Genomics data analysis
基因组学数据分析

RNA-seq表达谱分析

李程
北京大学生命科学学院
生物信息中心
统计科学中心
邮件: cheng_li@pku.edu.cn

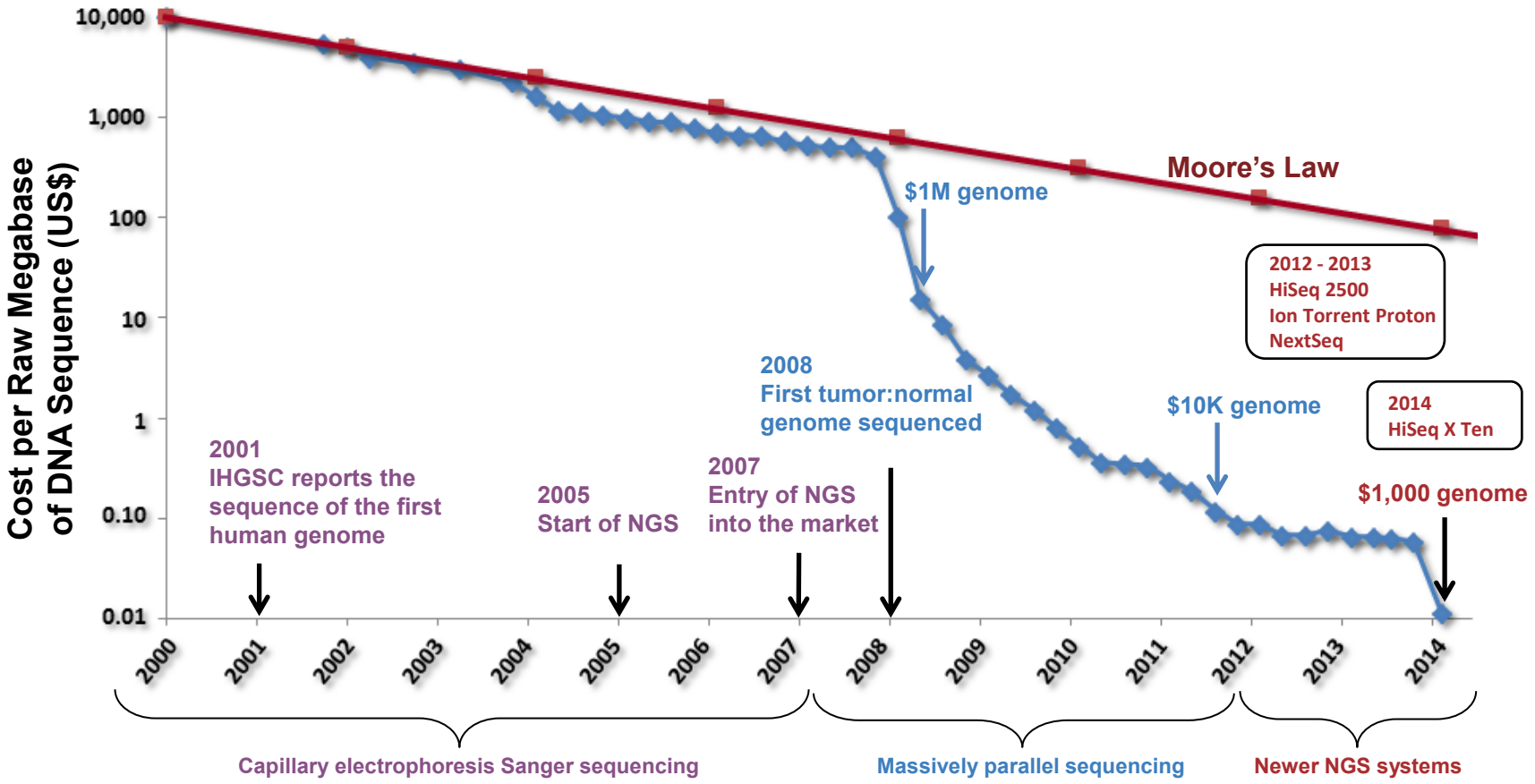
2020年秋季助教:
徐子晗，xzh0910@pku.edu.cn
崔泽嘉，18202782627@163.com

课程材料：http://3d-genome.life/
联系助教加入课程微信群

本节内容

- 高通量测序技术
- 测序原始数据
- RNA-seq分析流程
- 统计**模型**：数据归一化
- **演示：**Linux、云计算、分析资源

NGS (next generation sequencing) drives down the cost of sequencing

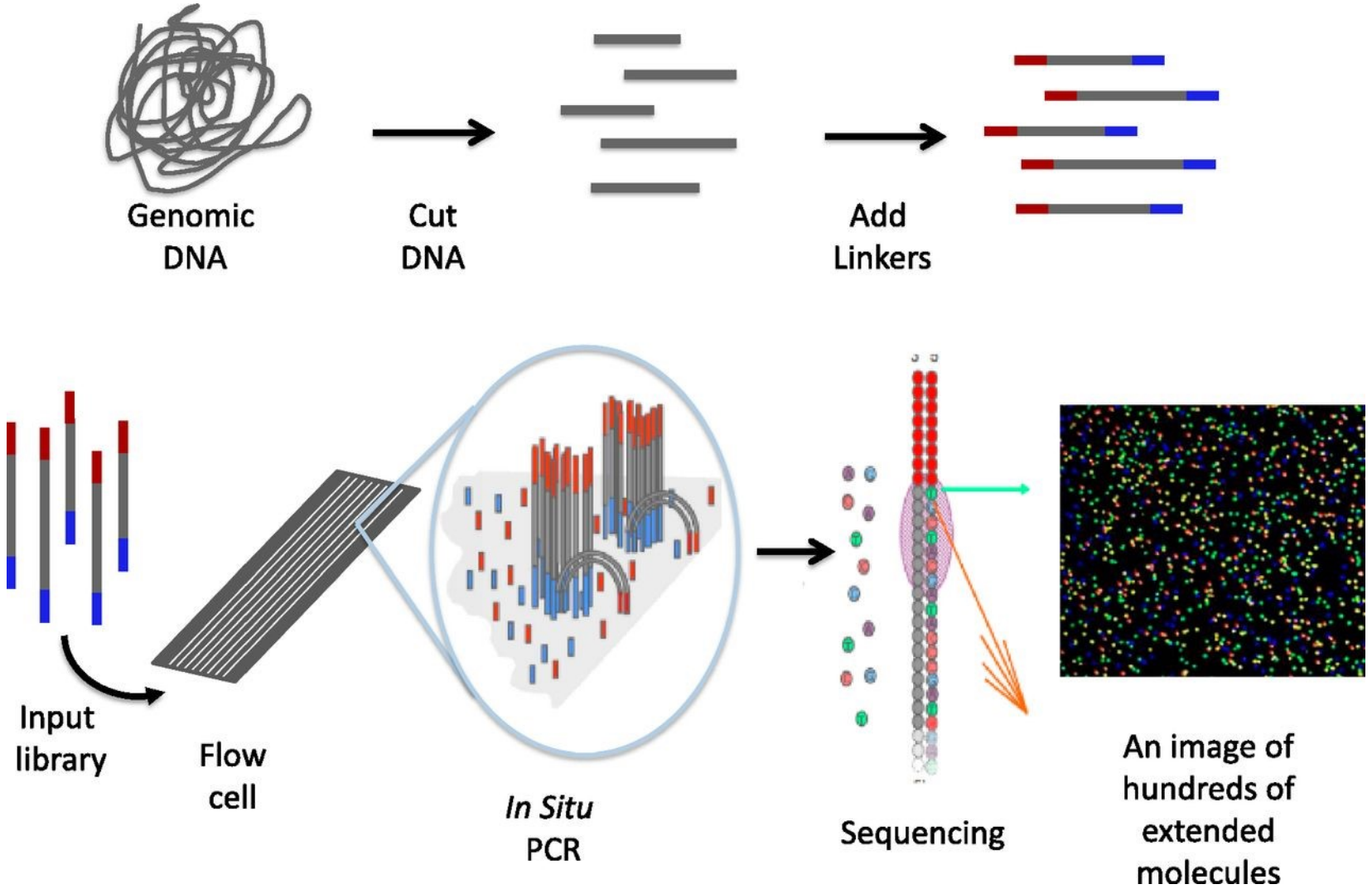


2005	2006	2007	2008	2009	2010	2011
454 pyrosequencing GS-20	Solexa/Illumina sequencer	ABI/SOLiD sequencer	Helicos BioSciences	Illumina GAIIX, SOLiD 3.0	Illumina HiSeq 2000 Oxford Nanopore	Ion Torrent PGM PacBio RS Illumina MiSeq

Adapted from: MacConaill LE. Existing and emerging technologies for tumor genomic profiling. *Journal of Clinical Oncology*, 31(15), 1815-1824 (2013). (slide by **Erick Lin**)


https://www.yourgenome.org/facts/what-is-the-illumina-method-of-dna-sequ

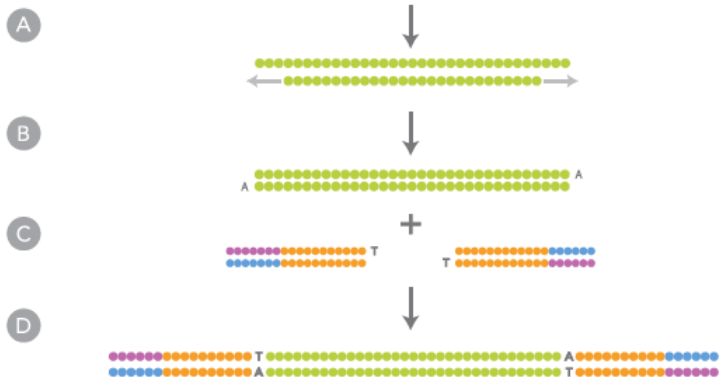
Illumina Sequencing by Synthesis



https://www.bilibili.com/video/av16817060/ (Video introduction)
https://www.illumina.com/science/technology/next-generation-sequencing/sequencing-technology.html

