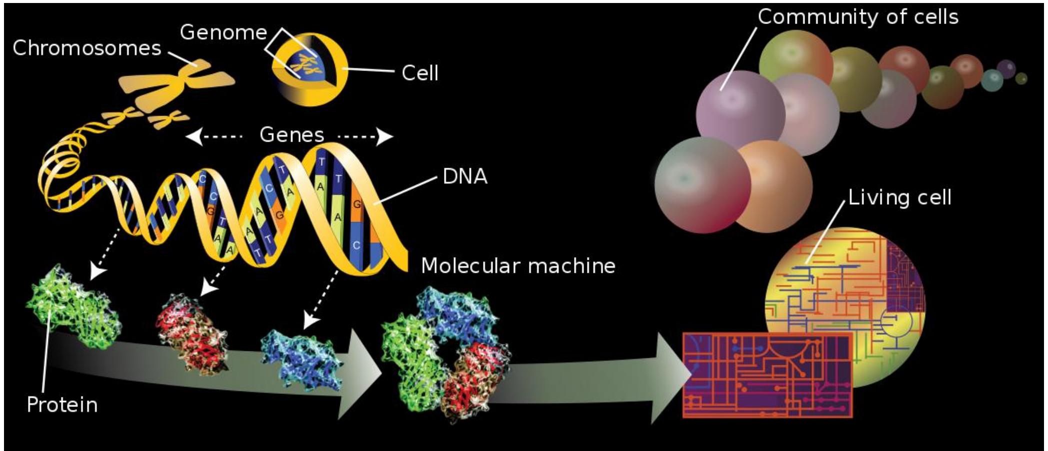
生物信息学：
组学时代的生物信息数据挖掘和理解

2020年秋



有关信息

- 授课教师： 宁康， 张礼斌， 陈鹏
  - Email: ningkang@hust.edu.cn
  - Office: 华中科技大学东十一楼504室
  - Phone: 87793041, 18627968927

- 课程网页
  - http://www.microbioinformatics.org/Bioinformatics.html
  - QQ群：

课程安排

（生物信息中的算法设计与概率统计模型）

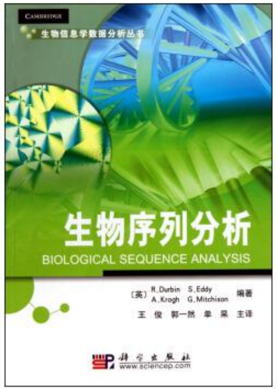
- 生物背景和课程简介
- 生物信息学和生物数据挖掘
  - 生物数据的格式及其意义
    - 序列数据
    - 树状数据
    - 网络数据
    - 表达数据等
  - 生物数据库及其用法
  - 生物信息基本算法
    - 双序列联配
    - 多序列联配
    - 基因组组装算法
    - 基因预测和功能注释
    - 系统发育树构建
    - 蛋白质结构预测
    - 生物调控网络解析
  - 组学数据分析方法
    - 基因组变异分析
    - 基因表达和比较分析
    - 非编码RNA分析
    - 蛋白组分析
    - 宏基因组分析
  - 系统生物学与交叉科学
- 面向生物大数据挖掘的深度学习

研究对象：
生物序列，
进化树，
生物网络，
基因表达

…

方法：
生物计算与生物信息

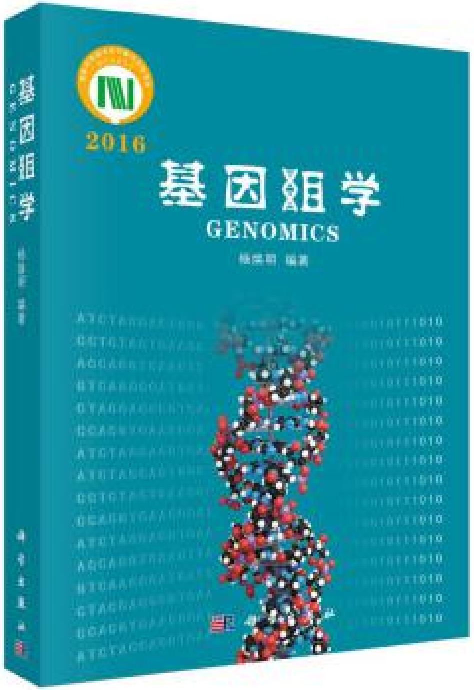
教材及参考书目

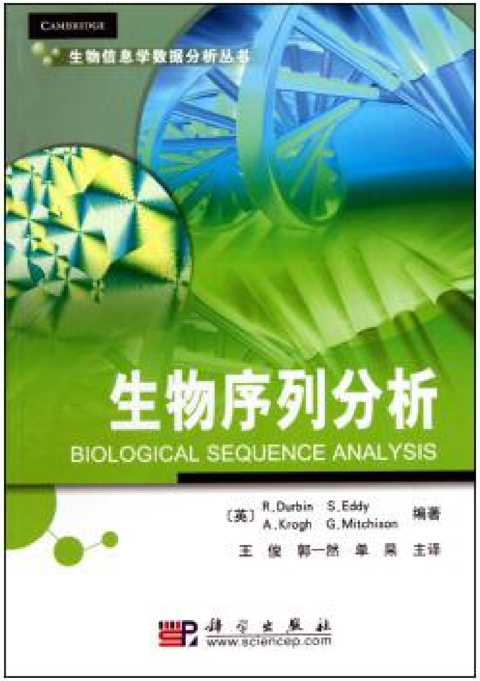


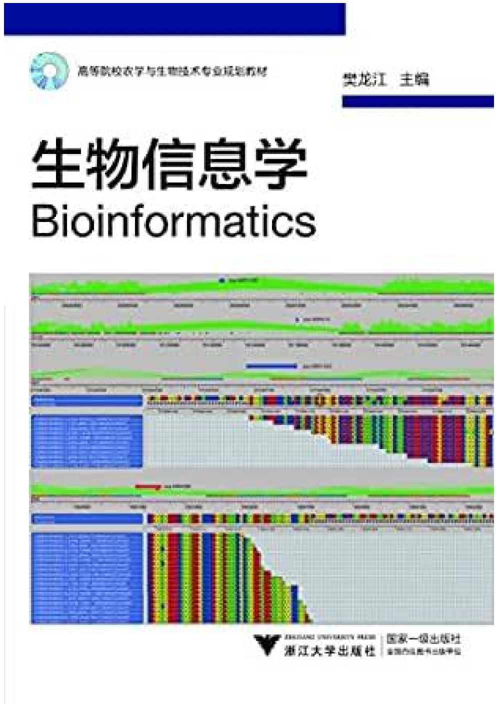
- **教学参考书：**
- 《生物序列分析》（第1版）.科学出版社. 2010年8月出版. R. Durbin等编著， 王俊等主译.

- **课外文献阅读：**
- 《生物信息学》（第1版）. 浙江大学出版社. 2017年3月出版. 樊龙江主编.
- 《基因组学》（第1版）. 科学出版社. 2016年10月出版. 杨焕明主编.

References


2016
基因组学
GENOMICS
杨焕明 编著
科学出版社


CAMBRIDGE
生物信息学数据分析丛书
生物序列分析
BIOLOGICAL SEQUENCE ANALYSIS
〔英〕 R.Durbin S.Eddy A.Krogh G.Mitchison 编著
王 俊 郭一然 单 杲 主译
科学出版社
www.sciencep.com


高等院校农学与生物技术专业规划教材
樊龙江 主编
生物信息学
Bioinformatics
ZHEJIANG UNIVERSITY PRESS
浙江大学出版社
国家一级出版社
全国百佳图书出版单位

Slides credits

——生物信息学研究方法概述：北京大学生物信息中心

——生物统计学：卜东波@中国科学院计算技术研究所，邓明华@北京大学

——神经网络与深度学习: 邱锡鹏@复旦大学

——Introduction to Computational Biology and Bioinformatics: Xiaole Shirley Liu Lab@Havard University

——Combinatorial Methods in Computation Biology: Ken Sung Lab@NUS

——Deep Learning in the Life Sciences: MIT

Transcriptomics

转录组学

Transcriptome: An evolving definition

❍(The population of) mRNAs expressed by a genome at any given time (Abbott, 1999)

- **转录组（Transcriptom）：细胞所包含mRNA的总和。**与基因组不同的是，转录组的定义中包含了时间和空间的限定。

- **转录组学（Transcriptomics）：**研究细胞在某一功能状态下所含**mRNA**的类型与拷贝数；比较不同功能状态下**mRNA**表达的变化，搜寻与功能状态变化紧密相关的重要基因群。

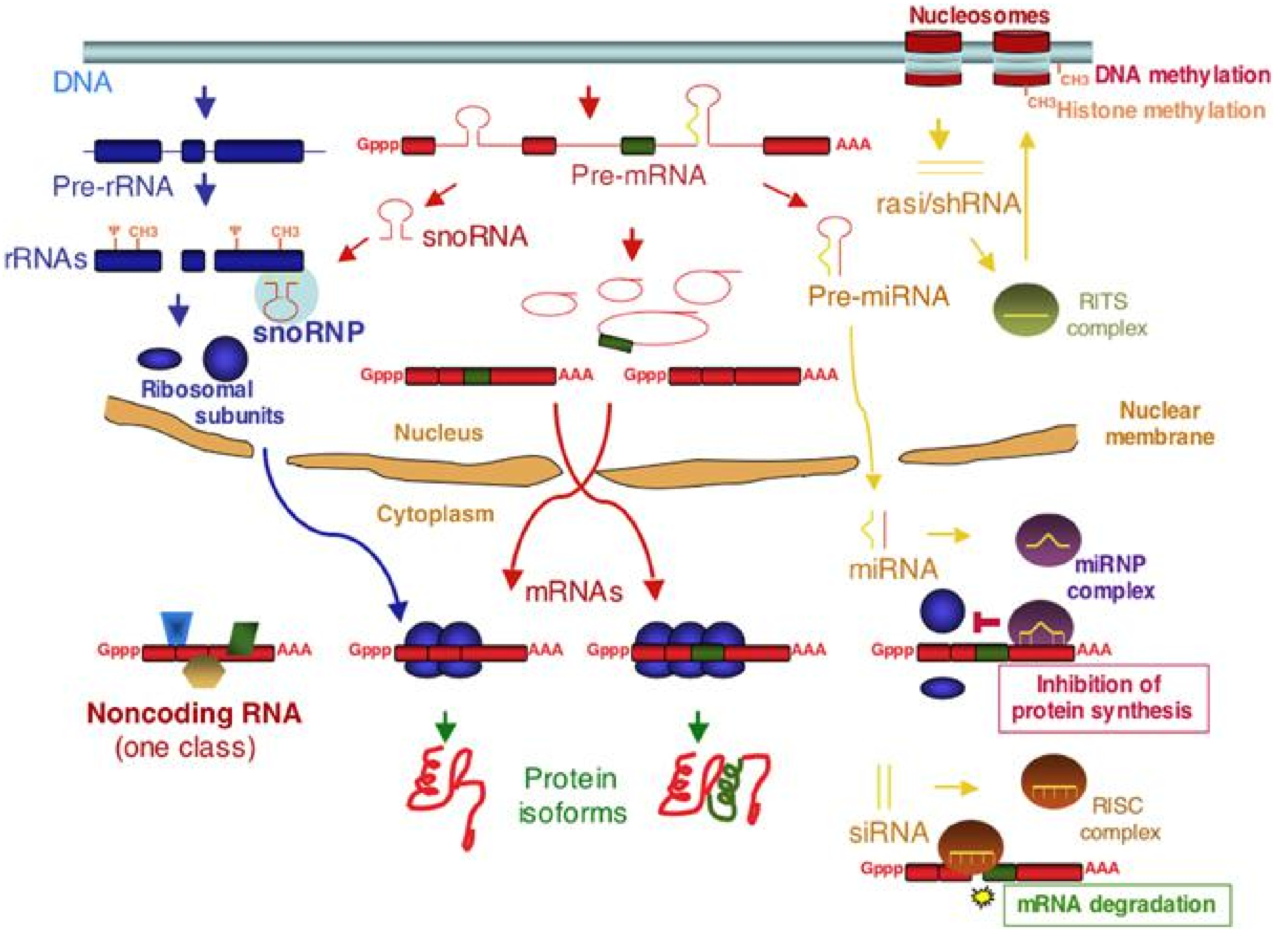
新定义

- The complete collection of transcribed elements of the genome.   (Affymetrix, 2004)
- mRNA
- rRNA, tRNA
- snmRNAs (small non-messenger RNAs)
  - microRNAs and siRNAs (small interferring RNAs)
  - snoRNAs (small nucleolar RNAs) 核仁小分子RNA
  - snRNAs (small nuclear RNAs)
  - Other non-coding RNAs
    - Long non-coding RNA (lncRNA)

转录本

All transcripts

All mRNAs



Transcriptomics

Definition

The study of characteristics and regulation of the functional RNA transcript population of a cell/s or organism at a specific time. Scope

- the population of functional RNA transcripts.
- the mechanisms that regulate the production of RNA transcripts
- dynamics of the trancriptome (time, cell type, genotype, external stimuli)

