# Design of RNA-Seq and Result Interpretation (I)

Sanzhen Liu
Department of Plant Pathology
Kansas State University
@K-State IGF RNA-Seq Workshop (PLPTH885)

6/7/2018

# Schedule (6/7)

- 9:30 – 10:50  Lecture I (RNA-Seq)
- 11:00 – 12:00 Lab I (R introduction)

Lunch break

- 1:00 – 1:30 Lecture II (DE analysis)
- 1:45 – 3:30 Lab II (DE analysis)

# Outline

**Review of RNA-Seq procedure**

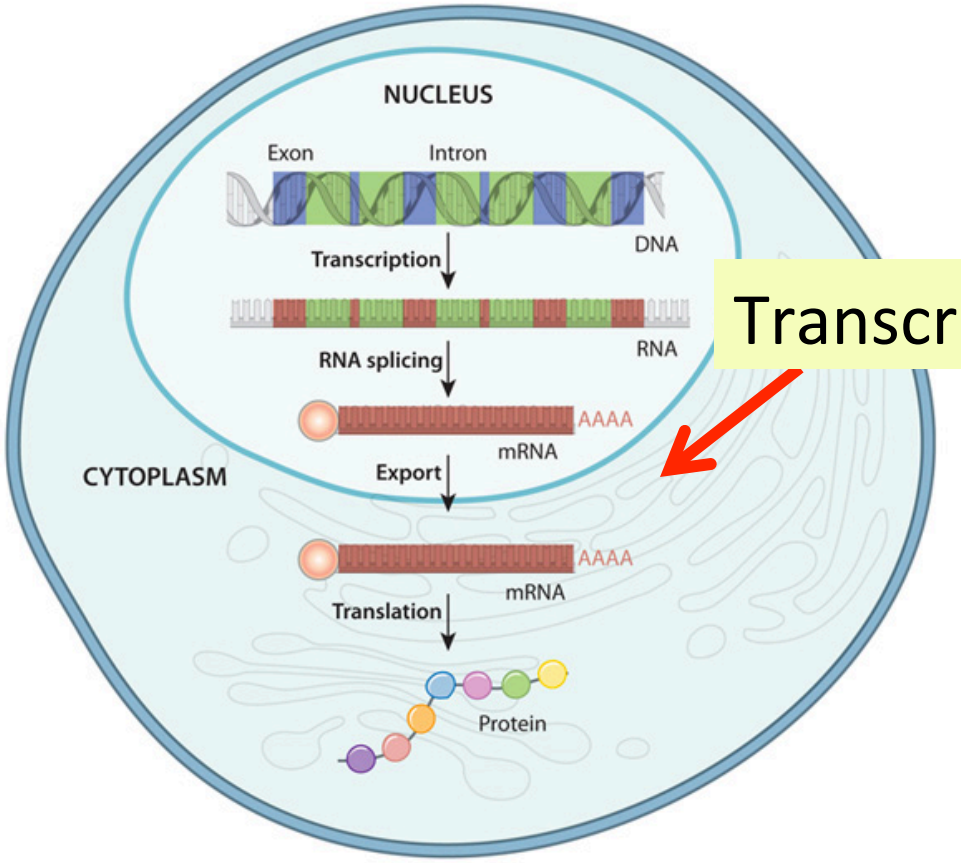
**Design of DE experiments and results**

- Experimental design
- Multiple test correction

**New technologies**

**Other applications**

# Gene expression

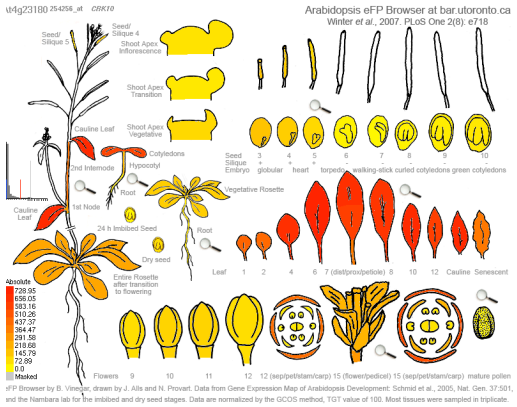



Transcripts


Response to biotic stress

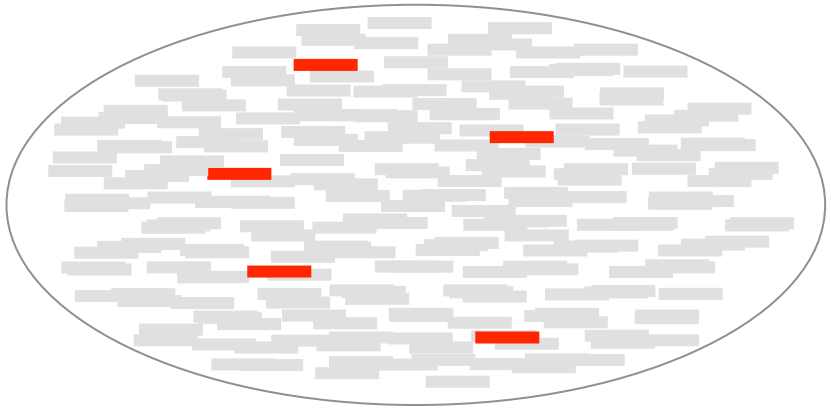
DNA to protein in eukaryote

nature.com/scitable/topicpage/gene-expression-14121669



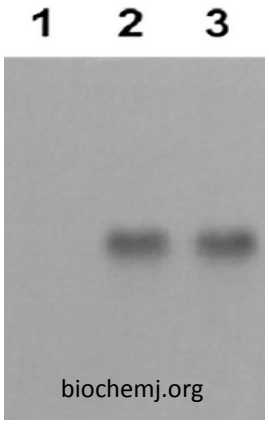

Expression profiles in different tissues

# Approaches for quantification of gene expression

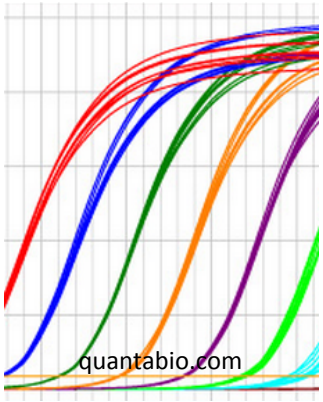


How can we measure the accumulative level of transcripts of **a given gene** in millions/ billions of transcripts?

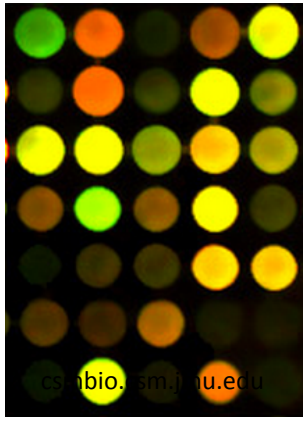
Northern blot




qRT-PCR

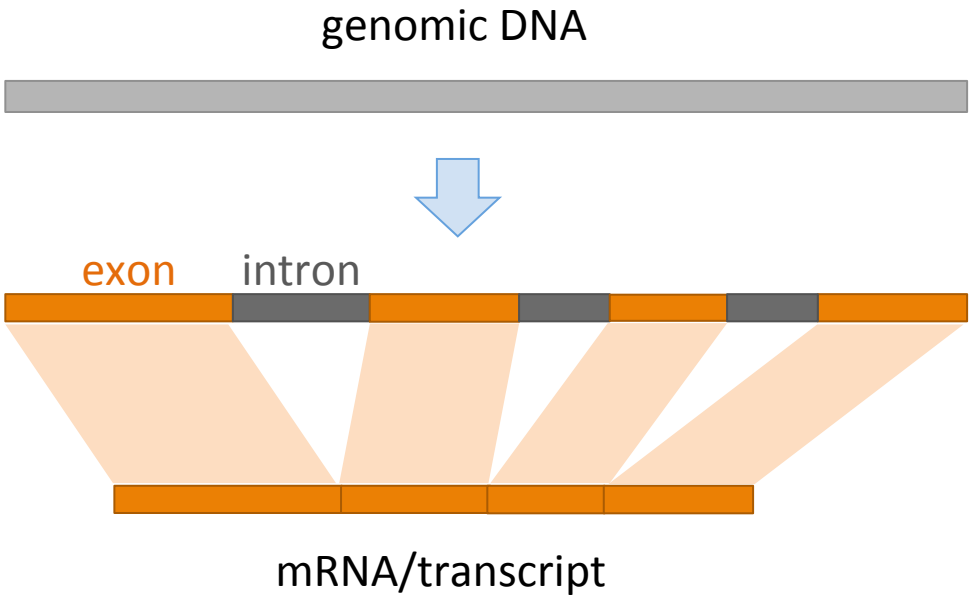



microarray



RNA-Seq

# Rationale of RNA-Seq for **differential expression (DE)**



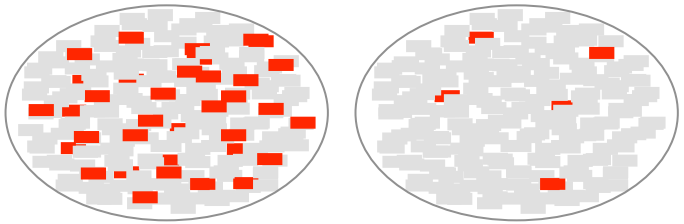

Essentially, RNA-Seq is designed to measure mRNA accumulation levels of genes by

**1) recognizing mRNA based on sequences**

**2) quantifying mRNA of each gene**

Millions times of sampling to quantify each component (transcript) in tissue samples.

10 millions of transcripts in each



100 gene of interest 5

sequence **1,000** transcripts

0 0

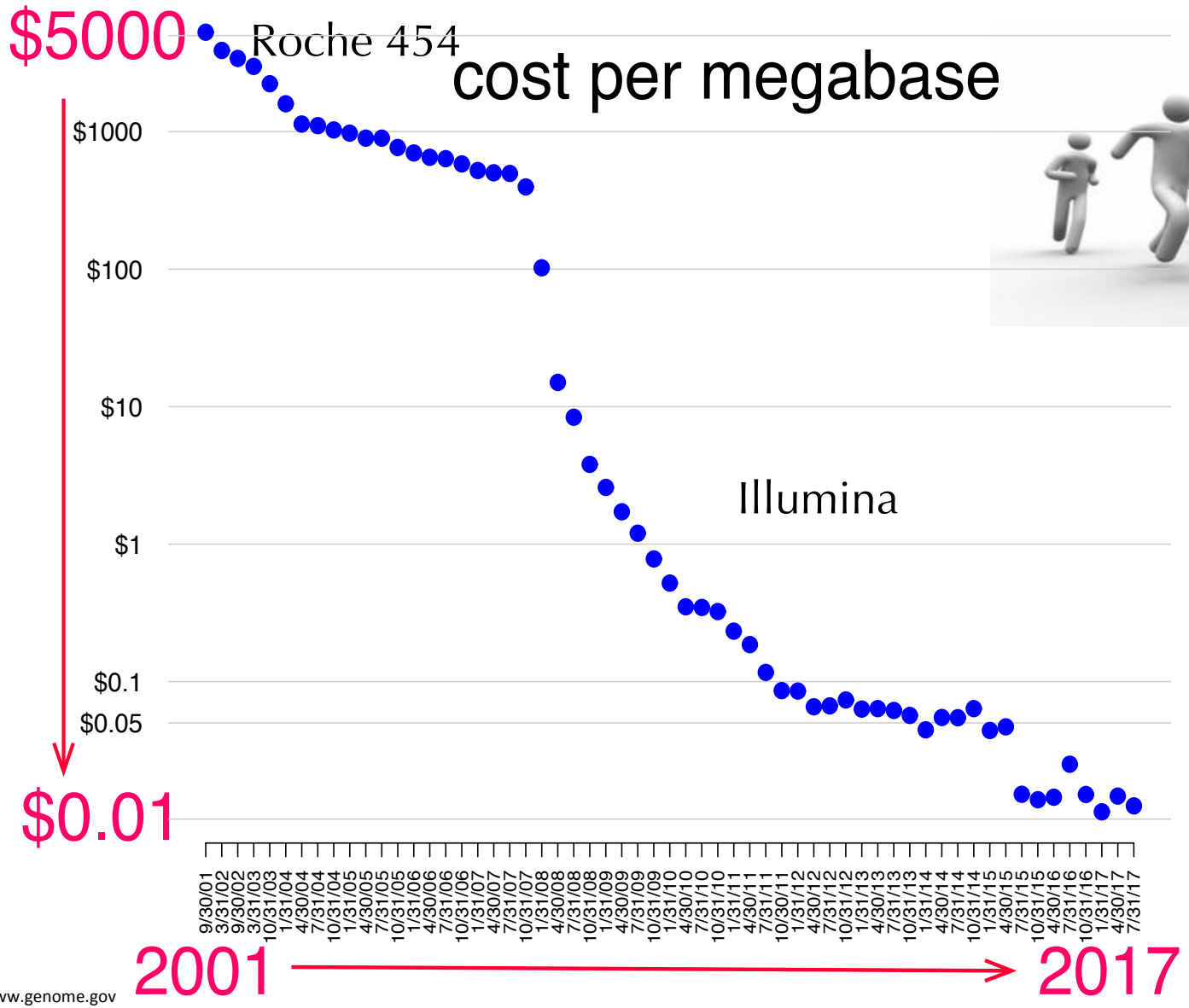
sequence **1 million** transcripts

10 1

Differential expression (DE)?

1970's Sanger sequencing

# Race of sequencing technologies



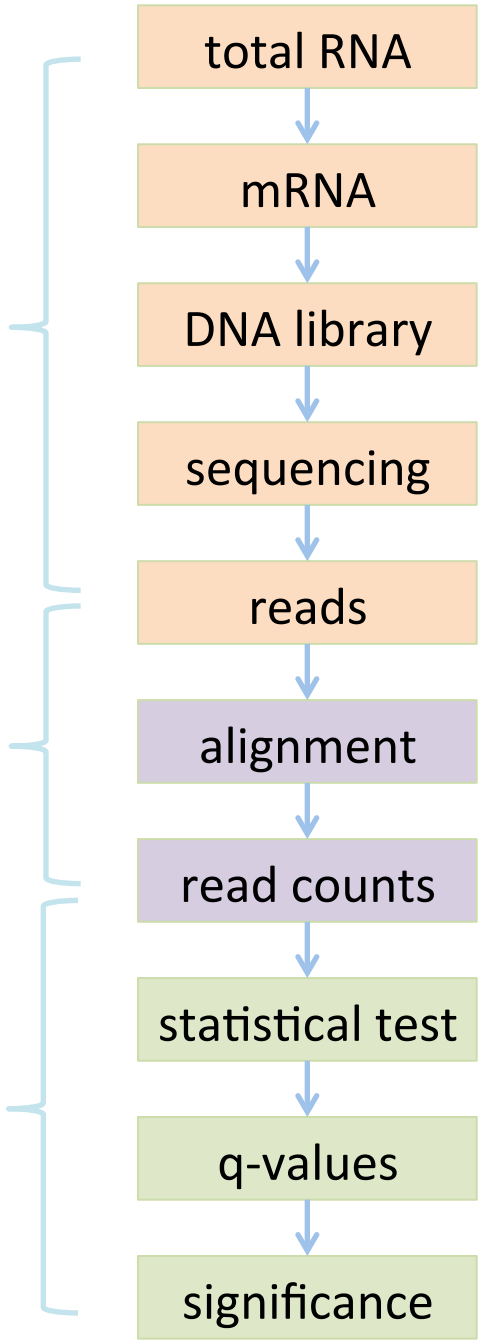


www.genome.gov

Data source: genome.gov/sequencingcosts

# RNA-Seq procedure for DE analysis

1. RNA to sequencing reads
2. reads to read counts per gene
3. read counts to significant genes



total RNA

mRNA

DNA library

sequencing

reads

alignment

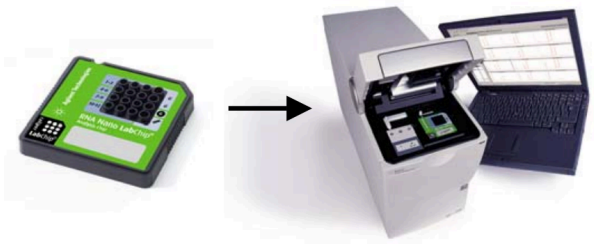
read counts

statistical test

q-values

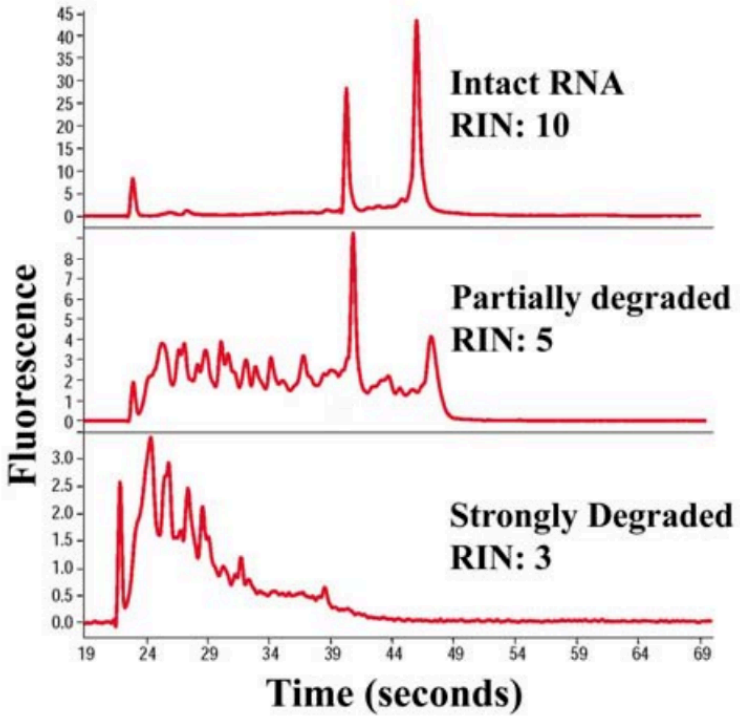
significance

# total RNA



RIN: RNA Integrity Number

a value from 1 to 10, with 10 being the least degraded.