1 LIBRARY PREPARATION

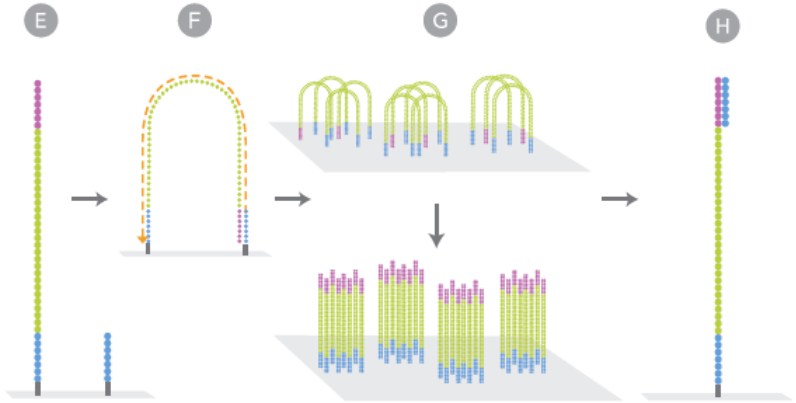
6 hours
3 hours hands-on time



- A Fragment DNA
- B Repair ends
  Add A overhang
- C Ligate adapters
- D Select ligated DNA

2 CLUSTER GENERATION

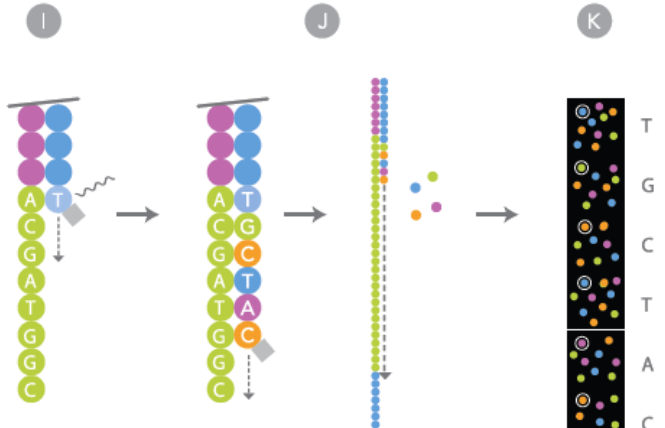
4 hours
< 10 minutes hands-on time
1–96 samples



- E Attach DNA to
  flow cell
- F Perform bridge
  amplification
- G Generate clusters
- H Anneal sequencing
  primer

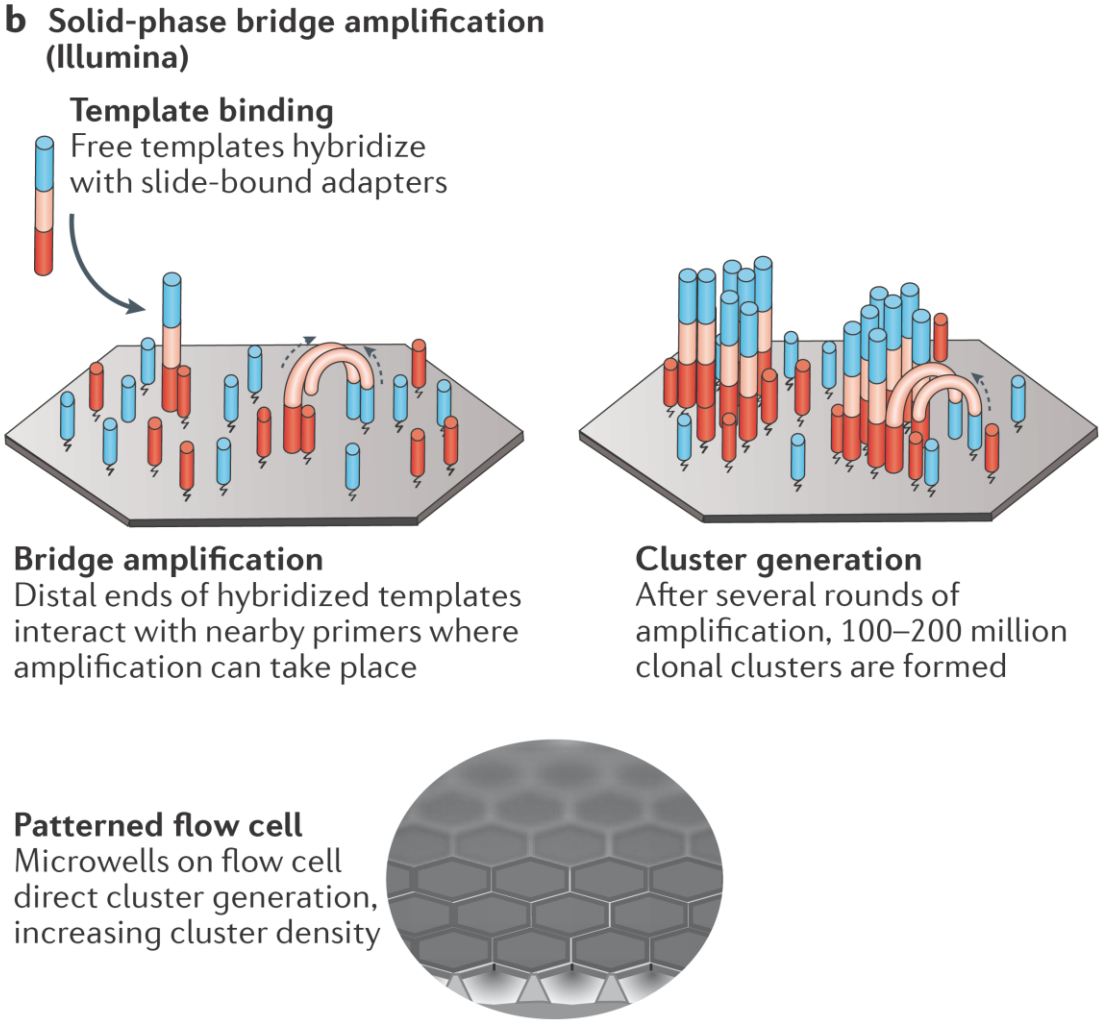
3 SEQUENCING

1–3 days single-read run
3–9 days paired-end run
30 minutes hands-on time
8 lanes, up to 96 samples
per flow cell (run)



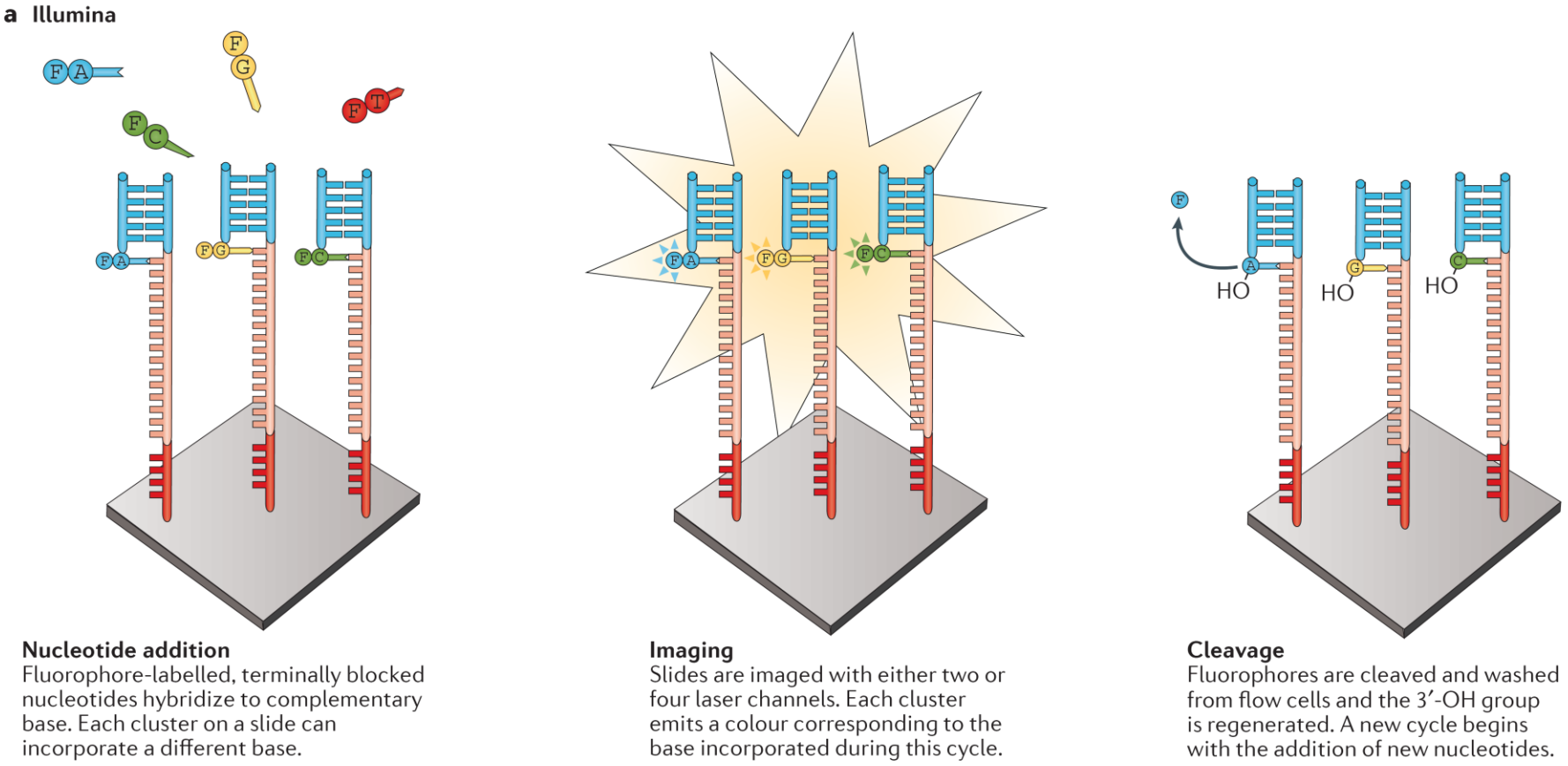
- I Extend first base,
  read, and deblock
- J Repeat step above
  to extend strand
- K Generate base calls