一、转录组学研究全部RNA的表达及功能

- 转录组（transcriptome）指特定状态下一种细胞或组织所能转录出来的所有RNA的总和。——包括编码RNA，即mRNA和非编码RNA (non-coding RNA, ncRNA)
- 转录组学（transcriptomics）：是在整体水平上研究细胞基因转录情况及转录调控规律的科学。
- RNA组学（RNomics）：是分析、鉴定非信使小RNA（small non-messenger RNA, snmRNA）在特定状态下表达情况、功能及其与蛋白质的相互作用。

- 转录组的特点：受到内外多种因素的调节，因而是动态可变的。能够揭示不同物种、不同个体、不同细胞、不同发育阶段及不同生理病理状态下的基因差异表达信息。

转录组学的研究方法

- **基于测序**：cDNA文库、illumina测序
- **基于杂交**：cDNA芯片（GeneChip，microarray）
- **基因表达聚类**

（一）微阵列是大规模基因组表达谱研究的早期主要技术

- 大规模表达谱或全景式表达谱（global expression profile）：是生物体（组织、细胞）在某一状态下基因表达的整体状况。
- 微阵列或基因芯片（DNA chip）:利用光导化学合成、照相平板印刷以及固相表面化学合成等技术，在固相表面合成成千上万个寡核苷酸探针，并与放射性同位素或荧光物标记的来自不同细胞、组织或整个器官的DNA或mRNA反转录生成的第一链cDNA进行杂交，然后用特殊的检测系统对每个杂交点进行定量分析。

不同的生物芯片技术平台

Spotted Microarrays

点样芯片

ø cDNA Arrays

ø Oligo Arrays

In Situ Oligo Synthesis

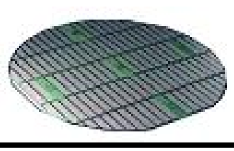
原位合成芯片

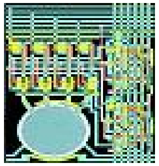
ø Photosynthesis

  ø Planer surface

  ø Microfluidics chip

ø E-field synthesis





Microfluidics

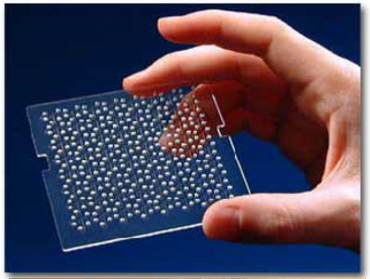
微流体芯片

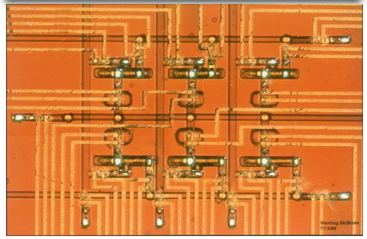
ø Plastics

ø Ceramics

ø Silicon

ø Other materials





Integrated Chips

整合型芯片

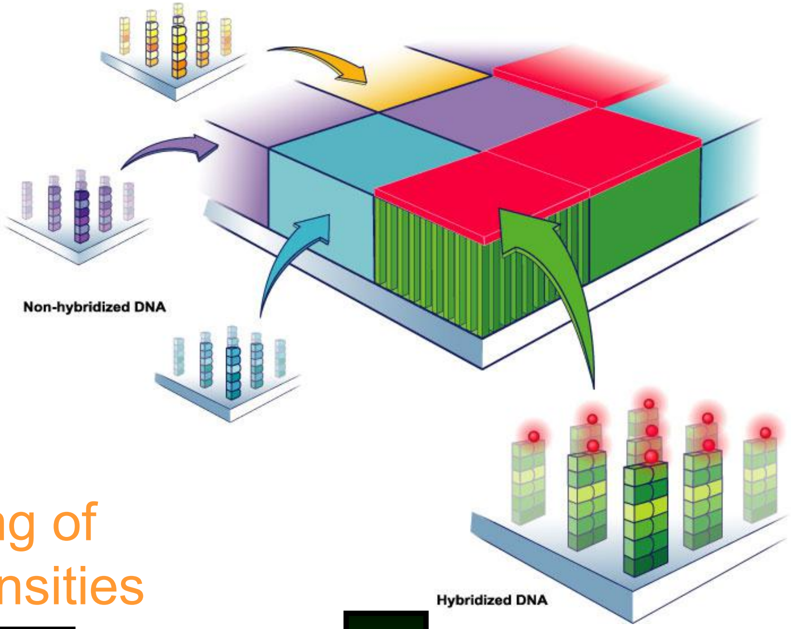
ø Integrated uF, microarray and detection chips with PCR, fluorescence or e-detection

基因芯片的探针

基因芯片的杂交实验

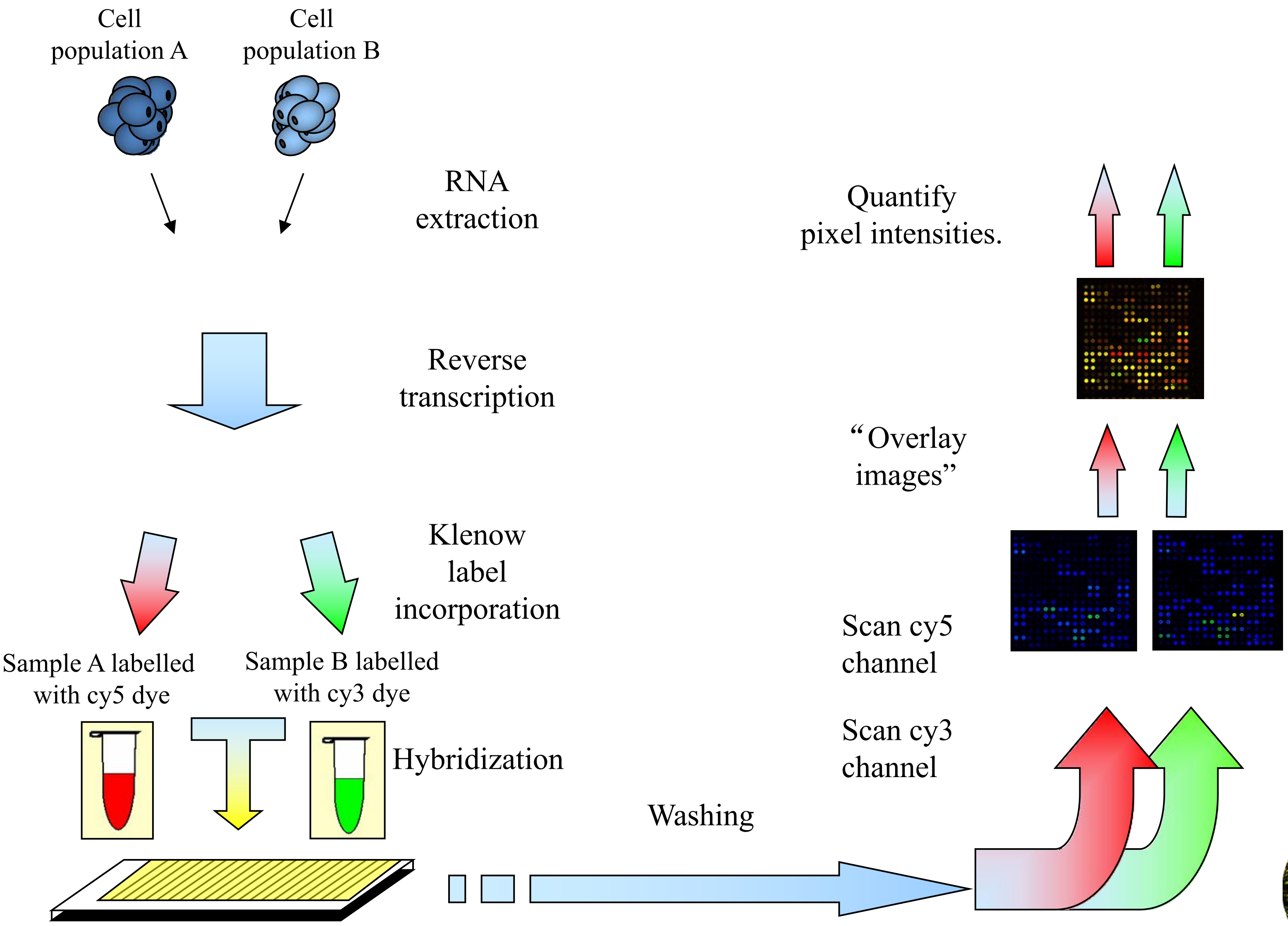
Tagged RNA fragments flushed over array

Laser activation of fluorescent tags



Optical scanning of hybridization intensities

Experimental overview:



Cell population A

Cell population B

RNA extraction

Reverse transcription

Klenow label incorporation

Sample A labelled with cy5 dye

Sample B labelled with cy3 dye

Hybridization

Washing

Scan cy3 channel

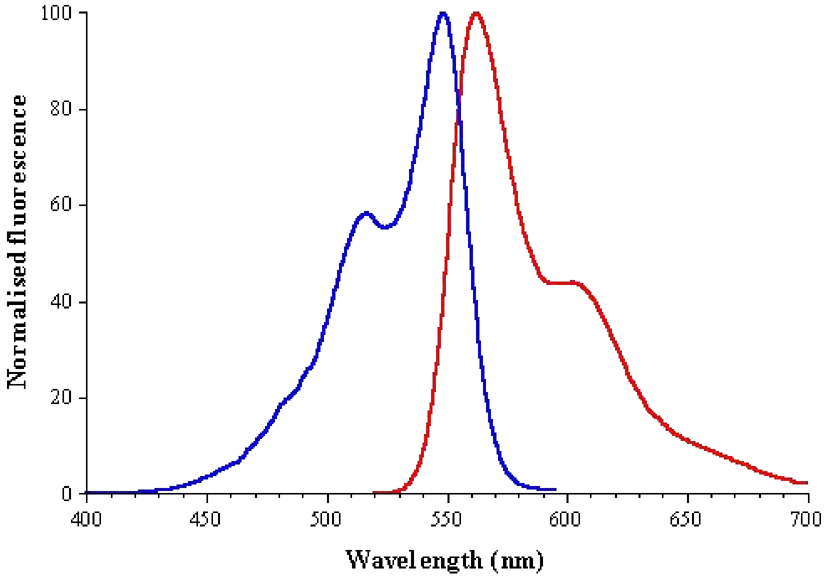
Scan cy5 channel

“Overlay images”

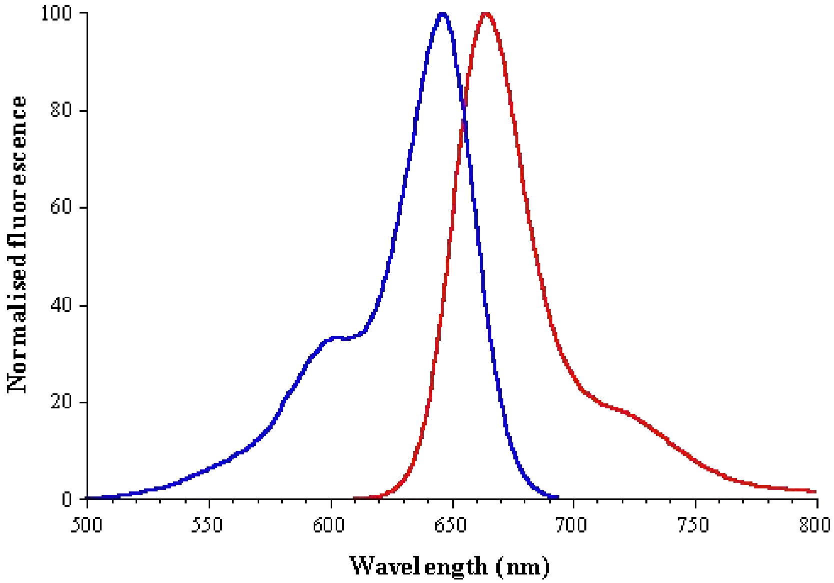
Quantify pixel intensities.

