

# KSBi-BIML 2024

Bioinformatics & Machine Learning(BIML)  
Workshop for Life and Medical Scientists

생명정보학 & 머신러닝 워크샵 (온라인)



## Introduction to Gene Expression Profiling

장혜식 \_ 서울대학교



**KSBI**  
KOREAN SOCIETY FOR  
BIOINFORMATICS

| 한국생명정보학회



본 강의 자료는 한국생명정보학회가 주관하는 BIML 2024 워크샵 온라인 수업을 목적으로 제작된 것으로 해당 목적 이외의 다른 용도로 사용할 수 없음을 분명하게 알립니다.

이를 다른 사람과 공유하거나 복제, 배포, 전송할 수 없으며 만약 이러한 사항을 위반할 경우 발생하는 **모든 법적 책임은 전적으로 불법 행위자 본인에게 있음을 경고**합니다.

# KSBI-BIML 2024

## Bioinformatics & Machine Learning(BIML) Workshop for Life and Medical Scientists

안녕하십니까?

한국생명정보학회가 개최하는 동계 교육 워크숍인 BIML-2024에 여러분을 초대합니다. 생명정보학 분야의 연구자들에게 최신 동향의 데이터 분석기술을 이론과 실습을 겸비해 전달하고자 도입한 전문 교육 프로그램인 BIML 워크숍은 2015년에 시작하여 올해로 벌써 10년 차를 맞이하게 되었습니다. BIML 워크숍은 국내 생명정보학 분야의 최초이자 최고 수준의 교육프로그램으로 크게 인공지능과 생명정보분석 두 개의 분야로 구성되어 있습니다. 올해 인공지능 분야에서는 최근 생명정보 분석에서도 응용이 확대되고 있는 다양한 인공지능 기반 자료모델링 기법들에 대한 현장 강의를 진행될 예정이며, 관련하여 심층학습을 이용한 단백질구조예측, 유전체분석, 신약개발에 대한 이론과 실습 강의를 함께 제공될 예정입니다. 또한 단일세포오믹스, 공간오믹스, 메타오믹스, 그리고 롱리드염기서열 자료 분석에 대한 현장 강의는 많은 연구자의 연구 수월성 확보에 큰 도움을 줄 것으로 기대하고 있습니다.

올해 BIML의 가장 큰 변화는 최근 연구 수요가 급증하고 있는 의료정보자료 분석에 대한 현장 강의를 추가하였다는 것입니다. 특히 의료정보자료 분석을 많이 수행하시는 의과학자 및 의료정보 연구자들께서 본 강좌를 통해 많은 도움을 받으실 수 있기를 기대하고 있습니다. 또한 다양한 생명정보학 분야에 대한 온라인 강좌 프로그램도 점차 증가하고 있는 생명정보 분석기술의 다양화에 발맞추기 위해 작년과 비교해 5강좌 이상을 신규로 추가했습니다. 올해는 무료 강좌 5개를 포함하여 35개 이상의 온라인 강좌가 개설되어 제공되며, 연구 주제에 따른 연관된 강좌 추천 및 강연료 할인 프로그램도 제공되며, 온라인을 통한 Q&A 세션도 마련될 예정입니다. BIML-2024는 국내 주요 연구 중심 대학의 전임 교원이자 각 분야 최고 전문가들의 강의로 구성되었기에 해당 분야의 기초부터 최신 연구 동향까지 포함하는 수준 높은 내용의 강의를 될 것이라 확신합니다.

BIML-2024을 준비하기까지 너무나 많은 수고를 해주신 운영위원회의 정성원, 우현구, 백대현, 김태민, 김준일, 김상우, 장혜식, 박종은 교수님과 KOBIC 이병욱 박사님께 커다란 감사를 드립니다. 마지막으로 부족한 시간에도 불구하고 강의 부탁을 흔쾌히 허락하시고 훌륭한 현장 강의와 온라인 강의를 준비하시는데 노고를 아끼지 않으신 모든 강사분들께 깊은 감사를 드립니다.

2024년 2월

한국생명정보학회장 이 인 석

# Introduction to Gene Expression Profiling

어떤 유전자나 물질이 세포에서 어떤 일을 할까? 어떤 가설을 갖고있던 가장 먼저 RNA-seq 먼저 해 보고 시작하게 된지 이제 15년이 되었다. 유전자 발현의 중심축인 DNA-RNA-단백질에서 RNA는 가장 노력과 비용에 비해 많은 정보를 얻을 수 있는 효율적인 물질이다. 최근 인기있는 단일세포 RNA-seq은 세포형과 세포의 위치까지 고려하는 막강한 도구로 떠올랐지만, 여전히 bulk RNA-seq은 가격과 시료 요구사항, 기술난이도와 간편함 면에서 비할 수 없이 좋은 접근성을 제공한다.

이 강의에서는 시퀀싱을 이용한 유전자 발현량 분석에 대해 전혀 경험이 없는 입문자를 위하여 라이브러리 준비 과정, 유전자 발현량 분석의 기초 개념에 대해 소개한다. 또한, 발현량 및 전사체 분석을 통해 달성하고자 하는 여러 생물학적 가설에서 사용되는 전형적인 분석 순서에 대해 입문자가 전체 윤곽을 쉽게 잡을 수 있도록 기본 시나리오들을 몇 가지 소개한다.

강의는 다음의 내용을 포함한다:

- 시퀀싱과 RNA-seq 라이브러리 준비 방법
- 유전자 발현량 추정 방법과 정규화
- 변하는 유전자 분석
- 주요 가설과 데이터 형태에 따른 이후 분석 기법

\* 강의 난이도: 초급 (일반생물학 수준의 지식이 필요함)

\* 강의: 장혜식 교수 (서울대학교 생명과학부)

# Curriculum Vitae

**Speaker Name: Hyesik Chang, Ph.D.**



## ► Personal Info

Name Hyesik Chang  
Title Assistant Professor  
Affiliation Seoul National University

## ► Contact Information

Address 1 Gwanak-ro Gwanak-gu, Seoul, 08826  
Email [hyeshik@snu.ac.kr](mailto:hyeshik@snu.ac.kr)  
Website <https://qbio.io>

---

## Research Interest

High-throughput sequencing, post-transcriptional regulation, RNA-protein interaction

## Educational Experience

1998–2007 B.S.E. in Information and Industrial Engineering, Yonsei University, Korea  
2007–2009 M.S.E. in Bio and Brain Engineering, KAIST, Korea  
2009–2014 Ph.D. in Biological Sciences, Seoul National University, Korea

## Professional Experience

2001–2005 Software Developer, Solution Development Team, LinuxKorea, Inc.  
2014–2019 Research Assistant Professor, IBS Center for RNA Research, Seoul National University  
2018– Research Fellow, Center for RNA Research, Institute for Basic Science  
2019– Assistant Professor, School of Biological Sciences, Seoul National University

## Selected Publications (5 maximum)

1. D. Kim, J.-Y. Lee, J.-S. Yang, J. W. Kim, V. N. Kim, and H. Chang. (2020) "The Architecture of SARS-CoV-2 Transcriptome." *Cell*, 181(4):914–921.
2. H. Chang<sup>1</sup>, J. Yeo<sup>1</sup>, J.-G. Kim, H. Kim, M. Lee, J. Lim, H. H. Kim, J. Ohk, H.-Y. Jeon, H. Lee, H. Jung, K.-W. Kim, and V. N. Kim. (2018) "Terminal uridylyltransferases execute programmed clearance of maternal transcriptome in vertebrate embryos." *Molecular Cell*, 70:72–82.e7.
3. J. Lim<sup>1</sup>, M. Ha<sup>1</sup>, H. Chang<sup>1</sup>, S. C. Kwon, D. K. Simanshu, D. J. Patel, and V. N. Kim. (2014) "Uridylation by TUT4 and TUT7 marks mRNA for degradation." *Cell*, 159(6):1365–1376.
4. H. Chang<sup>1</sup>, J. Lim<sup>1</sup>, M. Ha, and V. N. Kim. (2014) "TAIL-seq: genome-wide determination of poly(A) tail length and 3' end modifications." *Molecular Cell*, 53(6):1044–1052.
5. J. Cho<sup>1</sup>, H. Chang<sup>1</sup>, S. C. Kwon, B. Kim, Y. Kim, J. Choe, M. Ha, Y. K. Kim, and V. N. Kim. (2012) "LIN28A is a suppressor of ER-associated translation in embryonic stem cells." *Cell*, 151(4):765–777.

<sup>1</sup>Co-first authors

# KSBi-BIML 2023

## Introduction to Gene Expression Analysis

장혜식

서울대학교 생명과학부

## 강의 대상과 목표

- 이 강의는 일반생물학 수준의 생물학 용어에 대한 이해가 필요합니다.
- 이 강의는 프로그래밍, 데이터 분석, 시퀀싱, 대규모 실험 경험 이 전혀 없는 분들을 위해 준비 되었지만, 종종 컴퓨터와 통계학 개념이 일부 필요할 수 있습니다.
- 이 강의를 수강하면 RNA-Seq을 실제로 배우기 전에 어떤 것인지 전체적 시야에서 얕게 이해할 수 있게 됩니다.



What are typical examples of applications of high-throughput gene expression analysis using RNA-seq?



High-throughput gene expression analysis using RNA-seq is a powerful tool for studying gene expression on a genome-wide scale. Some typical applications of RNA-seq include:

1. Identifying differentially expressed genes between different experimental conditions or samples, such as comparing cancerous and non-cancerous tissue.
2. Characterizing gene expression patterns in various cell types or tissues, such as identifying genes that are specifically expressed in brain cells.
3. Discovering novel transcripts and alternative splicing events, which can lead to the identification of new protein-coding genes and non-coding RNAs.
4. Quantifying gene expression levels for functional genomics studies, such as identifying genes involved in specific biological pathways or determining the effects of genetic mutations.
5. Assessing the quality and integrity of RNA samples, such as detecting contamination or degradation of samples.
6. Study of gene expression in non-model organisms, such as studying the gene expression of a plant species which is not well studied.
7. Study of the expression of viral genes in host organisms, such as studying the expression of viral genes in a host organism infected with a virus.

3



Can you provide more detailed use cases for the first item?



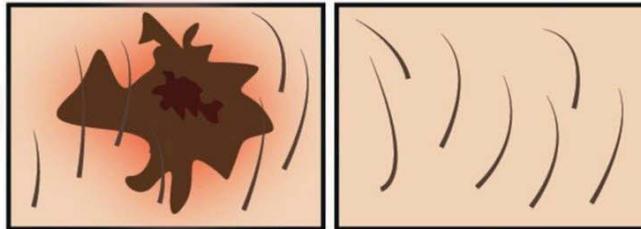
Sure, here are a few more detailed examples of applications of RNA-seq for identifying differentially expressed genes:

1. Cancer research: Comparing the gene expression profiles of cancerous and non-cancerous tissue samples can reveal genes that are specifically overexpressed or underexpressed in cancer cells. This can lead to the identification of novel biomarkers for cancer diagnosis and the discovery of new targets for cancer therapy.
2. Developmental biology: Comparing the gene expression profiles of different stages of development can reveal genes that are specifically expressed during certain stages of development, such as genes involved in the formation of specific organs or tissues.
3. Drug discovery: Comparing the gene expression profiles of cells treated with different drugs can reveal genes that are regulated by the drugs, which can aid in the discovery of new drug targets and the development of more effective drugs.
4. Immunology: Comparing the gene expression profiles of immune cells from healthy individuals and individuals with specific immune disorders can reveal genes that are specific to the disorder, which can aid in the discovery of new therapeutic targets.
5. Environmental studies: Comparing the gene expression profiles of organisms exposed to different environmental conditions can reveal genes that are specific to certain environmental conditions. This can aid in the discovery of new biomarkers for environmental exposure and the identification of new targets for environmental conservation.

4

1. Identifying differentially expressed genes between different experimental conditions or samples, such as comparing cancerous and non-cancerous tissue.

### Samples of interest

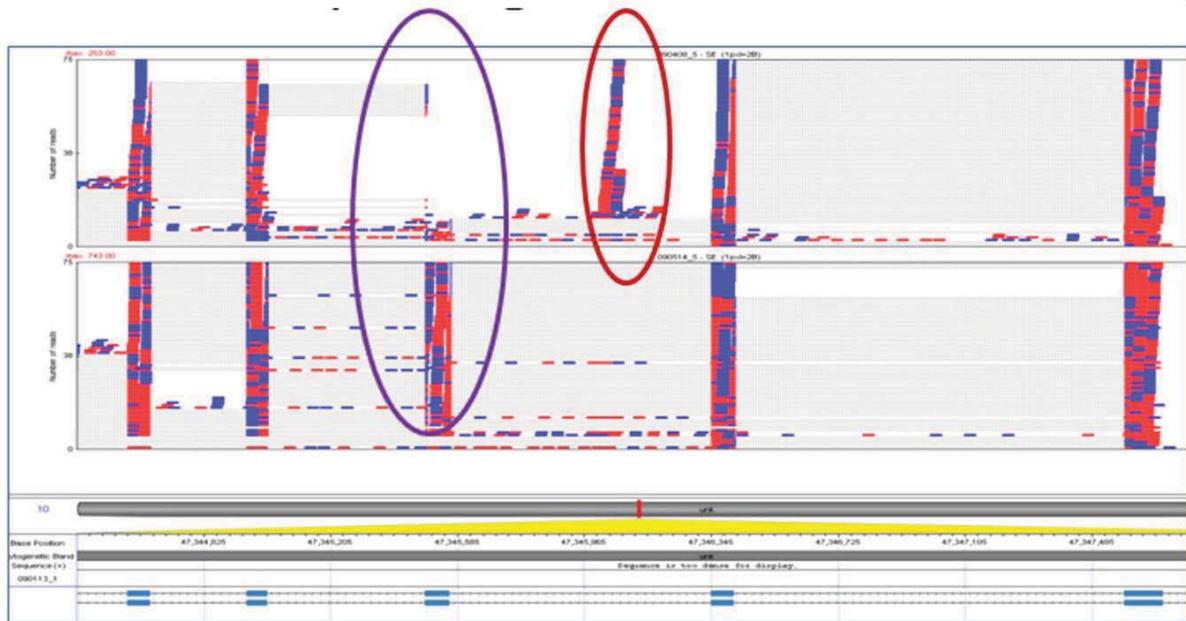


Condition 1  
(e.g. tumor)

Condition 2  
(e.g. normal)

2. Characterizing gene expression patterns in various cell types or tissues, such as identifying genes that are specifically expressed in brain cells.

3. Discovering novel transcripts and alternative splicing events, which can lead to the identification of new protein-coding genes and non-coding RNAs.