Template amplification



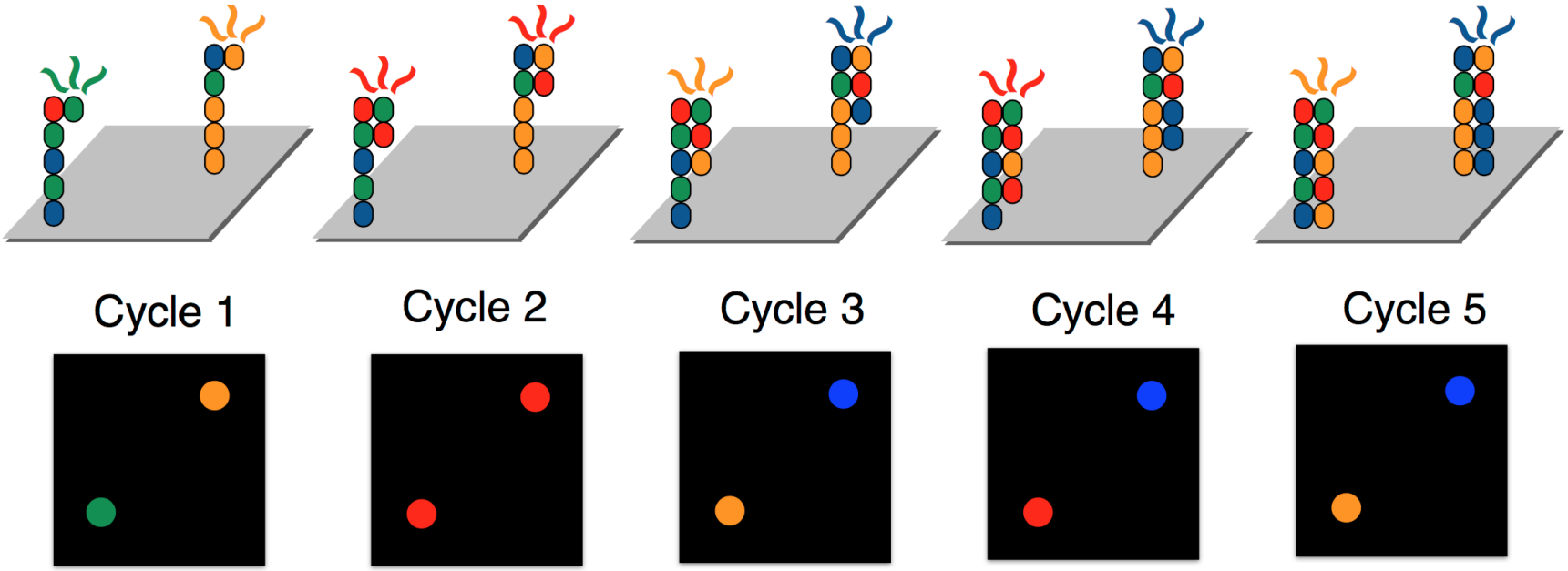
NatRev 16 Coming of age ten years of next-generation sequencing technologies

Cyclic reversible termination



NatRev 16 Coming of age ten years of next-generation sequencing technologies

Sequencing by Synthesis



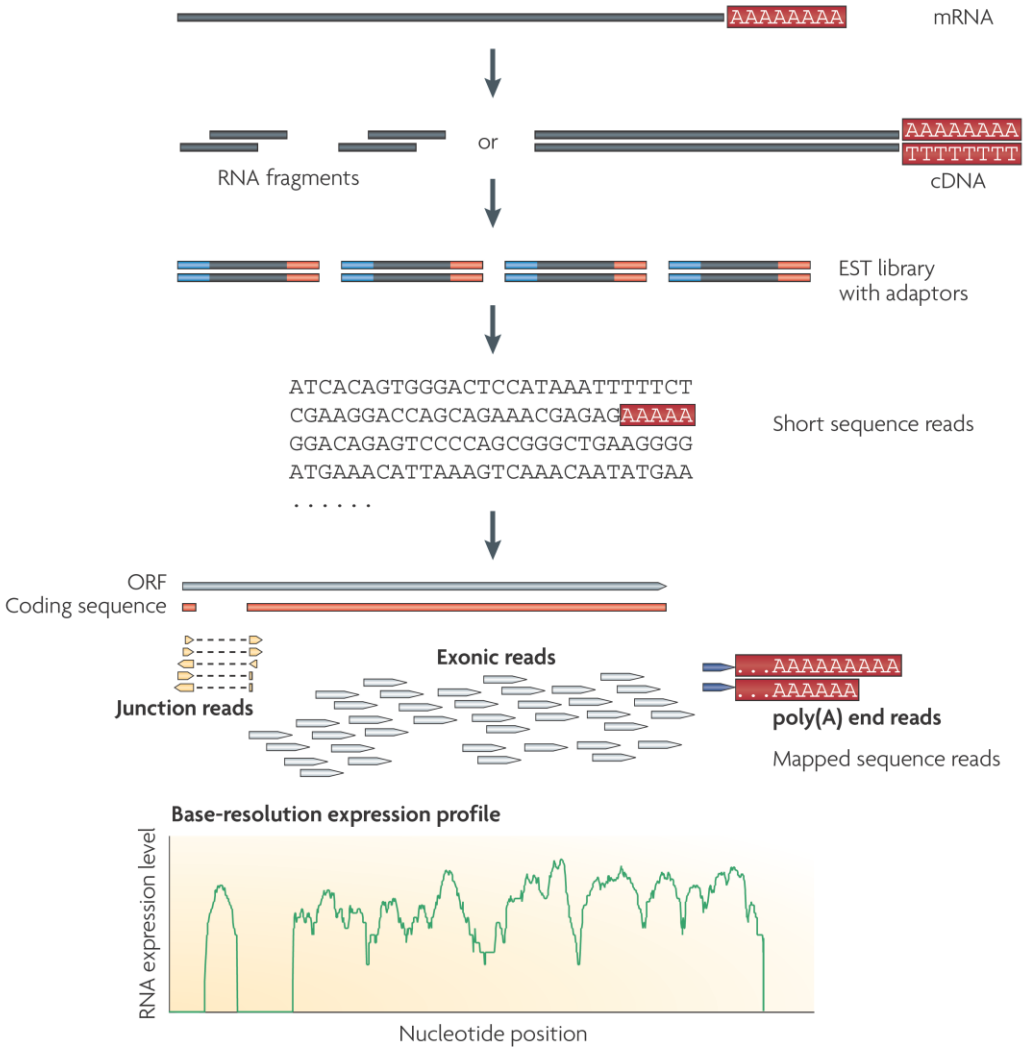
https://www.bilibili.com/video/av16817060/ (Video introduction)