Harwell µArray Facility

Cy3和Cy5

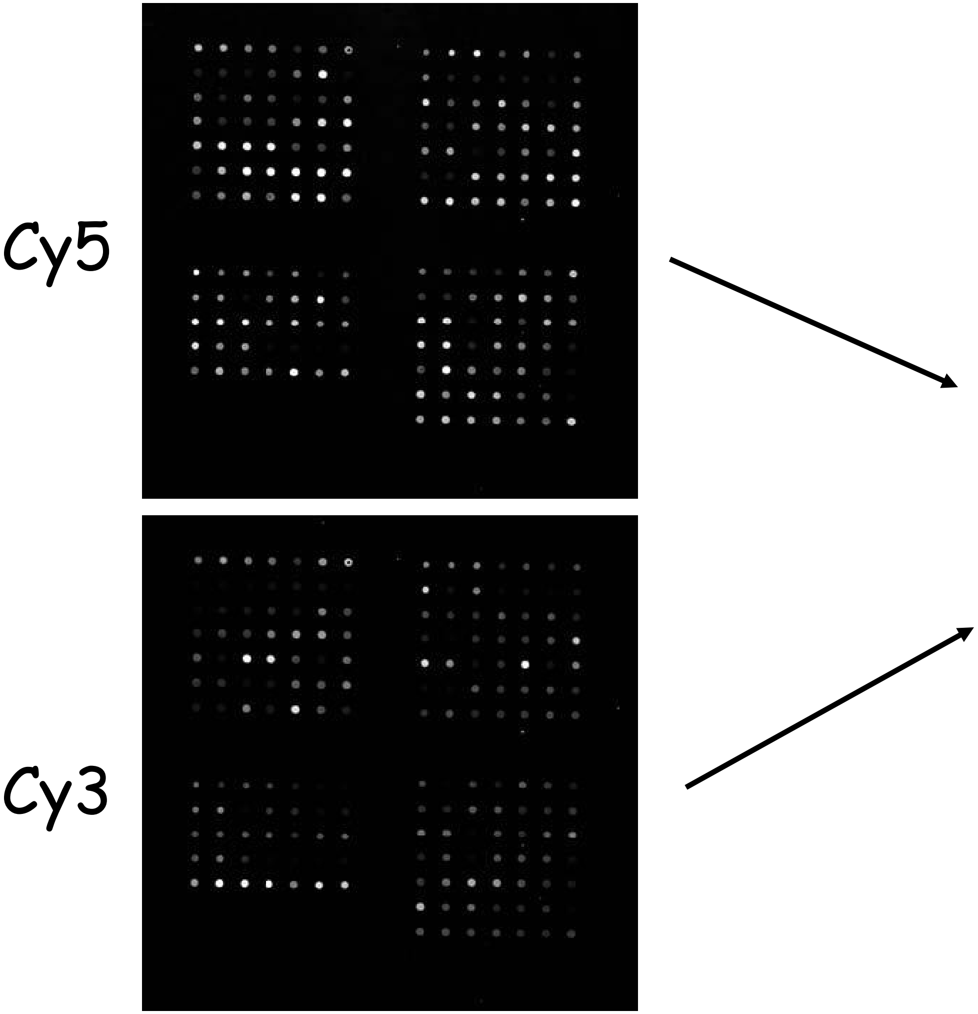


Normalised fluorescence
100
80
60
40
20
0
400
450
500
550
600
650
700
Wavelength (nm)



Normalised fluorescence
100
80
60
40
20
0
500
550
600
650
700
750
800
Wavelength (nm)

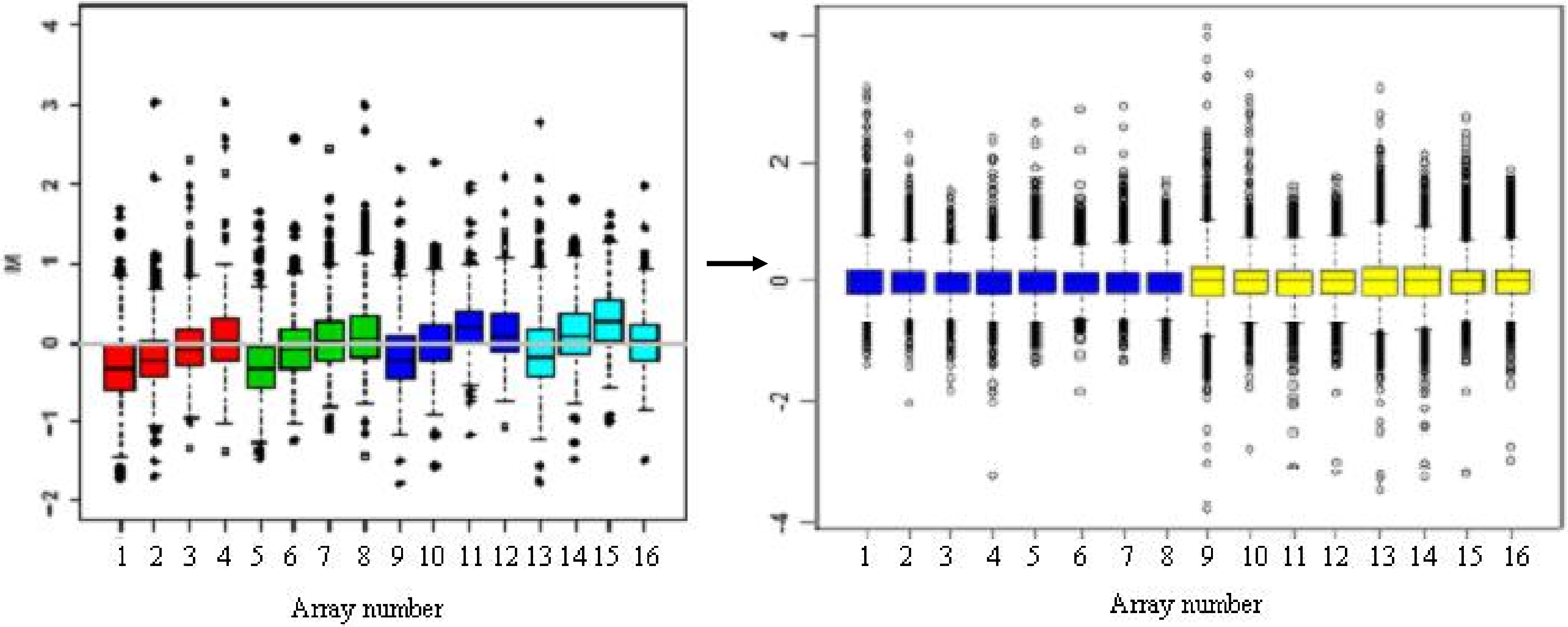
图像扫描

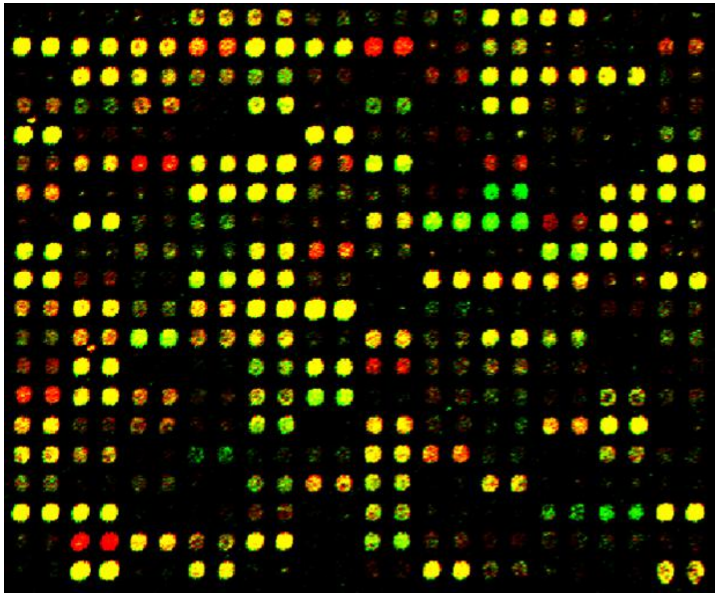


Cy5

Cy3

归一化


M
4
3
2
1
0
-1
-2
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16
Array number
4
2
0
-2
-4
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16
Array number



Red – increase of Cy5 sample transcripts

Green – increase of Cy3 sample transcripts

Yellow – equal abundance

Limit of Detection: 1 in 30,000 transcripts

~ 20 transcripts/cell

差异基因筛选

- 原理：采用cy3/cy5的ratio值对差异基因进行 判断，或采用统计方法对差异基因进行统计推断。

方法：倍数法：cy3/cy5比值大于2或者小于0.5

- Z值法： Z=(X-μ)/σ

作用：发现两个样本间的差异表达基因，便于后续分析。

Microarray and GeneChip Approaches

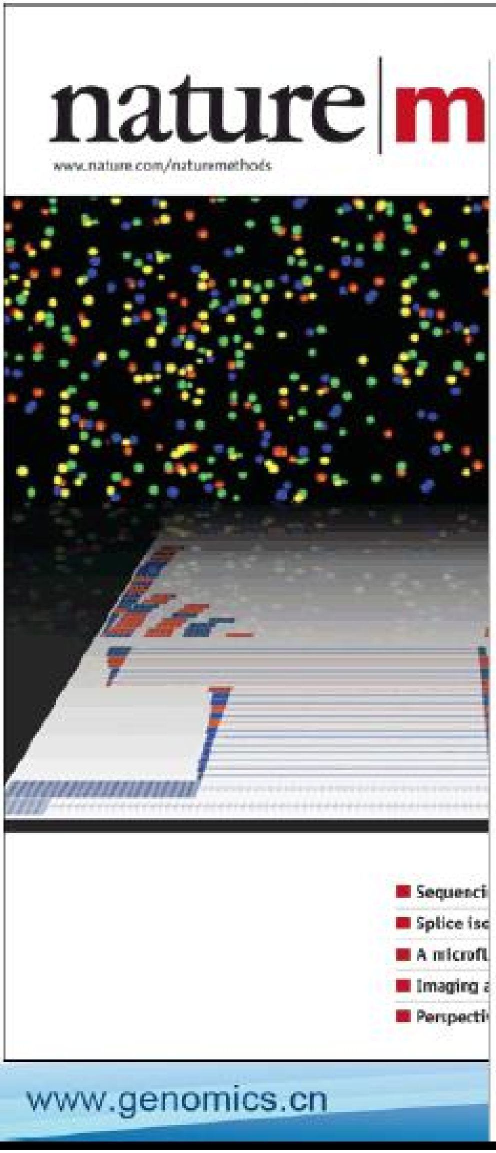
Advantages:

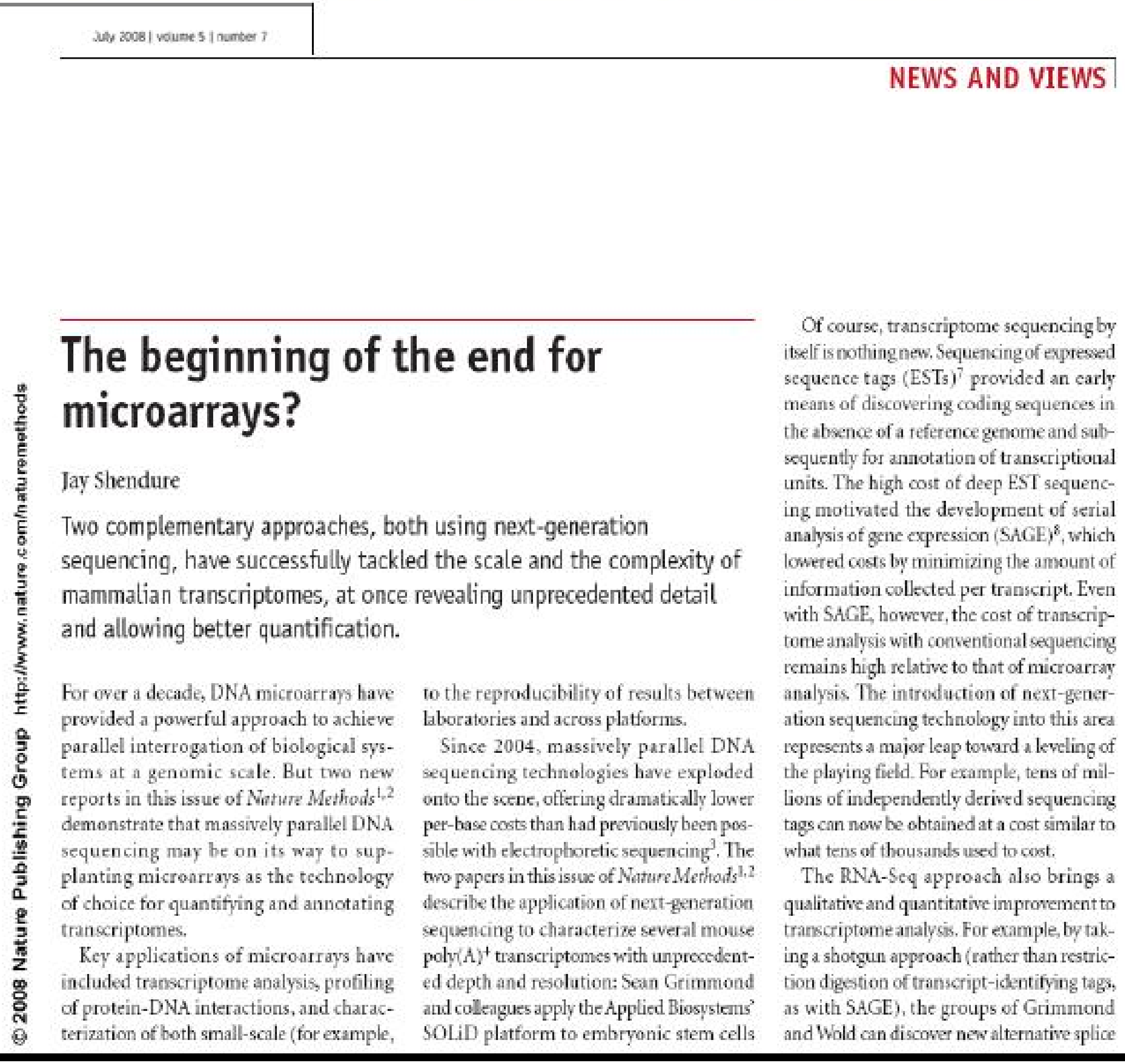
- ❍ Rapid
- ❍ Method and data analysis well described and supported
- ❍ Robust
- ❍ Convenient for directed and focussed studies

Disadvantages:

- ❍ Closed system approach
- ❍ Difficult to correlate with absolute transcript number
- ❍ Sensitive to alternative splicing ambiguities

The Beginning of a New Era?