http://www.agilent.com/about/newsroom/lsca/background/rna_integrity.pdf

# RNA to sequencing reads

total RNA

mRNA

DNA library

sequencing

reads

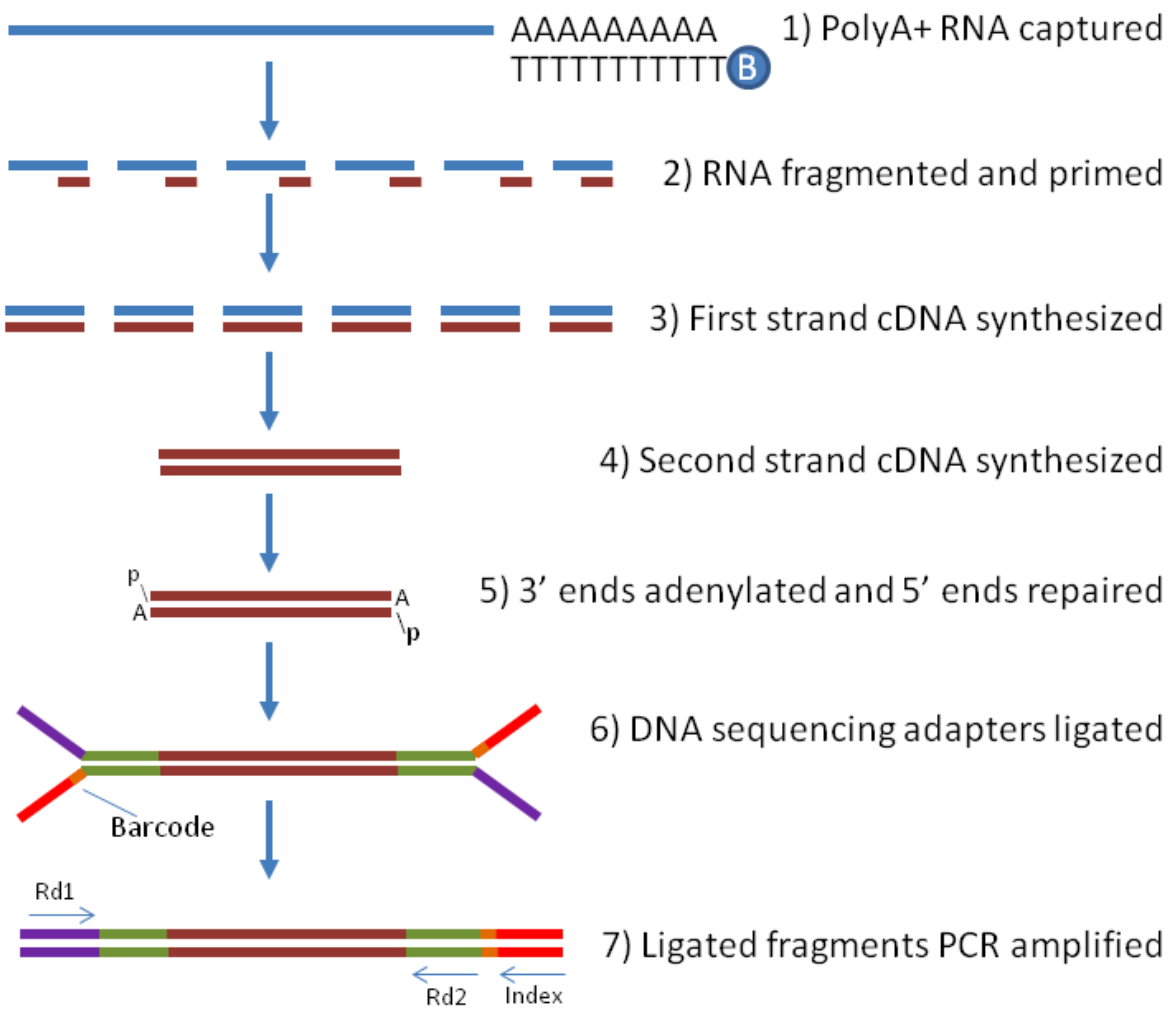
alignment

read counts

statistical test

q-values

significance




labome.com/method/RNA-seq-Using-Next-Generation-Sequencing.html

total RNA

mRNA

DNA library

sequencing

reads

alignment

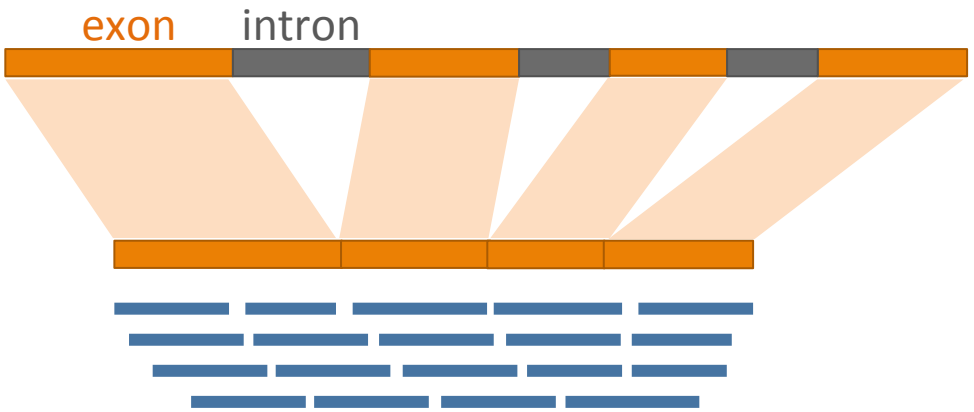
read counts

statistical test

q-values

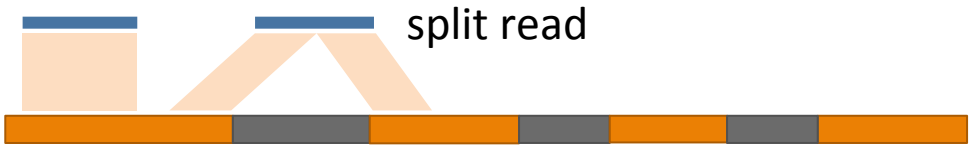
significance

# Reads to read counts per gene




1. reads

2. alignment to the reference genome



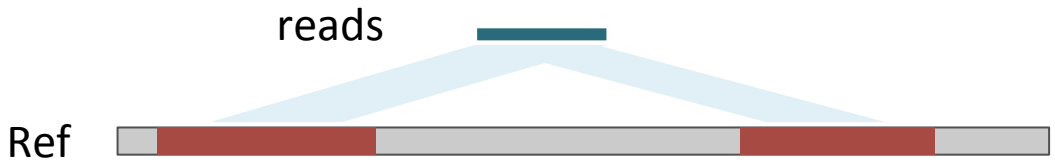

An **intron-aware** aligner is important for RNA-Seq reads alignment (e.g., Tophat, STAR, and GSNAP)

3. **read counts**

N = 19 if all reads can be confidently mapped to the reference genome

# Alignment issues

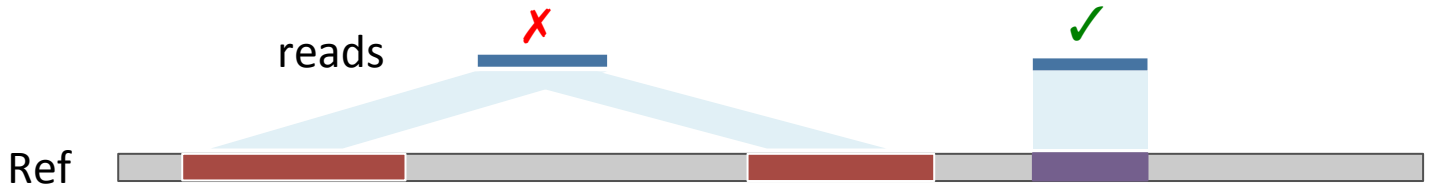
- Repeats



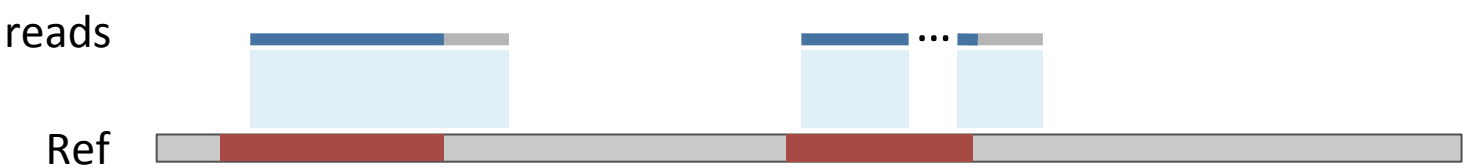

- Sequencing errors
- Polymorphisms (reference and sequenced sample)
- Quality of reference genomes (mis-assembly and incomplete genome)

# Solutions to mitigate problems

- Unique mapped reads




- Longer reads or Paired-end reads



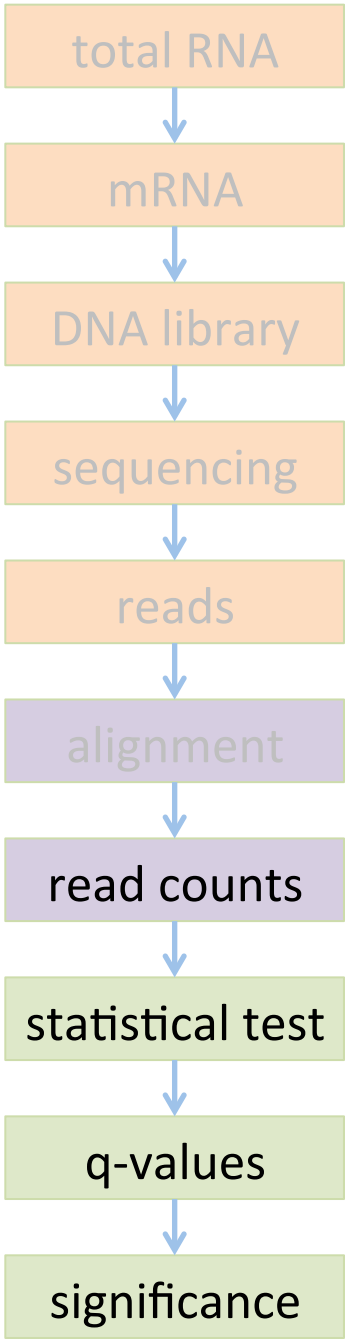

- Tolerance of mismatches or gaps for each alignment

- Better reference genome

# Count matrix: Read counts (Raw) per gene

| Gene | sample 1 | sample 2 | sample 3 |
|---|---|---|---|
| gene 1 | 6,075 | 5,934 | 3,370 |
| gene 2 | 295 | 377 | 169 |
| ... | ... | ... | ... |

# Read counts to significant genes





Control 1 TRT 1 (salt)

RNA
Library
Sequencing

N1 reads N2 reads

Trimming
Alignment
Read counting per gene

| Gene | Control 1 | TRT 1 |
|---|---|---|
| 1 | C1 | T1 |
| 2 | C2 | T2 |
| … | | |
| Sum | N1 | N2 |

2x2 Table for Gene 1

| | Gene 1 | Others |
|---|---|---|
| Control 1 | C1 | N1 – C1 |
| TRT 1 | T1 | N2 – T1 |

- Fisher's Exact Test or $\chi^2$ test on Gene 1

A p-value for Gene 1

- Repeat on all the genes

p-values

- Multiple testing correction

q-values

- Declaration of significance

a significant gene set

# Statistical test for differential expression

- Statistical test to discover differential expression (DE)
  - **Count data**: Generalized Linear Model (GLM) to deal with count data

  e.g., Poisson GLM could handle count data but overdispersion exits
  - **Dispersion issue**: Using **negative binomial GLM** to incorporate dispersion into the model

edgeR (Robinson and Smyth, 2007), **DESeq** (Anders and Huber, 2010), NBPSeq (Di et al., 2011), and QuasiSeq (Lund 2012)

Conesa *et al. Genome Biology* (2016) 17:13
DOI 10.1186/s13059-016-0881-8

Genome Biology

REVIEW Open Access

A survey of best practices for RNA-seq data analysis

Ana Conesa[1,2*], Pedro Madrigal[3,4*], Sonia Tarazona[2,5], David Gomez-Cabrero[6,7,8,9], Alejandra Cervera[10], Andrew McPherson[11], Michał Wojciech Szcześniak[12], Daniel J. Gaffney[3], Laura L. Elo[13], Xuegong Zhang[14,15] and Ali Mortazavi[16,17*]

# Outline

**Review of RNA-Seq procedure**

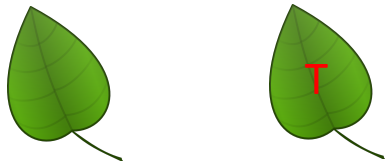
**Design of DE experiments and results**

- Experimental design
- Multiple test correction

**Other analyses**

- Visualization
- GO term enrichment analysis

# An RNA-Seq experiment – source of variance




Control 1 TRT 1 (salt)

| | | |
|---|---|---|
| Gene 1 | **C1** | **T1** |
| … | | |
| Sum | N1 | N2 |

Our interest:
the effect of the salt treatment on gene expression

**Question**: what would cause the difference between two values, C1 and T1?

- **Treatment effect**
- Plant difference
- RNA quality
- Library preparation
- Sequencing
- Sampling

- Sequencing depth

# Source of variance in RNA-Seq - sampling

- **Sampling variance** derived from the inherent nature of counting experiments

total molecules: $10^9$
gene X: 1000 molecules

Randomly sample $10^7$

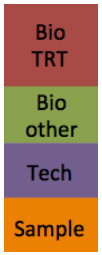
| | |
|---|---|
| First sampling | 6 |
| Second sampling | 13 |
| Third sampling | 8 |

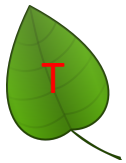
Randomly sample $10^8$

| | |
|---|---|
| First sampling | 102 |
| Second sampling | 93 |
| Third sampling | 97 |

Sequence depth (sampling number) matters.