http://data-science-sequencing.github.io/Win2018/lectures/lecture2/

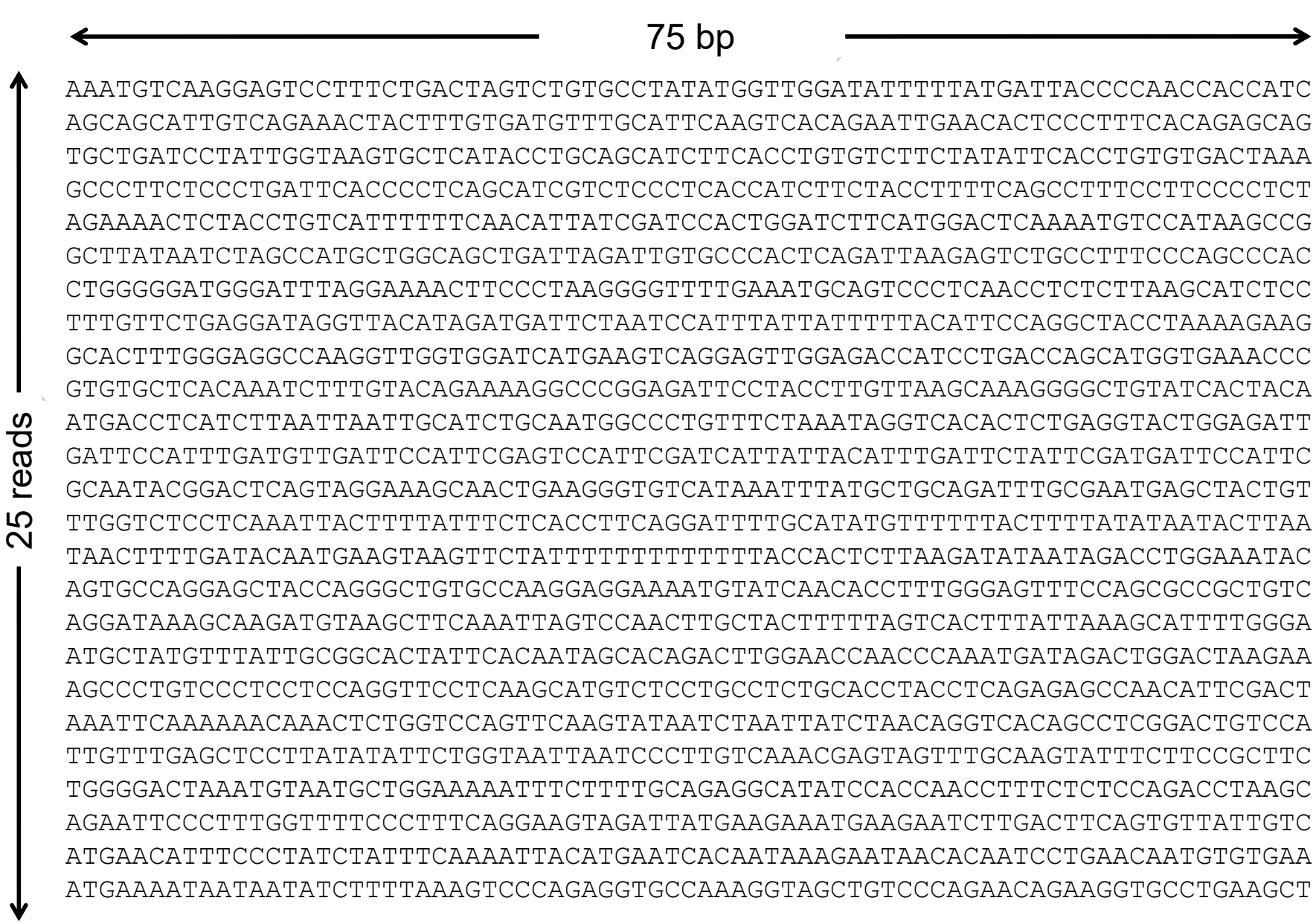
本节内容

- 高通量测序技术
- 测序原始数据
- RNA-seq分析流程
- 统计**模型**：数据归一化
- **演示：**Linux、云计算、分析资源

A typical RNA-seq experiment

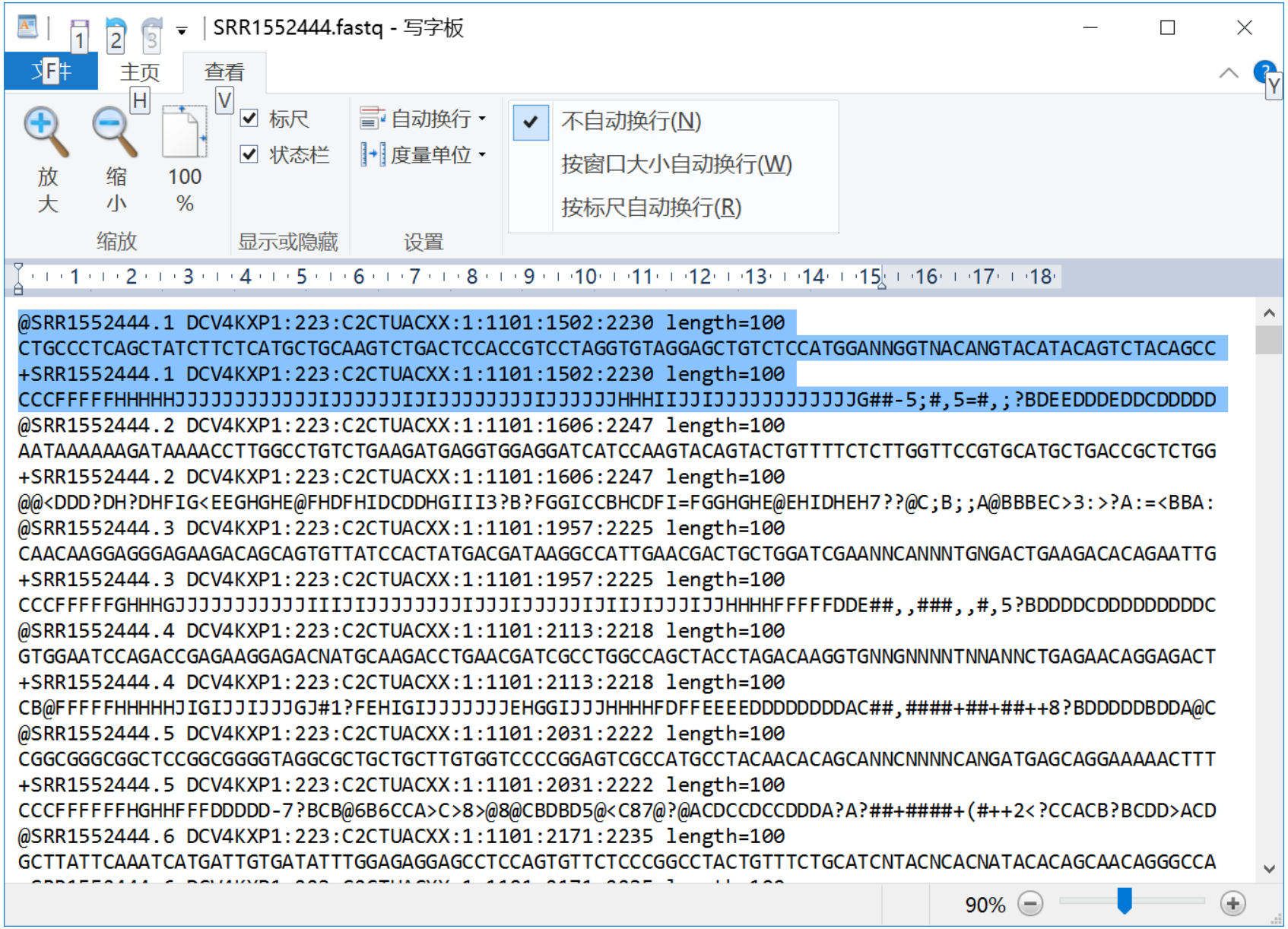


Nat Rev Genet 2009 - RNA-Seq: a revolutionary tool for transcriptomics

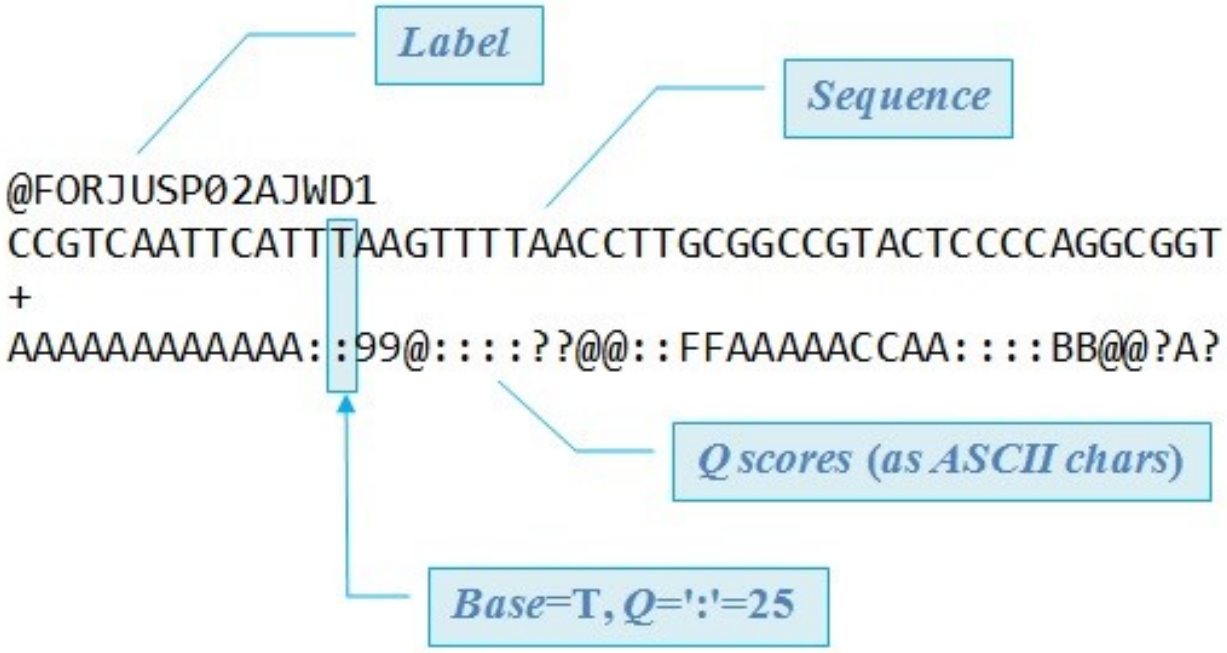


On an Illumina HiSeq 2500, one lane of an eight-lane flow cell generates >600 million reads (**45 GB**), slide by Trevor Pugh

A FASTQ file containing reads



Sequencing Reads (FASTQ format)



A Phred score of 40 can be represented as the ASCII char “I” (40+33= ASCII #73)

40 means 10^{-4} error probability.

A FASTQ file normally uses four lines per sequence.

- Line 1 begins with a ‘@’ character and is followed by a sequence identifier.
- Line 2 is the raw sequence letters.
- Line 3 begins with a ‘+’ character, is optionally followed by the same sequence identifier.
- Line 4 encodes the Phred quality values for the sequence in line 2, each value represents the error probability of a given base call. （从图像转换为碱基的错误概率）

Align reads to reference genome

1.1 An example

Suppose we have the following alignment with bases in lower cases clipped from the alignment. Read `r001/1` and `r001/2` constitute a read pair; `r003` is a chimeric read; `r004` represents a split alignment.

```
Coor     12345678901234  5678901234567890123456789012345
ref      AGCATGTTAGATAA**GATAGCTGTGCTAGTAGGCAGTCAGCGCCAT

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

The corresponding SAM format is:[1]

```
@HD VN:1.5 SO:coordinate
@SQ SN:ref LN:45
r001   99 ref  7 30 8M2I4M1D3M = 37  39 TTAGATAAAGGATACTG *
r002    0 ref  9 30 3S6M1P1I4M *  0   0 AAAAGATAAGGATA    *
r003    0 ref  9 30 5S6M       *  0   0 GCCTAAGCTAA       * SA:Z:ref,29,-,6H5M,17,0;
r004    0 ref 16 30 6M14N5M    *  0   0 ATAGCTTCAGC       *
r003 2064 ref 29 17 6H5M       *  0   0 TAGGC             * SA:Z:ref,9,+,5S6M,30,1;
r001  147 ref 37 30 9M         =  7 -39 CAGCGGCAT         * NM:i:1
```

https://samtools.github.io/hts-specs/SAMv1.pdf

Alignment against reference genome (SAM/BAM)

```
Coor     12345678901234  5678901234567890123456789012345
ref      AGCATGTTAGATAA**GATAGCTGTGCTAGTAGGCAGTCAGCGCCAT

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

This alignment contains the following features: (1) bases in lower cases are clipped from the alignment; (2) read r001/1 and r001/2 constitute a read pair; (3) r003 is a chimeric read; (4) r004 represents a split alignment.

```
@HD VN:1.5 SO:coordinate
@SQ SN:ref LN:45
r001   99 ref  7 30 8M2I4M1D3M = 37  39 TTAGATAAAGGATACTG *
r002    0 ref  9 30 3S6M1P1I4M *  0   0 AAAAGATAAGGATA    *
r003    0 ref  9 30 5S6M       *  0   0 GCCTAAGCTAA       * SA:Z:ref,29,-,6H5M,17,0;
r004    0 ref 16 30 6M14N5M    *  0   0 ATAGCTTCAGC       *
r003 2064 ref 29 17 6H5M       *  0   0 TAGGC             * SA:Z:ref,9,+,5S6M,30,1;
r001  147 ref 37 30 9M         =  7 -39 CAGCGGCAT         * NM:i:1
```

Header section

Alignment section

Optional fields in the format of TAG:TYPE:VALUE

QUAL: read quality; * meaning such information is not available

SEQ: read sequence

TLEN: the number of bases covered by the reads from the same fragment. Plus/minus means the current read is the leftmost/rightmost read. E.g. compare first and last lines.

PNEXT: Position of the primary alignment of the NEXT read in the template. Set as 0 when the information is unavailable. It corresponds to POS column.

RNEXT: reference sequence name of the primary alignment of the NEXT read. For paired-end sequencing, NEXT read is the paired read, corresponding to the RNAME column.