July 2008 | volume 5 | number 7

NEWS AND VIEWS

The beginning of the end for microarrays?

Jay Shendure

Two complementary approaches, both using next-generation sequencing, have successfully tackled the scale and the complexity of mammalian transcriptomes, at once revealing unprecedented detail and allowing better quantification.

For over a decade, DNA microarrays have provided a powerful approach to achieve parallel interrogation of biological systems at a genomic scale. But two new reports in this issue of *Nature Methods*[1,2] demonstrate that massively parallel DNA sequencing may be on its way to supplanting microarrays as the technology of choice for quantifying and annotating transcriptomes.

Key applications of microarrays have included transcriptome analysis, profiling of protein-DNA interactions, and characterization of both small-scale (for example,

to the reproducibility of results between laboratories and across platforms.

Since 2004, massively parallel DNA sequencing technologies have exploded onto the scene, offering dramatically lower per-base costs than had previously been possible with electrophoretic sequencing[3]. The two papers in this issue of *Nature Methods*[1,2] describe the application of next-generation sequencing to characterize several mouse poly(A)^+ transcriptomes with unprecedented depth and resolution: Sean Grimmond and colleagues apply the Applied Biosystems' SOLiD platform to embryonic stem cells

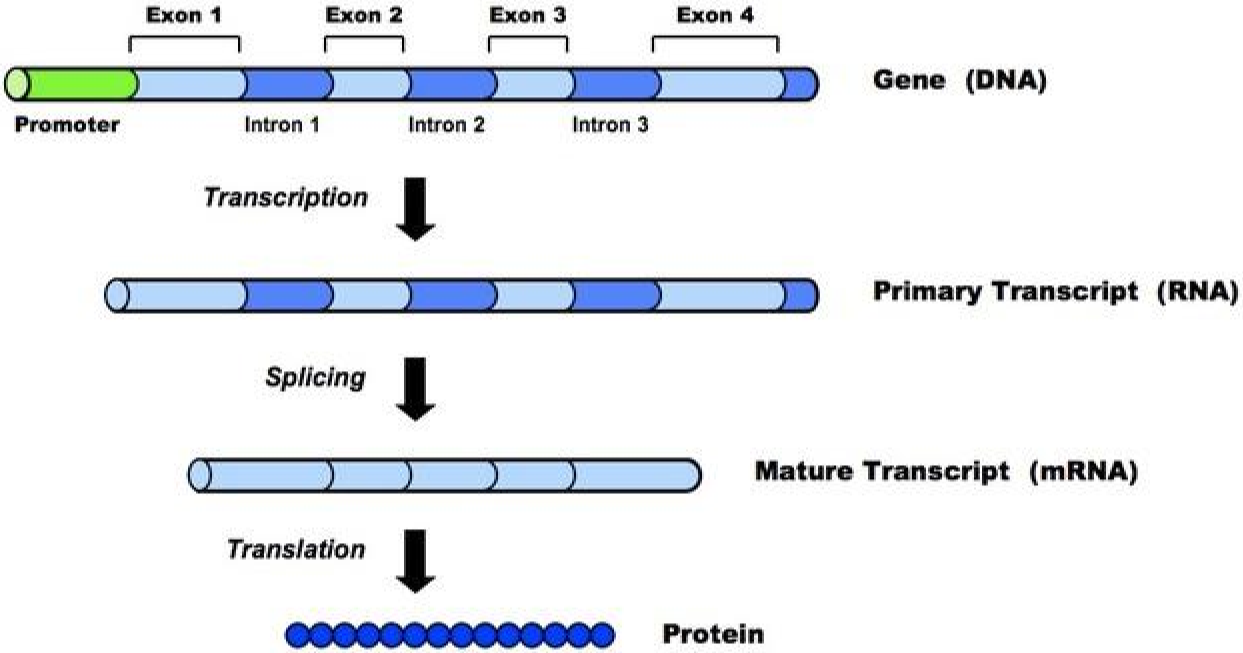
Of course, transcriptome sequencing by itself is nothing new. Sequencing of expressed sequence tags (ESTs)[7] provided an early means of discovering coding sequences in the absence of a reference genome and subsequently for annotation of transcriptional units. The high cost of deep EST sequencing motivated the development of serial analysis of gene expression (SAGE)[8], which lowered costs by minimizing the amount of information collected per transcript. Even with SAGE, however, the cost of transcriptome analysis with conventional sequencing remains high relative to that of microarray analysis. The introduction of next-generation sequencing technology into this area represents a major leap toward a leveling of the playing field. For example, tens of millions of independently derived sequencing tags can now be obtained at a cost similar to what tens of thousands used to cost.

The RNA-Seq approach also brings a qualitative and quantitative improvement to transcriptome analysis. For example, by taking a shotgun approach (rather than restriction digestion of transcript-identifying tags, as with SAGE), the groups of Grimmond and Wold can discover new alternative splice

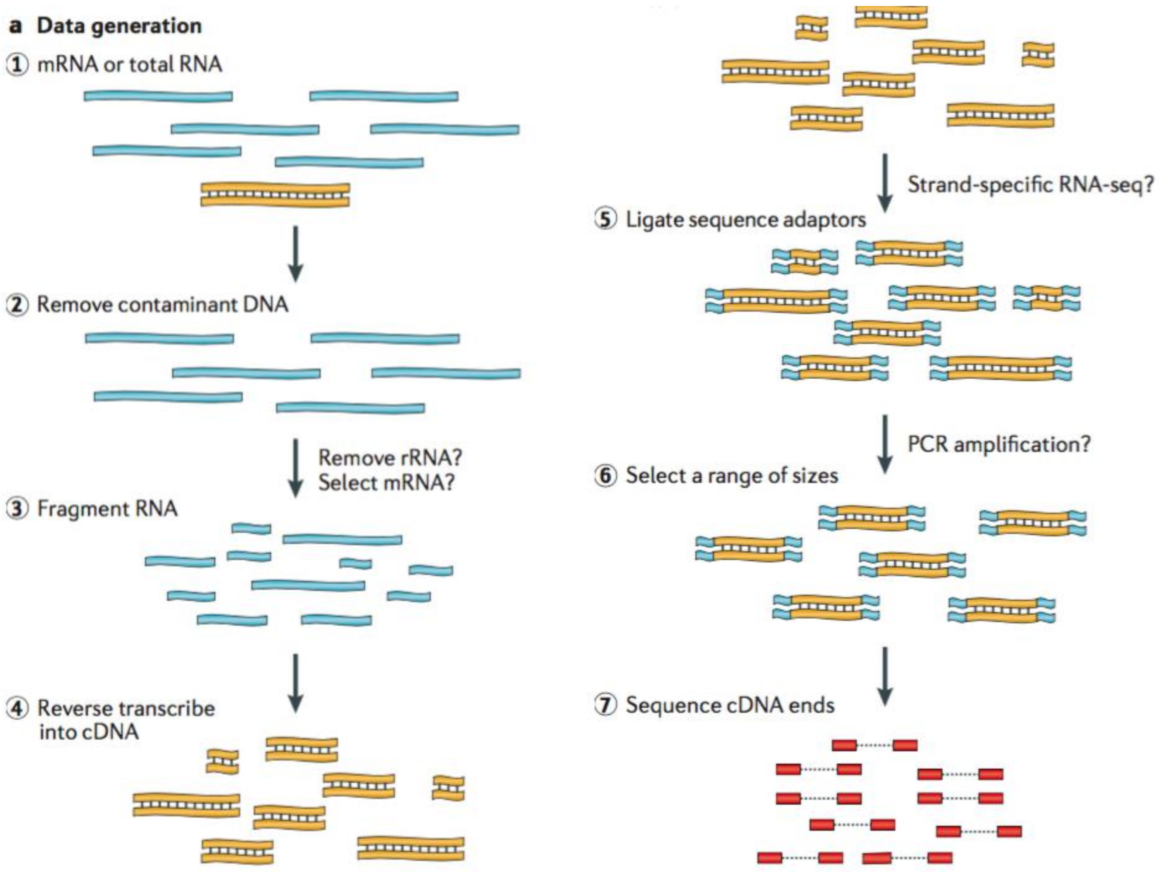
高通量测序

- 高通量测序技术（High-throughput sequencing）是指能够一次并行对几十万到几百万条DNA分子进行序列测定，每一次序列测定的读长一般较短的测序技术。
- 高通量测序技术是对传统测序一次革命性的改变，一次对几十万到几百万条DNA分子进行序列测定，因此在有些文献中称其为下一代测序技术(next generation sequencing)足见其划时代的改变，同时高通量测序使得对一个物种的转录组和基因组进行细致全貌的分析成为可能，所以又被称为深度测序(deep sequencing)。

Transcription and Splicing



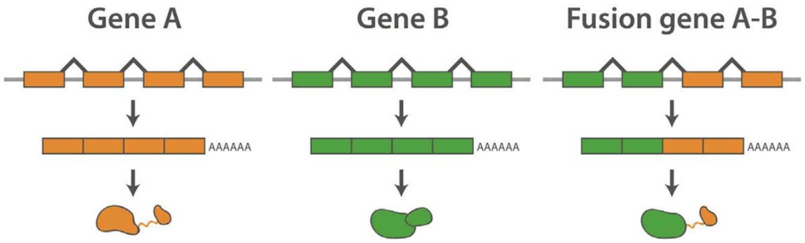
RNA-seq Protocol



Martin and Wang Nat. Rev. Genet. (2011)

RNA-seq Applications

- Examine the expression of all the genes in specific conditions (developmental stages, different tissues, normal vs disease, drug treatment)
- Find novel genes
- Find alternative splicing
- Find gene mutations or gene fusion



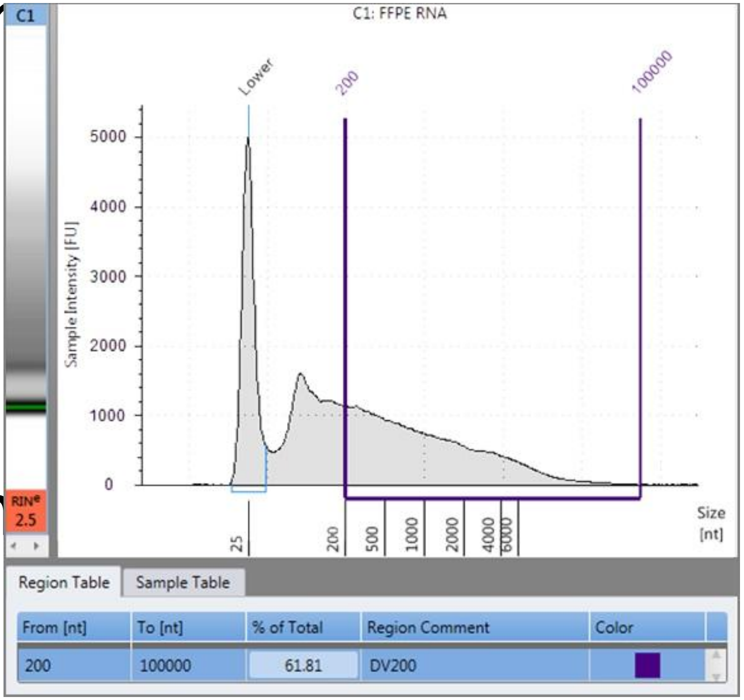
Latysheva et al, Mol Cell, 2016

- Do not have to know the genome sequence or predict genes
- Digital representation of gene expression
- Good detection range from low to high expression level

RNA Quality Control

- Degraded RNA (shorter) are hard to be made into a sequencing librar

- DV200 is important for
- DV200 > 30% is recomm



From [nt]	To [nt]	% of Total	Region Comment	Color
200	100000	61.81	DV200	■

Experimental Design

- Ribo-minus (remove too abundant r/tRNA transcripts)
- PolyA (mRNA after splicing, enrich for exons, no tRNA/rRNA)
- Strand specific (directionality of RNA, useful for novel lncRNA)

- Sequencing: what does $200 / sample mean?
  - SE or PE: PE getting more popular
  - Depth: 20-50M differential expression, deeper transcript assembly or splicing
  - Read length: longer for transcript assembly, splicing mutation calls

Experimental Design

- Assessing biological variation requires replicates:
  - Technical (not needed): same RNA, library prep and sequencing separately
  - Bio1: Cells grown in different dishes, processed on different days
  - Bio2: Cells / tissues from different lab animals (genetically identical, similar age and lifestyle)
  - Bio3: Cells / tissues from different human individuals
- How many replicates are good enough:
  - 1 only for exploratory assays, not good for publications
  - 2 OK for cell line samples
  - 3 preferred for animal samples
  - More for human samples