(c) SEQOME

7

4. Quantifying gene expression levels for functional genomics studies, such as identifying genes involved in specific biological pathways or determining the effects of genetic mutations.



freepik, used under a premium license

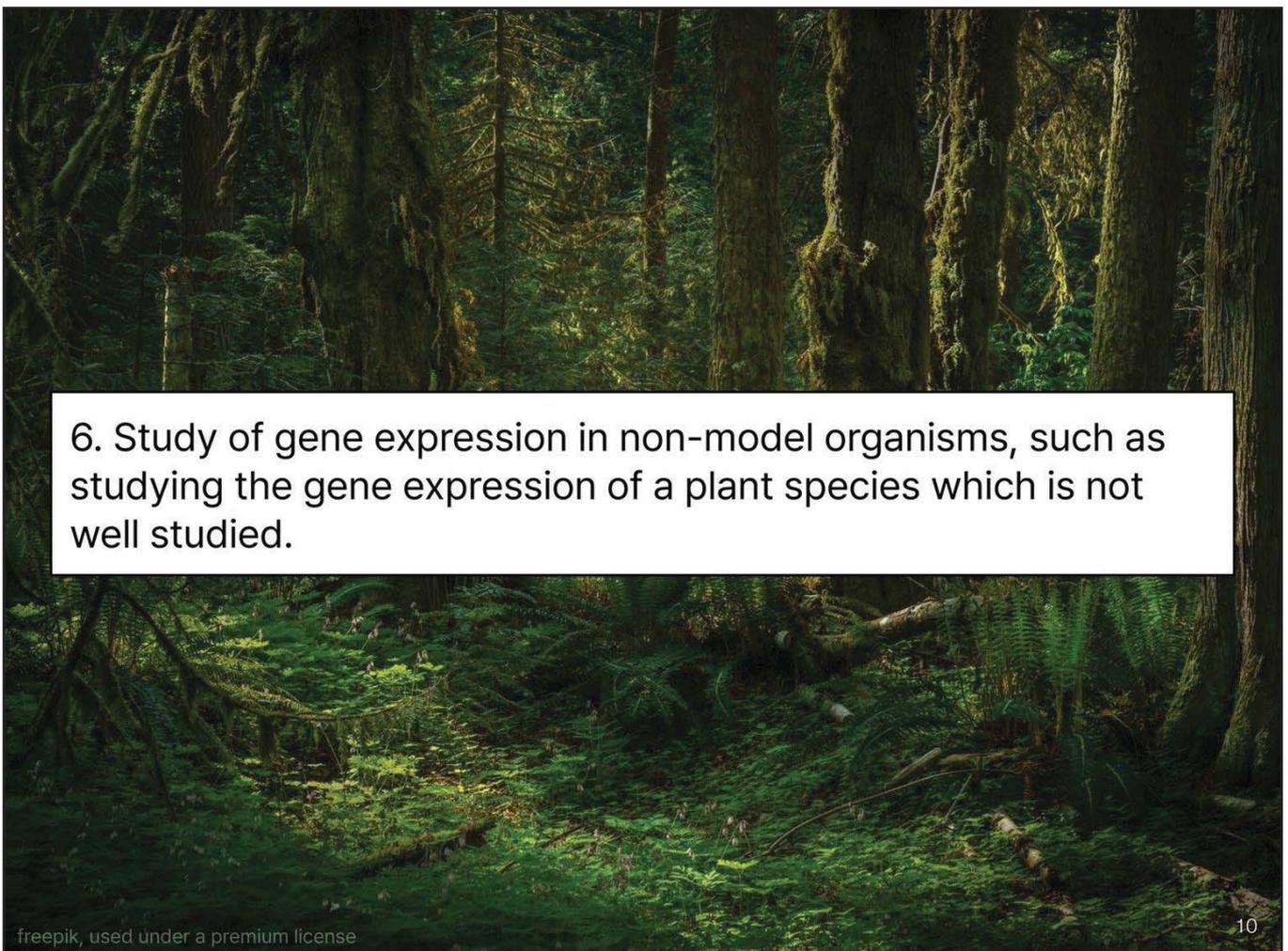
8



5. Assessing the quality and integrity of RNA samples, such as detecting contamination or degradation of samples.

freepik, used under a premium license

9



6. Study of gene expression in non-model organisms, such as studying the gene expression of a plant species which is not well studied.

freepik, used under a premium license

10

7. Study of the expression of viral genes in host organisms, such as studying the expression of viral genes in a host organism infected with a virus.

(c) NIAID

**왜 DNA가 아니라 RNA를?**

# 왜 DNA가 아니라 RNA를?

- [기능 연구] Genome은 대체로 변치 않지만 유전자 발현은 실험 조건에 따라 크게 변한다.
  - 약물을 처리한 것과 처리하지 않은 셀라인
  - Wild-type과 knock-out 마우스
- Genome 서열에서 transcript 서열을 정확히 예측하기는 어렵다.
  - RNA-seq의 등장으로 gene annotation이 혁명적으로 바뀌었다.
- 어떤 분자적 특징들은 RNA에서만 관찰 가능하다.
  - Alternative isoforms, fusion transcripts, RNA editing
- 돌연변이 중 단백질 서열에 변화를 주지 않는 것을 해석할 때
  - 조절 요소의 돌연변이는 어떤 mRNA 아형이 얼마나 발현될 지 영향을 준다.
- 많은 단백질 코딩 체세포 돌연변이 중 우선순위를 매길 때
  - 발현되지 않는 유전자에 있는 돌연변이는 재미가 덜할 가능성이 높다.
  - Wild-type allele에서만 발현이 되면 loss-of-function mutation일 가능성이 높다.
  - 돌연변이 allele이 발현되면, drug target 후보로 고려해 볼 수 있다.

From a slide by M. Griffith, O. Griffith, and F. Yousif

13



(c) Phenomenex

14

# 왜 단백질이 아니라 RNA를?

	RNA 프로파일링	단백질 프로파일링
포괄하는 범위	거의 발현되는 전체 RNA를 어렵지 않게 볼 수 있음	일반적으로 부분적으로만 관찰 가능
Alternative splicing	세세히 살펴볼 수 있음	훨씬 어렵다
실험적 문제	실험적 편차, 편향성, 오류가 적다	상대적으로 편차, 편향성, 오류가 크다
Scalability	가격이 싸고, 대량으로 하기 쉽다	대량으로 늘리기에 부담이 된다
Novel transcript	심지어 genome이 없더라도 가능한 경우가 있다	높은 품질의 reference가 필요하다

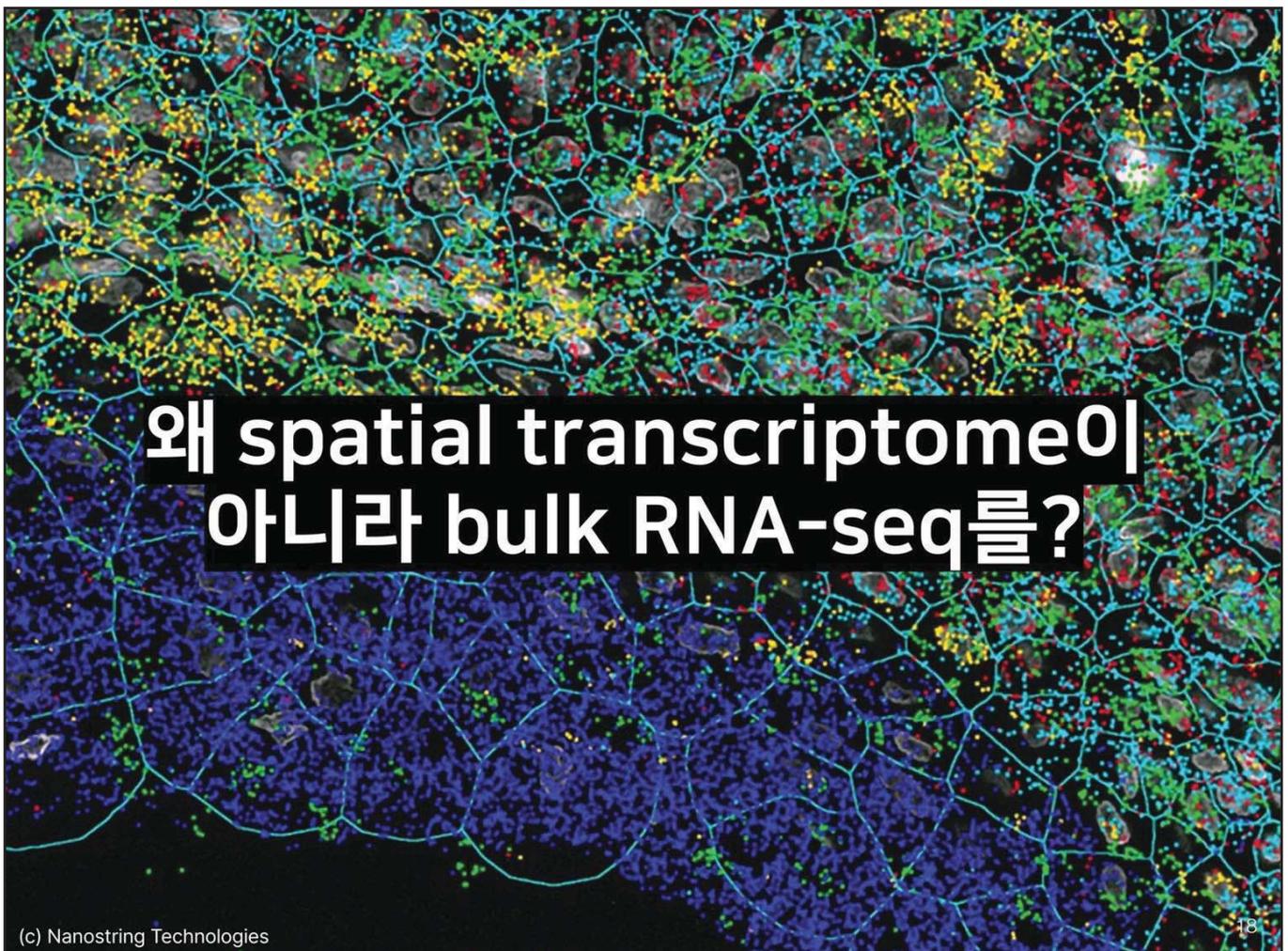
15

# 왜 single-cell이 아니라 bulk를?

# 왜 single-cell이 아니라 bulk를?

	Bulk RNA-seq	Single-cell RNA-seq
비용, 수고	간편하고 저렴한 편	여러모로 부담됨
시료의 품질과 양	시료 품질이 안 좋거나 양이 적을 때에도 가능	품질과 양에서 비교적 까다로운 기준을 충족해야 함
Profile에 반영되는 상태	외부 영향이 적은 스냅샷을 얻을 수 있음	Sorting하지 않더라도 lysis까지 최소한 시간 단위 지연이 있음
Scalability	샘플 수가 많을 때도 다루기 쉬움	샘플 수를 늘리는 데 한계가 있음
Customization	보고자 하는 대상에 따라 변화를 주기 쉬움	프로토콜 변경에 한계가 있음

17



# 왜 spatial transcriptome이 아니라 bulk RNA-seq를?

(c) Nanostring Technologies

18

# 왜 spatial transcriptome이 아니라 bulk RNA-seq을?

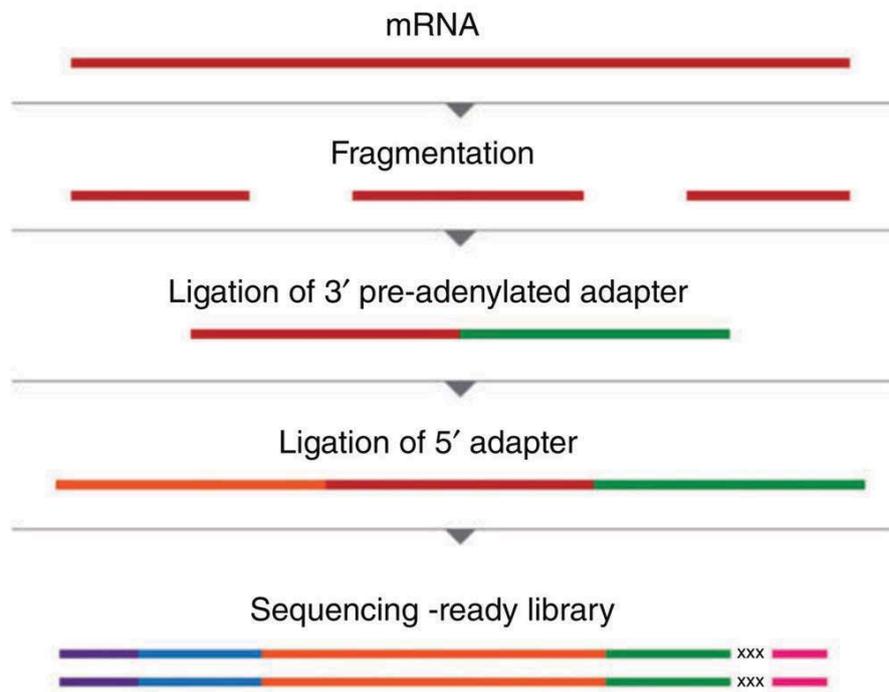
	Bulk RNA-seq	Spatial Transcriptomics
비용, 수고	간편하고 저렴한 편	여러모로 부담됨
시료의 품질과 양	시료 품질이 안 좋거나 양이 적을 때에도 가능	품질과 양에서 비교적 까다로운 기준을 충족해야 함
포괄하는 범위	거의 발현되는 전체 RNA를 어렵지 않게 볼 수 있음	양, 길이, 위치, 서열 등에 따라 일부만 관찰 가능함
Scalability	샘플 수가 많을 때도 다루기 쉬움	샘플 수를 늘리는 데 한계가 있음
Customization	요리 조리 바꾸기 쉬움	까다로운 제약 사항이 많음