CIGAR: summary of alignment, e.g. insertion, deletion

MAPQ: mapping quality

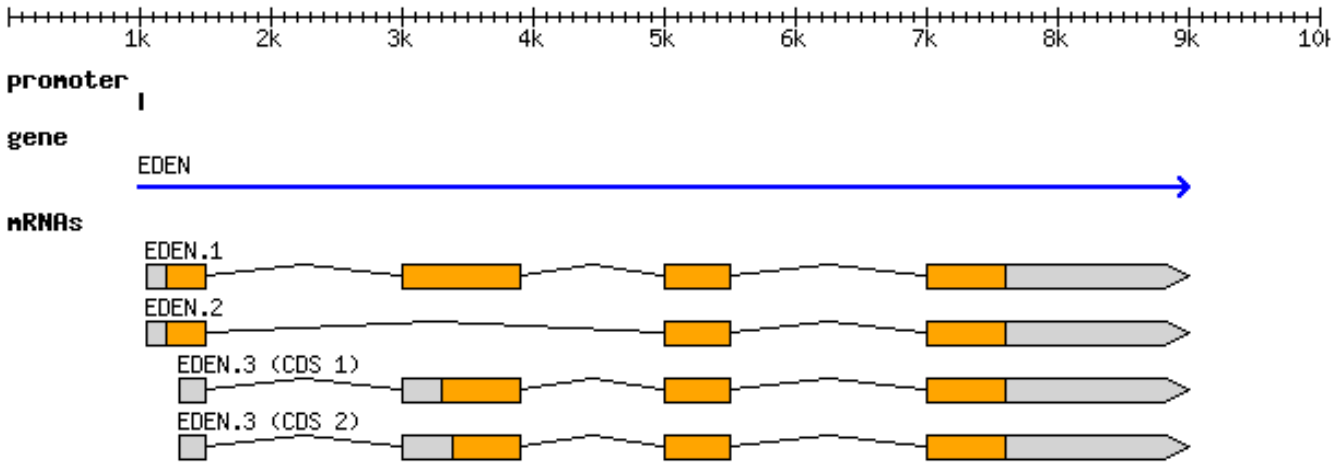
POS: 1-based position

RNAME: reference sequence name, e.g. chromosome/transcript id

FLAG: indicates alignment information about the read, e.g. paired, aligned, etc.

QNAME: query template name, aka. read ID

Gene structure annotation (GFF / GTF format)

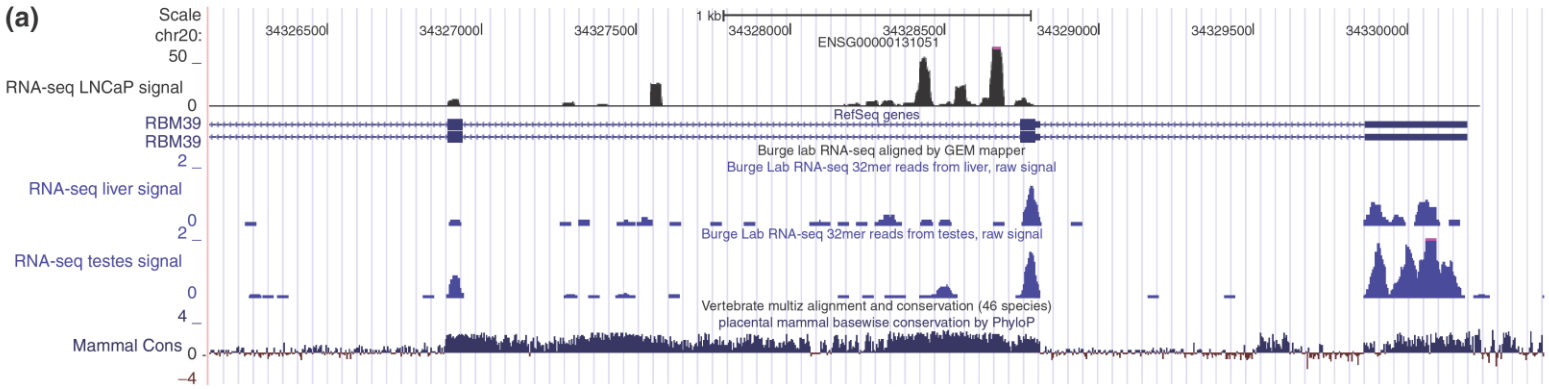



```
0  ##gff-version 3.2.1
1  ##sequence-region ctg123 1 1497228
2  ctg123 . gene            1000  9000  .  +  .  ID=gene00001;Name=EDEN
3  ctg123 . TF_binding_site 1000  1012  .  +  .  ID=tfbs00001;Parent=gene00001
4  ctg123 . mRNA            1050  9000  .  +  .  ID=mRNA00001;Parent=gene00001;Name=EDEN.1
5  ctg123 . mRNA            1050  9000  .  +  .  ID=mRNA00002;Parent=gene00001;Name=EDEN.2
6  ctg123 . mRNA            1300  9000  .  +  .  ID=mRNA00003;Parent=gene00001;Name=EDEN.3
7  ctg123 . exon            1300  1500  .  +  .  ID=exon00001;Parent=mRNA00003
8  ctg123 . exon            1050  1500  .  +  .  ID=exon00002;Parent=mRNA00001,mRNA00002
9  ctg123 . exon            3000  3902  .  +  .  ID=exon00003;Parent=mRNA00001,mRNA00003
10 ctg123 . exon            5000  5500  .  +  .  ID=exon00004;Parent=mRNA00001,mRNA00002,mRNA00003
11 ctg123 . exon            7000  9000  .  +  .  ID=exon00005;Parent=mRNA00001,mRNA00002,mRNA00003
12 ctg123 . CDS             1201  1500  .  +  0  ID=cds00001;Parent=mRNA00001;Name=edenprotein.1
13 ctg123 . CDS             3000  3902  .  +  0  ID=cds00001;Parent=mRNA00001;Name=edenprotein.1
```

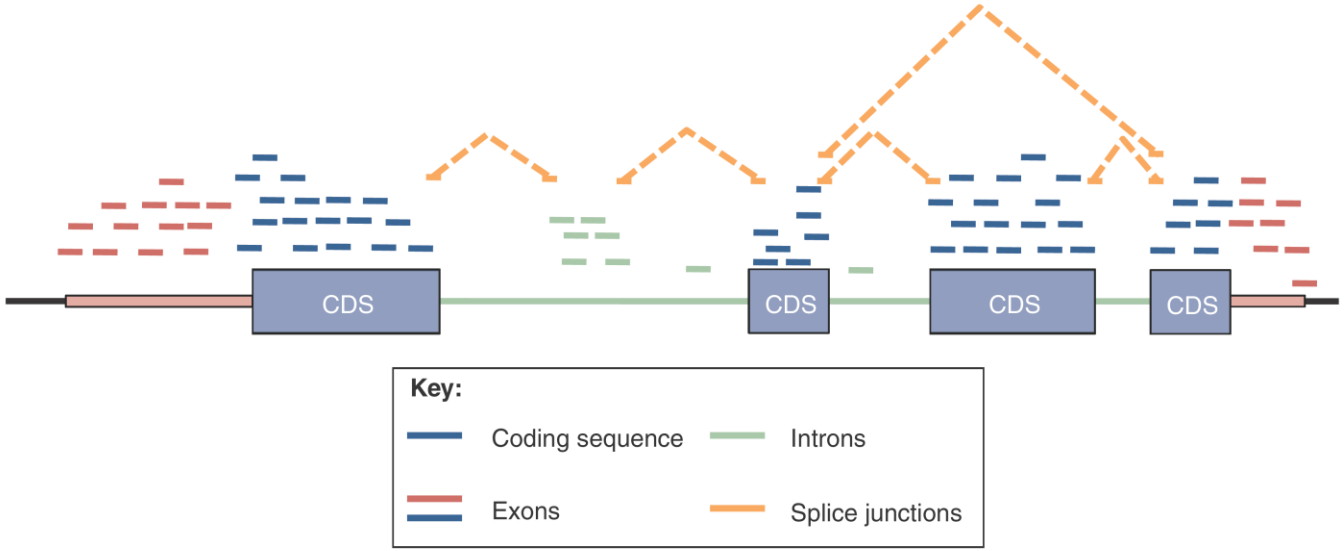
http://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/1000/index.php?manual=GFF3_format.html

Summarizing mapped reads into a gene level count

Y-axis: the number of mapped reads that overlap a location



Different summarization strategies will result in the inclusion or exclusion of different sets of reads in the table of counts



