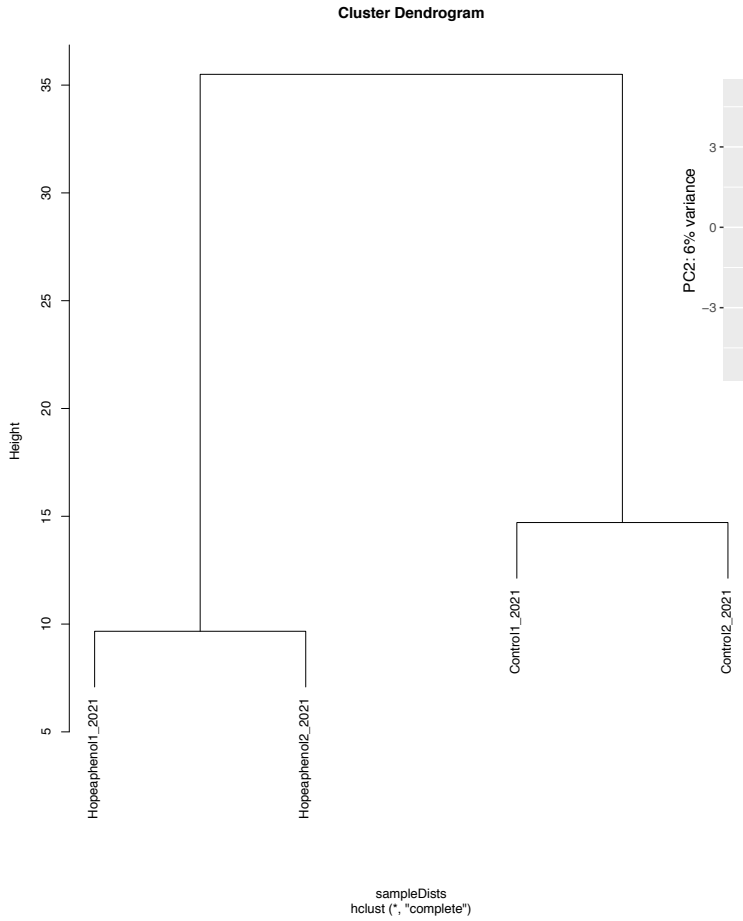
Data analysis

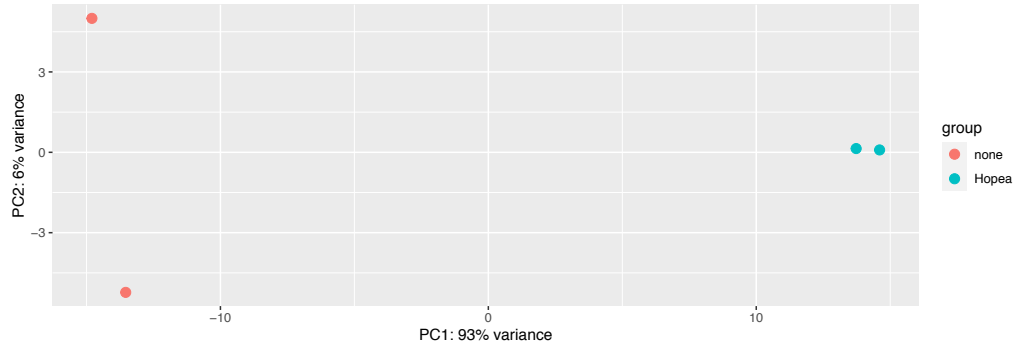
-Alignment: Bowtie2

-Read counting: HTSeq

-Statistics: DESeq2





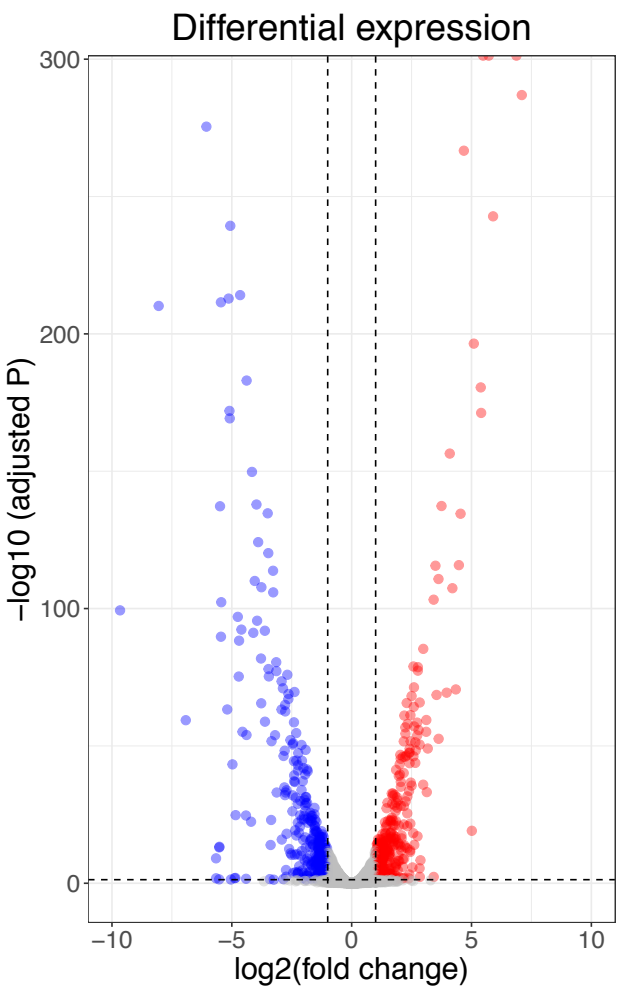


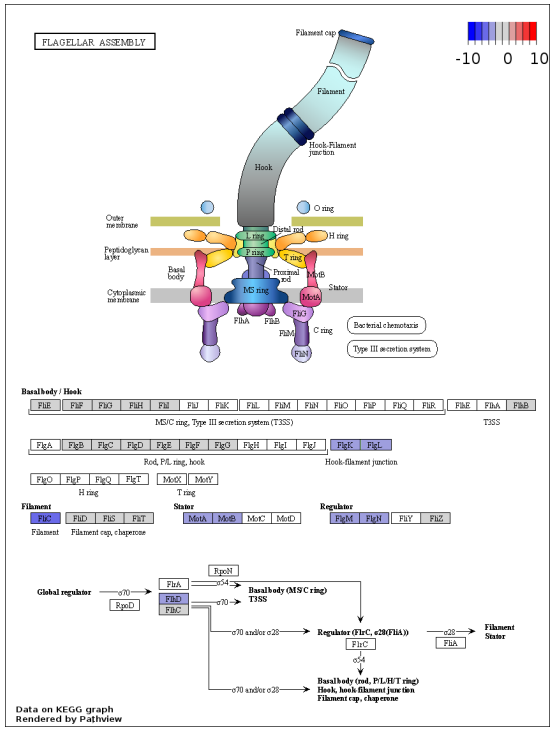
Unpublished data from Chung lab

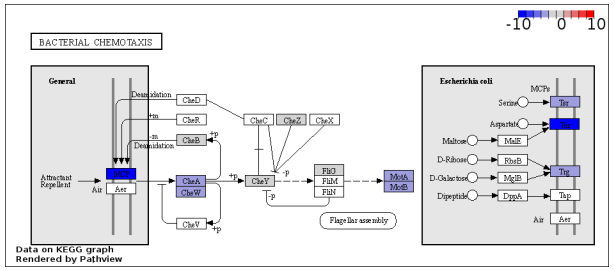