19

## RNA 프로파일링 라이브러리 제작 단계

20

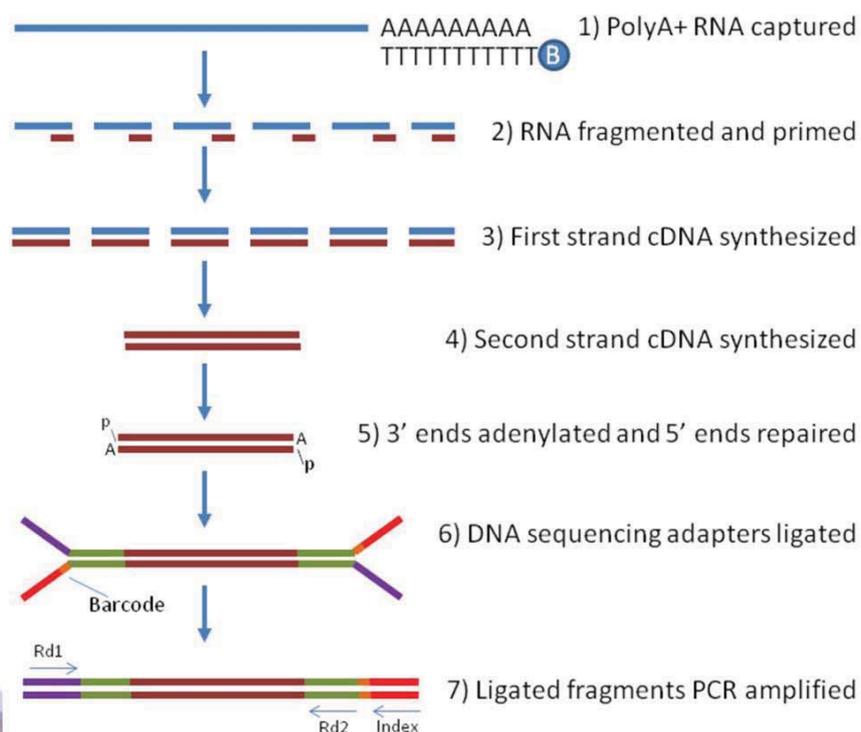
# Illumina TruSeq Small RNA



Hrdlickova, Toloue, Tian (2017) doi:10.1002/wrna.1364

21

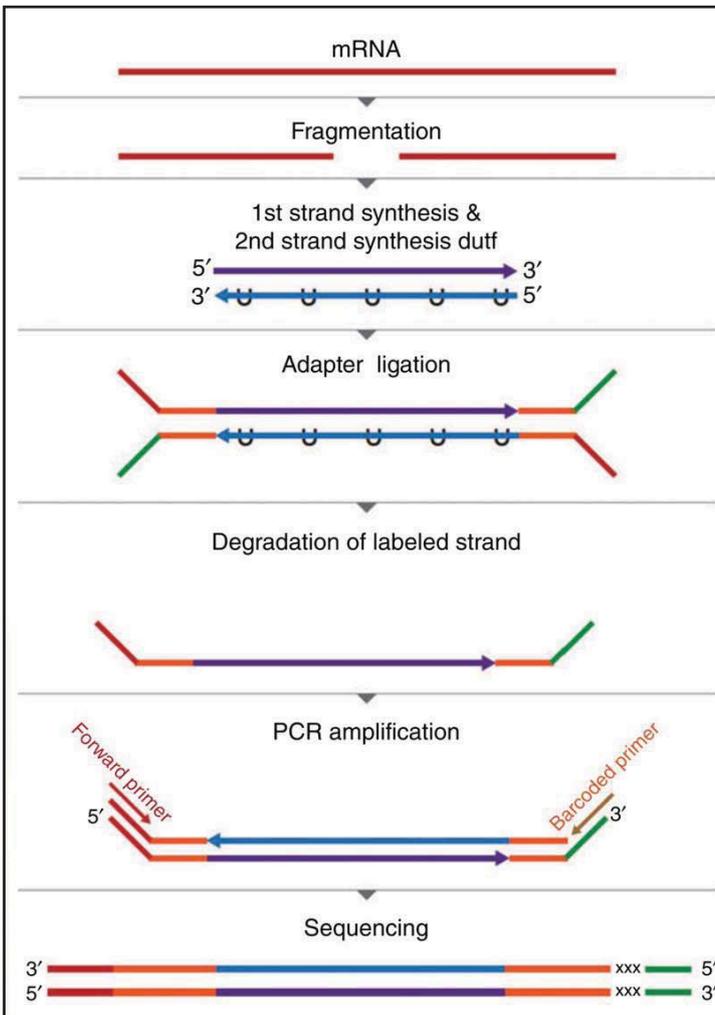
# Vanilla RNA-seq



Corney, Basturea (2013) doi:10.13070/mm.en.3.203

22

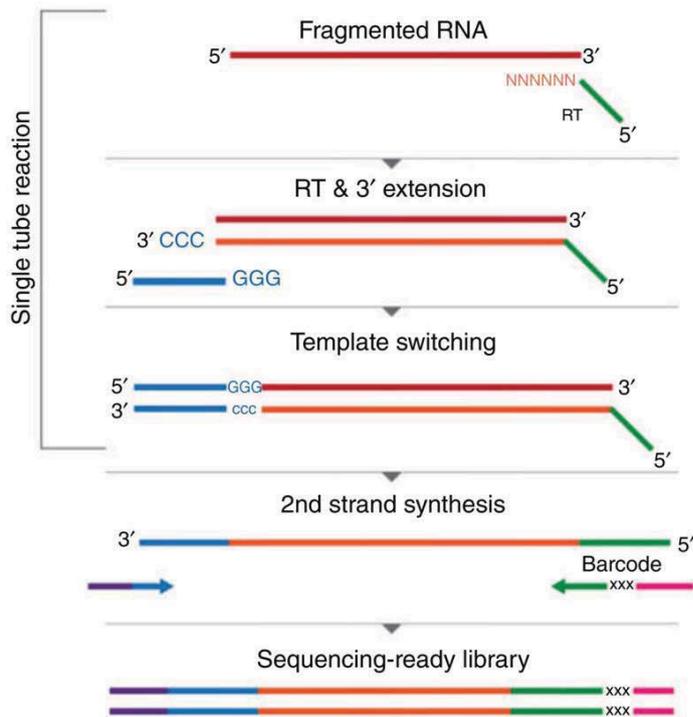
# Stranded RNA-seq (dUTP-based)



Hrdlickova, Toloue, Tian (2017) doi:10.1002/wrna.1364

23

# The Peregrine Method



Hrdlickova, Toloue, Tian (2017) doi:10.1002/wrna.1364

24



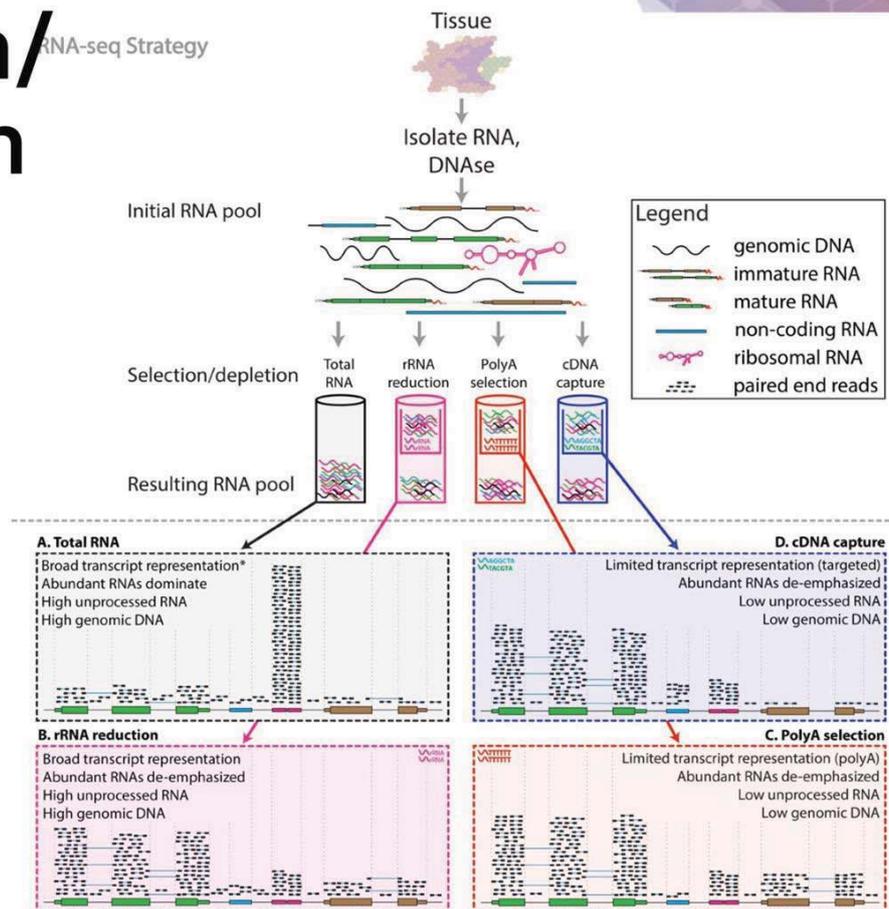
# RNA 분석에서 신경쓰는 점

- 샘플의 상태
  - 순도: 화학적 순도, 세포형의 순도 등
  - 품질: RNA integrity
  - 양: < 1ng?, 10-50ng, 50-500ng, ≥500ng
- RNA서열은 유전체에서 긴 intron이 사이에 들어있는 짧은 exon들이 연결되어 나타남.
  - 유전체에 매핑가능한 실험 설계인지 미리 고려해야 함.
- RNA의 양은 매우 크게 차이남.
  - 대략  $10^5$ 에서  $10^7$  정도의 차이
  - RNA 시퀀싱은 random sampling로 돌아간다. 매우 많이 발현되는 유전자가 대부분의 리드를 차지한다.
    - Ribosomal RNA와 mitochondrial RNA
    - Ribosomal protein mRNA
  - RNA는 크기도 다양하다.
    - Small RNA는 완전히 다른 방법으로 준비한다.
    - 긴 RNA를 poly(A) selection하면 3' 편향성이 발생한다.
- RNA는 DNA보다 잘 깨진다.

From a slide by M. Griffith, O. Griffith, and F. Yousif

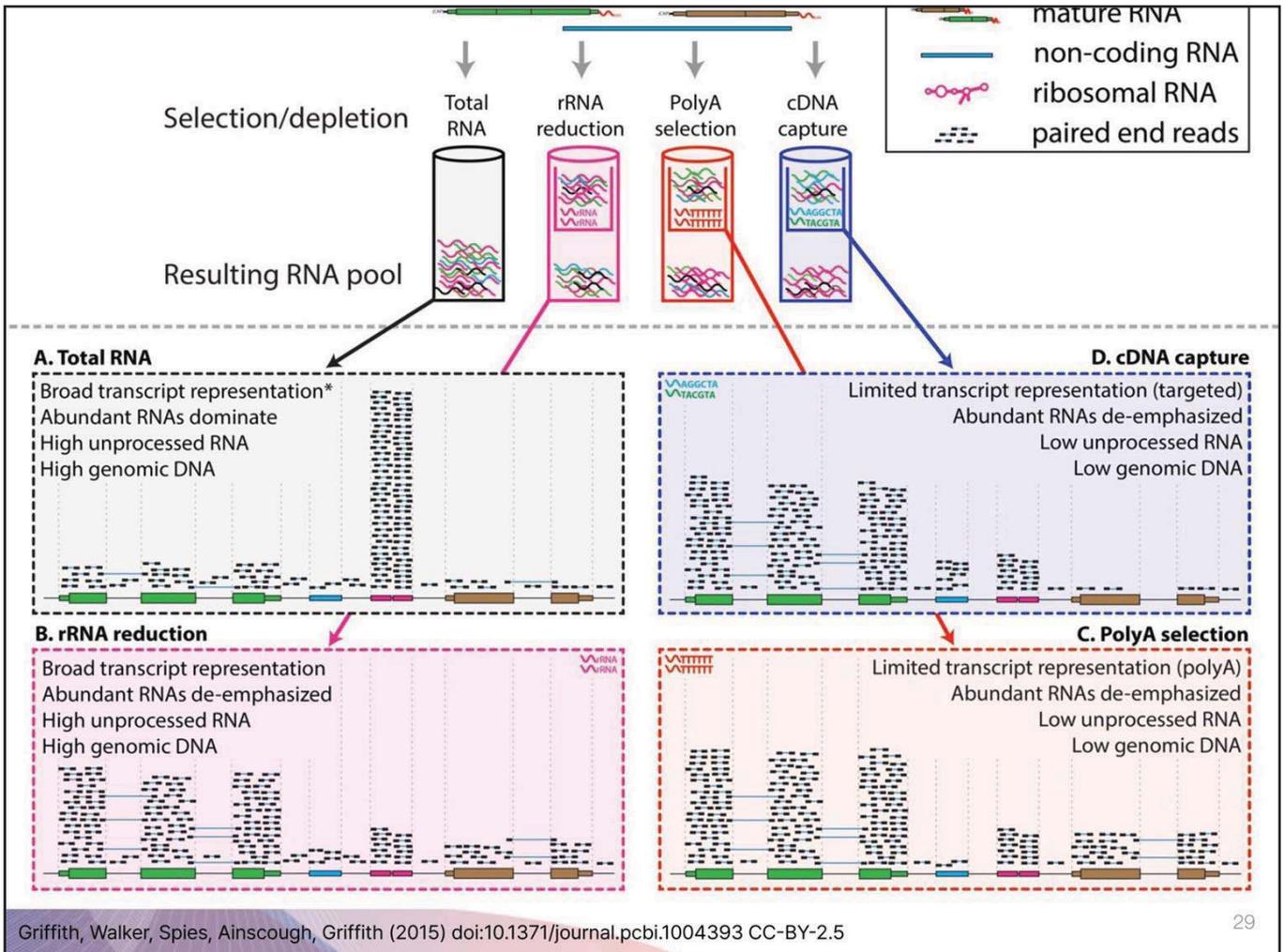
27

## Selection/Depletion RNA-seq Strategy



Griffith, Walker, Spies, Ainscough, Griffith (2015) doi:10.1371/journal.pcbi.1004393 CC-BY-2.5

28

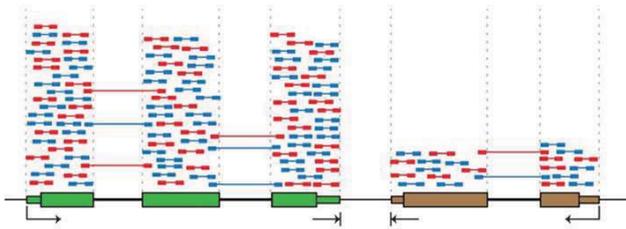


# Stranded & Unstranded

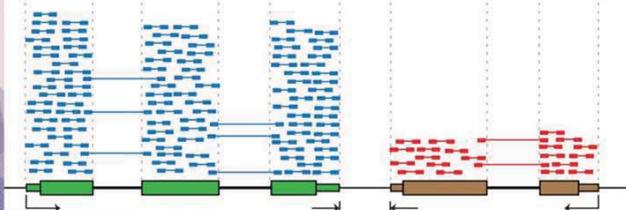
## A. Depiction of cDNA fragments from an unstranded library