# Technical replication




Control 1 Control 2 Control 3 TRT 1 TRT 2 TRT 3

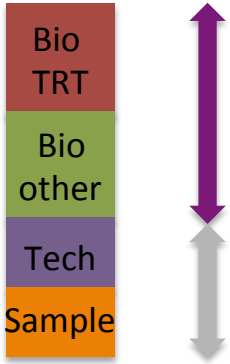
within groups within groups

between groups

***Technical replication*** refers to the sequencing of multiple libraries derived from **the same biological sample**.

Technical replicate


within groups

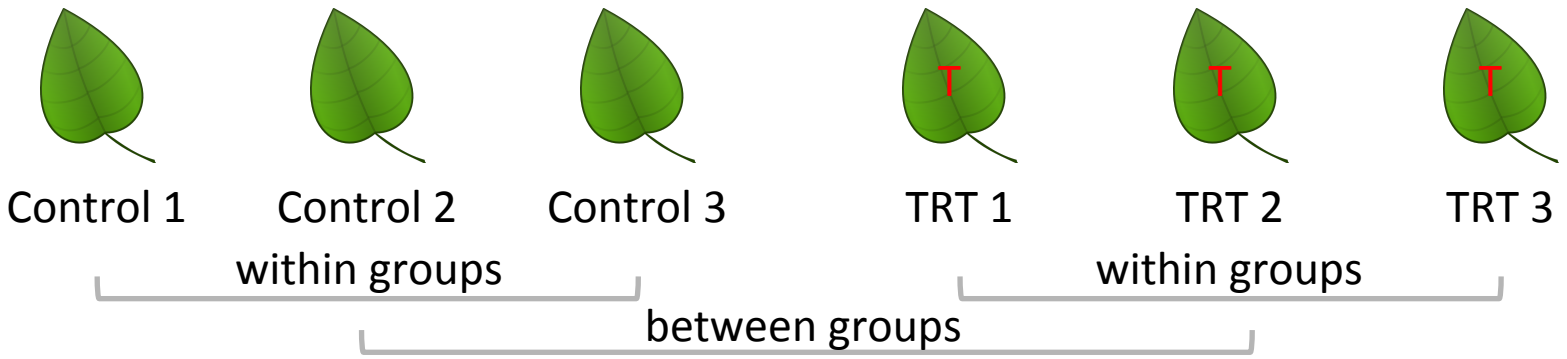



between groups

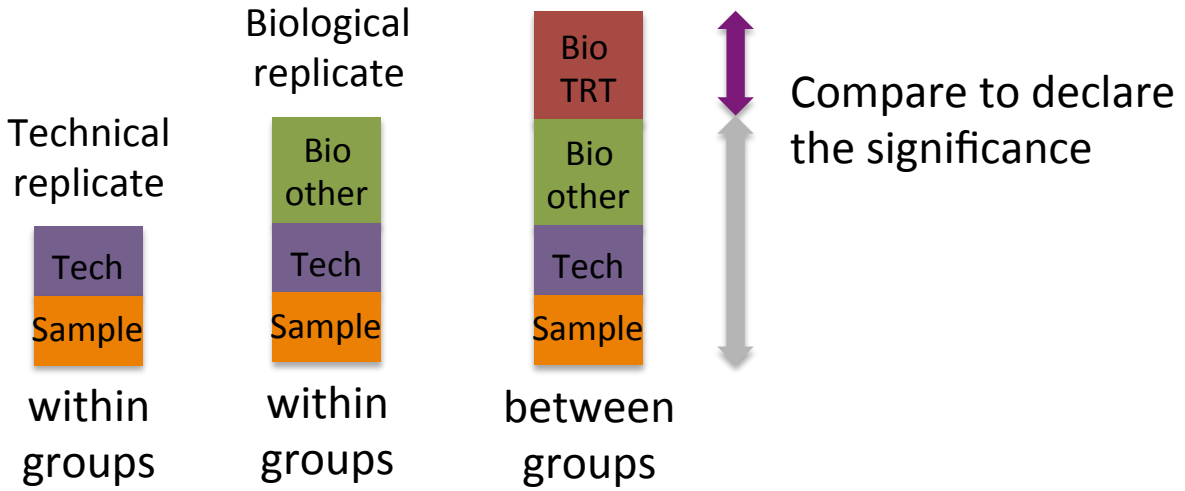
Compare to declare the significance

False power

# Biological replication




***Biological replication*** refers to the sequencing of multiple libraries derived from **different biological samples**.

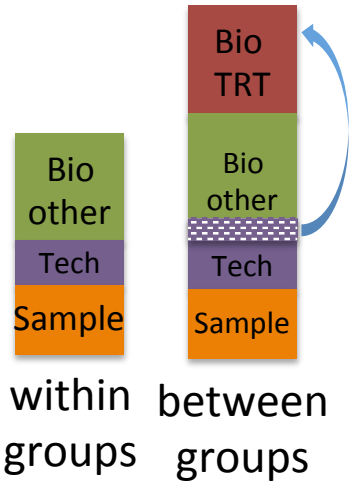



1. Use biological replication instead of technical replication unless you have your own interest.
2. More replicates increase the power to detect small treatment effect

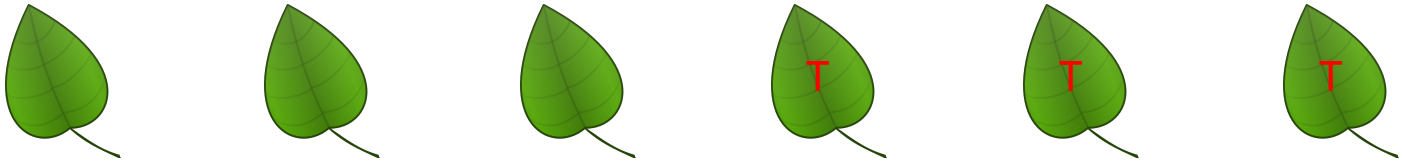
# Question I

My lab conducted an RNA-Seq experiment to identify the DEs between two biological groups to examine a treatment of great interest. Each group has five biological replicates. I told my graduate student to perform the experiment of each group separately (then I don't need to worry that the samples from two groups are messed up).

Is this a sound experimental design? Why?



# Comparison of read counts among different samples




| | Control 1 | Control 2 | Control 3 | TRT 1 | TRT 2 | TRT 3 |
|---|---|---|---|---|---|---|
| Gene 1 | C1 | C2 | C3 | T1 | T2 | T3 |
| … | | | | | | |
| Sum | N1 | N2 | N3 | N4 | N5 | N6 |

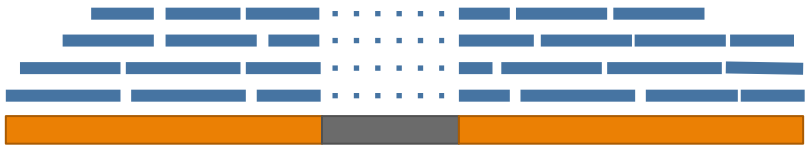
Sequence depth (total read number) influences read counts.
Therefore, raw read counts can not be compared directly.

Can we generate some comparable numbers among samples?

# A **normalization** method: RPKM and FPKM

- **RPKM**: **Read** number per kilobase of exons per million of total reads

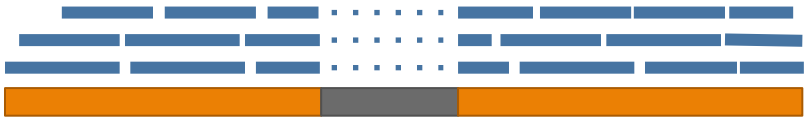
Control 1 read count = **23**



total reads: **15 millions** of total reads

RPKM of X = ? = **3.1**

Treatment 1 read count = **18**



exon 1 (**220 bp**) exon 2 (**280 bp**)

gene X

total reads: **10 millions** of total reads

RPKM of X = ? = **3.6**

- **FPKM**: **Fragment** number per kilobase per million of total reads.

Fragment = one pair of paired-end reads or one single-end read

# More about RPKM


Can we say that the gene B has higher expression than the gene A?

- RPKM is not an ideal indicator to compare the expression/ accumulation levels between two genes

1. amplification bias
2. alignment efficiency

# Experimental Design

- **Sequencing depth**

Increasing sequencing depth decreases sampling variance

- **Biological replication**

Reasonable number of biological replication helps accurately estimate variances to achieve reliable statistical inference.

- **Randomization and unbiasedness**

To avoid confounding effect

# Outline

**Review of RNA-Seq procedure**

**Design of DE experiments and results**

- Experimental design
- Multiple test correction

**New technologies**

**Other applications**

# DE result

| | DE Result | |
|---|---|---|
| GeneID | Log2FC* | p-value |
| 1 | -0.40 | 0.037 |
| 2 | 0.03 | 0.916 |
| 3 | -0.89 | 2.42E-05 |
| 4 | 0.30 | 0.130 |
| 5 | -0.36 | 0.140 |
| 6 | -0.07 | 0.811 |
| … | | |

* Log2FC: log2 of fold change (trt / control)

# single test vs. multiple tests

- **Single test:**

p = 0.03

At the 5% significant level (P-value threshold = 0.05), we can reject the null hypothesis.

- **Multiple tests:**

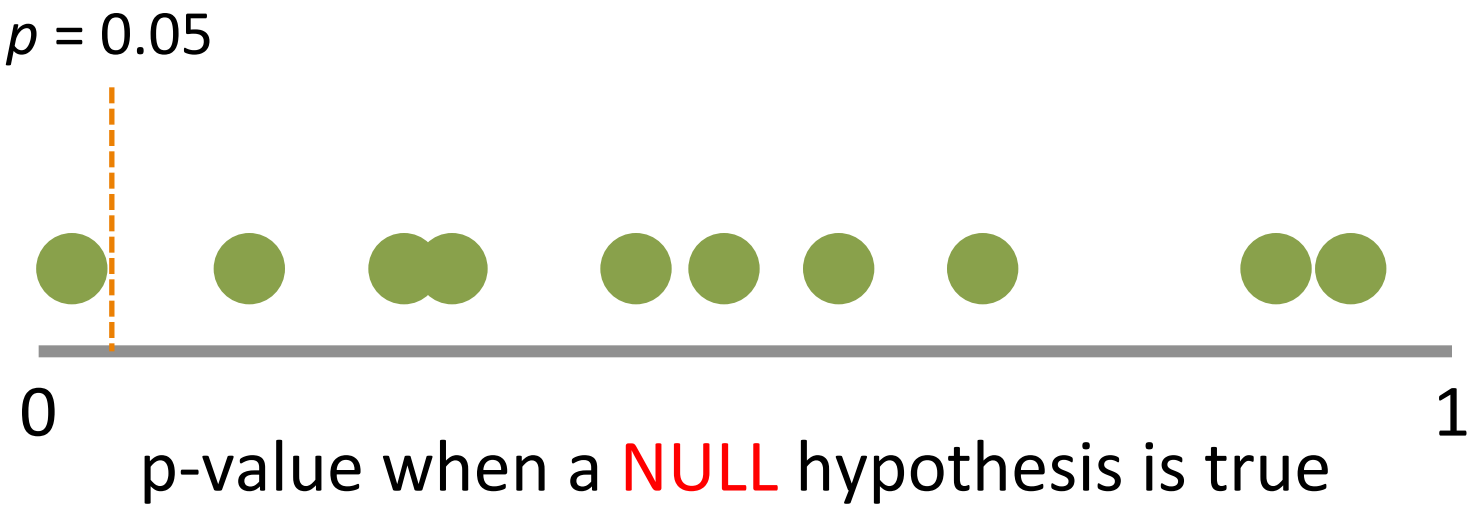
p1 = 0.8; p2 = 0.1; p3 = 0.3; p4 = 0.5; ...; p20 = 0.03

At the 5% significant level (P-value threshold = 0.05), we will reject the null hypothesis for p20.

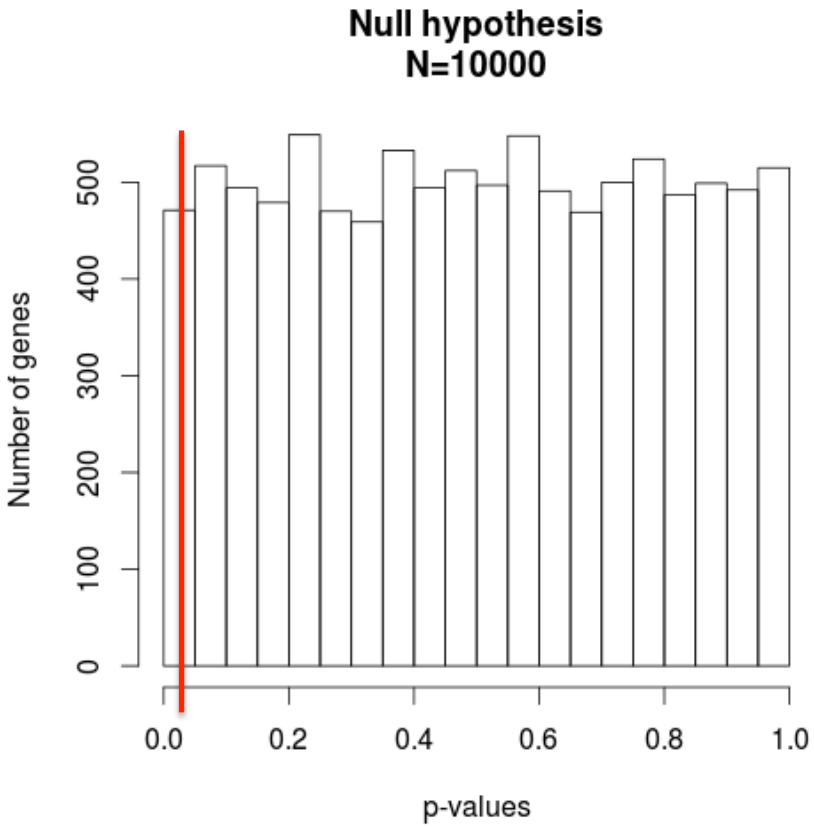
Anything wrong here?

# Multiple testing correction

"A p-value is only statistically valid when a single score is computed."



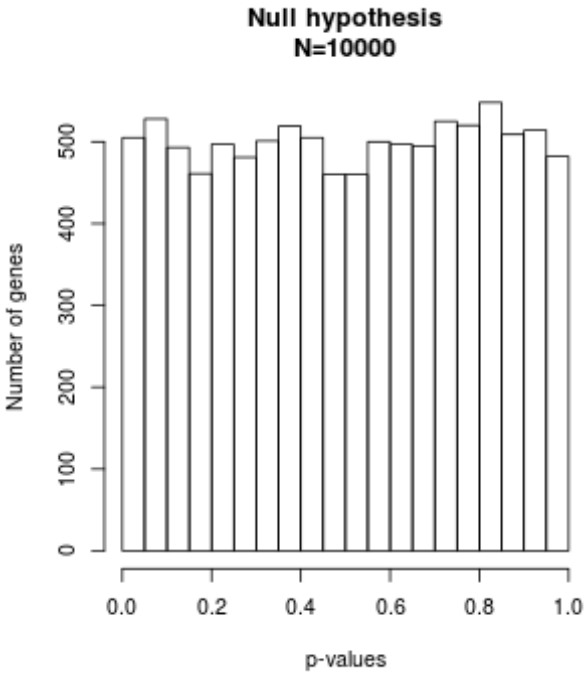
# P-value distribution under the null hypothesis (e.g., no treatment effect)

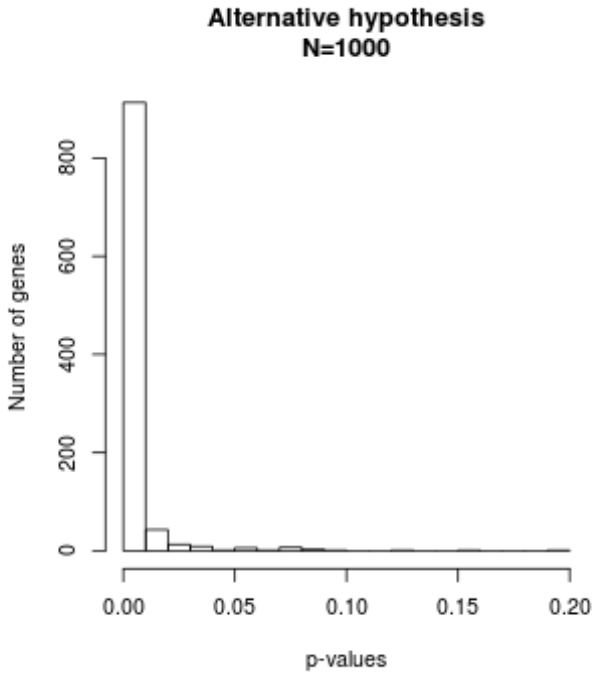


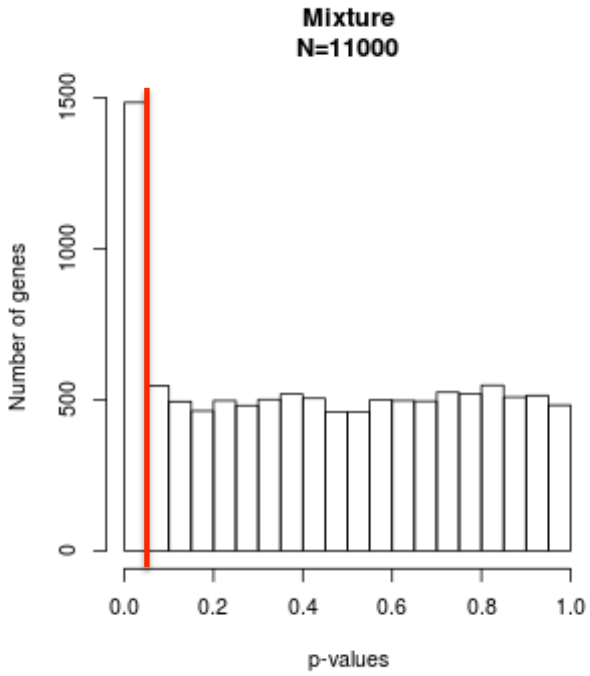

When the null hypothesis is true, a P-value is distributed uniformly from 0 to 1.

No matter how stringent the criteria are, you'll identify genes with very small p-values and the false discovery rate (FDR) is 100%.