Transcriptomics: Effect of hopeaphenol to *Pectobacterium atrosepticum* SCRI1043

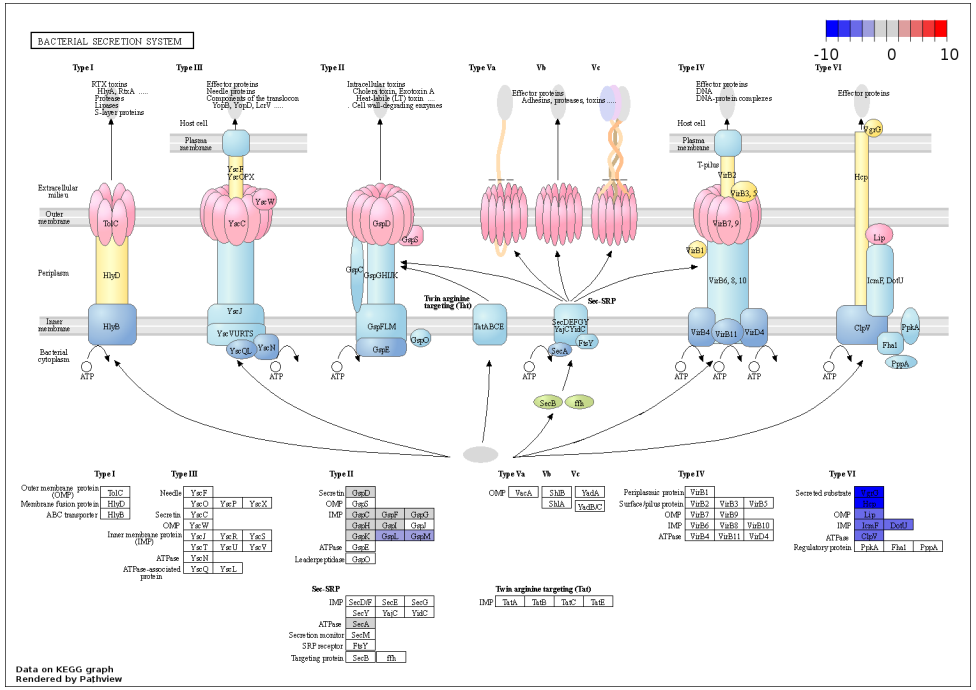
Hopea vs non-treated

689 DEGs (336 up, 353 down)









Unpublished data from Chung lab

Hand-on RNA-seq data analysis

GitHub repository

https://github.com/sungminhwang-duke/RNAseq_hand_on_work