**Legend**

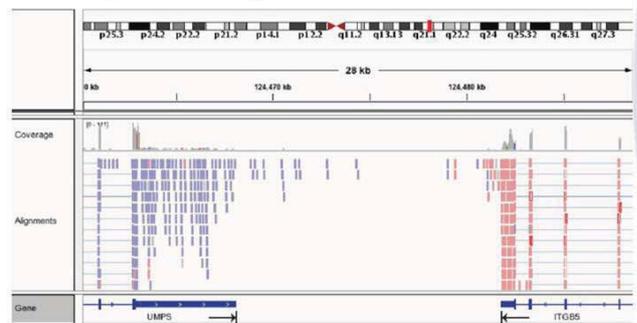
- Transcription start site and direction
- ⌊ PolyA site (transcription end)
- Read sequenced from positive strand (forward)
- Read sequenced from negative strand (reverse)

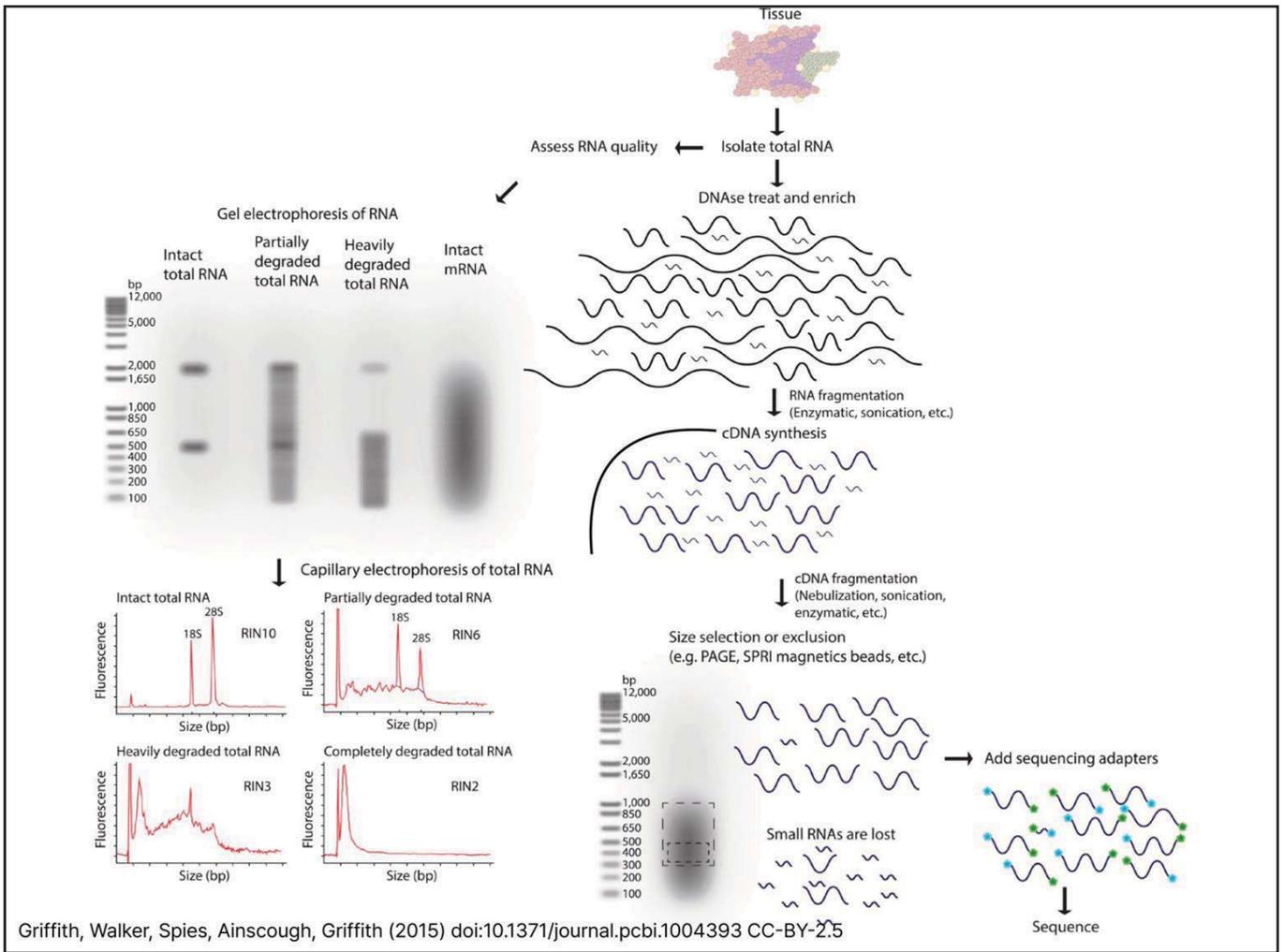


## B. Depiction of cDNA fragments from a stranded library



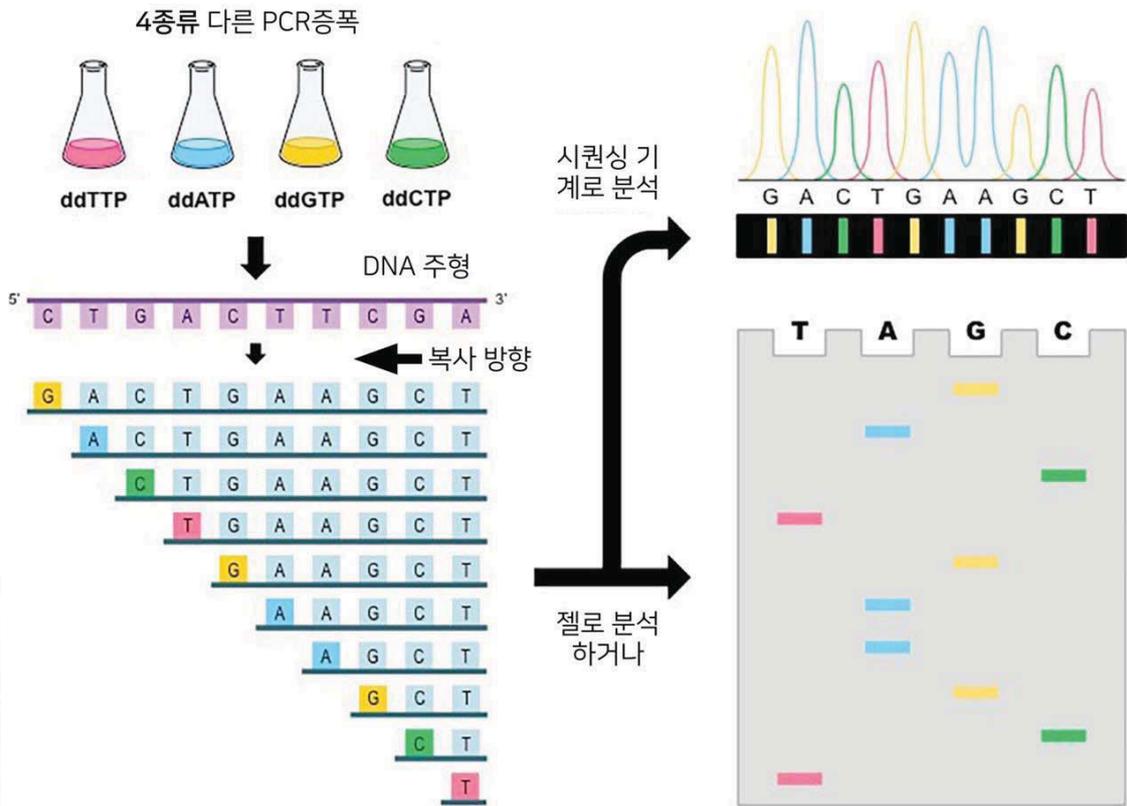
## C. Viewing strand of aligned reads in IGV





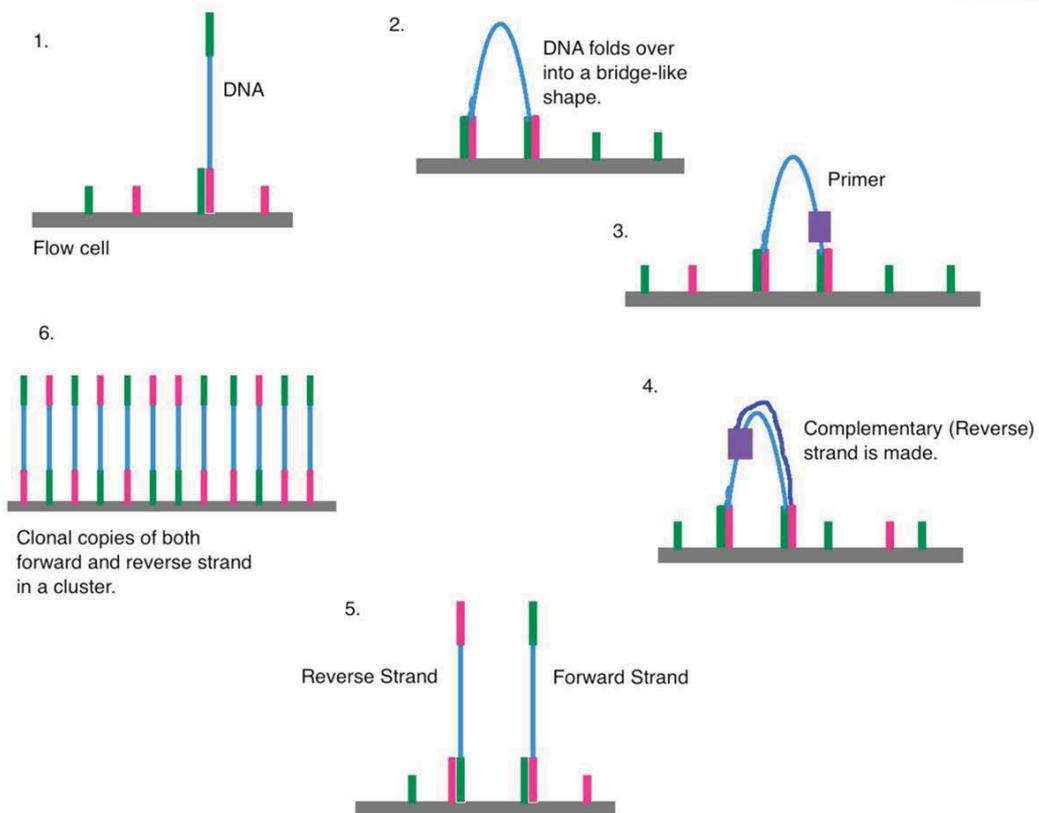
## 시퀀싱 방법

# Sanger Sequencing



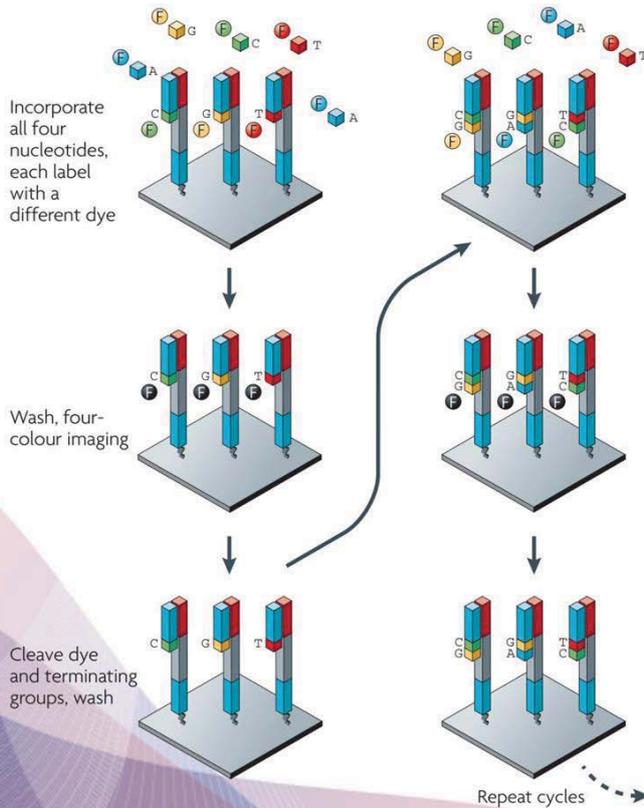
39

# Solexa: Cluster Generation

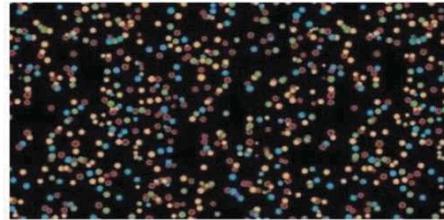


40

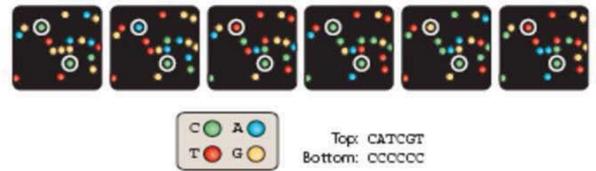
# Solexa: Sequencing Cycles



4 different images merged



6 cycles w/ base-calling



Metzker (2010) doi:10.1038/nrg2626

41

# Solexa: Paired-end Sequencing



(c) Illumina

42

# More about RNA-seq libraries...

(3) Adding TSO for second strand synthesis:

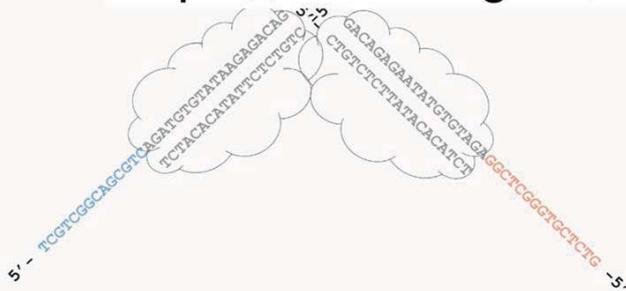
```
5' - AAGCAGTGGTATCAACGCAGAGTACATGGGXXXXXXXXXXXXXXXXXXXX(A)n
      <-----CCXXXXXXXXXXXXXXXXXXXX(T)36CATGAGACGCAACTATGGTGACGAA -5'
```

(4) Adding ISPCR for single primer cDNA amplification: (i.e. semi-suppressive PCR)

```
5' - AAGCAGTGGTATCAACGCAGAGT----->
5' - AAGCAGTGGTATCAACGCAGAGTACATGGGXXXXXXXX...XXXXXX(pA)GTACTCTGCGTTGATACCACTGCTT
      TTGTCACCATAGTTGCTCTCATGTACCCXXXXXXXX...XXXXXX(jT)CATGAGACGCAACTATGGTGACGAA -5'
      <-----TGAGACGCAACTATGGTGACGAA -5'
```

(5) Nextera tagmentation on amplified cDNA (will create 9-bp gap):