# P-value distribution under both the null and non-null hypotheses



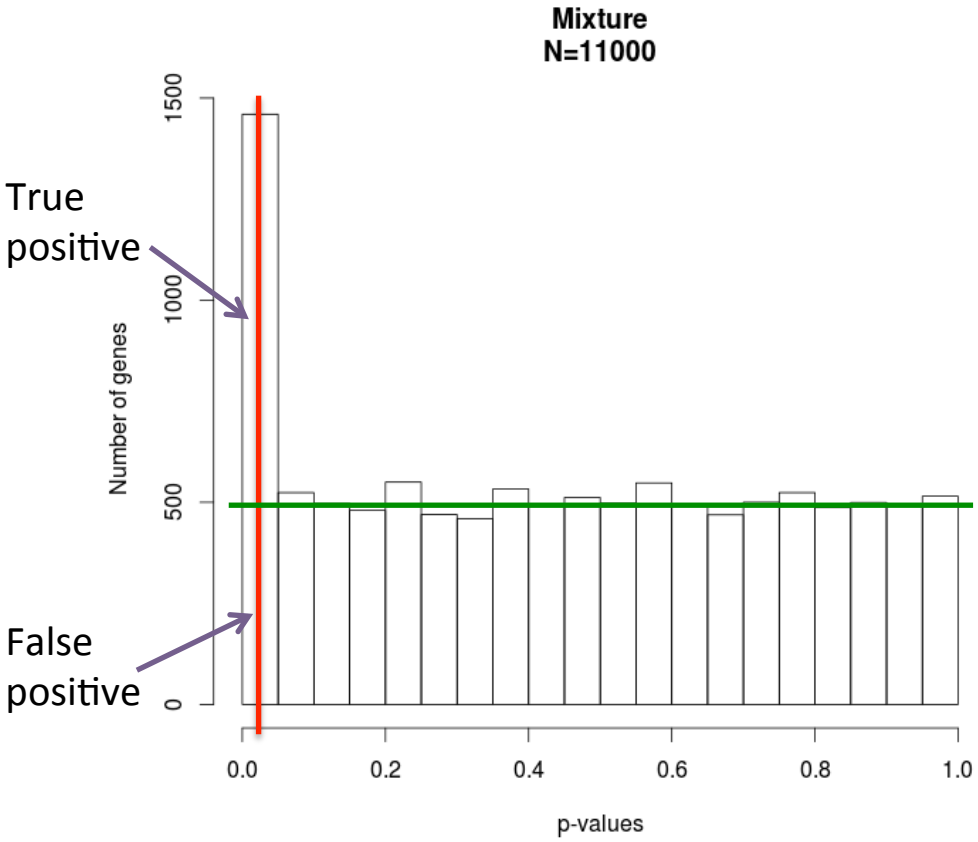

When the null hypothesis is true, a P-value is distributed uniformly.

When the null hypothesis is false, the P-value distribution is skewed toward 0.

Cutoff: p=0.05
FDR=471/(471+989)=32%

Cutoff: p=0.01
FDR=102/(102+912)=10%

# Multiple test correction – FDR method




P-values < 0.00009
DE=992
False DE=99

FDR 10%

# q-values

The **q-value** of a test in a set of tests is **the smallest FDR** for which we can reject the null hypothesis for that one test and all others with smaller p-values.

| Gene | p-values | q-values |
|---|---|---|
| 1 | 0.000 | 0.006 |
| 2 | 0.002 | 0.015 |
| 3 | 0.009 | 0.059 |
| 4 | 0.013 | 0.063 |
| 5 | 0.035 | 0.139 |
| 6 | 0.051 | 0.171 |
| 7 | 0.155 | 0.442 |
| 8 | 0.197 | 0.492 |
| 9 | 0.247 | 0.539 |
| 10 | 0.269 | 0.539 |
| 11 | 0.358 | 0.651 |
| 12 | 0.396 | 0.656 |
| 13 | 0.426 | 0.656 |
| 14 | 0.493 | 0.702 |
| 15 | 0.526 | 0.702 |
| 16 | 0.622 | 0.777 |
| 17 | 0.782 | 0.920 |
| 18 | 0.862 | 0.958 |
| 19 | 0.925 | 0.974 |
| 20 | 0.992 | 0.992 |

Total number of tests: m = 20

FDR method (BH) is a method to calculate q-values/adjusted p-values/ corrected p-values based on p-values

5% FDR, q-values < 0.05

10% FDR, q-values < 0.1

20% FDR, q-values < 0.2

# Question

If we identify 500 differential expression (DE) genes using the 5% FDR to account for multiple tests. Which one below is a better description?

1. I am 95% confident that 500 genes are DE.

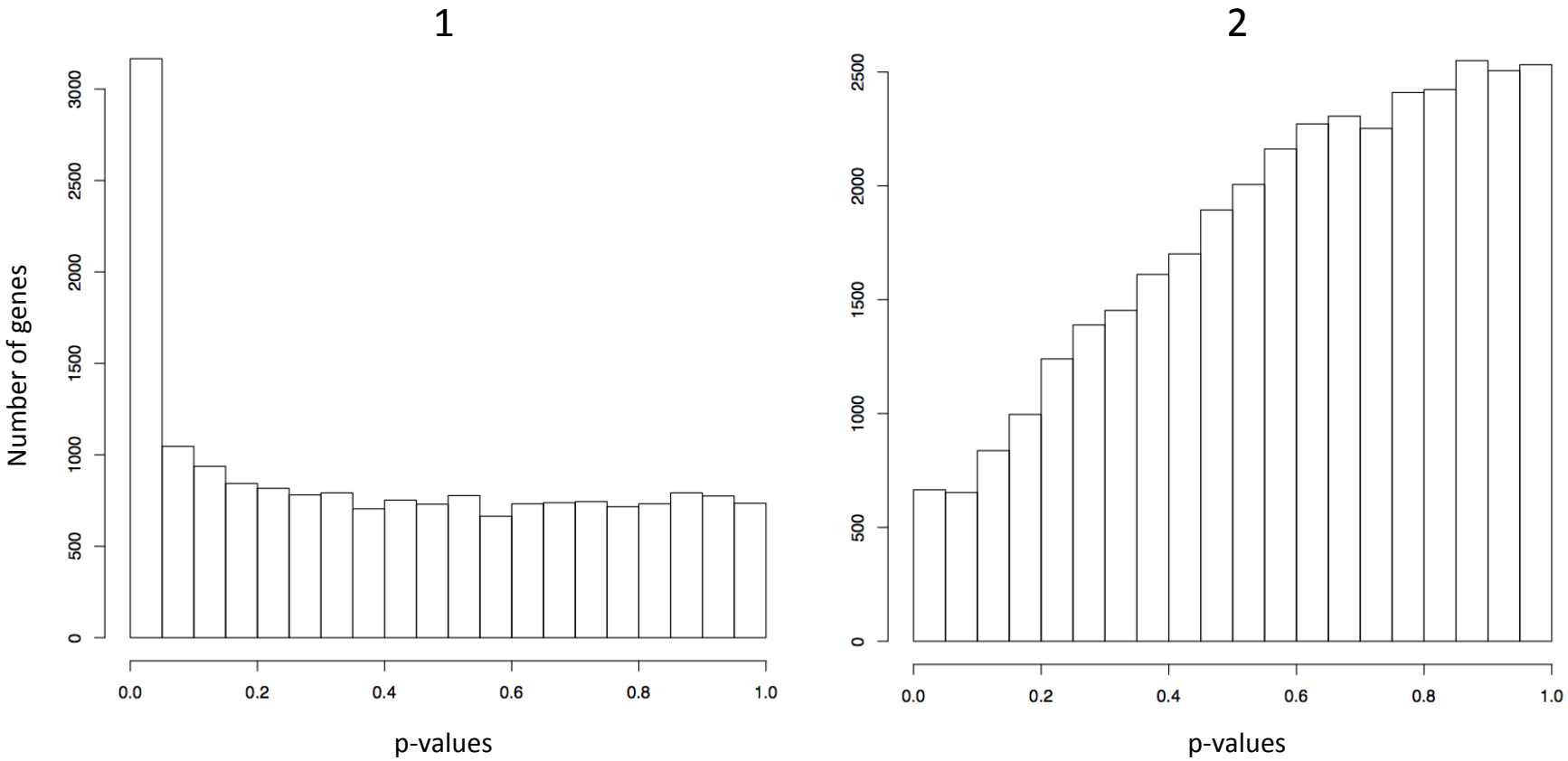
2. The 5% genes (25 genes) in the set are expected to be false DE genes.

# False discovery rate (concept)

For example, among 10,000 tests (10,000 genes), 100 significant genes are declared, in which 10 gene is falsely rejected. In this case, the false discovery rate is 10%.

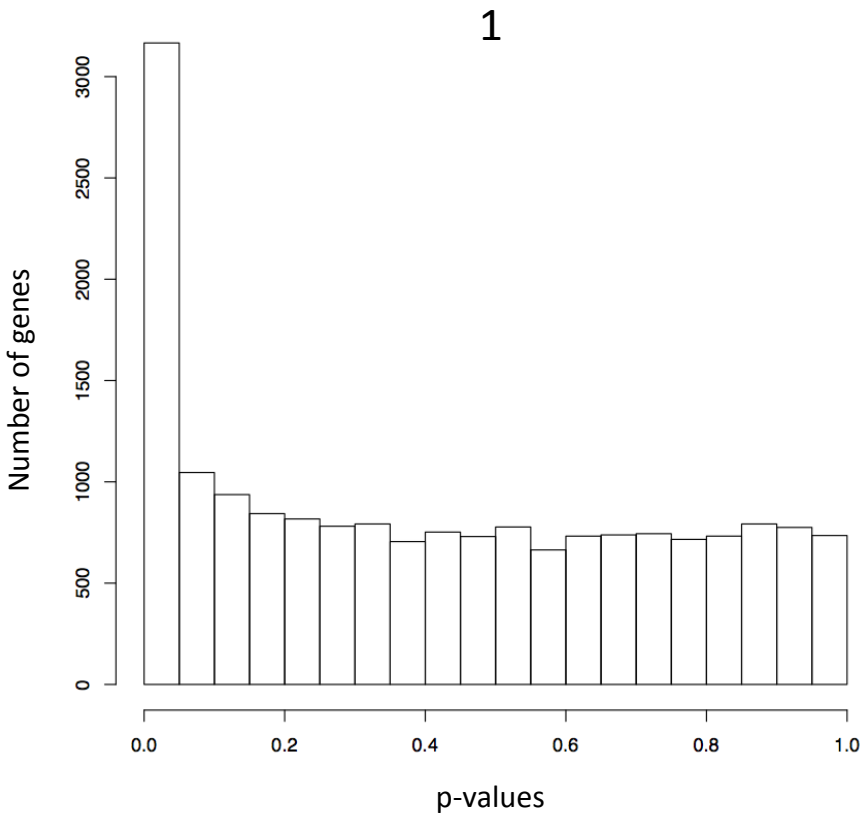
| | True null hypothesis ($H_0$) | False null hypothesis ($H_1$) | Total |
|---|---|---|---|
| Rejected (Declared significance) | **10** | 90 | 100 |

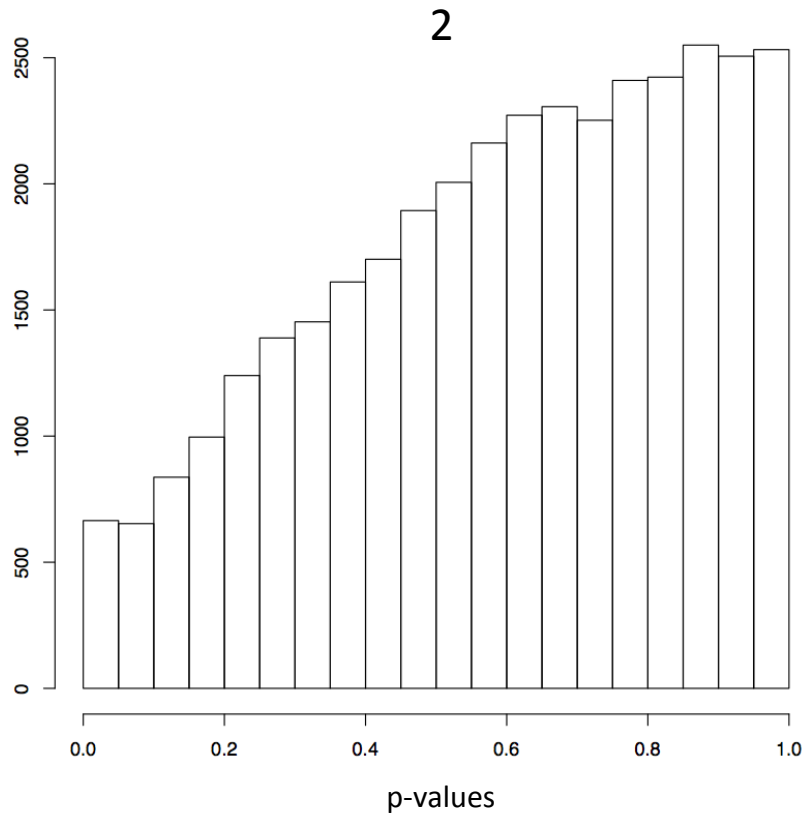
# P-value histograms from real studies




If you perform an RNA-Seq experiment, which one would you hope to obtain? Why?

# P-value histograms from real studies

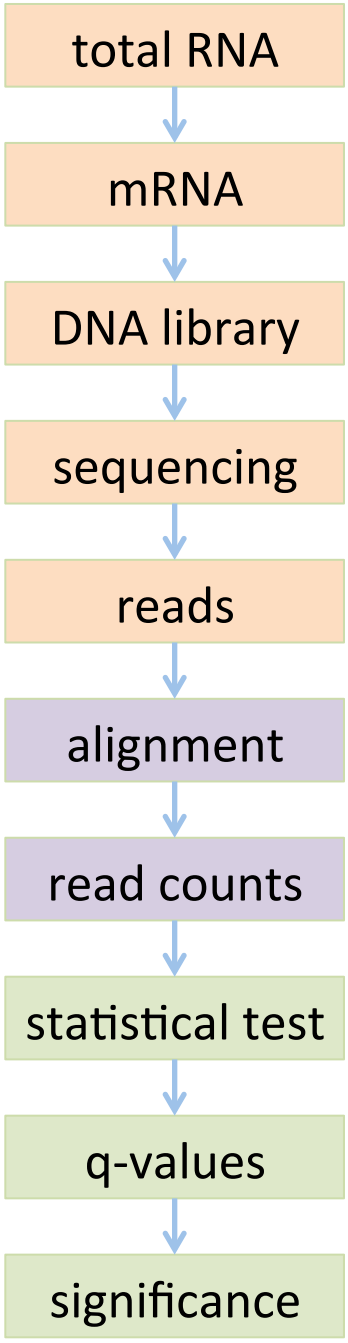



DE = 1,370, FDR=5%

DE = 0, FDR=20%

# Keywords




randomization, replication, RNA quality

short or long reads
single- or paired-end reads, read length
sequencing depths
(e.g., >20 million short reads for most experiments)

intron-spanning Aligner (e.g., GSNAP, STAR)

count data statistical analysis (DESeq2 & edgeR)
multiple test p-value adjustment

# Outline

**Review of RNA-Seq procedure**

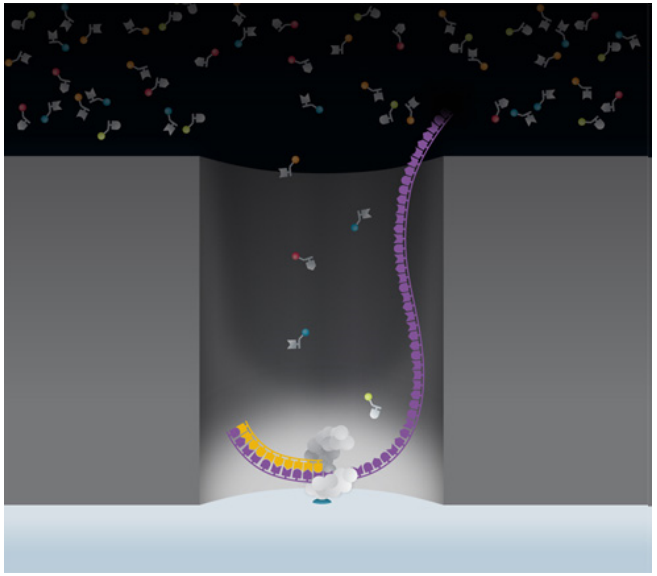
**Design of DE experiments and results**