Break

RNA-seq Alignment and QC

Alignment

- BWA or Bowtie?
- Prefer splice-aware aligners

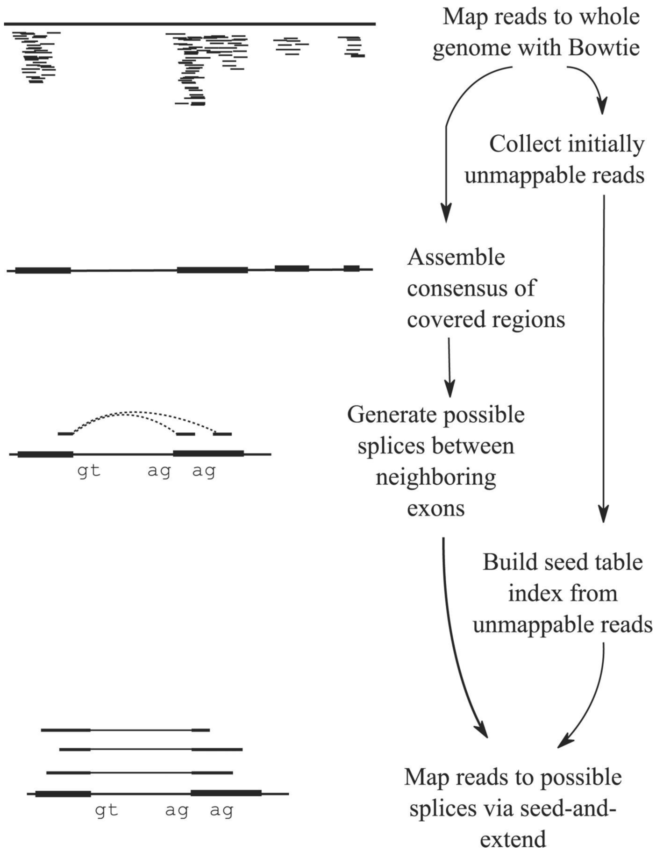
Splice Aware Alignment

- Need genome index file
- Transcript annotation file is optional but not required

- Map to exons first
- Create junction database
- Map unmapped reads to junctions

- For longer reads, map shorter segments (~25bp)

TopHat, Trapnell et al, Bioinfo 2009


Map reads to whole genome with Bowtie
Collect initially unmappable reads
Assemble consensus of covered regions
Generate possible splices between neighboring exons
gt
ag
ag
Build seed table index from unmappable reads
Map reads to possible splices via seed-and-extend
gt
ag
ag

BAM file for PE RNA-seq alignment from STAR

col 1: read id; col 2: binary encoding flags, e.g. first bit indicate PE, so all PE samples have odd number

col 3,4: chromosome and the beginning location of the read

col 6: cigar string: read matches 27bp at first, spans a 1099bp intron, then maps 48bp on the next exon

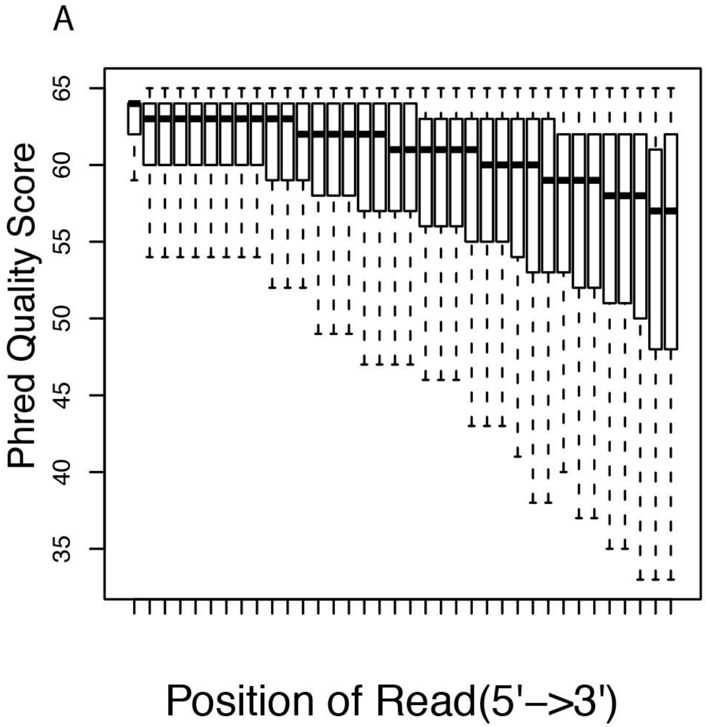
col 7, 8, 9: the location of its mate and how long the two mates spanned on the reference genome.

Optional fields:
NH: how many places this read aligned to; nM: number of mismatches for the pair;
XS: strand of the underlying transcript generating the read

Second line is the mapping of the mate pair, which span a 566bp intron

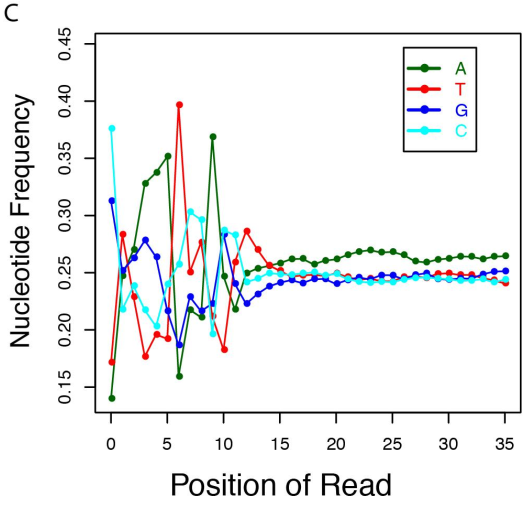
```
E00510:163:H2GYFCCXY:6:2117:28189:8341 163 1 497273 255 27M1099N48M = 501602 4970
TCCAGCCTTGGACAACAGAGGGAGACCCATAATGTAGAATCAGTGGGCGTGTTAAGCTTGTTTTCCTGCAACTGG
AAFFFJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJ NH:i:1
HI:i:1 AS:i:149 nM:i:1 XS:A:-

E00510:163:H2GYFCCXY:6:2117:28189:8341 83 1 501602 255 19M566N56M = 497273 -4970
TCTTTCATGGTAGATCCAGTATAACGTAGACTCAGTGGGATCTCTGAGCTTGTTTTCCTGCAACTAGACTGTCCA
JJJJFJJJJJFJJJAJJJJFJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJFFFAA NH:i:1
HI:i:1 AS:i:149 nM:i:1 XS:A:-
```

RNA-seq QC (RSeQC): FASTQC Read Quality



Wang et al, Bioinfo 2012

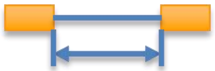
RSeQC: Nucleotide Compositions



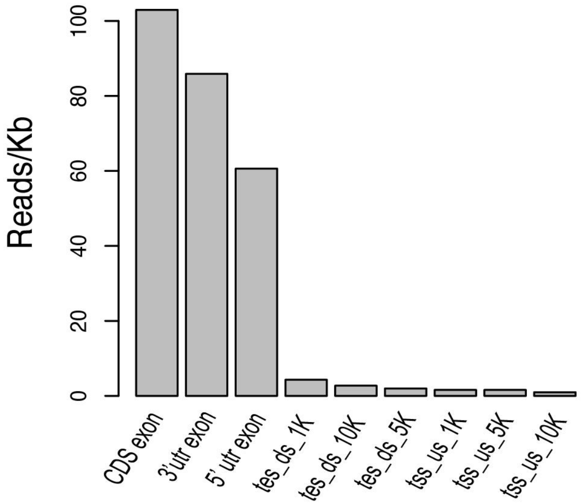
- Could trim first few bases of every read

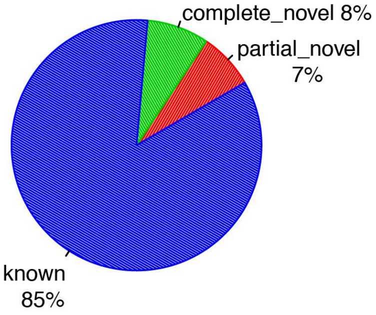
RSeQC: Insert Size and Read Distributions

Paired-end read



Insert size

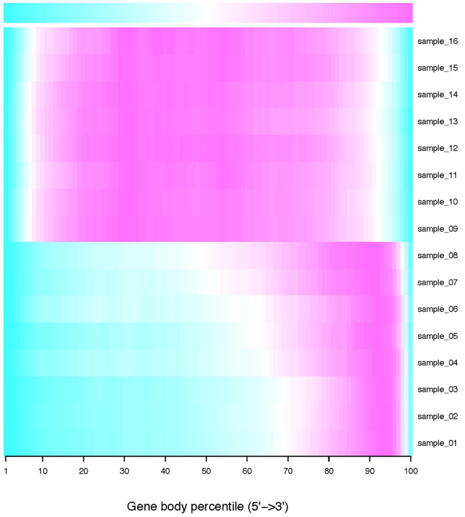




Splice junstion annotation

RSeQC: TIN and medTIN

- TIN (transcript integrity number) on each transcript
- medTIN (meidan TIN score across all the trans sure the RNA integrity at samp %)

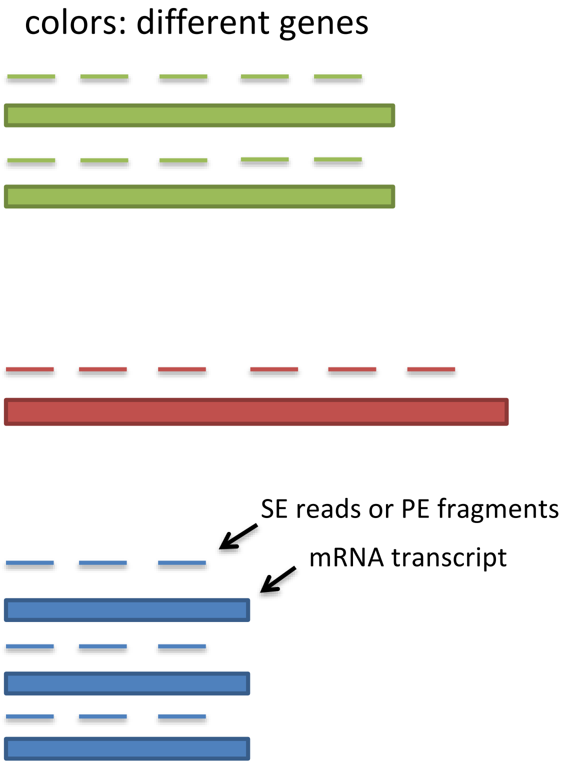


Break

http://rseqc.sourceforge.net/

RNA-seq Abundance

mRNAs to RNA-seq fragments



K_{ij} = count of fragments aligned to gene i, sample j

is proportional to:

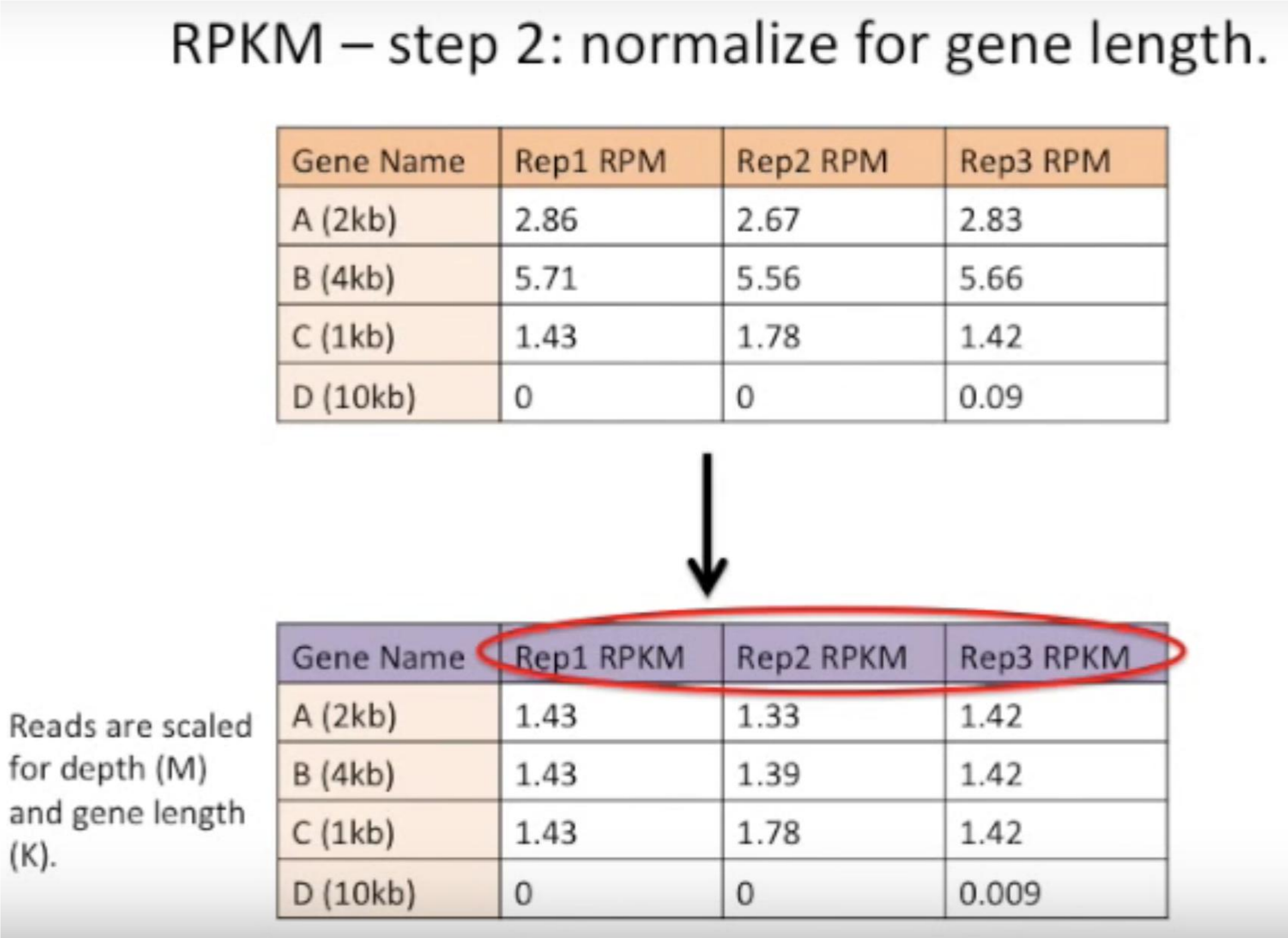
- expression of RNA
- length of gene
- sequencing depth
- lib. prep. factors (PCR)
- in silico factors (alignment)
- ...