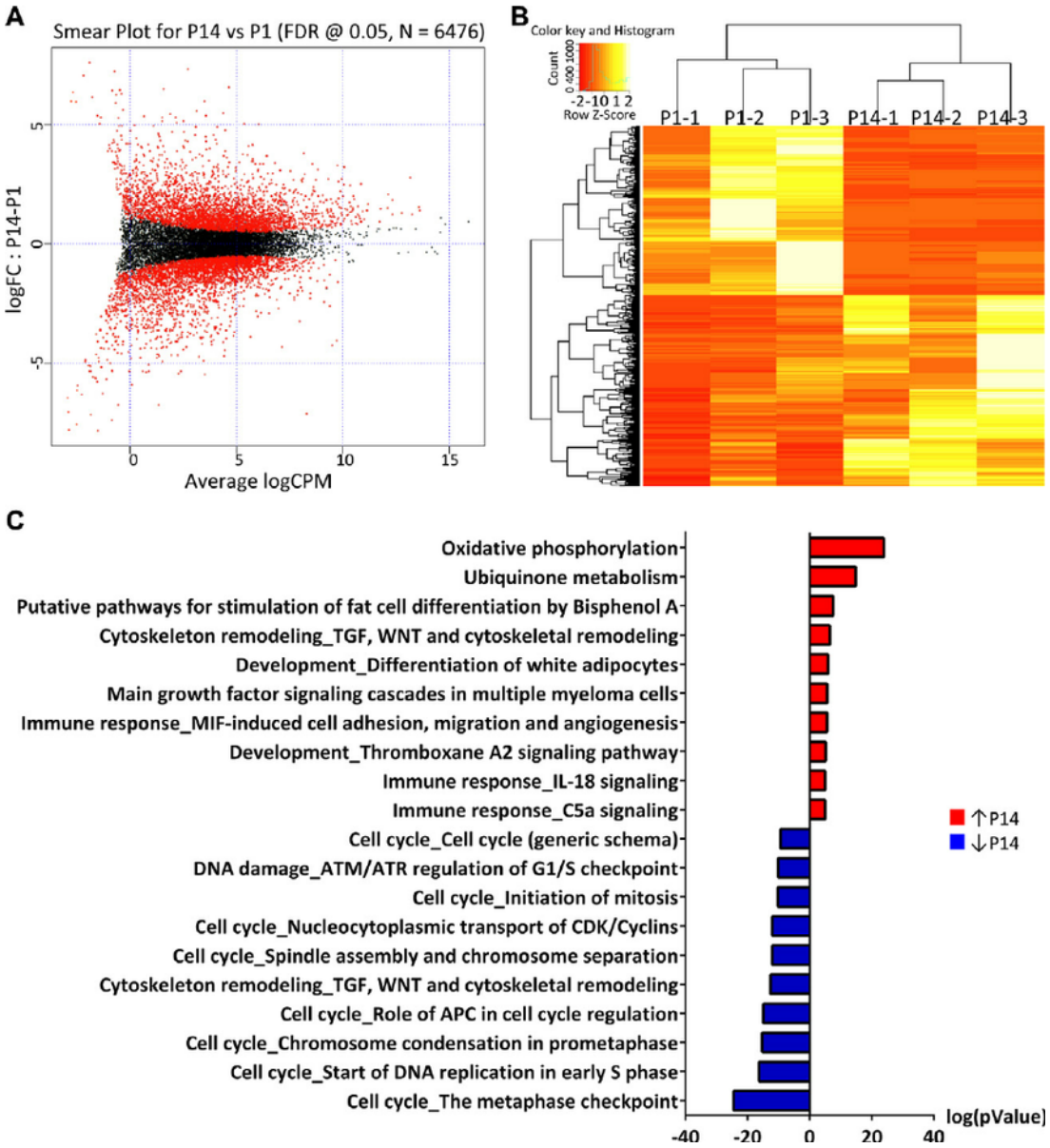
本节内容

- 高通量测序技术
- 测序原始数据
- RNA-seq分析流程
- 统计**模型**：数据归一化
- **演示：**Linux、云计算、分析资源

Analysis Workflow of RNA-Seq Gene Expression Data

- Alignment of RNA reads to reference
  - Reference can be genome or transcriptome.
- Count reads overlapping with annotation features of interest
  - Most common: counts for exonic gene regions, but many viable alternatives exist here: counts per exons, genes, introns, etc.
- Normalization
  - Main adjustment for sequencing depth and compositional bias.
- Identification of Differentially Expressed Genes (DEGs)
  - Identification of genes with significant expression differences.
  - Identification of strongly expressed genes.
- Special applications
  - Splice variant discovery (semi-quantitative), gene discovery, antisense expressions, etc.
- Cluster Analysis
  - Identification of genes with similar expression profiles across many samples.
- Enrichment Analysis of Functional Annotations
  - Gene ontology analysis of obtained gene sets from above.

RNA-seq analysis plots



Mark Ziemann

早期RNA-seq分析流程

Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks

Cole Trapnell[1,2], Adam Roberts[3], Loyal Goff[1,2,4], Geo Pertea[5,6], Daehwan Kim[5,7], David R Kelley[1,2], Harold Pimentel[3], Steven L Salzberg[5,6], John L Rinn[1,2] & Lior Pachter[3,8,9]

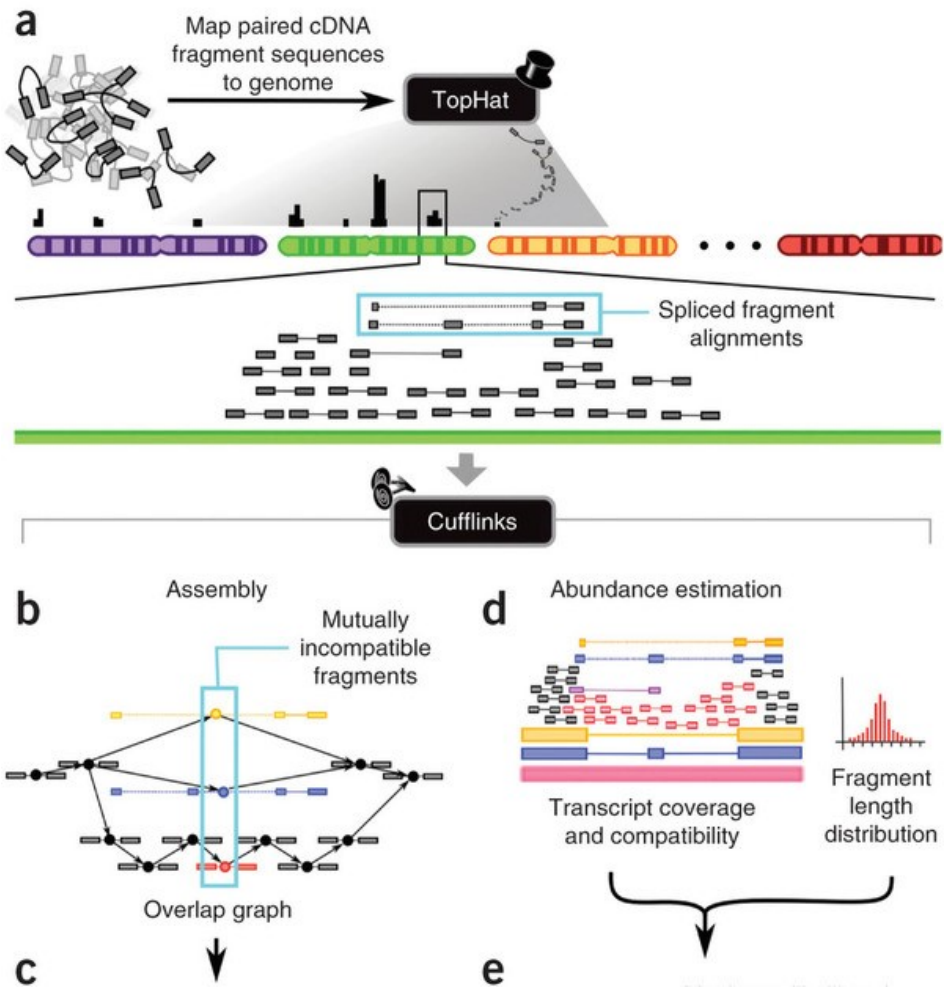
[1]Broad Institute of MIT and Harvard, Cambridge, Massachusetts, USA. [2]Department of Stem Cell and Regenerative Biology, Harvard University, Cambridge, Massachusetts, USA. [3]Department of Computer Science, University of California, Berkeley, California, USA. [4]Computer Science and Artificial Intelligence Lab, Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, Cambridge, Massachusetts, USA. [5]Department of Medicine, McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins University School of Medicine, Baltimore, Maryland, USA. [6]Department of Biostatistics, Johns Hopkins University, Baltimore, Maryland, USA. [7]Center for Bioinformatics and Computational Biology, University of Maryland, College Park, Maryland, USA. [8]Department of Mathematics, University of California, Berkeley, California, USA. [9]Department of Molecular and Cell Biology, University of California, Berkeley, California, USA. Correspondence should be addressed to C.T. (cole@broadinstitute.org).

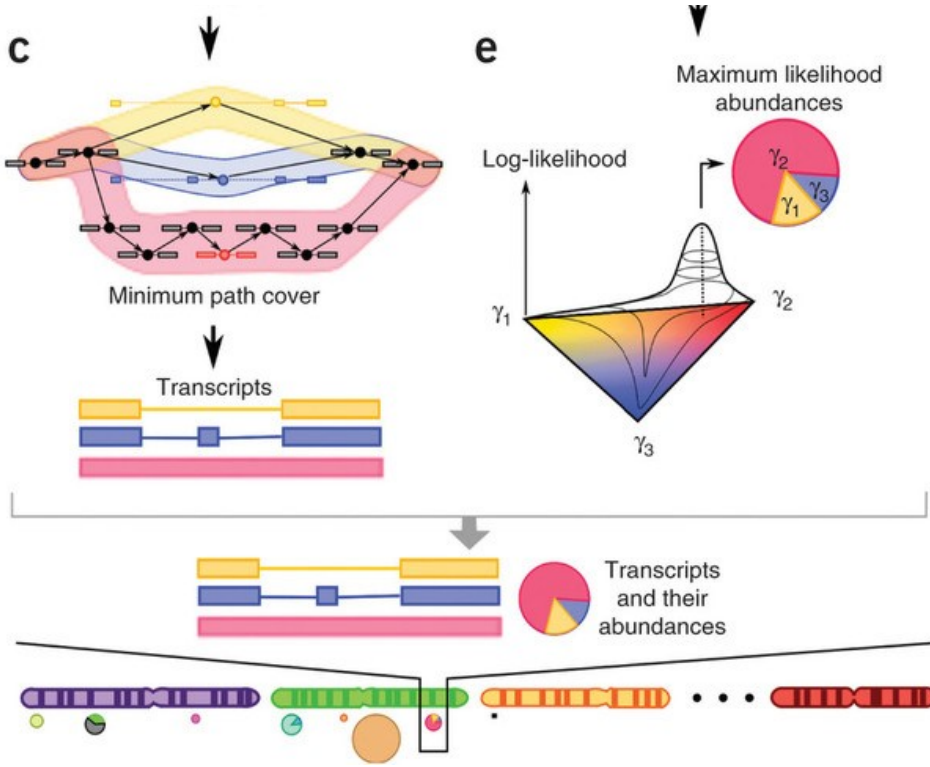
Published online 1 March 2012; corrected after print 7 August 2014; doi:10.1038/nprot.2012.016

Recent advances in high-throughput cDNA sequencing (RNA-seq) can reveal new genes and splice variants and quantify expression genome-wide in a single assay. The volume and complexity of data from RNA-seq experiments necessitate scalable, fast and mathematically principled analysis software. TopHat and Cufflinks are free, open-source software tools for gene discovery and comprehensive expression analysis of high-throughput mRNA sequencing (RNA-seq) data. Together, they allow biologists to identify new genes and new splice variants of known ones, as well as compare gene and transcript expression under two or more conditions. This protocol describes in detail how to use TopHat and Cufflinks to perform such analyses. It also covers several accessory tools and utilities that aid in managing data, including CummeRbund, a tool for visualizing RNA-seq analysis results. Although the procedure assumes basic informatics skills, these tools assume little to no background with RNA-seq analysis and are meant for novices and experts alike. The protocol begins with raw sequencing reads and produces a transcriptome assembly, lists of differentially expressed and regulated genes and transcripts, and publication-quality visualizations of analysis results. The protocol's execution time depends on the volume of transcriptome sequencing data and available computing resources but takes less than 1 d of computer time for typical experiments and ~1 h of hands-on time.

Nature Protocols 12 Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks （可按照流程学习分析）

Overview of Cufflinks





NBT 10 Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation

新RNA-seq分析算法

Transcript-level expression analysis of RNA-seq experiments with HISAT, StringTie and Ballgown

Mihaela Pertea[1,2], Daehwan Kim[1], Geo M Pertea[1], Jeffrey T Leek[3] & Steven L Salzberg[1–4]

[1]Center for Computational Biology, McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins School of Medicine, Baltimore, Maryland, USA. [2]Department of Computer Science, Whiting School of Engineering, Johns Hopkins University, Baltimore, Maryland, USA. [3]Department of Biostatistics, Bloomberg School of Public Health, Johns Hopkins University, Baltimore, Maryland, USA. [4]Department of Biomedical Engineering, Johns Hopkins University, Baltimore, Maryland, USA. Correspondence should be addressed to S.L.S. (salzberg@jhu.edu).