- Experimental design
- Multiple test correction

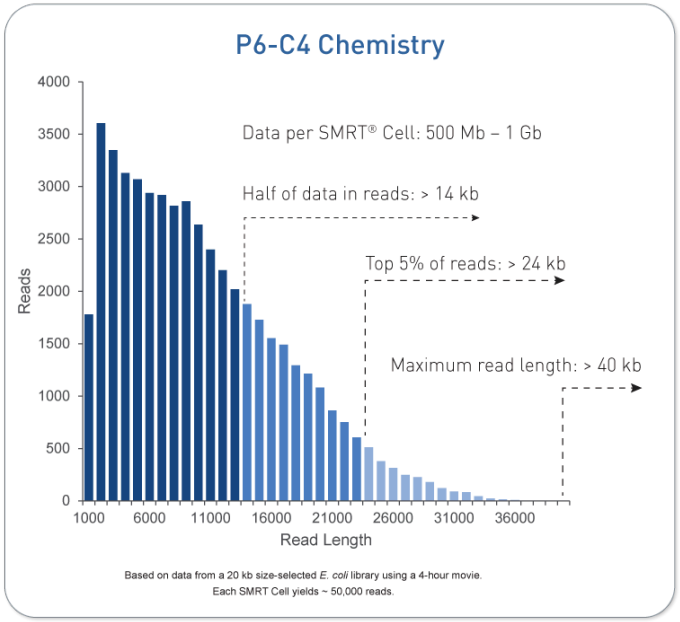
**New technologies**

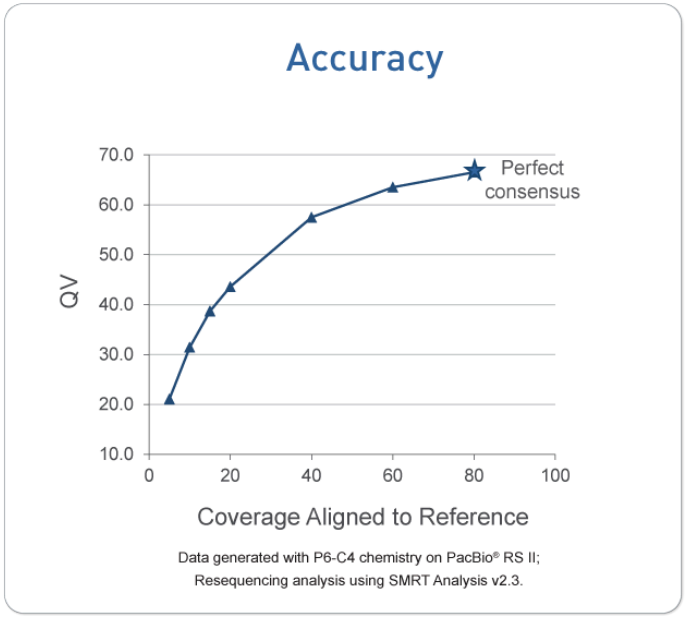
**Other applications**

# PacBio – Single Molecule Real Time (SMRT)



[PacBio tech video]



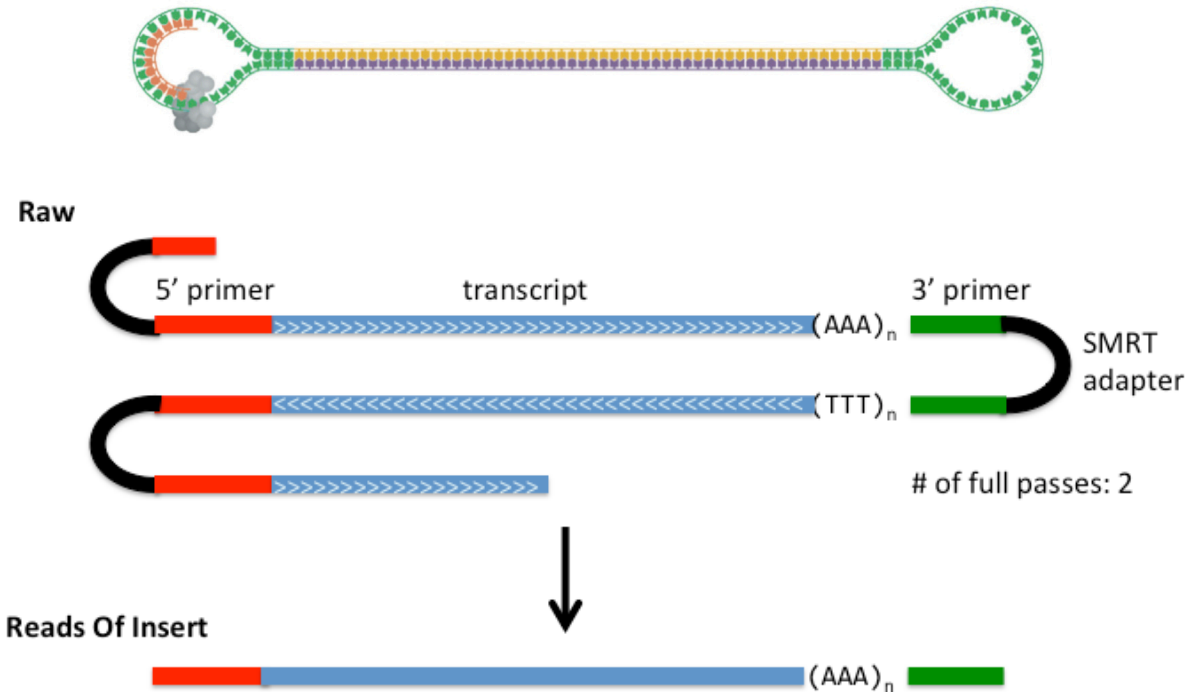




- Single molecule sequencing
- no amplifications required
- up to 70+ kbp sequencing
- Moderate sequencing throughput
- high sequencing error rate (~15%, random, no-context-specific errors)

pacbio.com

# PacBio long reads for RNA sequencing (Iso-Seq)

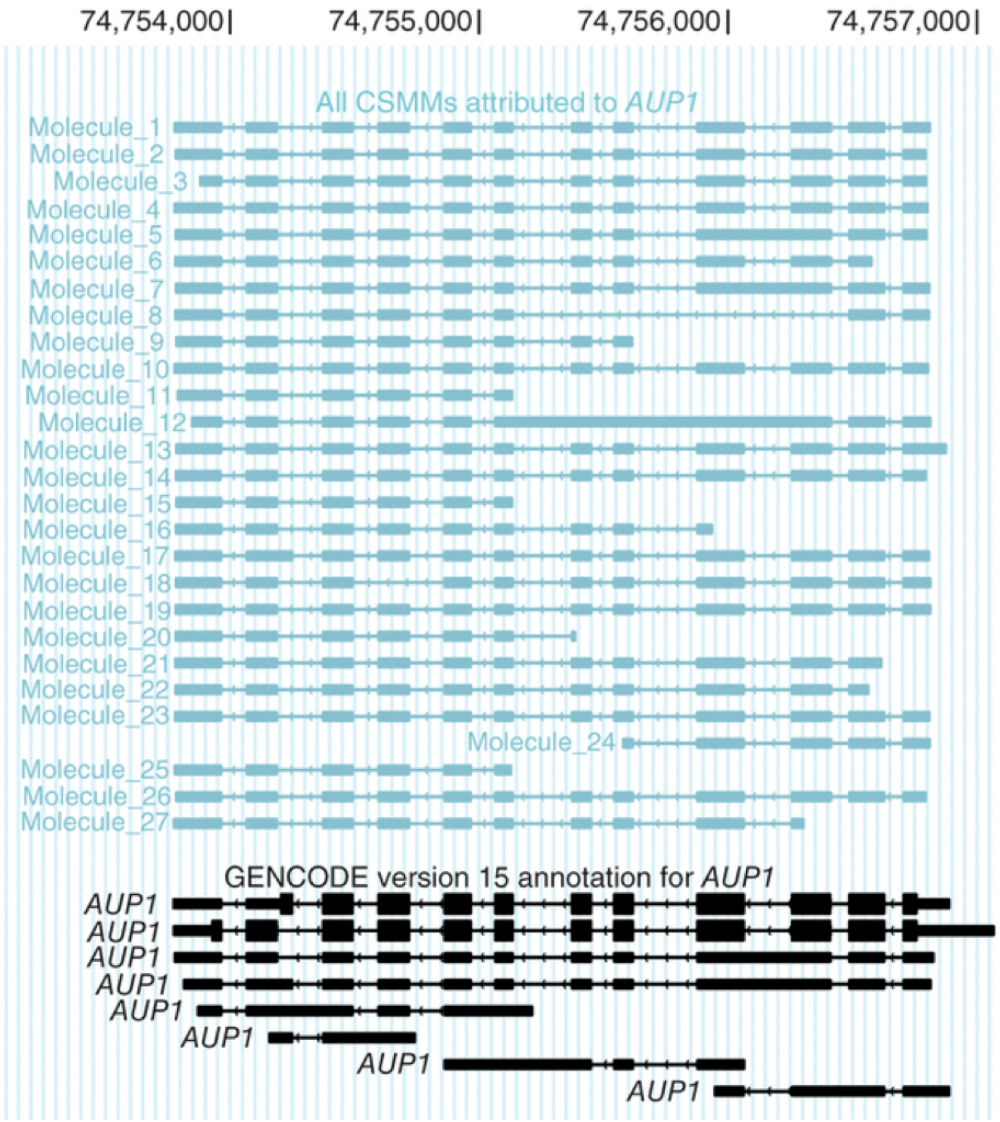
1. High-quality, single-molecule, circular-consensus (CCS)
2. Multiple passes improve sequence quality




https://github.com/PacificBiosciences/cDNA_primer/wiki/Understanding-PacBio-transcriptome-data

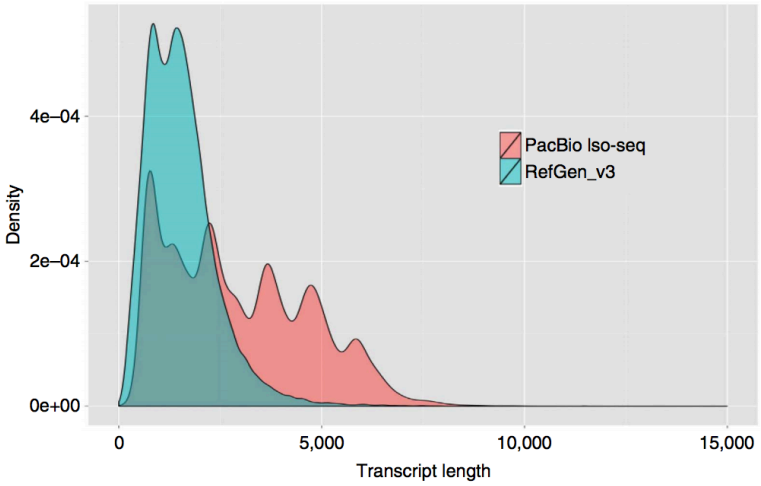
# Long reads to sequence full-length cDNA

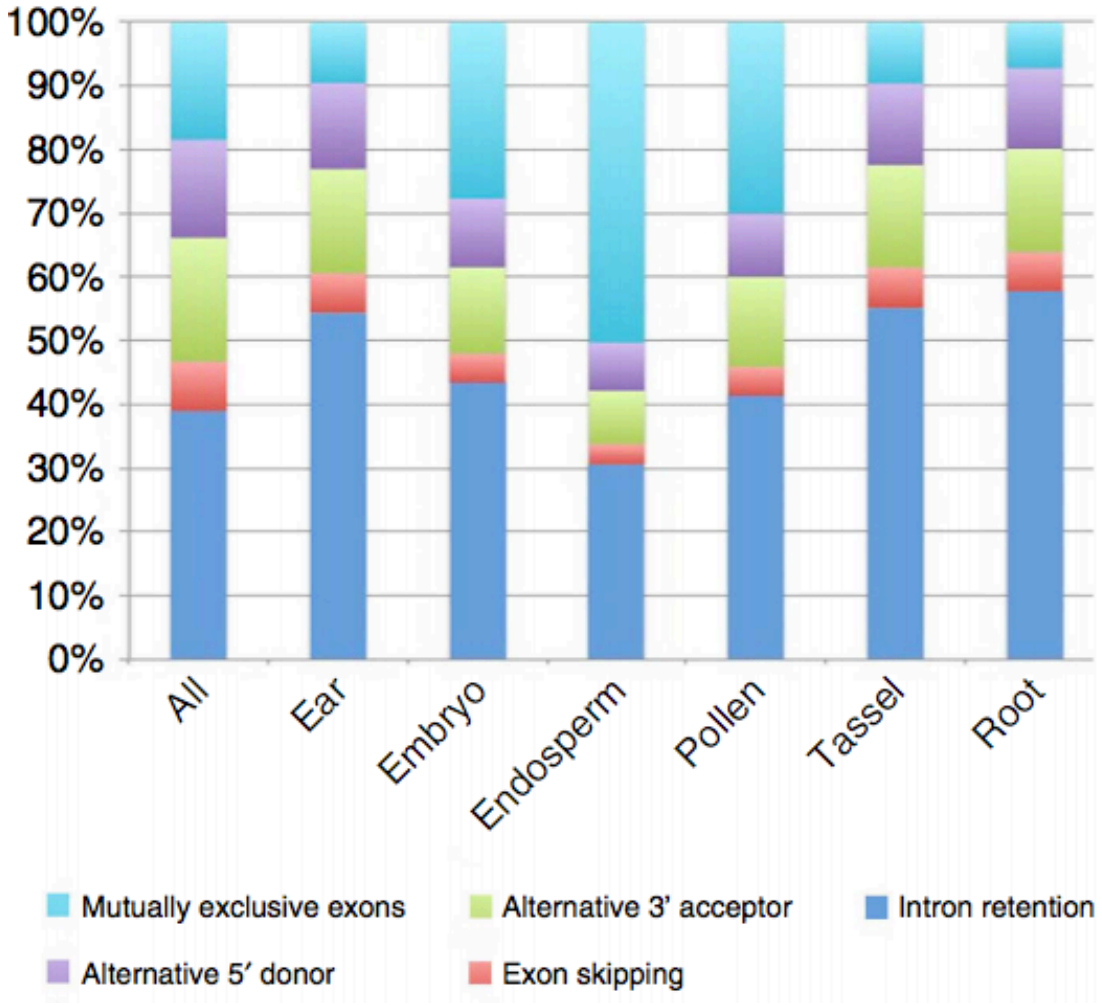
- The majority of reads represent all splice sites of original transcripts
- Isoforms can be monitored at a single-molecule level without amplification or fragmentation

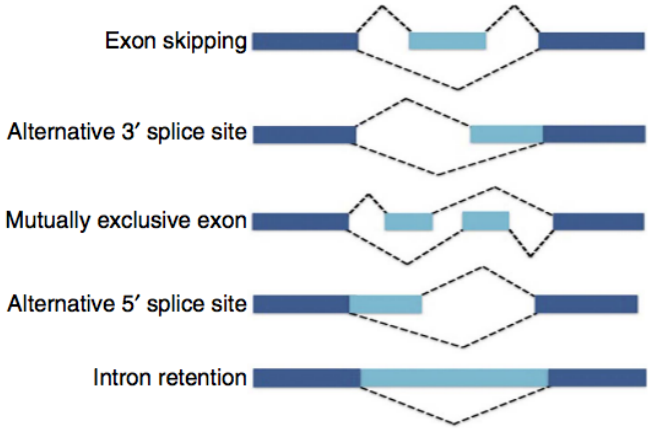



Sharon et al. Nature Biotechnology 31, 1009–1014 (2013)

# maize – Iso-Seq



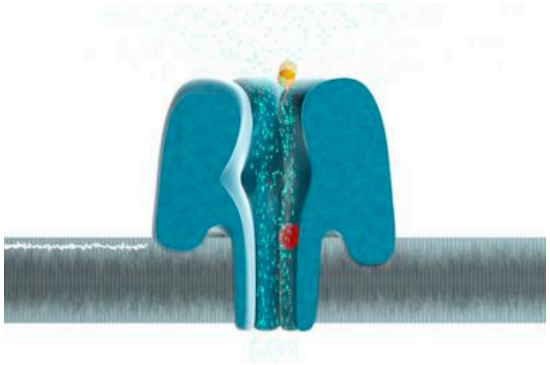




Wang et al., Nature Communications, 2016, 7:11708

# Oxford Nanopore

A promising technology

- Single molecular sequencing
- No amplifications
- **Long reads (up to hundreds of kb)**
- **Error rate is high (~10-30%)**



**MinION**

1. USB disposable sequencer
2. Hundreds of Mb in several hours

As each nucleobase passes through the pore the current is affected and this change allows sequence to be read out.