Expression Index

- CPM (count per million)

- RPKM (Reads per kilobase of transcript per million reads of library)
  - Total reads / 1M, divide by gene length in KB
  - Corrects for coverage, gene length
  - TopHat / Cufflinks
  - FPKM (Fragments), PE libraries, ~RPKM/2

- TPM (transcripts per million) RSEM (Li et al, Bioinfo 2011)
  - Divide read count by gene length in KB (RPK) FIRST, divide by scaling factor (sum of RKP across all genes / 1M)
  - Proportion of reads mapped to a gene in each sample is comparable



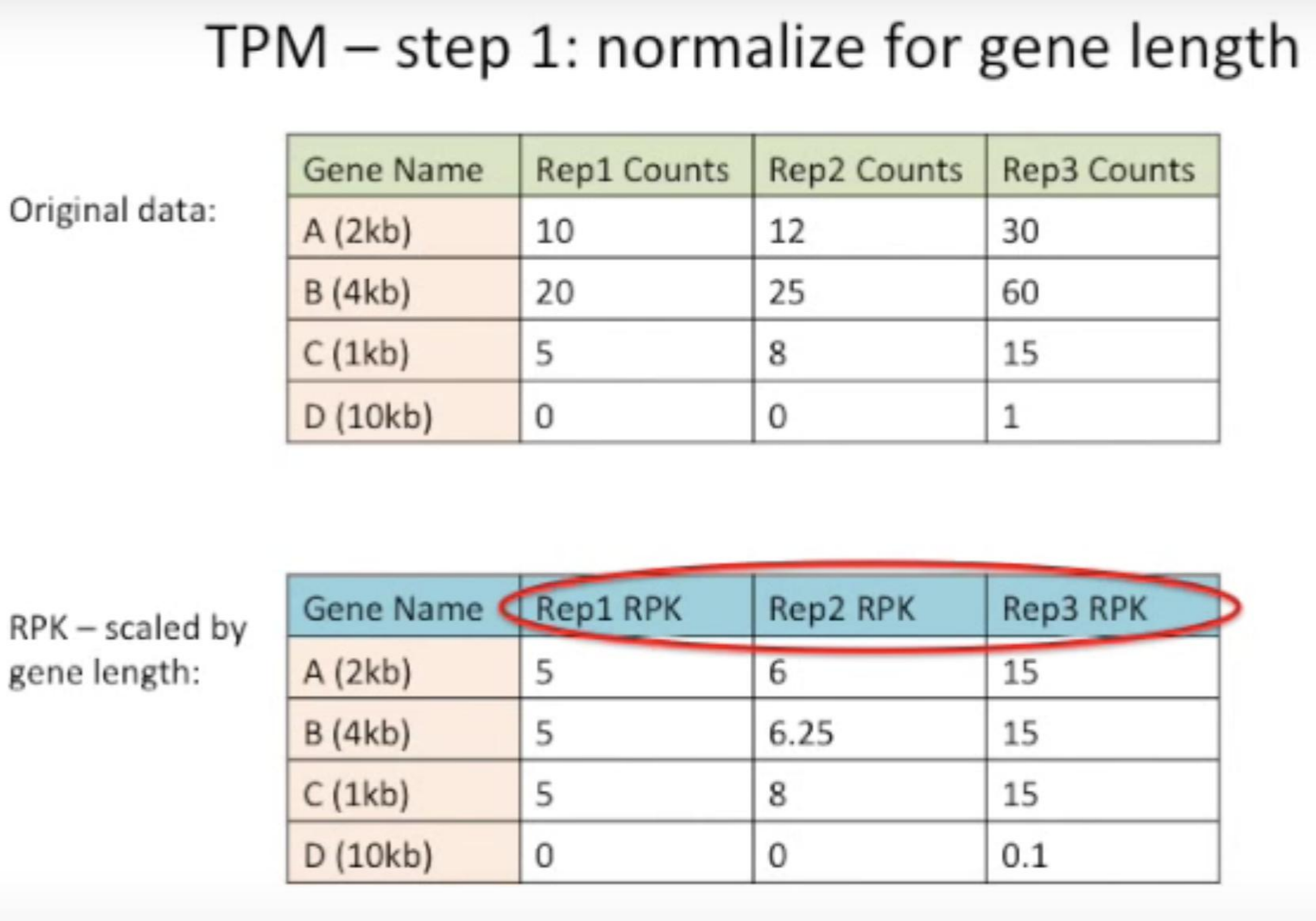
RPKM – step 2: normalize for gene length.

Gene Name	Rep1 RPM	Rep2 RPM	Rep3 RPM
A (2kb)	2.86	2.67	2.83
B (4kb)	5.71	5.56	5.66
C (1kb)	1.43	1.78	1.42
D (10kb)	0	0	0.09

↓

Gene Name	Rep1 RPKM	Rep2 RPKM	Rep3 RPKM
A (2kb)	1.43	1.33	1.42
B (4kb)	1.43	1.39	1.42
C (1kb)	1.43	1.78	1.42
D (10kb)	0	0	0.009

Reads are scaled for depth (M) and gene length (K).



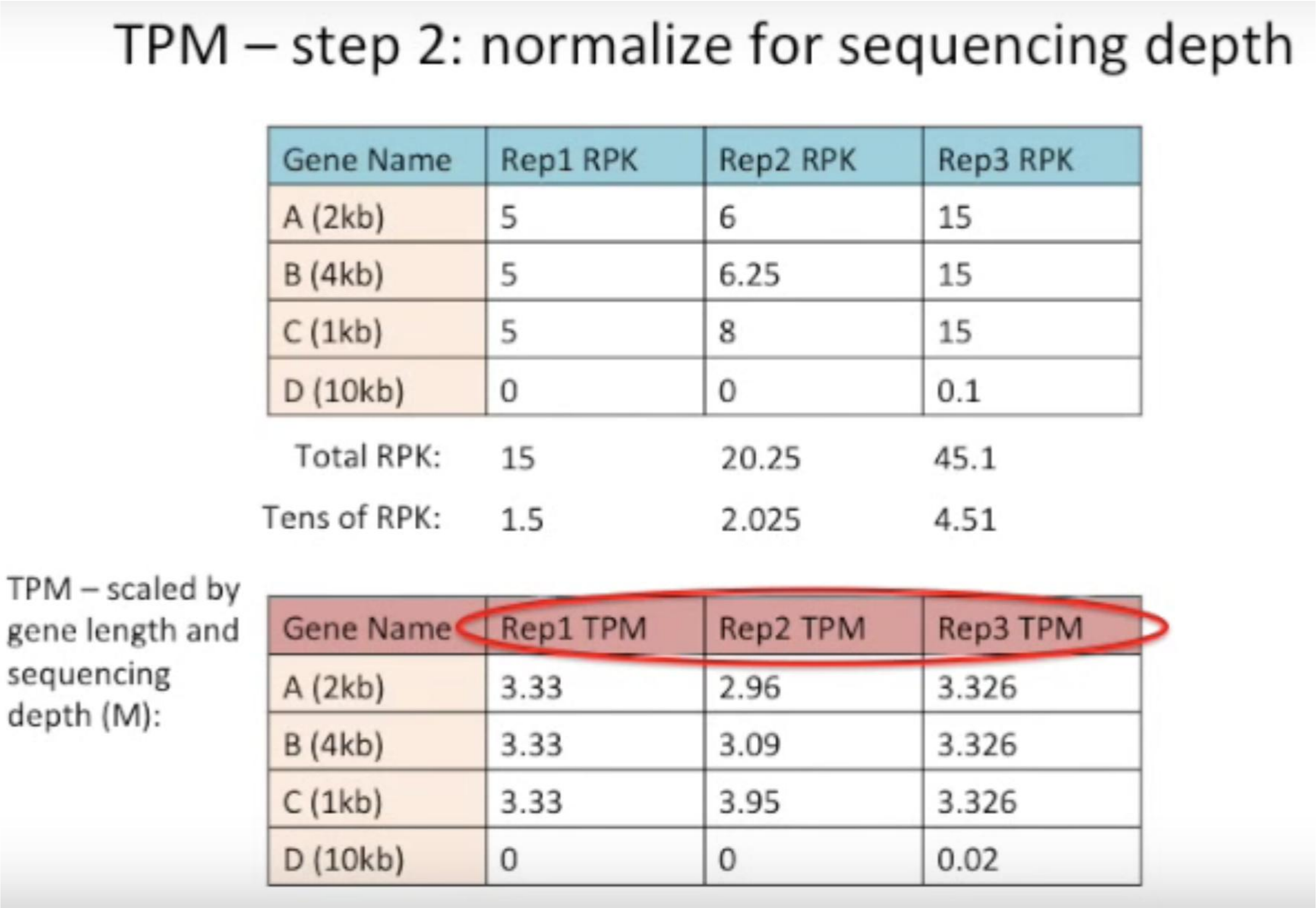
TPM – step 1: normalize for gene length

Original data:

Gene Name	Rep1 Counts	Rep2 Counts	Rep3 Counts
A (2kb)	10	12	30
B (4kb)	20	25	60
C (1kb)	5	8	15
D (10kb)	0	0	1

RPK – scaled by gene length:

Gene Name	Rep1 RPK	Rep2 RPK	Rep3 RPK
A (2kb)	5	6	15
B (4kb)	5	6.25	15
C (1kb)	5	8	15
D (10kb)	0	0	0.1

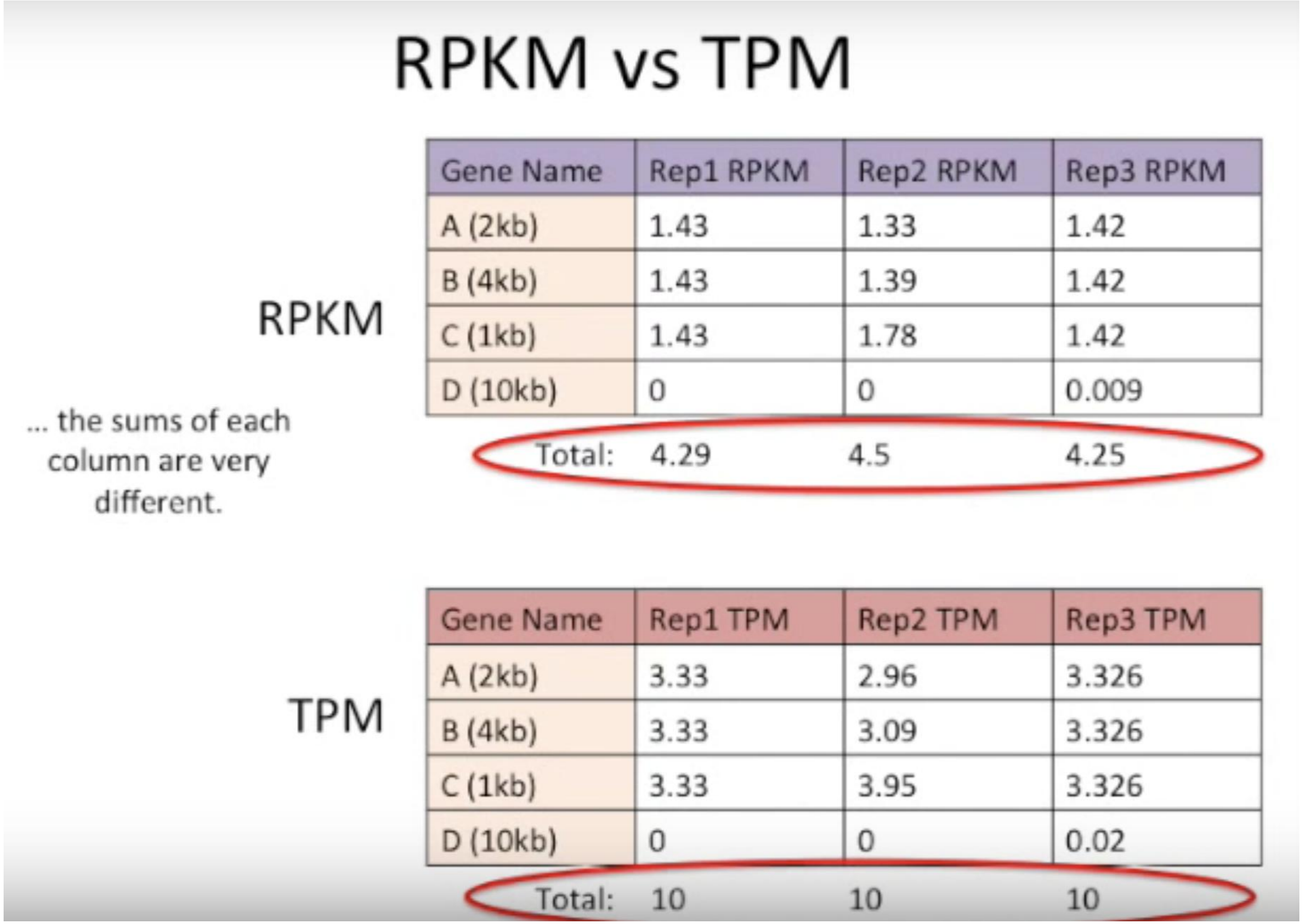


TPM – step 2: normalize for sequencing depth

Gene Name	Rep1 RPK	Rep2 RPK	Rep3 RPK
A (2kb)	5	6	15
B (4kb)	5	6.25	15
C (1kb)	5	8	15
D (10kb)	0	0	0.1
Total RPK:	15	20.25	45.1
Tens of RPK:	1.5	2.025	4.51

TPM – scaled by gene length and sequencing depth (M):

Gene Name	Rep1 TPM	Rep2 TPM	Rep3 TPM
A (2kb)	3.33	2.96	3.326
B (4kb)	3.33	3.09	3.326
C (1kb)	3.33	3.95	3.326
D (10kb)	0	0	0.02



RPKM vs TPM

RPKM

Gene Name	Rep1 RPKM	Rep2 RPKM	Rep3 RPKM
A (2kb)	1.43	1.33	1.42
B (4kb)	1.43	1.39	1.42
C (1kb)	1.43	1.78	1.42
D (10kb)	0	0	0.009
Total:	4.29	4.5	4.25

... the sums of each column are very different.