Published online 11 August 2016; doi:10.1038/nprot.2016.095

High-throughput sequencing of mRNA (RNA-seq) has become the standard method for measuring and comparing the levels of gene expression in a wide variety of species and conditions. RNA-seq experiments generate very large, complex data sets that demand fast, accurate and flexible software to reduce the raw read data to comprehensible results. HISAT (hierarchical indexing for spliced alignment of transcripts), StringTie and Ballgown are free, open-source software tools for comprehensive analysis of RNA-seq experiments. Together, they allow scientists to align reads to a genome, assemble transcripts including novel splice variants, compute the abundance of these transcripts in each sample and compare experiments to identify differentially expressed genes and transcripts. This protocol describes all the steps necessary to process a large set of raw sequencing reads and create lists of gene transcripts, expression levels, and differentially expressed genes and transcripts. The protocol's execution time depends on the computing resources, but it typically takes under 45 min of computer time. HISAT, StringTie and Ballgown are available from http://ccb.jhu.edu/software.shtml.

Nature Protocols 16 Transcript-level expression analysis of RNA-seq experiments with HISAT, StringTie and Ballgown

Commonly used RNA-seq analysis methods

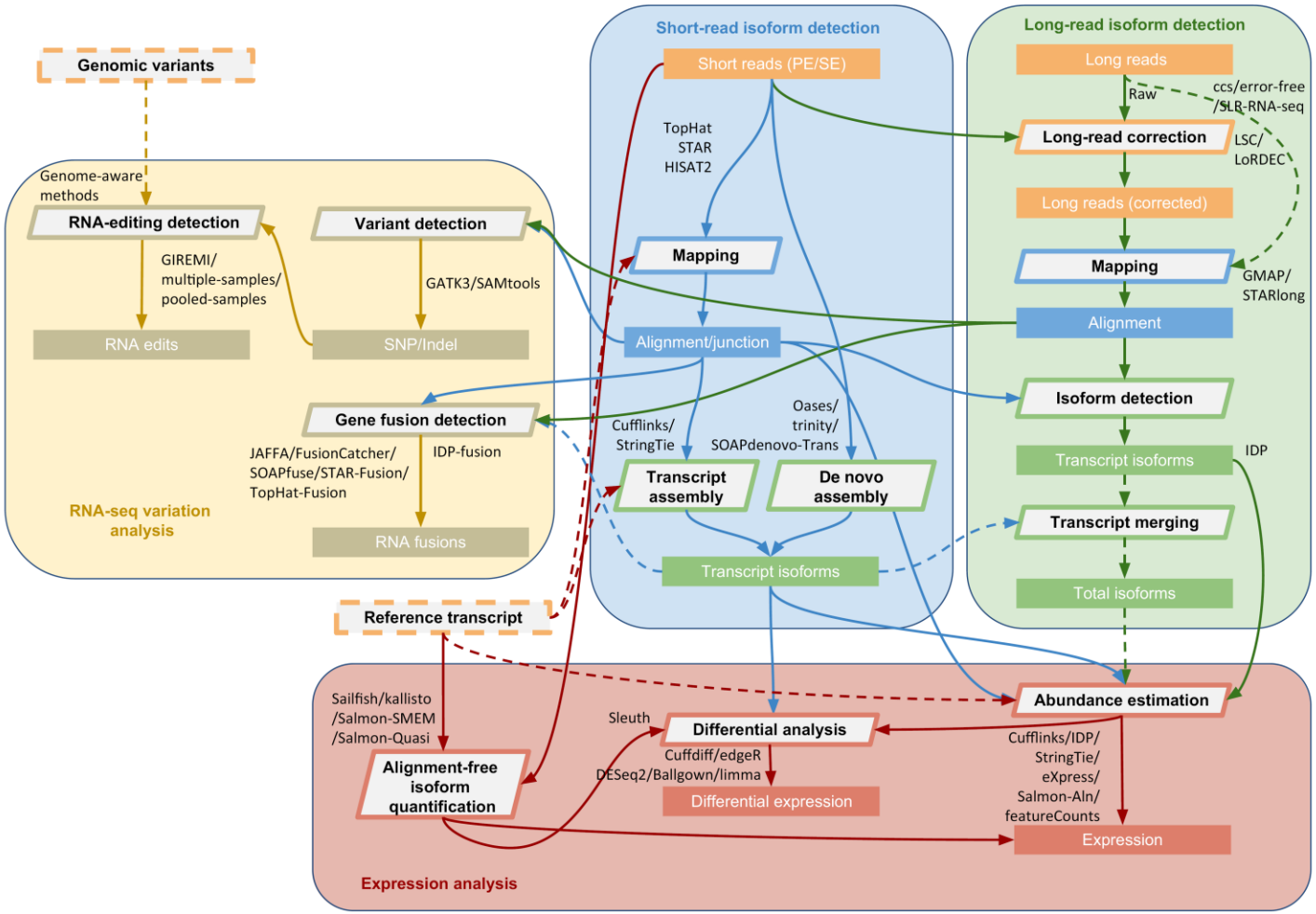
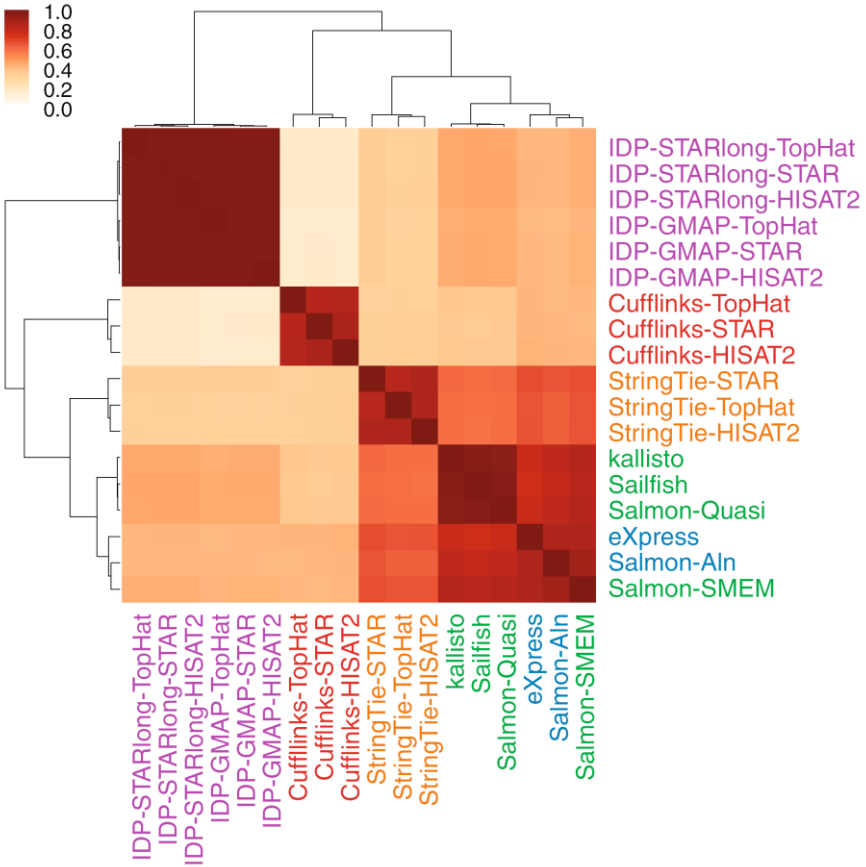
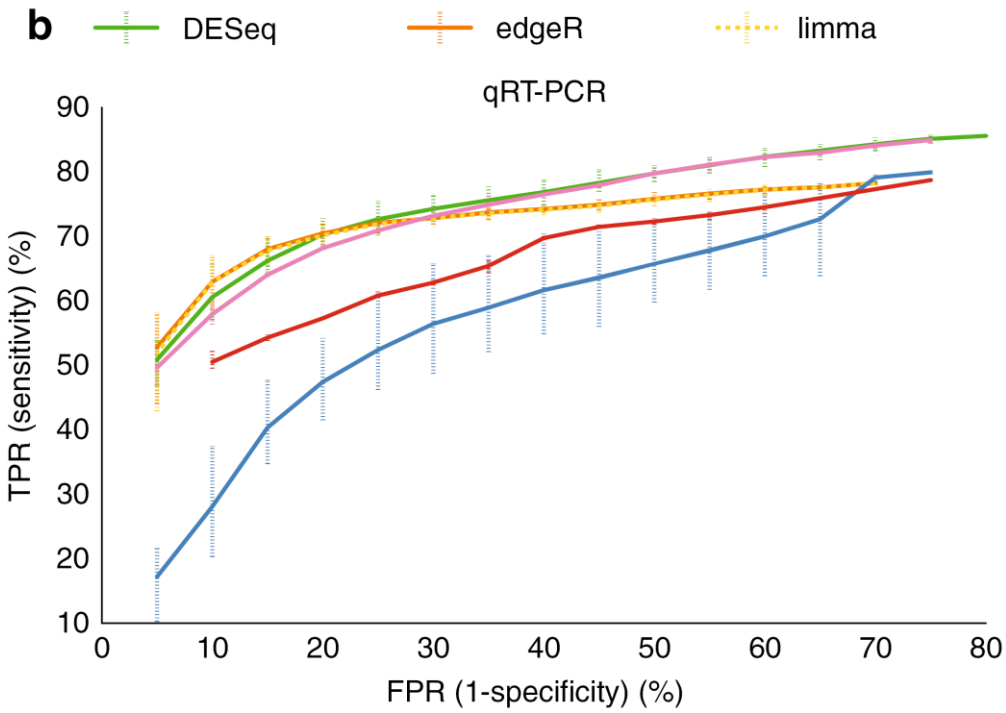


Fig. 1 The RNACocktail analysis protocol. RNACocktail is a comprehensive protocol of RNA-seq data analysis. The figure summarizes the widely used approaches for the key steps over the broad spectrum of RNA-seq analysis and also succinctly captures the possible workflows one can use to analyse RNA-seq data

NC 17 Gaining comprehensive biological insight into the transcriptome by performing a broad-spectrum RNA-seq analysis