nanoporetech.com

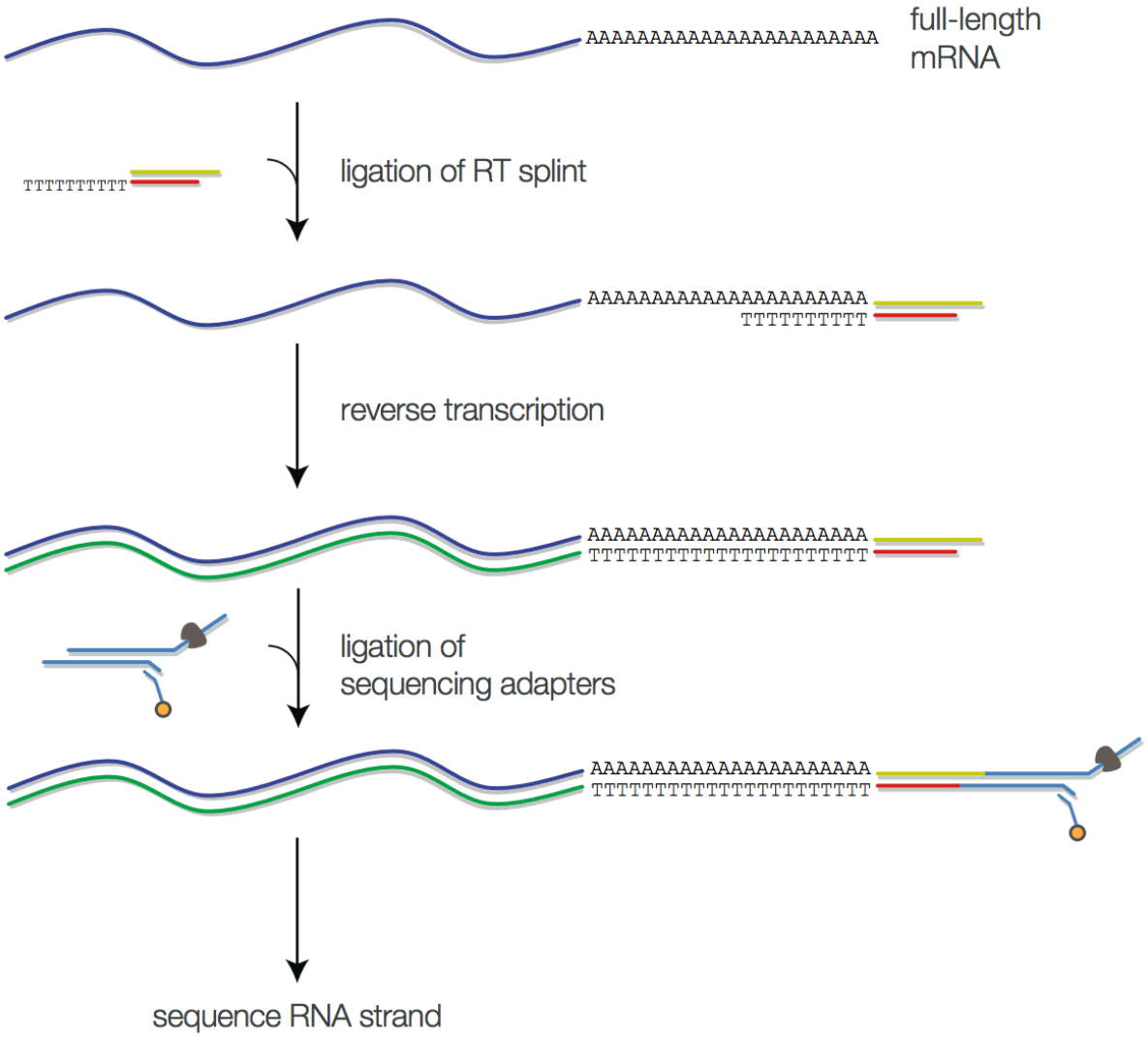
## Nanopore long-read RNAseq reveals widespread transcriptional variation among the surface receptors of individual B cells

Ashley Byrne, Anna E. Beaudin, Hugh E. Olsen, Miten Jain, Charles Cole, Theron Palmer, Rebecca M. DuBois, E. Camilla Forsberg, Mark Akeson & Christopher Vollmers ✉

*Nature Communications* **8**,
Article number: 16027 (2017)

Received: 24 April 2017
Accepted: 23 May 2017

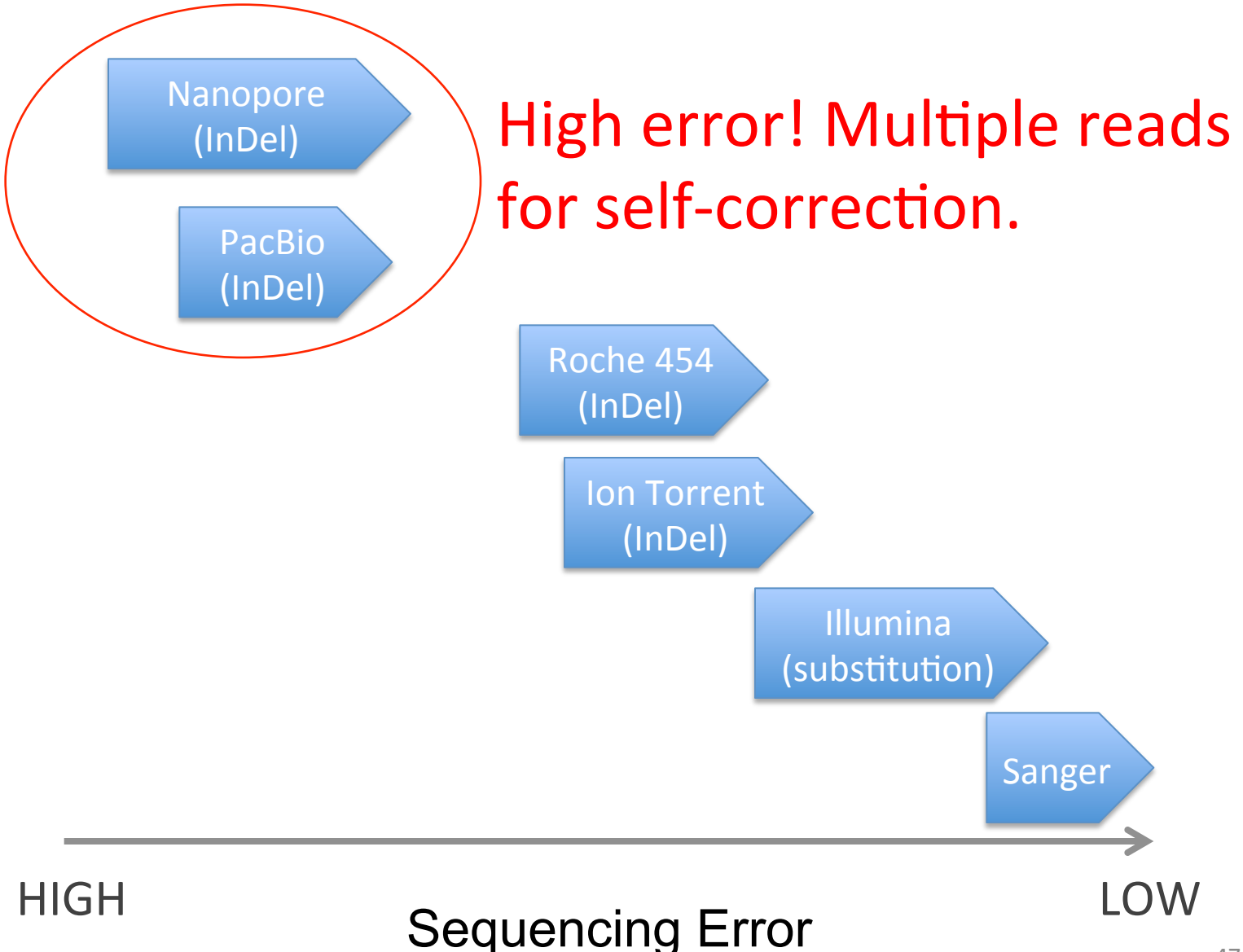
# Nanopore – **direct** RNA sequencing



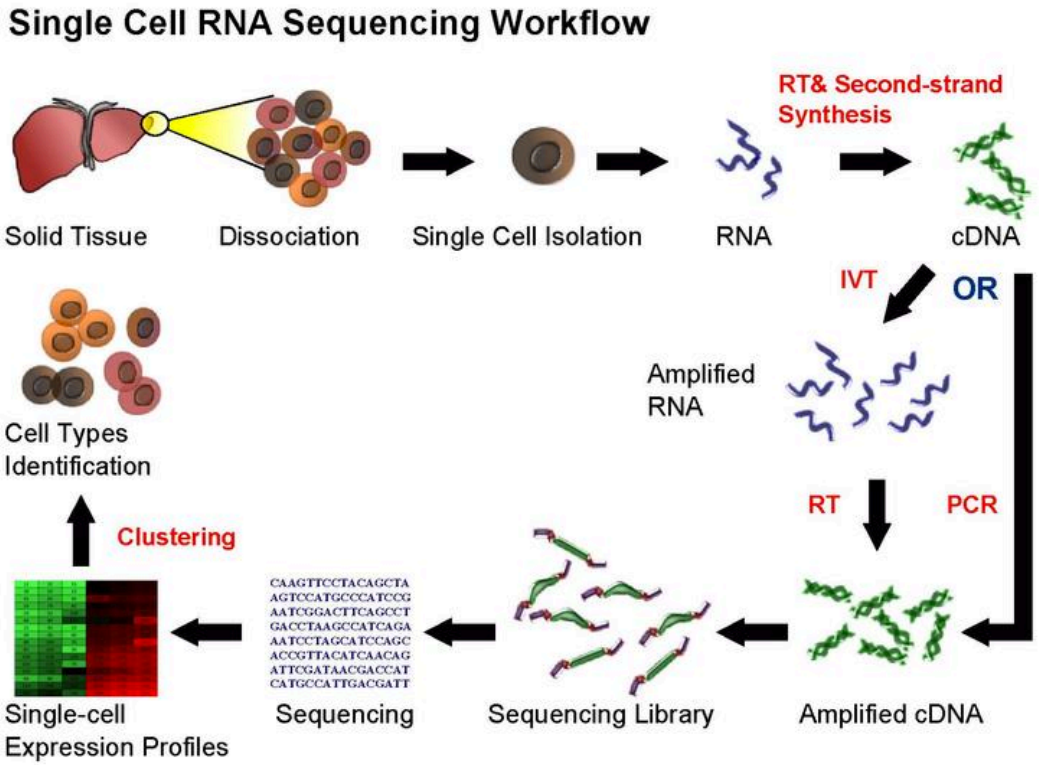

http://biorxiv.org/content/early/2016/08/12/068809

Garalde et al., Highly parallel direct RNA sequencing on an array of nanopores, 2018, Nature Methods 15:201–206

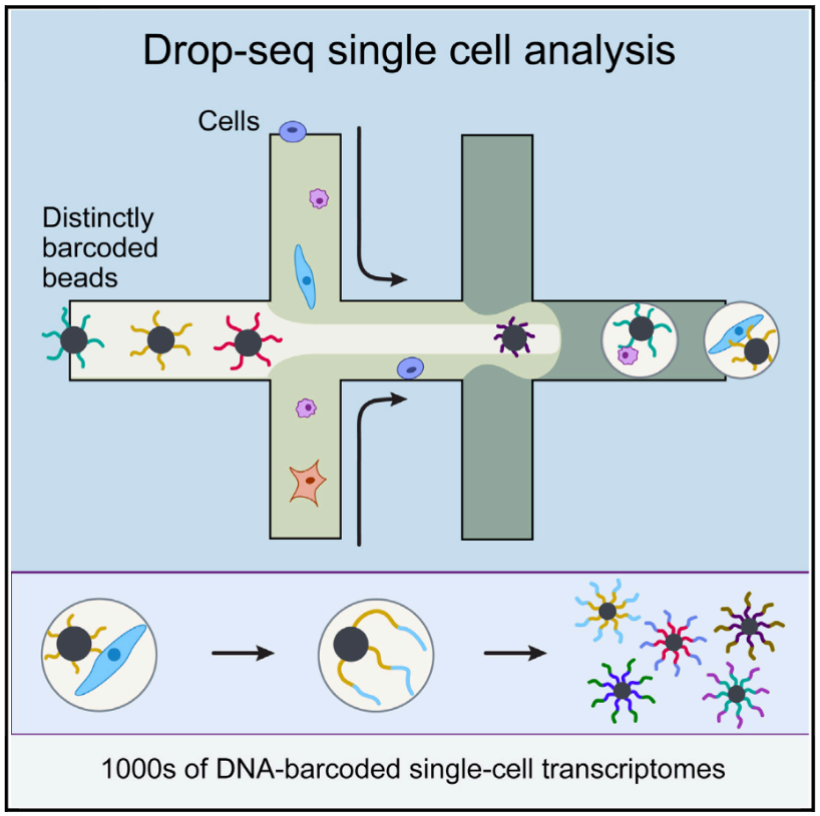
# High errors associated with single molecule sequencing



# single-cell RNA-Seq (scRNA-Seq)




wikipedia




Macosko et al., Cell, 2015

IVT: in vitro transcription

RT: reverse transcription

# Outline

**Review of RNA-Seq procedure**

**Design of DE experiments and results**

- Experimental design
- Multiple test correction

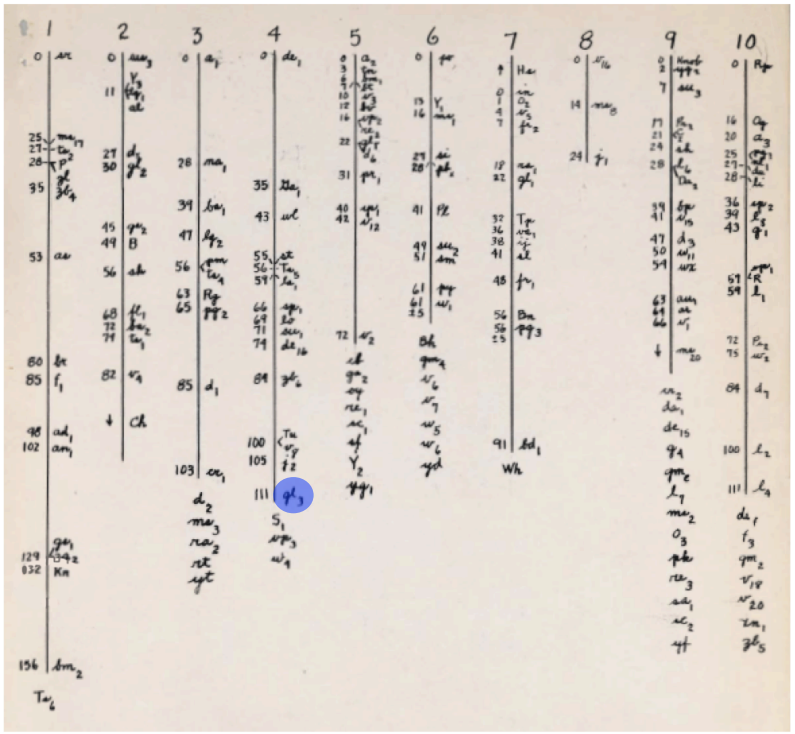
**New technologies**

**Other applications (an example)**

# BSR-Seq, an RNA-Seq based approach for genetic mapping of a mutant gene

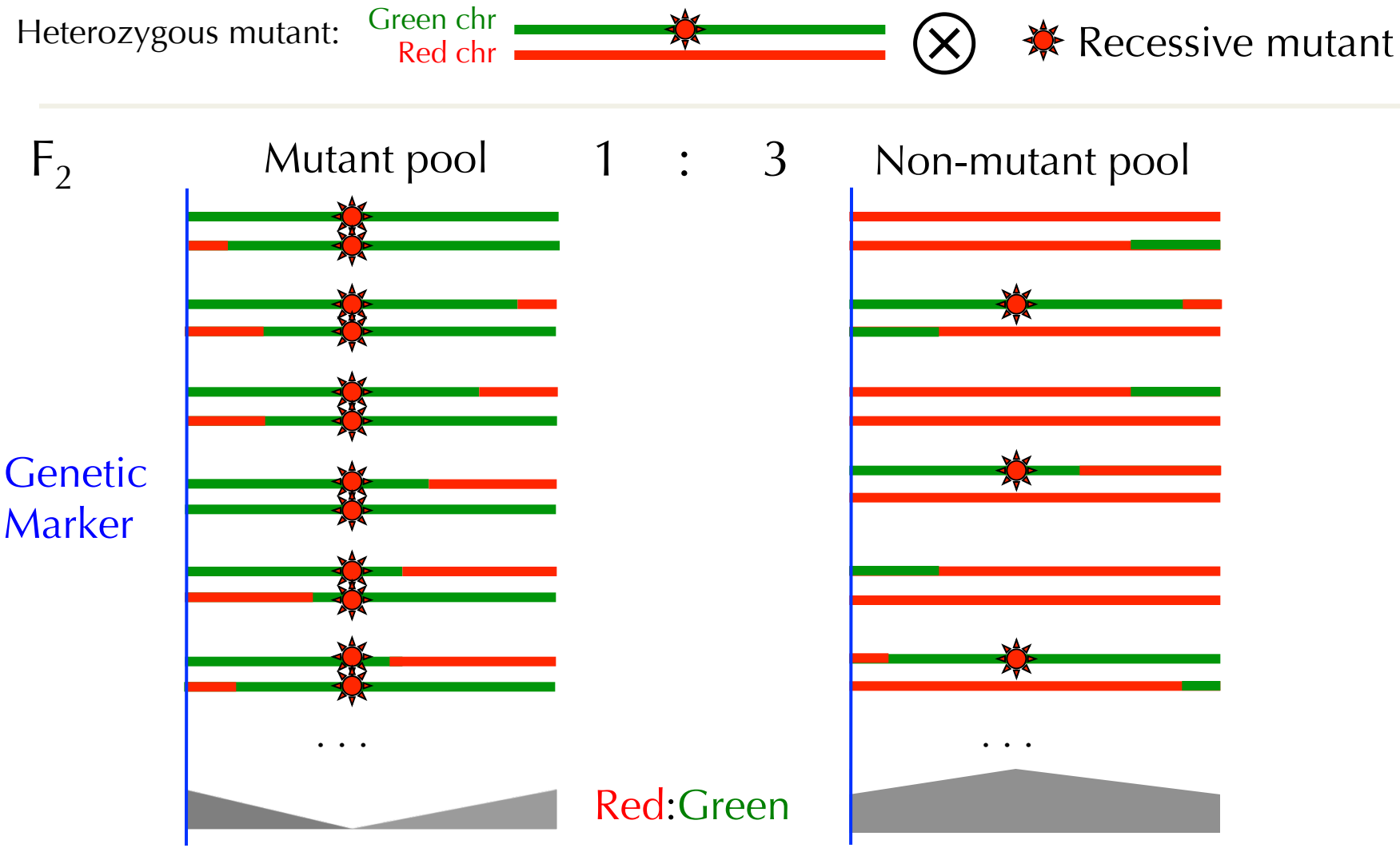
*gl3* mutant

*gl3* wt



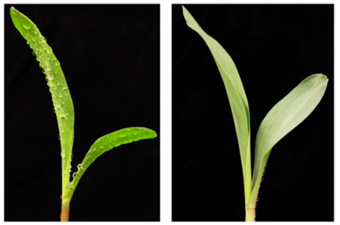
An early maize genetic map (1937) – image from maizeGDB.org