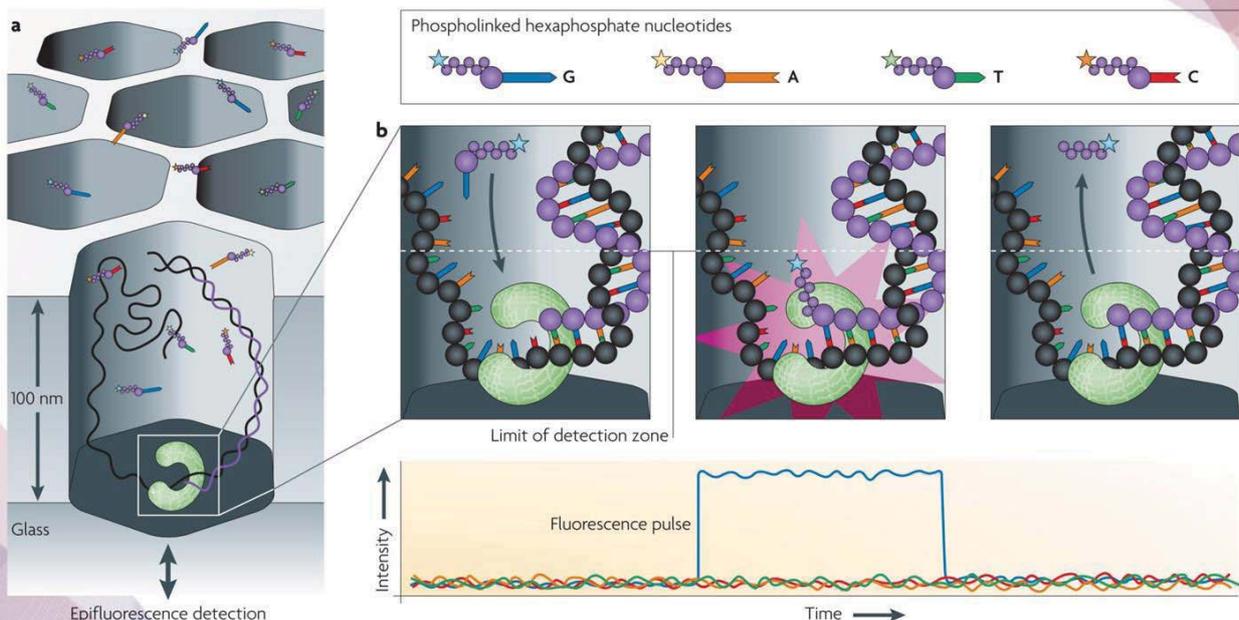
[https://teichlab.github.io/scg\\_lib\\_structs/](https://teichlab.github.io/scg_lib_structs/)



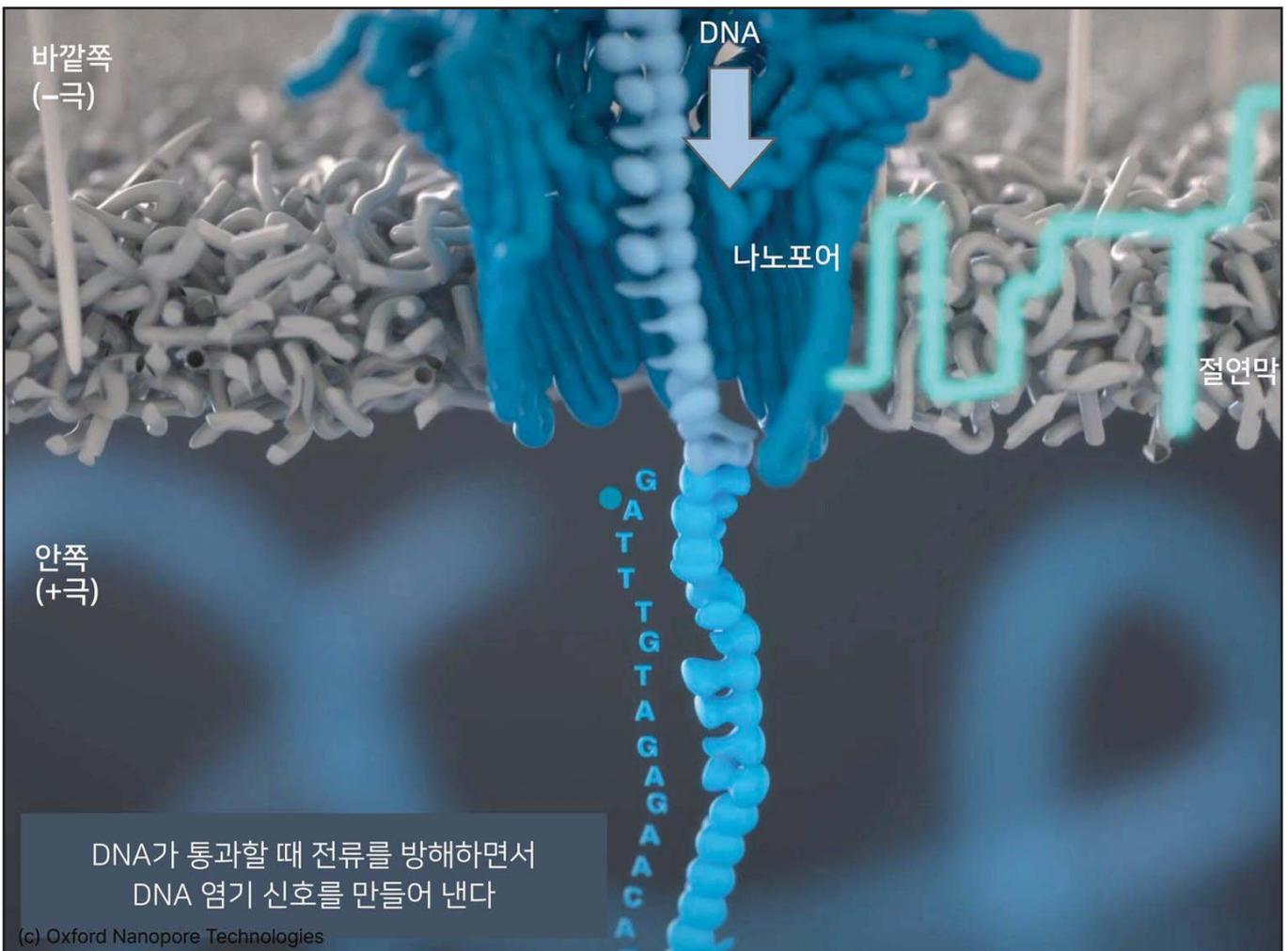
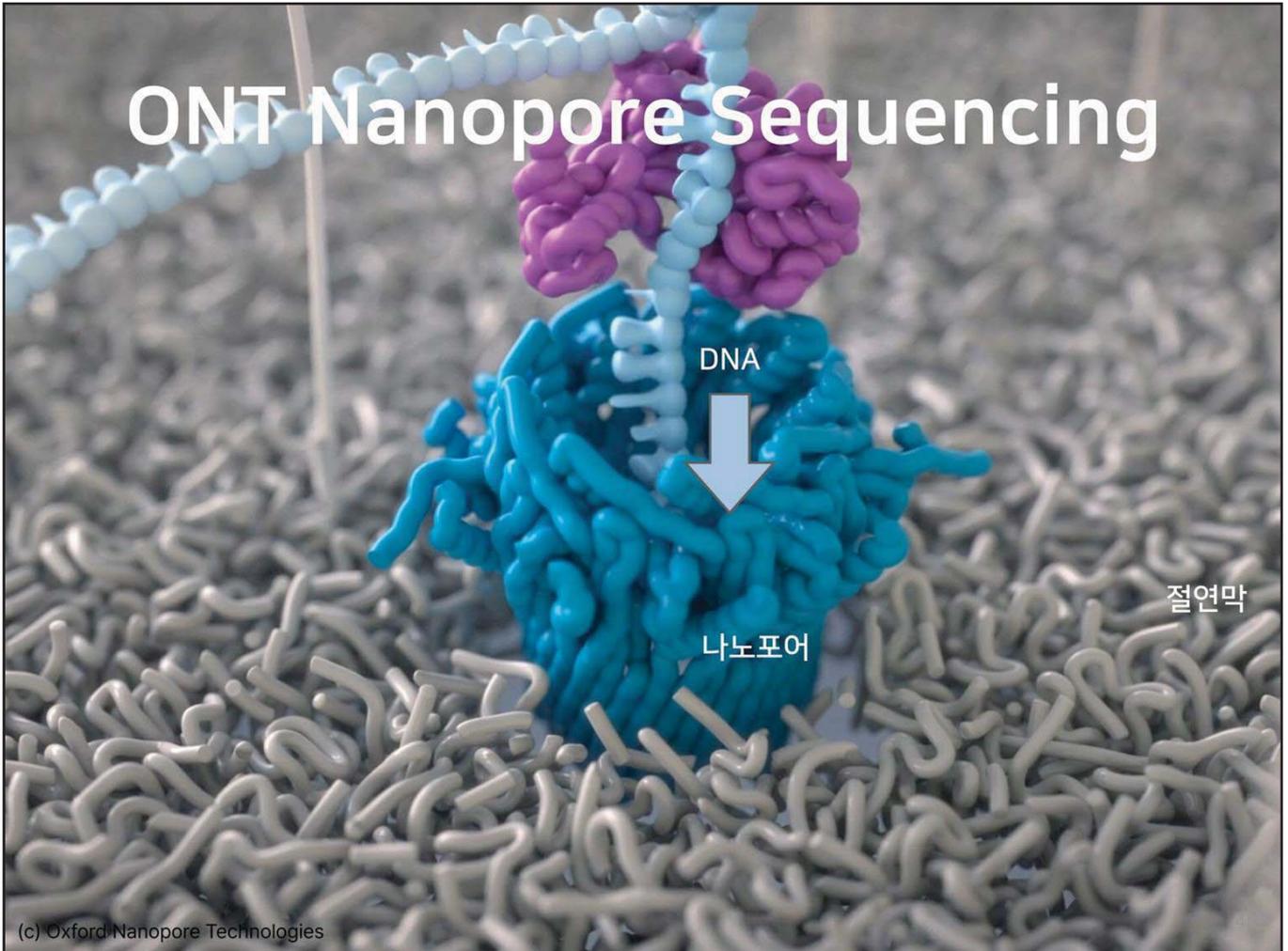
Product 1 (s5 at both ends, not amplifiable due to semi-suppressive PCR):

```
5' - TCCTCGGCAGCGCTCAGAGTGTGTATAAGAGACAGXXXXXXXXXXXX...XXX CTGTCTTTATACATCT
      TCTACACATATTCTCTGTC XXX...XXXXXXXXXXXXGACAGAGAATATGTGTAGACTGCGACGGCTGCT -5'
```

# PacBio Real-time Sequencing



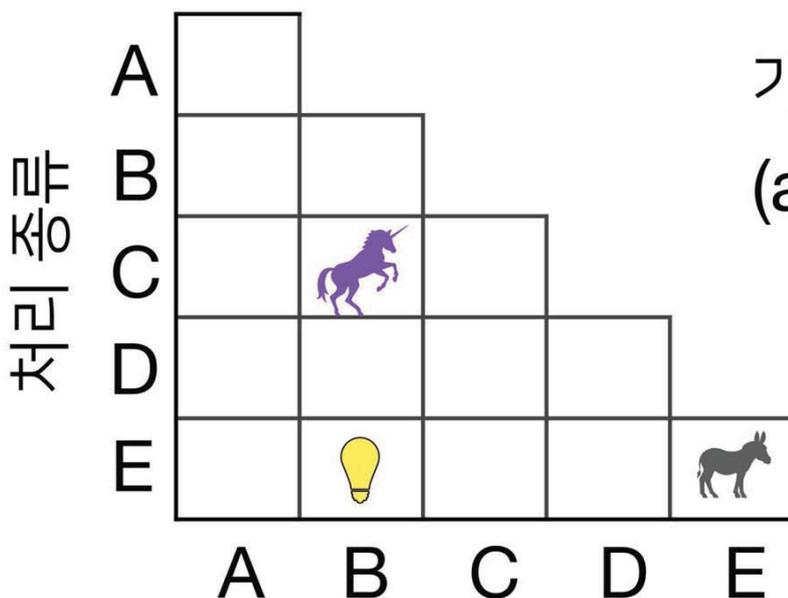
# ONT Nanopore Sequencing



# RNA-Seq 실험 설계

47

## 어떤 통찰을 얻는 것이 목표인가?



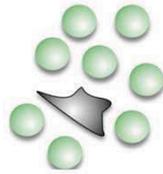
가능한 총 비교쌍

$$(a)(a-1)/2$$

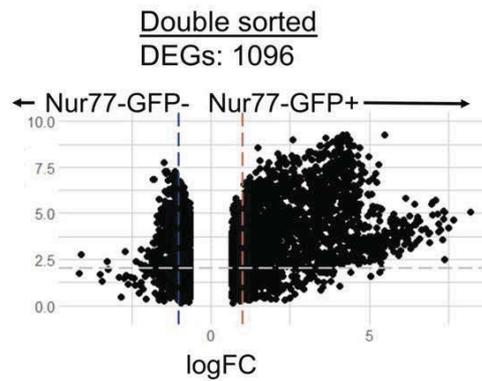
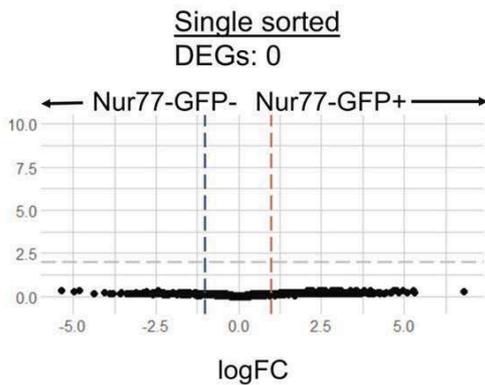
$$(5)(5-1)/2 = 10$$

# 신호 대 잡음비를 고려하자

Condition A



Condition B



Data courtesy of Lindsey Shallberg (Hunter lab)  
Berry, Amorim, Berry, Syrett, English, Beiting (2021) doi:10.1128/mBio.01214-21

# 더 많이 읽을까, 반복을 늘릴까?

*BIOINFORMATICS* **DISCOVERY NOTE**

Vol. 30 no. 3 2014, pages 301–304  
doi:10.1093/bioinformatics/btt688

Gene expression

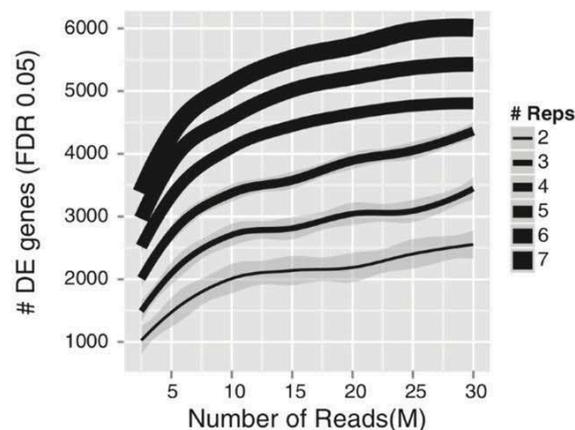
Advance Access publication December 6, 2013