TPM

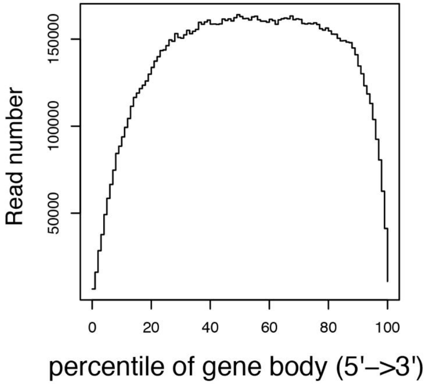
Gene Name	Rep1 TPM	Rep2 TPM	Rep3 TPM
A (2kb)	3.33	2.96	3.326
B (4kb)	3.33	3.09	3.326
C (1kb)	3.33	3.95	3.326
D (10kb)	0	0	0.02
Total:	10	10	10

RSEM for Quantification

- Input: FASTQ or BAM files, reference transcript annotation file
- Output: transcript-level gene expression (read count, TPM, FPKM) calculated on effective transcript length
- Effective length: given the sequence composition of these transcripts, you'd expect a priori to sample more reads from them

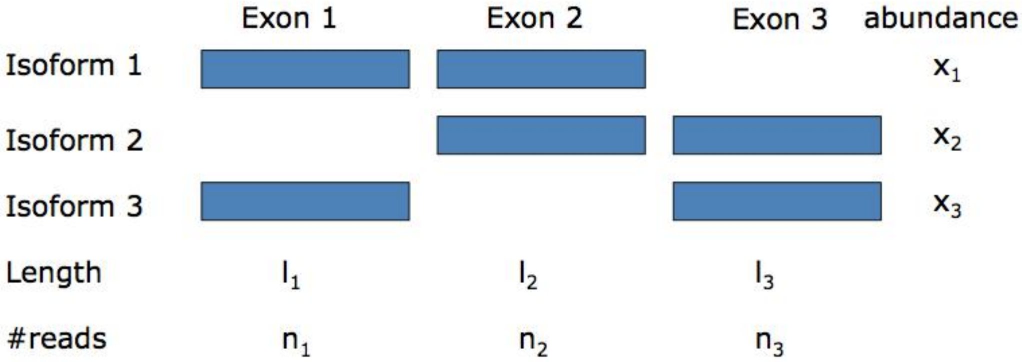
$$\tilde{l}_i = l_i - \mu + 1$$

- l_i is the length of transcript
- μ is the average fragment length



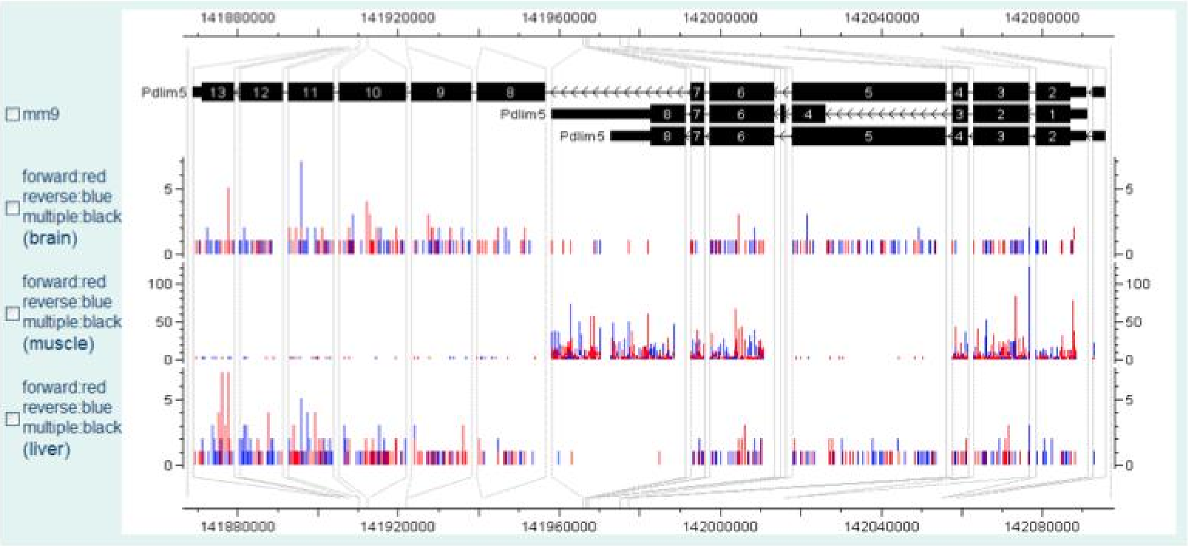
Isoform Inference

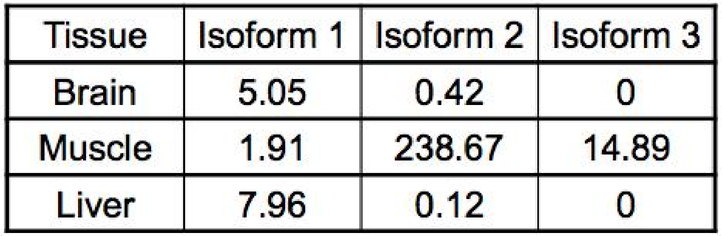
- Given known set of isoforms



- Estimate *x* to maximize the likelihood of observing *n*

Isoform Abundance Inference





Tissue	Isoform 1	Isoform 2	Isoform 3
Brain	5.05	0.42	0
Muscle	1.91	238.67	14.89
Liver	7.96	0.12	0

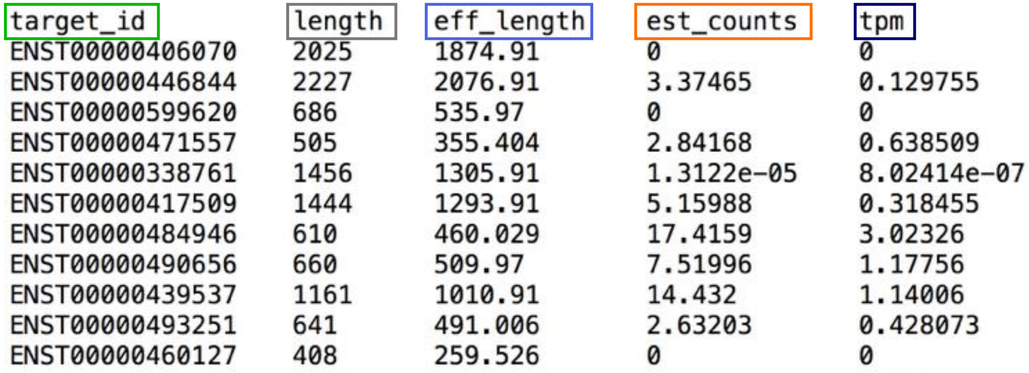
Pseudoalignment

- Kallisto (Bray et al, Nat Biotech 2016)
- Salmon (Patro et al, Nat Meth 2017)
- Do not provide “full” alignment (i.e. no exact base-by-base alignment)
- Need reference transcript annotation files
- Find all transcripts (and positions) that a read is compatible with
- Salmon also corrects for sequence-specific and GC biases
- Can map 10M reads in a few min

output

kallisto

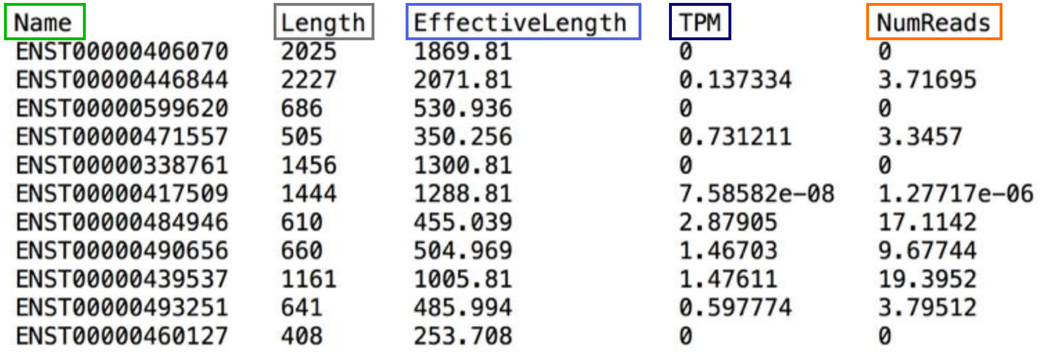
[abundance.tsv]



```
target_id         length    eff_length    est_counts    tpm
ENST00000406070   2025      1874.91       0             0
ENST00000446844   2227      2076.91       3.37465       0.129755
ENST00000599620   686       535.97        0             0
ENST00000471557   505       355.404       2.84168       0.638509
ENST00000338761   1456      1305.91       1.3122e-05    8.02414e-07
ENST00000417509   1444      1293.91       5.15988       0.318455
ENST00000484946   610       460.029       17.4159       3.02326
ENST00000490656   660       509.97        7.51996       1.17756
ENST00000439537   1161      1010.91       14.432        1.14006
ENST00000493251   641       491.006       2.63203       0.428073
ENST00000460127   408       259.526       0             0
```

Salmon

[quant.sf]



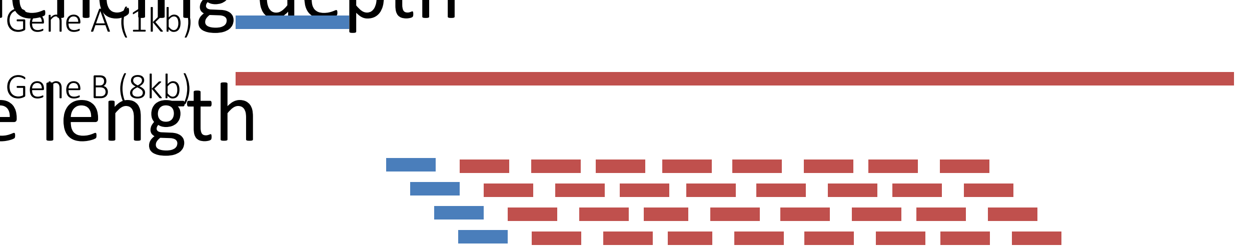
```
Name              Length    EffectiveLength    TPM            NumReads
ENST00000406070   2025      1869.81            0              0
ENST00000446844   2227      2071.81            0.137334       3.71695
ENST00000599620   686       530.936            0              0
ENST00000471557   505       350.256            0.731211       3.3457
ENST00000338761   1456      1300.81            0              0
ENST00000417509   1444      1288.81            7.58582e-08    1.27717e-06
ENST00000484946   610       455.039            2.87905        17.1142
ENST00000490656   660       504.969            1.46703        9.67744
ENST00000439537   1161      1005.81            1.47611        19.3952
ENST00000493251   641       485.994            0.597774       3.79512
ENST00000460127   408       253.708            0              0
```

Break

Differential RNA-seq with DESeq2

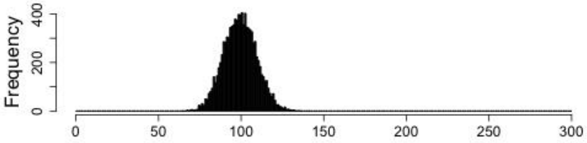
Statistical Power of Detecting Differential Expression

- Sample size
- Expression level
- True fold change
- Dispersion / variability
- Sequencing depth
- Gene length

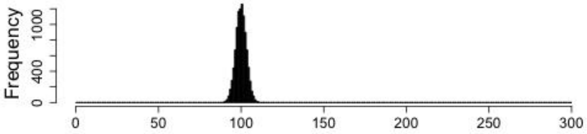


Raw Counts vs. Normalized Counts

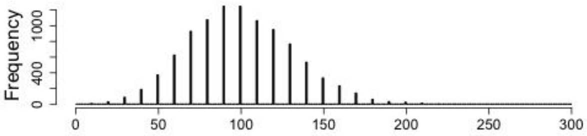
Raw count with mean of 100
Poisson sampling, so SD=10



Raw count mean = 1000
Scaled by 1/10
SD = ?