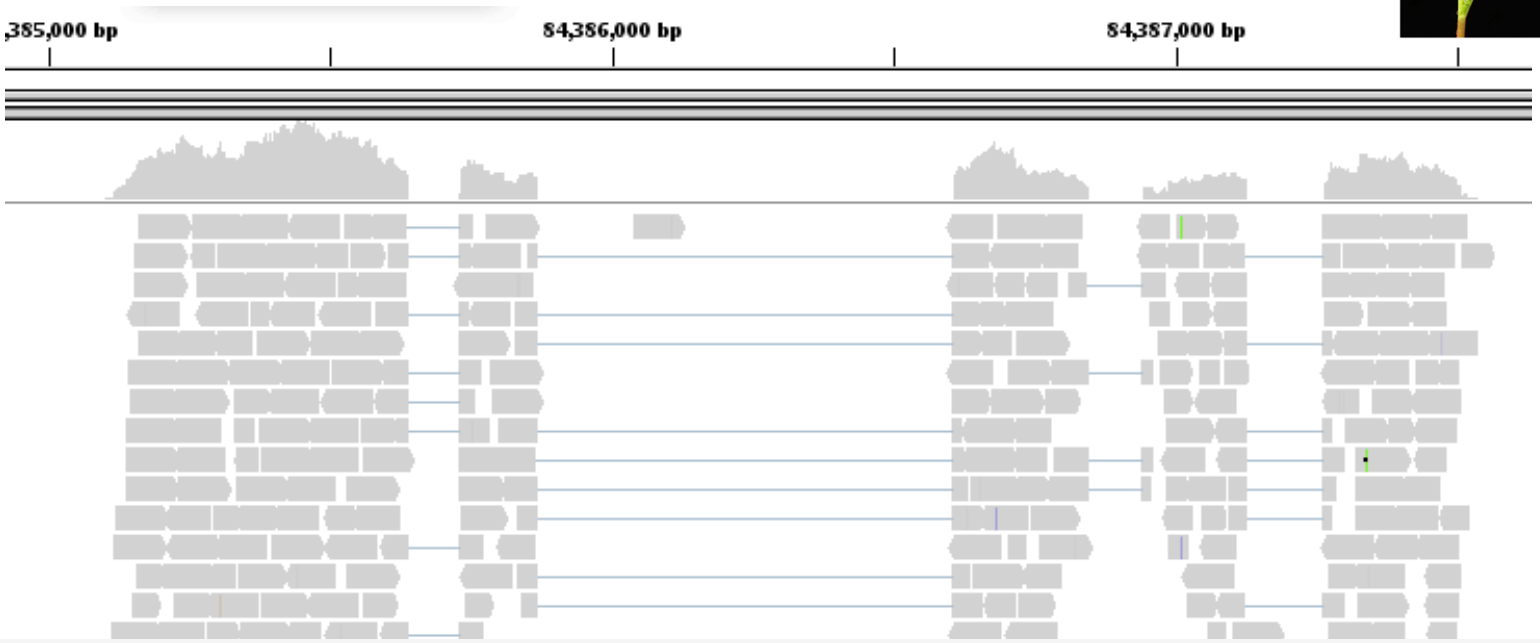
# Example of a mutant mapping using BSA



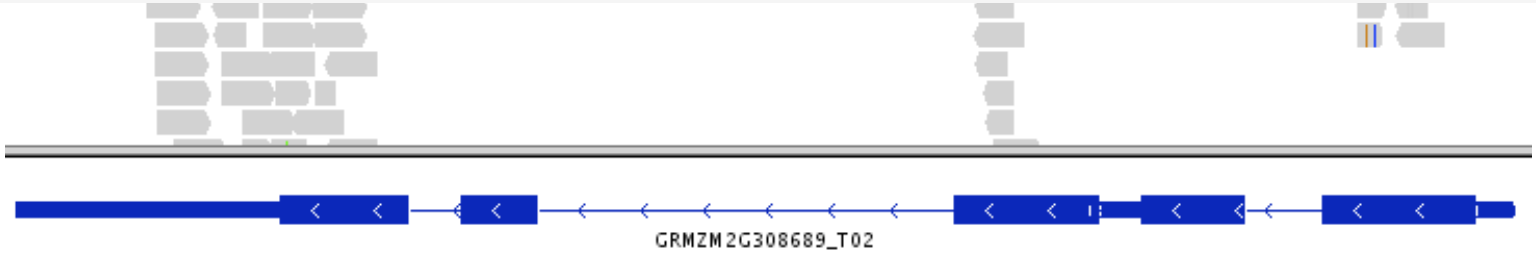

The ratio of Red:Green is proportional to the genetic distance between the mutant gene and the genetic marker.

# RNA-Seq



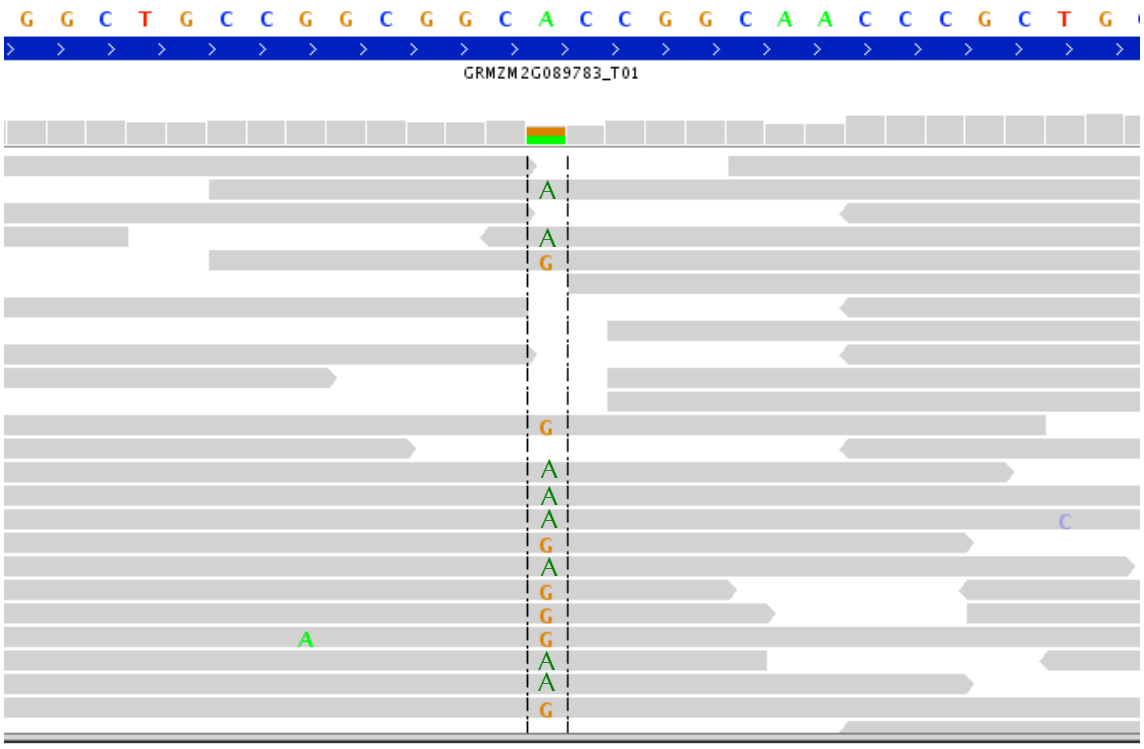



- RNA-Seq generated ~13 millions of reads per sample
- 64,852 SNPs were identified and quantified in both the mutant pool and the non-mutant pool



# Example of a SNP UNLINKED to a mutant gene

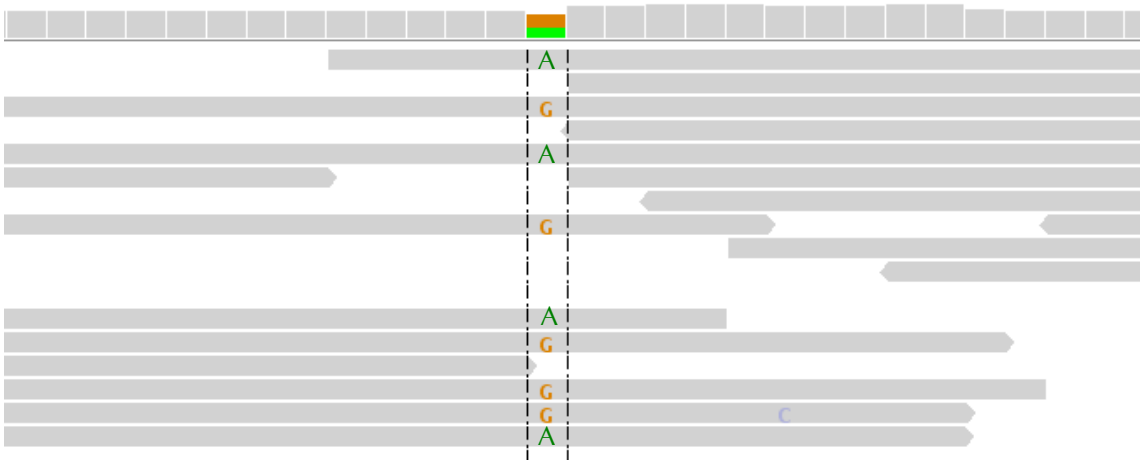
Reference




Reads from Mutants

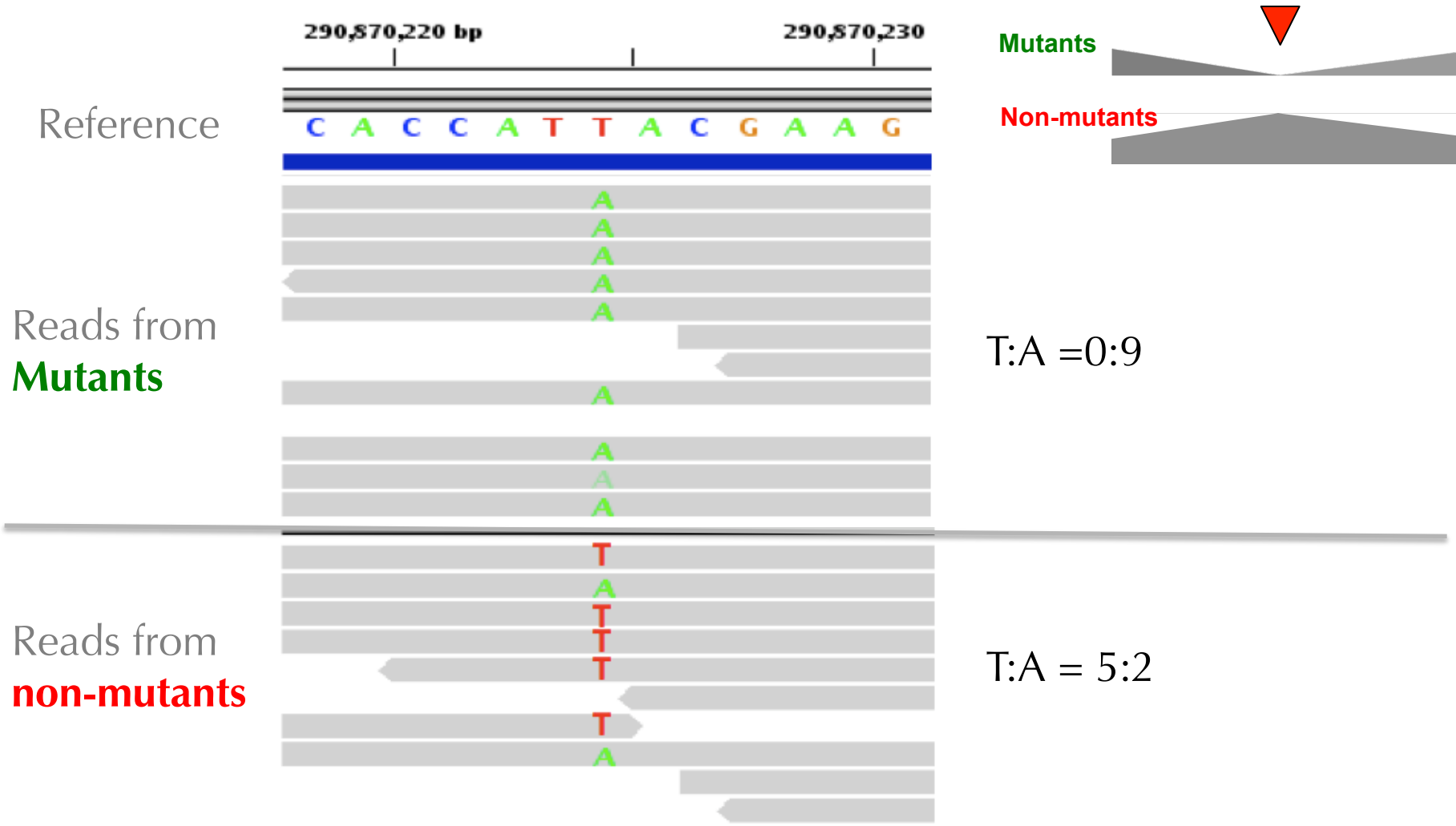
**A:G = 8:7**

Reads from non-mutants



**A:G = 4:5**

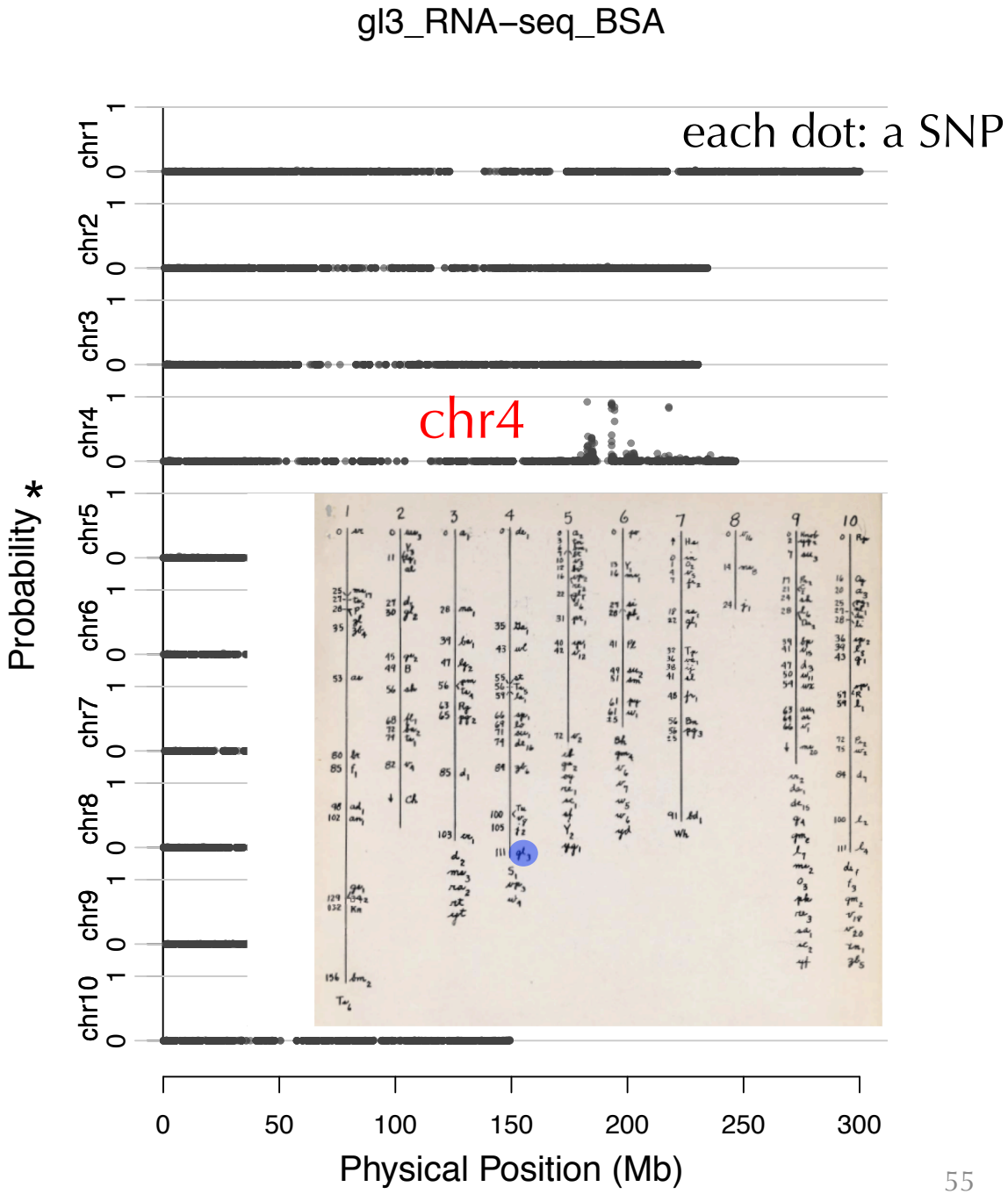
# Example of a SNP Completely LINKED to a mutant gene