## RNA-seq differential expression studies: more sequence or more replication?

Yuwen Liu<sup>1,2</sup>, Jie Zhou<sup>1,3</sup> and Kevin P. White<sup>1,2,3,\*</sup>

<sup>1</sup>Institute of Genomics and Systems Biology, <sup>2</sup>Committee on Development, Regeneration, and Stem Cell Biology and <sup>3</sup>Department of Human Genetics, University of Chicago, Chicago, IL 60637, USA

Associate Editor: Janet Kelso



Berry, Amorim, Berry, Syrett, English, Beiting (2021) doi:10.1128/mBio.01214-21

# 더 많이 읽을까, 반복을 늘릴까?

BIOINFORMATICS DISCOVERY NOTE

Vol. 30 no. 3 2014, pages 301–304  
doi:10.1093/bioinformatics/btt688

Gene expression

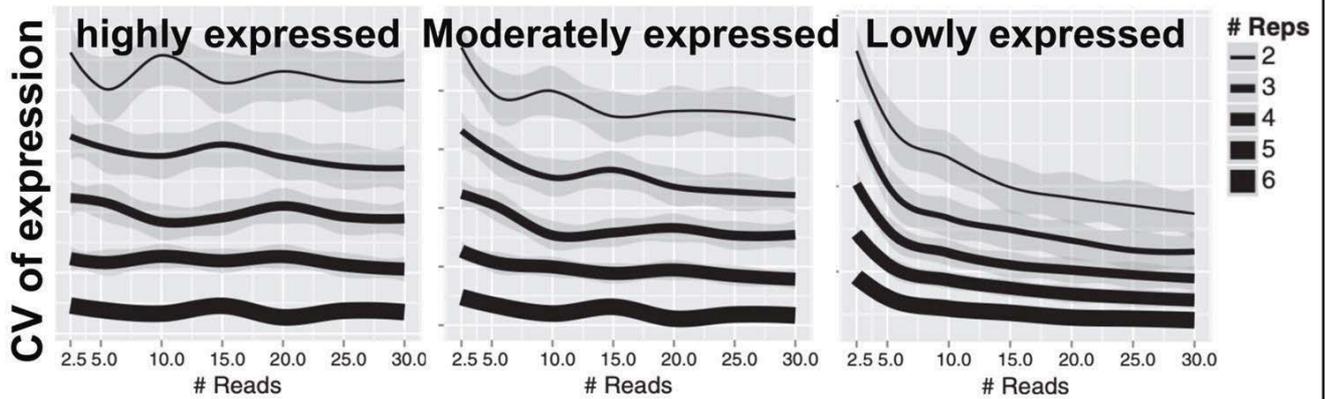
Advance Access publication December 6, 2013

## RNA-seq differential expression studies: more sequence or more replication?

Yuwen Liu<sup>1,2</sup>, Jie Zhou<sup>1,3</sup> and Kevin P. White<sup>1,2,3,\*</sup>

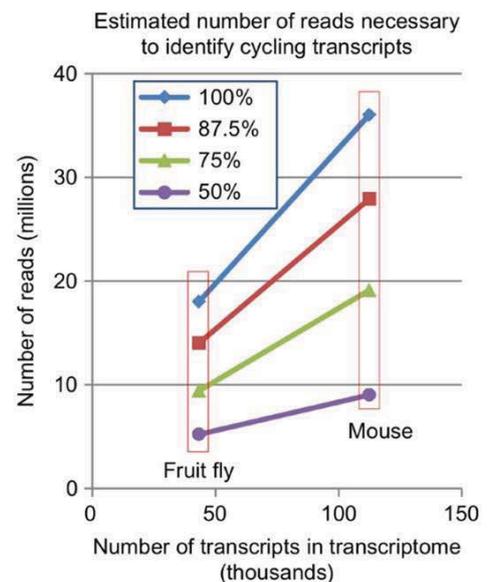
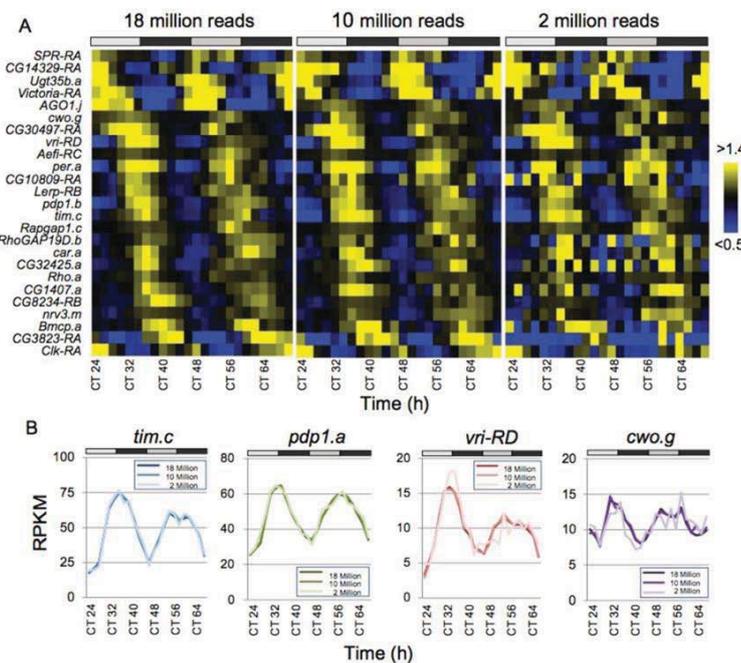
<sup>1</sup>Institute of Genomics and Systems Biology, <sup>2</sup>Committee on Development, Regeneration, and Stem Cell Biology and <sup>3</sup>Department of Human Genetics, University of Chicago, Chicago, IL 60637, USA

Associate Editor: Janet Kelso



Berry, Amorim, Berry, Syrett, English, Beiting (2021) doi:10.1128/mBio.01214-21

# 얇아도 분석할 수 있는 것이 있다



Li, Grant, Hogenesch, Hughes (2015) doi:10.1016/bs.mie.2014.10.020

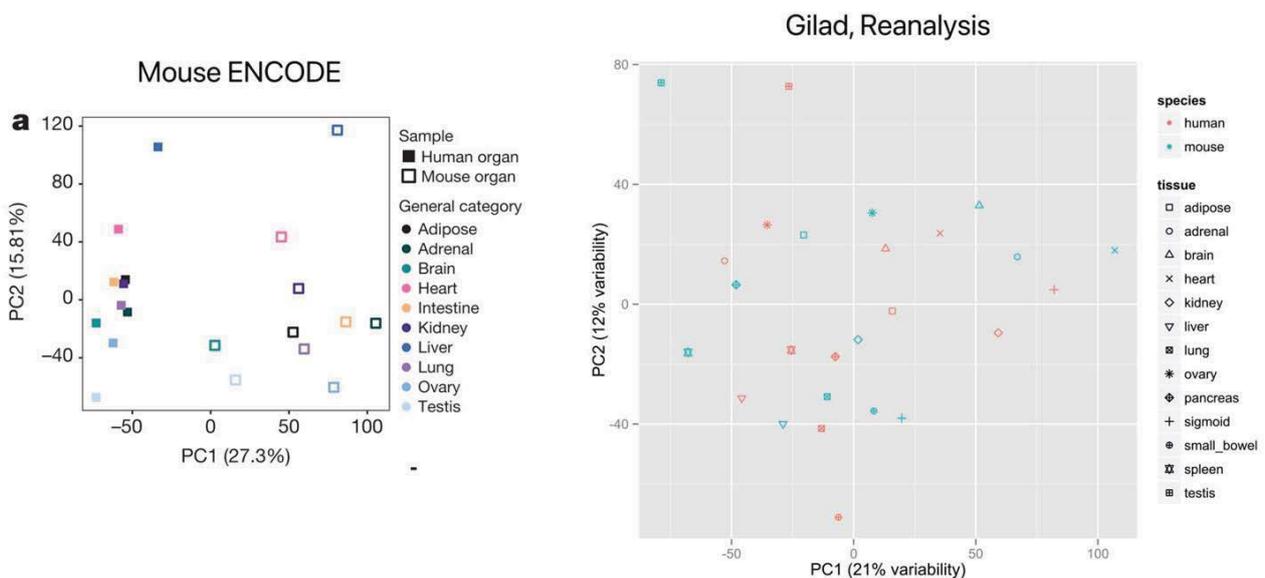
## How many biological replicates are needed in an RNA-seq experiment and which differential expression tool should you use?

TABLE 2. A summary of the recommendations of this paper

	Agreement with other tools <sup>a</sup>	WT vs. WT FPR <sup>b</sup>	Fold-change threshold (T) <sup>c</sup>	Tool recommended for: (# good replicates per condition) <sup>d</sup>		
				≤3	≤12	>12
DESeq	Consistent	Pass	0	-	-	Yes
			0.5	-	Yes	Yes
			2.0	Yes	Yes	Yes
EBSeq	Consistent	Pass	0	-	-	Yes
			0.5	-	Yes	Yes
			2.0	Yes	Yes	Yes
edgeR (exact)	Consistent	Pass	0	-	-	Yes
			0.5	Yes	Yes	Yes
			2.0	Yes	Yes	Yes
Limma	Consistent	Pass	0	-	-	Yes
			0.5	-	Yes	Yes
			2.0	Yes	Yes	Yes
cuffdiff	Consistent	Fail				
DESeq2	Consistent	Fail				
BaySeq	Inconsistent	Pass				
edgeR (GLM)	Inconsistent	Pass				
DEGSeq	Inconsistent	Fail				
NOISeq	Inconsistent	Fail				
PoissonSeq	Inconsistent	Fail				
SAMSeq	Inconsistent	Fail				

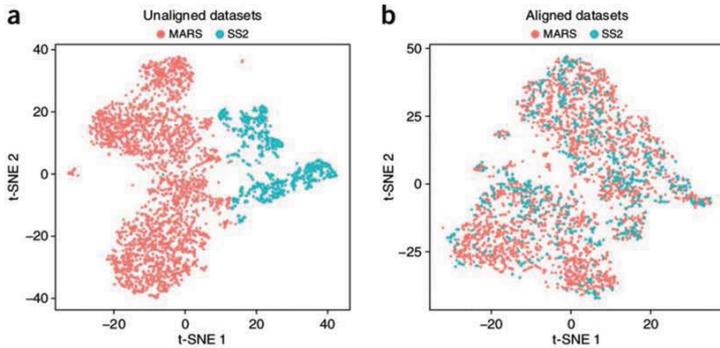
Schurch, Schofield, Gierlinski, Cole, Sherstnev, Singh, Wrobel, Gharbi, Simpson, Owen-Hughes, Blaxter, Barton (2016)  
doi:10.1261/rna.053959.115

## 같이 실험할 것들을 잘 정해야 한다



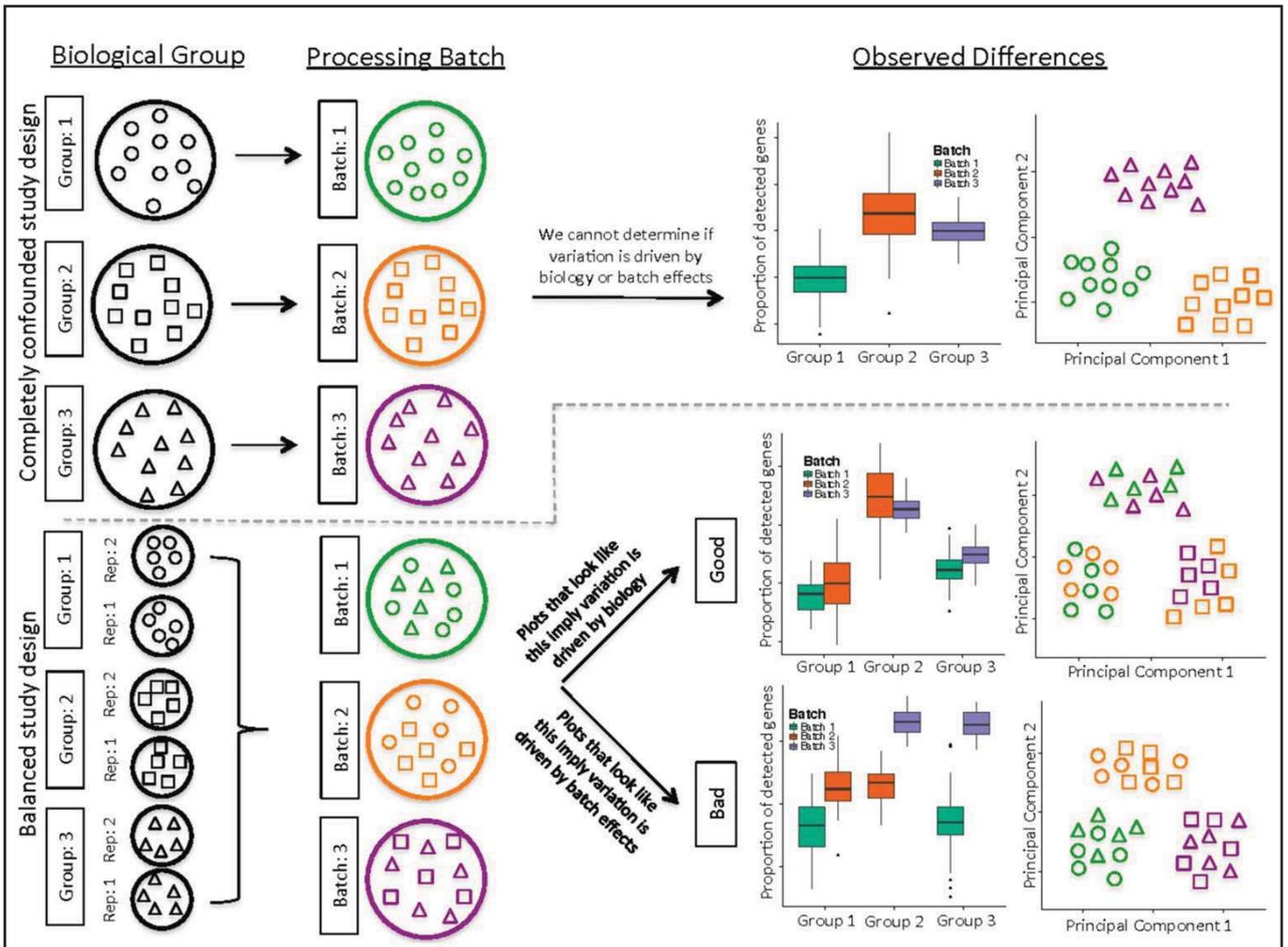
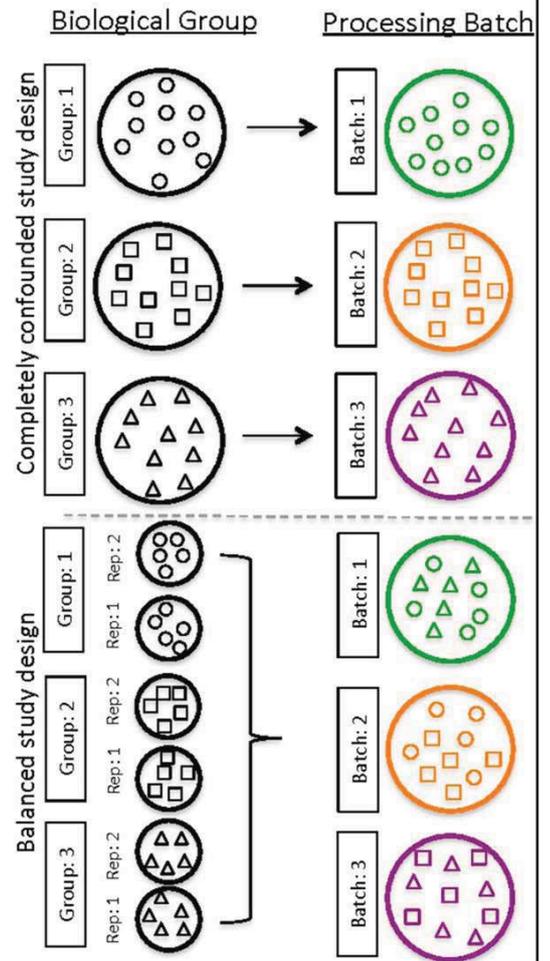
Yue et al. (2014) doi:10.1038/nature13992  
Gilad and Mizrahi-Man (2015) doi:10.12688/f1000research.6536.1