Raw count mean = 10
Scaled by 10
SD = ?



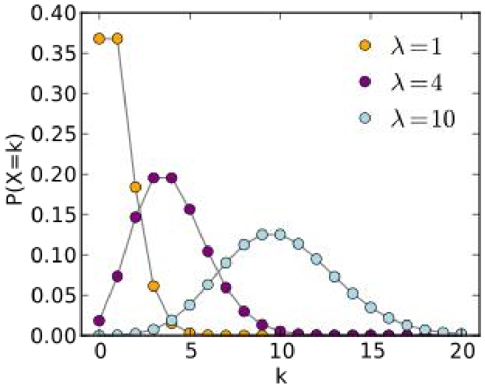
- Our ability to detect differential genes are better for longer genes and deeper sequencing because of more reads

Sequencing Read Distribution

- The number of patients arriving in an emergency room between 10 and 11 pm
- # Reads mapped to a gene with 3KB effective length

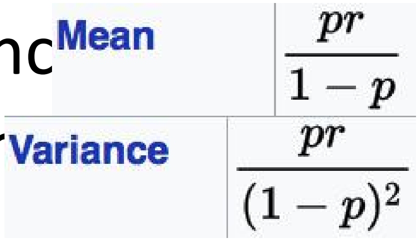
$$P(k \text{ events in interval}) = \frac{\lambda^k e^{-\lambda}}{k!}$$

- Poisson distribution
  - λ average events per interval
  - K # events in an interval
  - Var = mean = λ

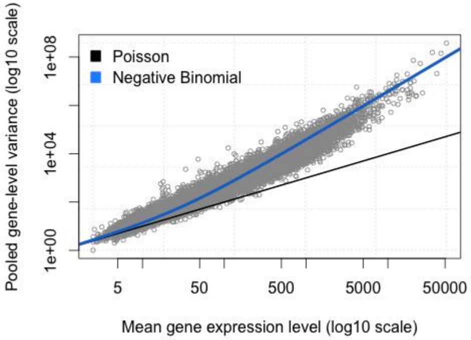


Sequencing Read Distribution

- In reality, sequencing data is
  - (Mean < Varianc

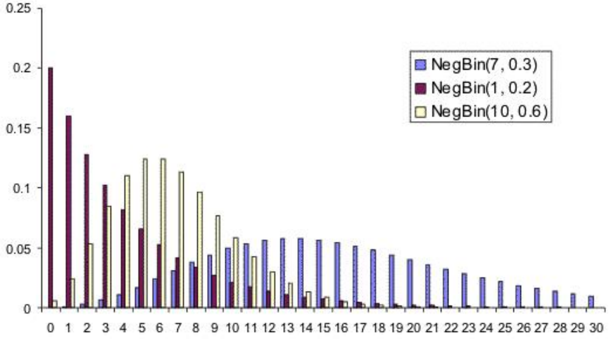


Mean	$\frac{pr}{1-p}$
Variance	$\frac{pr}{(1-p)^2}$



- Negative binom
  - NB(r, p)
  - # of success before the first r failure, if Pb(succ) i
- Probability mass functi

$$k \mapsto \binom{k+r-1}{k} \cdot (1-p)^r p^k,$$



DESeq2: Modeling RNA-seq Read Over Dispersion

raw count for gene i, sample j

normalization factor

quantity of interest

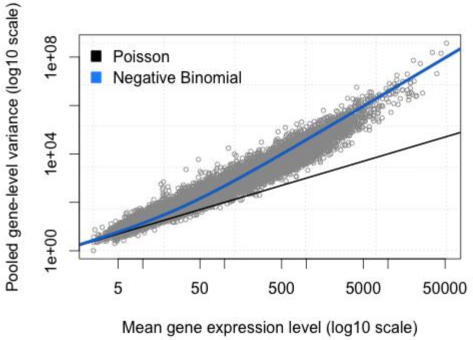
dispersion for gene i

$$K_{ij} \sim \mathrm{NB}(s_{ij} q_{ij}, \alpha_i)$$

$$\mathrm{Var}(K_{ij}) = \mu_{ij} + \hookleftarrow_i \mu_{ij}^2$$

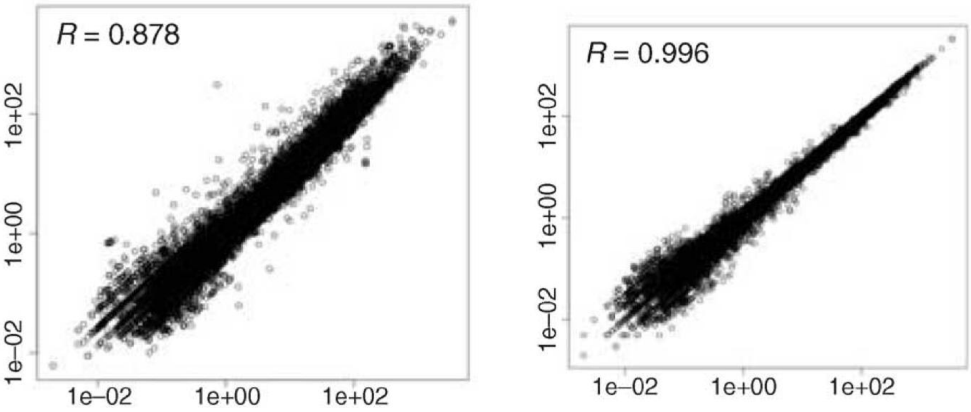
Poisson from sampling fragments

Extra variation due to biological variance



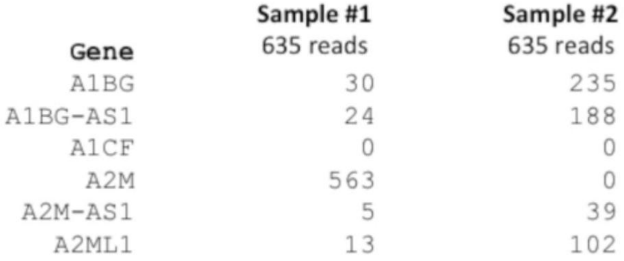
Model Variance from Limited Replicates

- Problem with estimating variance when the sample size is small (e.g. 2-3 replicates in each condition)
- Smooth gene-wise variance towards a
  comm ted way by
  borrow genes, but
  allow f



DESeq2 Differential Expression

- Normalize raw counts in different samples (prefer similar depth)



Gene	Sample #1 635 reads	Sample #2 635 reads
A1BG	30	235
A1BG-AS1	24	188
A1CF	0	0
A2M	563	0
A2M-AS1	5	39
A2ML1	13	102

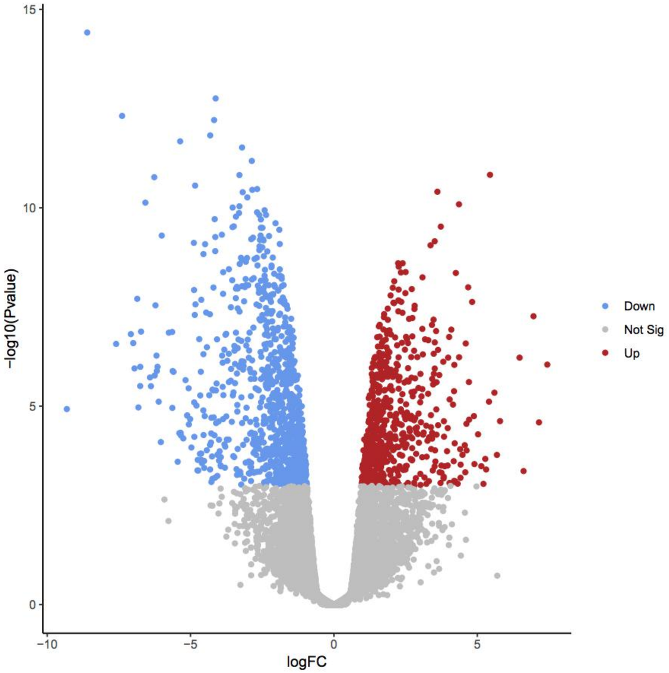
- Stabilize / shrink va [illegible] ation from other genes
- Differential expression: test whether gene i expression follows same NB()

$$k \mapsto \binom{k+r-1}{k} \cdot (1-p)^r p^k,$$

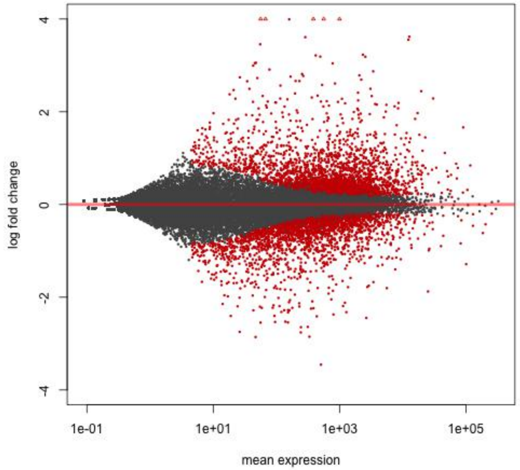
$$\Pr(X=k) = \binom{k + \frac{\mu^2}{\sigma^2-\mu} - 1}{k} \left(\frac{\sigma^2-\mu}{\sigma^2}\right)^k \left(\frac{\mu}{\sigma^2}\right)^{\mu^2/(\sigma^2-\mu)}$$

Visualize Differential Expression

- Volcano plot

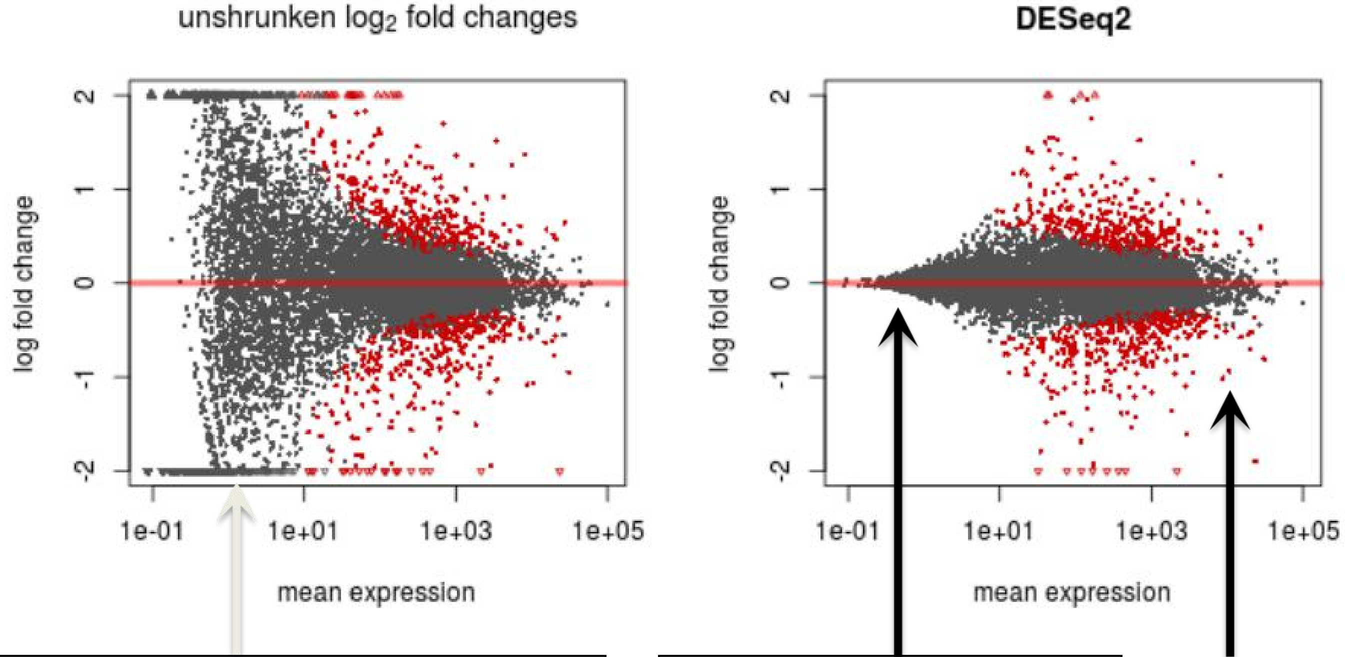


- MA plot



M: log ratio
A: mean average
Values should scatter around 0

Fold Change with Var Stabilization



low count genes have noisy variance and FC estimates

low count genes need strong variance moderation

almost no effect

Summary

- RNA-seq design considerations
- Splice aware read mapping: HISAT, STAR
- Quality control: RSeQC
- Quantification: RSEM
- Expression index: R/FPKM, TPM, and CPM
- Pseudo aligners: Kallisto and Salmon
- Differential expression DESeq2
  - NB distribution with over dispersion estimates
  - Hierarchical modeling to estimate stable variances
  - Volcano plot and MA plot visualization