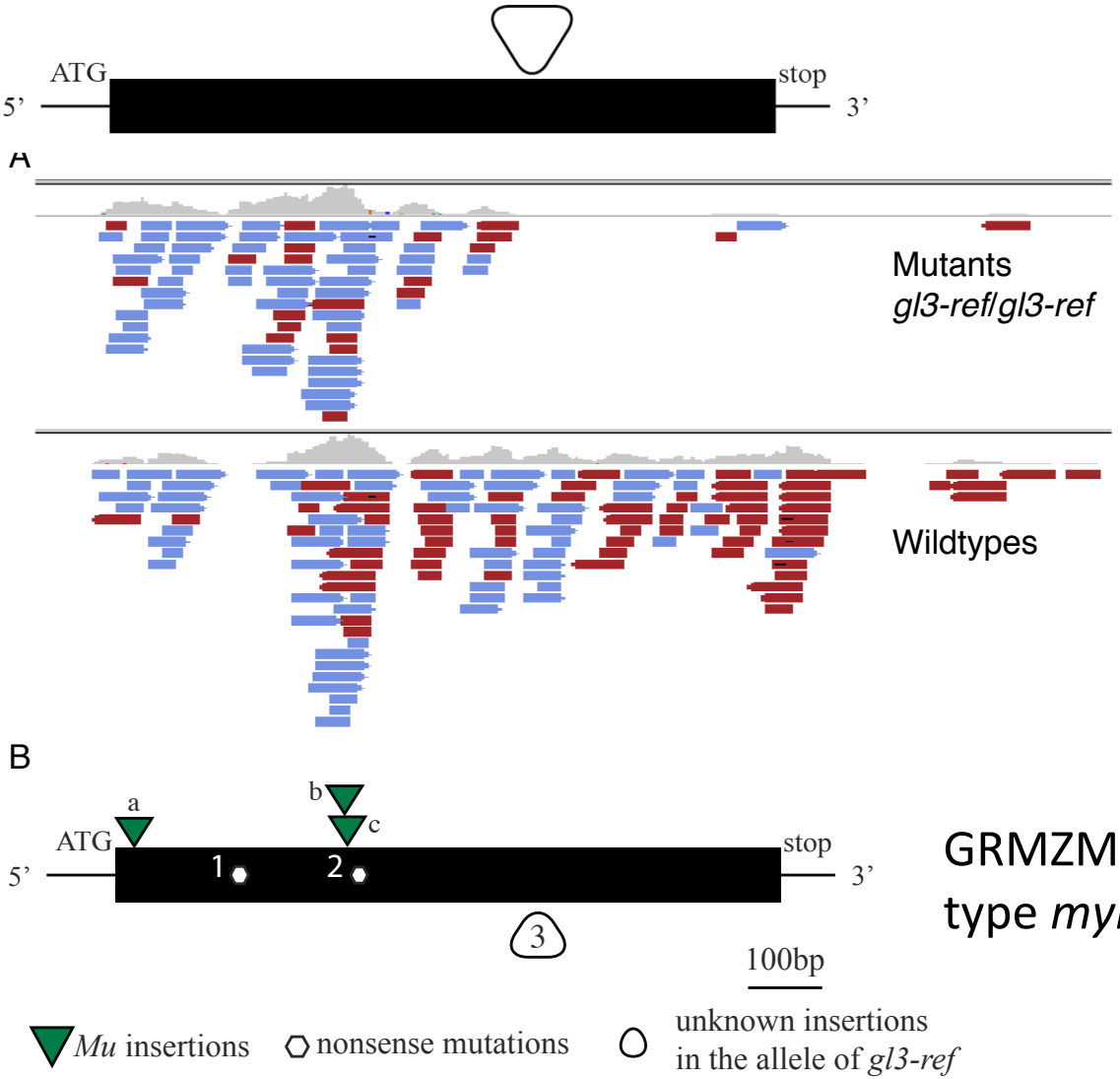

A Bayesian approach was developed to calculate the probability of complete linkage between the SNP site and the mutant gene.

# BSR-Seq mapping result

* Probability of complete linkage between the SNP and the *gl3* gene



# *gl3* cloning

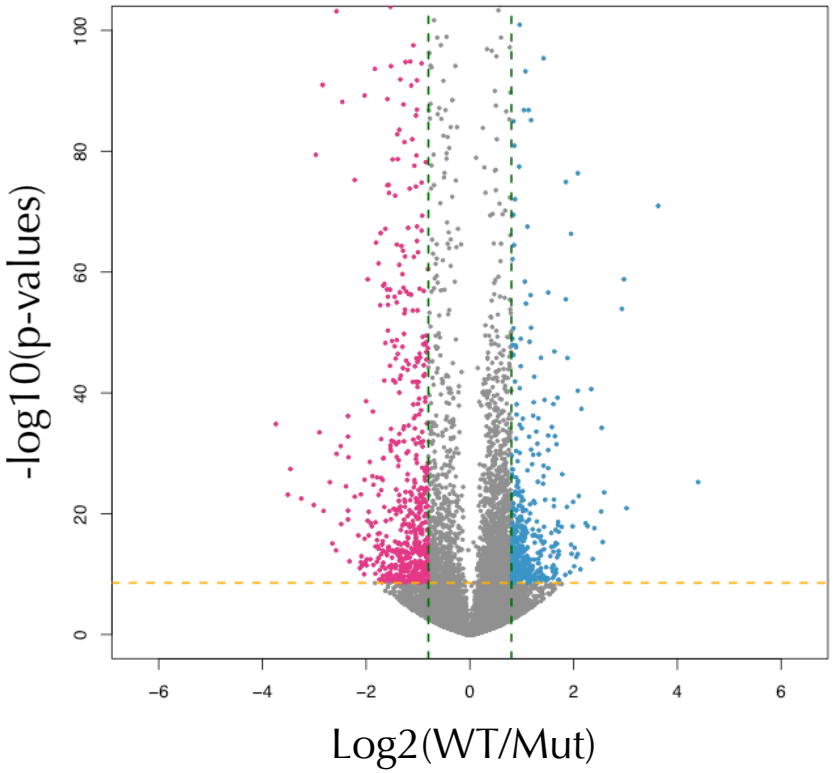


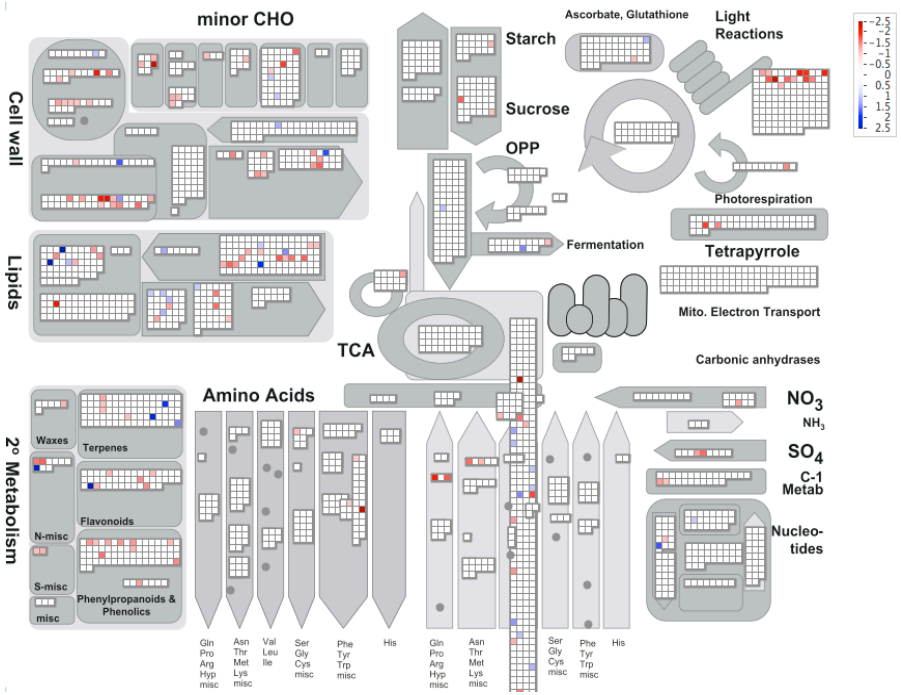

*Down-regulated*
*Partially expressed*

GRMZM2G162434, an R2R3 type *myb* transcription factor

Figure 2

# Differential expression analysis of *gl3* using the SAME RNA-Seq data used for BSR-Seq



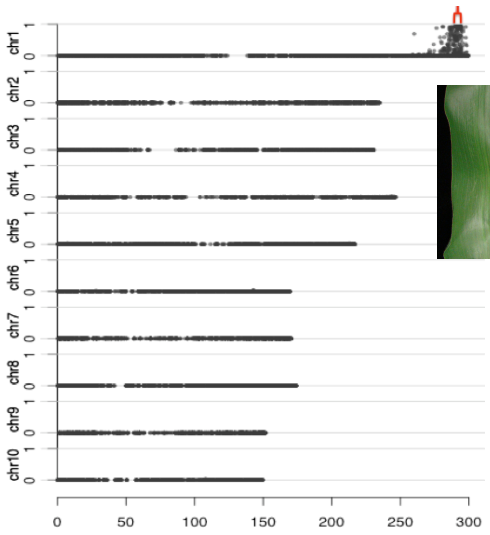

MapMan view

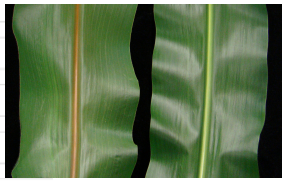
1,095 genes with significantly differential expression

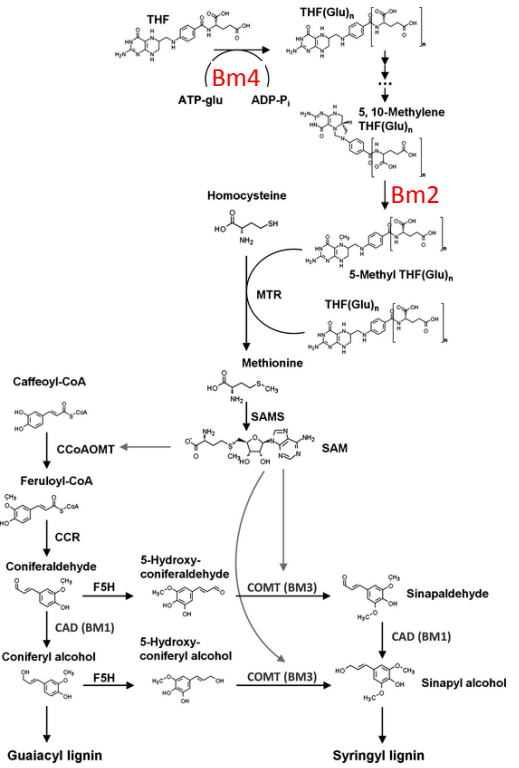
# Other successful cases in maize with BSR-Seq

1. cloning of lignin related genes (*bm2* and *bm4*)

- Tang *et al*. Plant J., 2: 380–392 (2013)
- Li *et al*. Plant J., 81: 493-504 (2015)

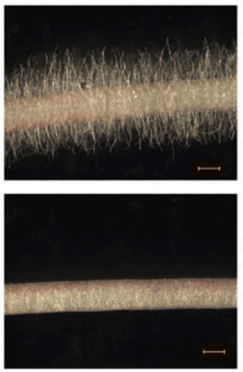


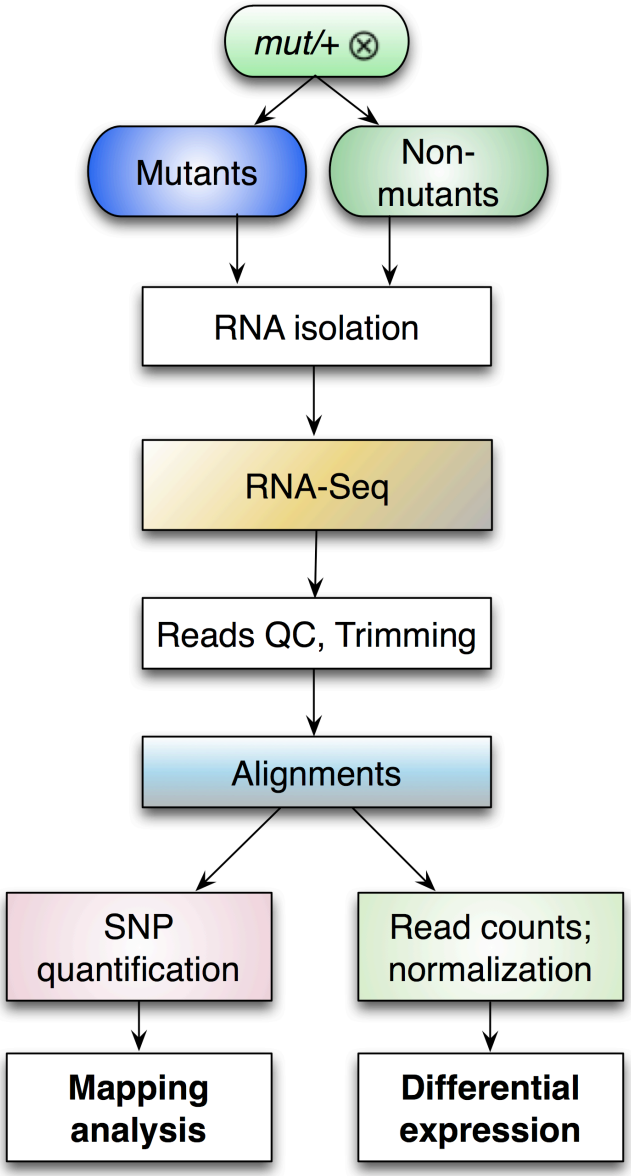



2. cloning of another glossy gene and and a root hairless gene (*gl13, rth6*)

- Li *et al*. PLoS One, 8: e82333 (2013)
- Li *et al*. Sci Rep, srep34395 (2016)




# Summary (BSR-Seq)




1. Define genetic markers
2. Map the causal gene or identify trait-associated genetic markers
3. Genome-wide gene expression

Liu, S *et al.*, 2012 PLoS ONE, 7(5): e36406.

# Summary

- Randomization of experimental units is needed and each sample requires certain sequencing depth.
- Biological replication rather than technical replication is typically needed for an RNA-Seq experiment.
- P-values need to be corrected to account for multiple tests. The FDR method is a reliable approach for the correction.
- Many bioinformatics pipelines and statistical methods have been developed. Methods and parameters need to be carefully selected.

# Selected references


- Benjamini Y, Hochberg Y. (FDR co-authors). 1995. Controlling the False Discovery Rate - a Practical and Powerful Approach to Multiple Testing. Journal of the Royal Statistical Society Series B-Methodological 57:289-300.
- Conesa A, Madrigal P, Tarazona S, Gomez-Cabrero D, Cervera A, McPherson A, Szczesniak MW, Gaffney DJ, Elo LL, Zhang X, et al. 2016. A survey of best practices for RNA-seq data analysis. Genome Biol 17:13.
- Liu S, Yeh CT, Tang HM, Nettleton D, Schnable PS. 2012. Gene mapping via bulked segregant RNA-Seq (BSR-Seq). PLoS One 7:e36406.
- Love MI, Huber W, Anders S. (DeSeq2 co-authors). 2014. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol 15:550.
- Macosko EZ, Basu A, Satija R, Nemesh J, Shekhar K, Goldman M, Tirosh I, Bialas AR, Kamitaki N, Martersteck EM, et al. 2015. Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets. Cell 161:1202-1214.
- Robinson MD, McCarthy DJ, Smyth GK. 2010. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics 26:139-140.
- Wang B, Tseng E, Regulski M, Clark TA, Hon T, Jiao Y, Lu Z, Olson A, Stein JC, Ware D. 2016. Unveiling the complexity of the maize transcriptome by single-molecule long-read sequencing. Nat Commun 7:11708.