# Batch Effect



t-SNE plots from two datasets using different technologies (MARS and SS2)

Butler et al. (2018) doi:10.1038/nbt.4096  
Hicks, Teng, Irizarry (2015) doi:10.1093/biostatistics/kxx053



# RNA-Seq 데이터 처리 파이프라인

57

## Mapping/Alignment

### De novo assembly

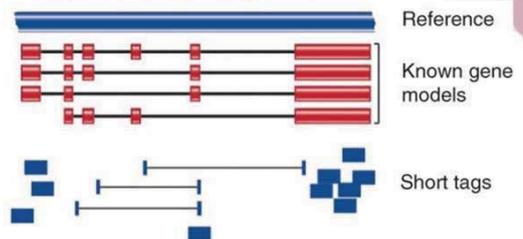


Assemble transcripts from overlapping tags



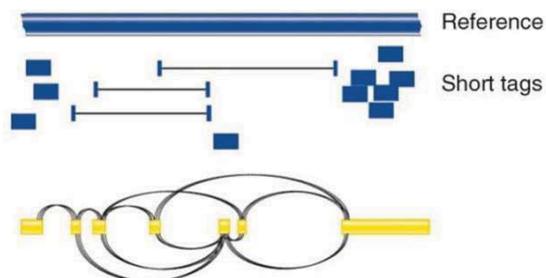
Optional: align to genome to get exon structure

### Align to transcriptome



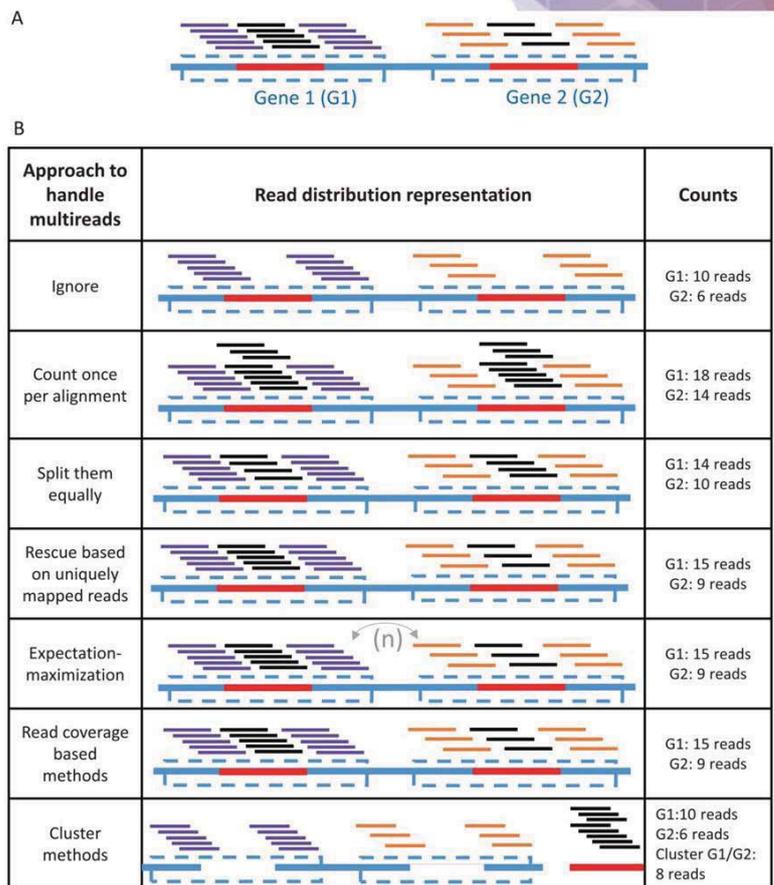
Use known and/or predicted gene models to examine individual features

### Align to reference genome



Infer possible transcripts and abundance

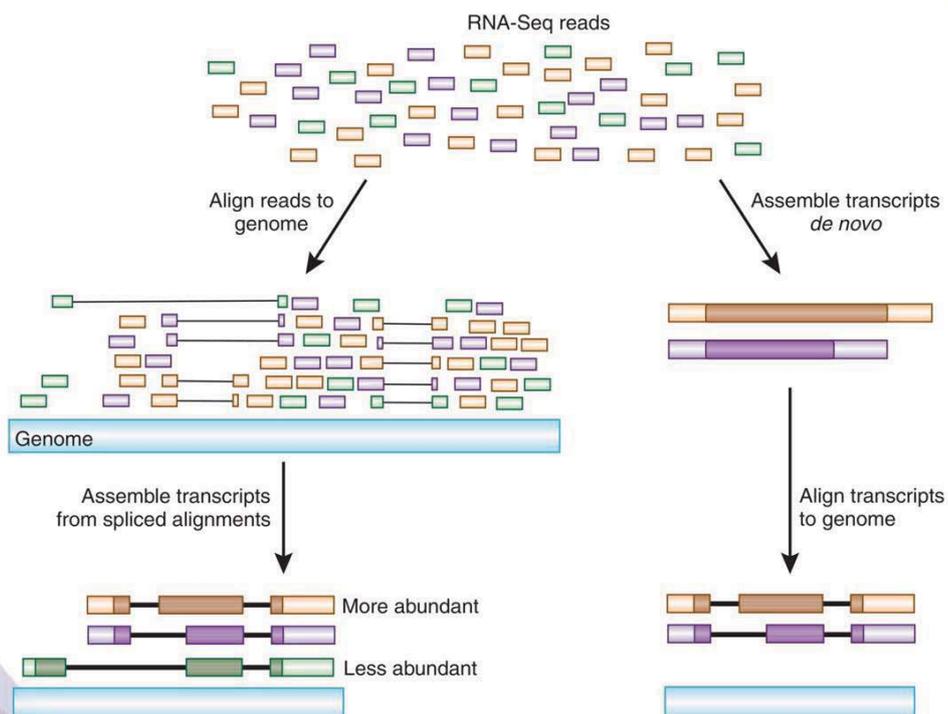
# Handling Multi-mapping



Deschamps-Francoeur, Simoneau, Scott (2020) doi:10.1016/j.csbj.2020.06.014

59

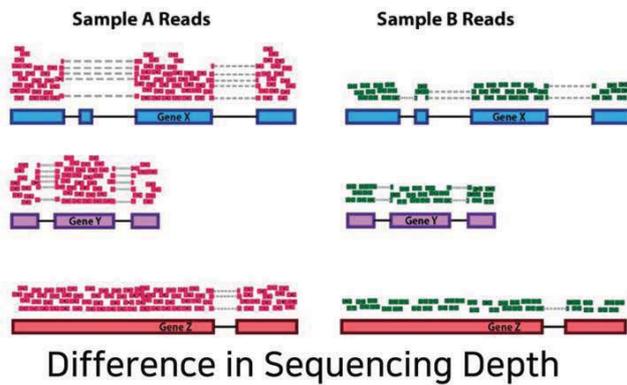
# Transcript Assembly



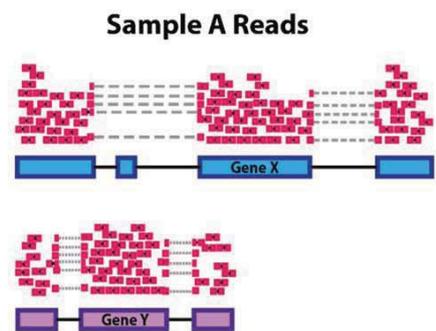
Haas, Zody (2010) doi:10.1038/nbt0510-421

60

# Expression Estimation



Difference in Gene Length

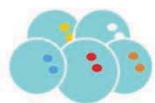


(c) Harvard Chan Bioinformatics Core

61

# Differential Expression Analysis

Condition A

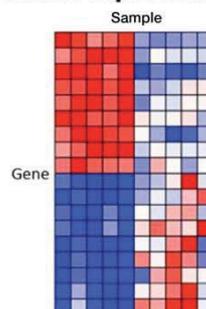


Condition B



RNA-seq

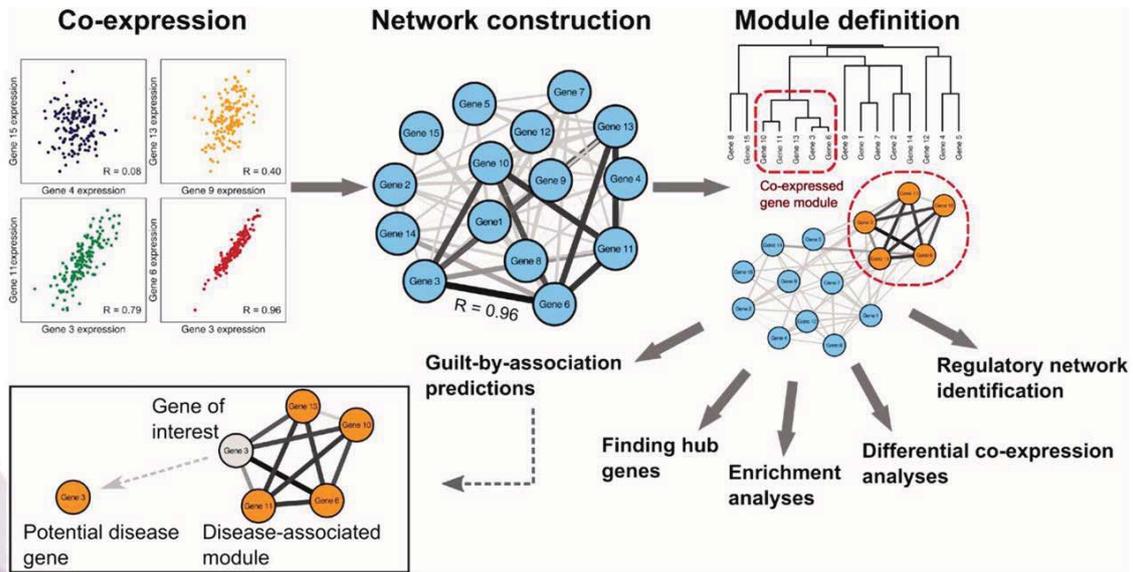
Differential expression analysis



Miao, Zhang (2016) doi:10.1007/s40484-016-0089-7

62

# Downstream Analyses



Van Dam, Vösa, van der Graaf, Franke, de Magalhães (2017) doi:10.1093/bib/bbw139

63

RNA-Seq  
정량 기법

64

“일반적으로 사용되는 모든 mRNA 정량 기술(qPCR, microarray, RNA-seq의 RPKM등)은 모두 **상대 몰 농도 (molar concentration)**에 최대한 비례관계를 가지는 것을 목표로 한다.”

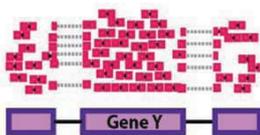
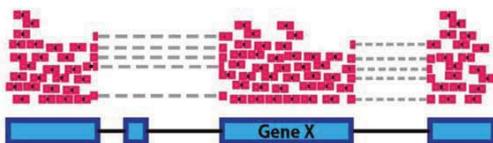
- Wagner, *Theories in Biosci.*, 2012

Recited from a slide by Daniel Beiting

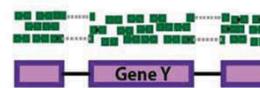
65

## 총 리드수 차이 보정

Sample A Reads



Sample B Reads



(c) Harvard Chan Bioinformatics Core

66

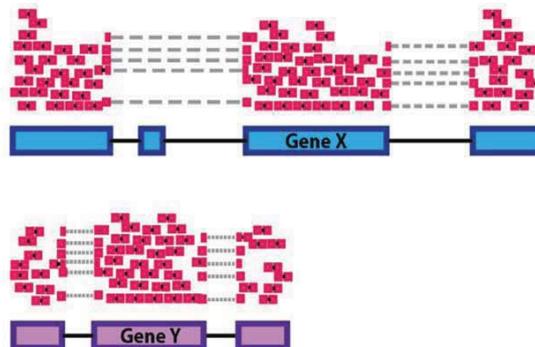
# 총 리드수 차이 보정

$$\text{RPM} = \frac{\text{\# reads mapped to genomic region}}{\text{(total \# reads)}} \times 10^6$$

(reads per million)

# 길이 차이 보정

## Sample A Reads



# 길이 차이 보정

$$\text{RPKM} = \frac{\text{\# reads mapped to genomic region}}{(\text{region length in kb}) (\text{total \# reads})} \times 10^6$$

(reads per kilobase, per million reads)

# RNA 길이와 정말 비례할까?

$$L_{\text{effective}} = L_{\text{actual}} - L_{\text{fragment}} + 1$$

<i>transcript</i>	ATGCGTAACATG	$L_{\text{actual}}=12$	$L_{\text{effective}} = 10$
<i>fragment</i>	NNN	$L_{\text{fragment}}=3$	



# 길이 차이 보정

$$\text{RPKM} = \frac{\text{\# reads mapped to genomic region}}{(\text{region length in kb}) (\text{total \# reads})} \times 10^6$$

(reads per kilobase, per million reads) effective

(c) Harvard Chan Bioinformatics Core

71

## RPKM의 특이한 문제

### read counts from each gene

	Gene A 100kb	Gene B 50kb	Gene C 25kb	Gene D 5kb	Gene E 1kb	total reads
Sample 1	80	10	6	3	1	100
Sample 2	20	20	10	50	400	500

### RPKM for each gene

	Gene A 100kb	Gene B 50kb	Gene C 25kb	Gene D 5kb	Gene E 1kb	total RPKM
Sample 1	8000	2000	2400	6000	10000	28400
Sample 2	400	800	800	20000	800000	822000

Berry, Amorim, Berry, Syrett, English, Beiting (2021) doi:10.1128/mBio.01214-21

# RPKM 문제 고치기

## read counts

	Gene A <i>100kb</i>	Gene B <i>50kb</i>	Gene C <i>25kb</i>	Gene D <i>5kb</i>	Gene E <i>1kb</i>	total reads
Sample 1	80	10	6	3	1	100
Sample 2	20	20	10	50	400	500

유전자 길이로 먼저 정규화



이 숫자로 추가로 정규화



	Gene A <i>100kb</i>	Gene B <i>50kb</i>	Gene C <i>25kb</i>	Gene D <i>5kb</i>	Gene E <i>1kb</i>	total RPK
Sample 1	0.8	0.2	0.24	0.6	1	2.84
Sample 2	0.2	0.4	0.4	10	400	4.11

Berry, Amorim, Berry, Syrett, English, Beiting (2021) doi:10.1128/mBio.01214-21

# RPKM 문제 고치기

## TPM

	Gene A <i>100kb</i>	Gene B <i>50kb</i>	Gene C <i>25kb</i>	Gene D <i>5kb</i>	Gene E <i>1kb</i>	total TPM
Sample 1	281690	70423	84507	211268	352113	1000000
Sample 2	487	973	973	24331	973236	1000000



	Gene A <i>100kb</i>	Gene B <i>50kb</i>	Gene C <i>25kb</i>	Gene D <i>5kb</i>	Gene E <i>1kb</i>	total RPK
Sample 1	0.8/2.84	0.2/2.84	0.24/2.84	0.6/2.84	1/2.84	2.84
Sample 2	0.2/4.11	0.4/4.11	0.4/4.11	10/4.11	400/4.11	4.11

Berry, Amorim, Berry, Syrett, English, Beiting (2021) doi:10.1128/mBio.01214-21

# Transcripts per million (TPM)

$$\text{TPM} = \frac{\text{reads per Kb}}{\text{total RPK in sample}} \times 10^6$$

Berry, Amorim, Berry, Syrett, English, Beiting (2021) doi:10.1128/mBio.01214-21

75

## TPM도 샘플간 비교에는 적당하지 않다

### TPM

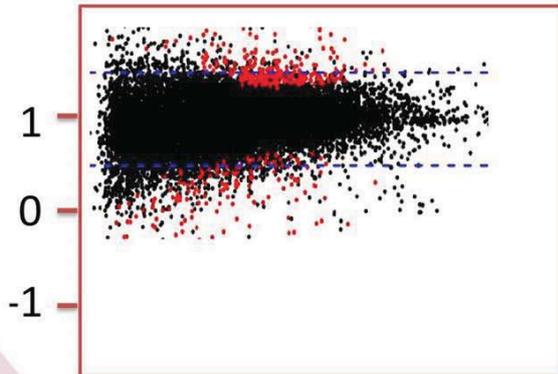
	Gene A <i>100kb</i>	Gene B <i>50kb</i>	Gene C <i>25kb</i>	Gene D <i>5kb</i>	Gene E <i>1kb</i>	total TPM
Sample 1	281690	70423	84507	211268	352113	1000000
Sample 2	487	973	973	24331	973236	1000000

Berry, Amorim, Berry, Syrett, English, Beiting (2021) doi:10.1128/mBio.01214-21

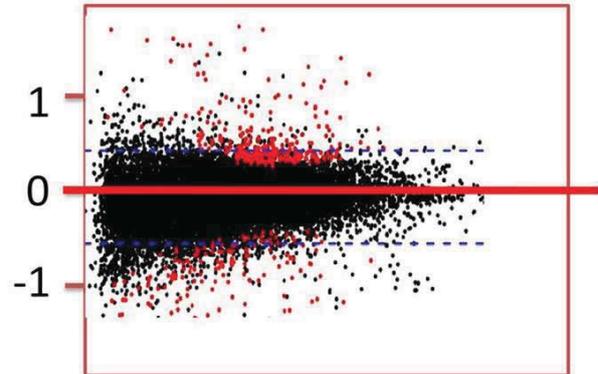
# 샘플간 비교를 위한 정규화

## MA Plots

Before normalization



After normalization



- Y axis: log ratio of expression level between two conditions;
- With the assumption that most genes are expressed equally, the log ratio should mostly be close to 0

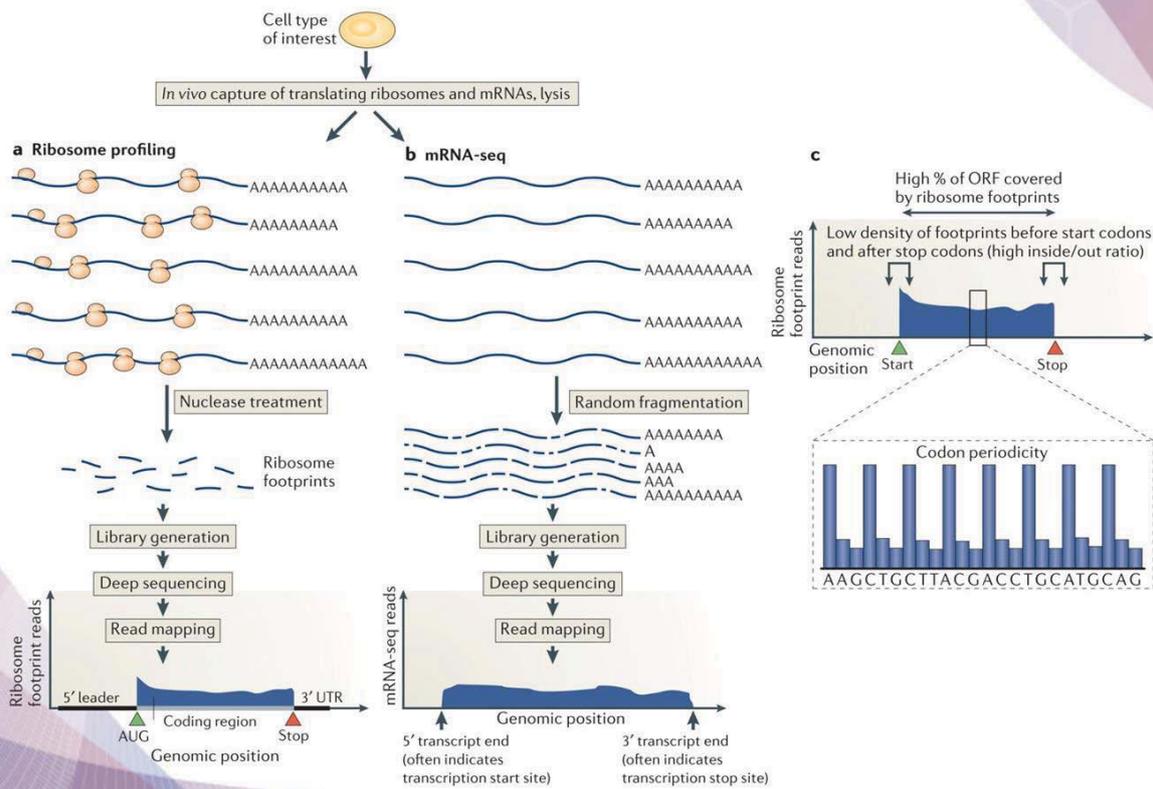
Slide from Minghui Wang, Qi Sun (Cornell University)

77

RNA-Seq  
응용 기법

78

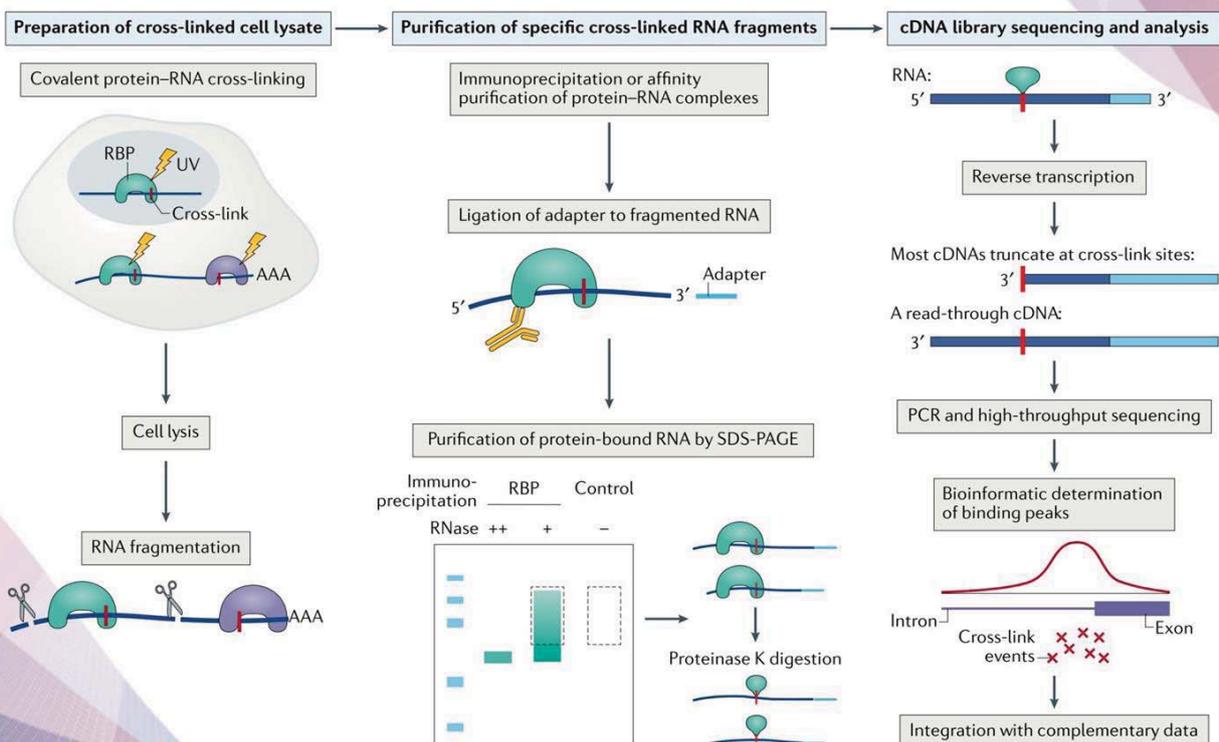
# Ribo-Seq / Ribosome Profiling



Brar, Weissman (2015) doi:10.1038/nrm4069

79

# CLIP-seq / HITS-CLIP



Hafner, Katsantoni, Köster, Marks, Mukherjee, Staiger, Ule, Zavolan (2021) doi:10.1038/s43586-021-00018-1

80

“일반적으로 사용되는 모든 mRNA 정량 기술(qPCR, microarray, RNA-seq의 RPKM등)은 모두 상대 몰 농도 (**molar concentration**)에 최대한 비례관계를 가지는 것을 목표로 한다.”

- Wagner, *Theories in Biosci.*, 2012

Recited from a slide by Daniel Beiting

81

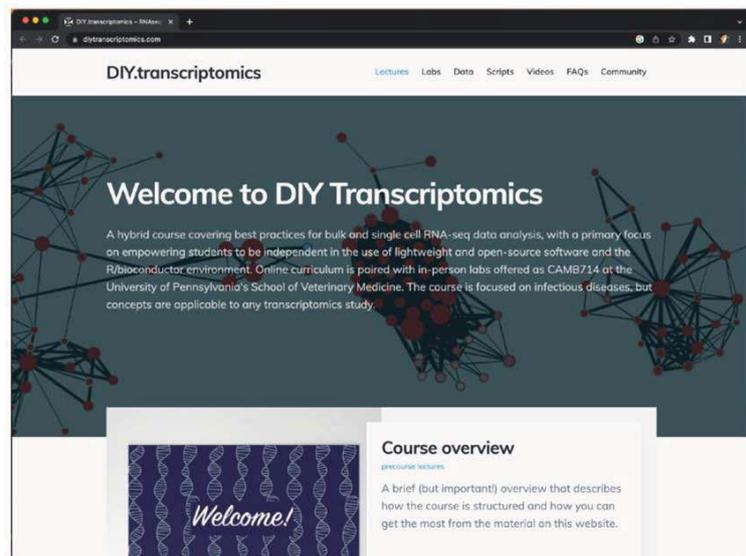
## 더 공부해 보기

DIY.transcriptomics

By Daniel Beiting

아주 가벼운 이론+  
R기반의 실습

완전 초급



<https://diytranscriptomics.com>

82

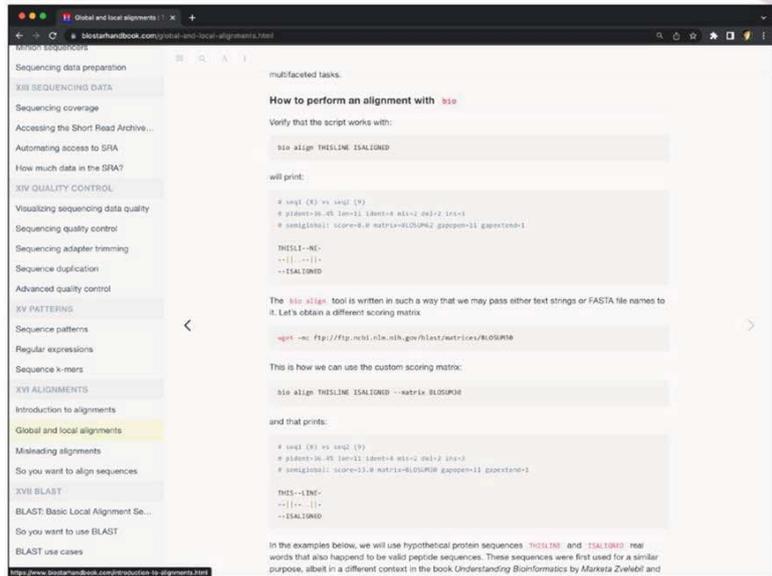
# 더 공부해 보기

Biostar Handbook  
By István Albert

아주 가벼운 이론+  
Python기반의 실습

Bioinformatics 전반

완전 초급~중급



<https://www.biostarhandbook.com/>

83

SBI 한국생명정보학회  
Korean Society for Bioinformatics

## KSBI-BIML 2023

### Introduction to Gene Expression Analysis

장혜식

서울대학교 생명과학부