Benchmark comparisons





RNACocktail protocol

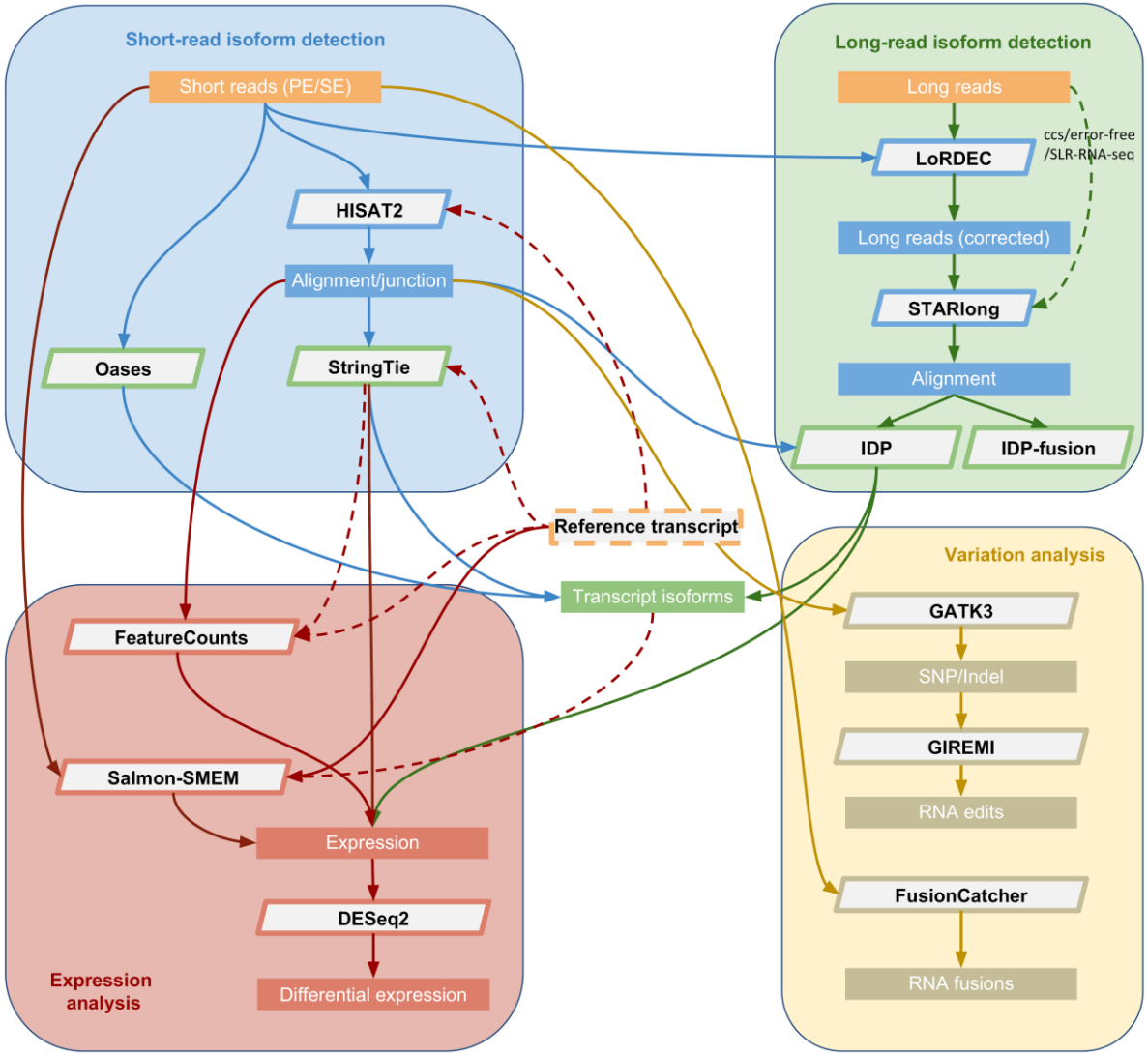


Fig. 8 The current RNACocktail computational pipeline. The pipeline is composed of high-accuracy tools in each step for general-purpose RNA-seq analysis

www.bioconductor.org

- **Bioconductor** provides R packages for the analysis of genomic data, tools for analysis of data from next-generation high-throughput sequencing methods
  - 1104 packages for bioscience data analysis (*as of March 10, 2016*)
  - active user community, open source and open development
  - two releases each year
- Statistical analysis: large data, technological artifacts, designed experiments; rigorous
- Comprehension: biological context, visualization, reproducibility
- High-throughput
  - Sequencing: RNASeq, ChIPSeq, variants, copy number, …
  - Microarrays: expression, SNP, …
  - Flow cytometry, proteomics, images, …

```
source("http://bioconductor.org/biocLite.R")
biocLite()  # install core packages
```

http://biocluster.ucr.edu/~rkaundal/workshops/R_mar2016/RNAseq.html

《Bioconductor Case Studies》

Bioconductor的教材，可了解组学数据常用的分析方法和数据结构

Software for RNA-Seq DEG (differentially expressed genes) Analysis

- edgeR (Robinson et al., 2010)
- DESeq / DESeq2 (Anders and Huber, 2010, 2014)
- DEXSeq (Anders et al., 2012)
- limmaVoom
- Cuffdiff / Cuffdiff2 (Trapnell et al., 2013)
- PoissonSeq
- baySeq

Published online 20 January 2015

Nucleic Acids Research, 2015, Vol. 43, No. 7 ***e47***
doi: 10.1093/nar/gkv007

limma powers differential expression analyses for RNA-sequencing and microarray studies

Matthew E. Ritchie[1,2], Belinda Phipson[3], Di Wu[4], Yifang Hu[5], Charity W. Law[6], Wei Shi[5,7] and Gordon K. Smyth[2,5,*]

Data example

EGF-mediated induction of Mcl-1 at the switch to lactation is essential for alveolar cell survival

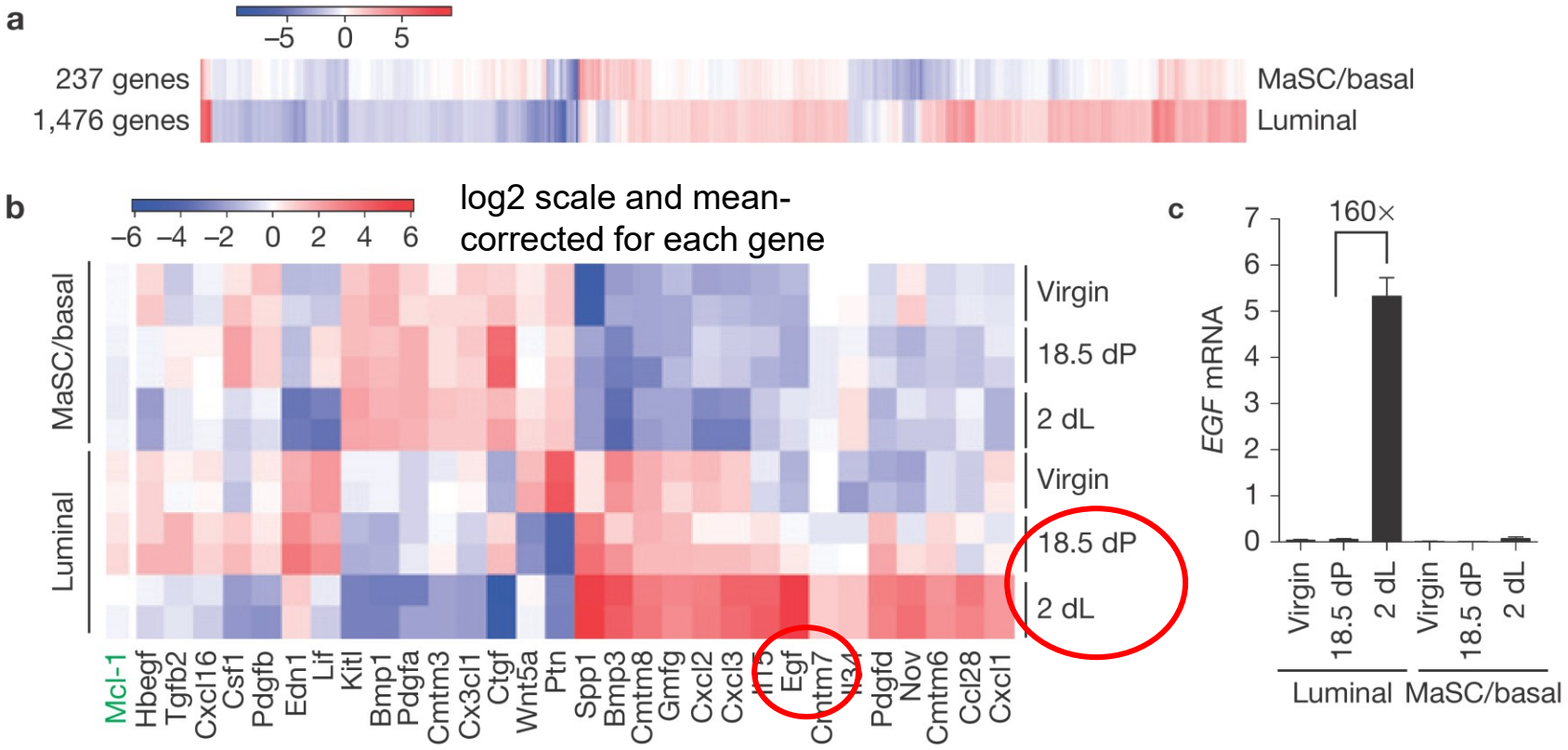
Nai Yang Fu[1,2], Anne C. Rios[1,2,10], Bhupinder Pal[1,2,10], Rina Soetanto[3], Aaron T. L. Lun[2,4], Kevin Liu[1,2], Tamara Beck[1,2], Sarah A. Best[1,2], François Vaillant[1,2], Philippe Bouillet[2,5], Andreas Strasser[2,5], Thomas Preiss[3,6], Gordon K. Smyth[4,7], Geoffrey J. Lindeman[1,8,9,10] and Jane E. Visvader[1,2,10,11]

Expansion and remodelling of the mammary epithelium requires a tight balance between cellular proliferation, differentiation and death. To explore cell survival versus cell death decisions in this organ, we deleted the pro-survival gene *Mcl-1* in the mammary epithelium. *Mcl-1* was found to be essential at multiple developmental stages including morphogenesis in puberty and alveologenesis in pregnancy. Moreover, *Mcl-1*-deficient basal cells were virtually devoid of repopulating activity, suggesting that this gene is required for stem cell function. Profound upregulation of the Mcl-1 protein was evident in alveolar cells at the switch to lactation, and *Mcl-1* deficiency impaired lactation. Interestingly, *EGF* was identified as one of the most highly upregulated genes on lactogenesis and inhibition of EGF or mTOR signalling markedly impaired lactation, with concomitant decreases in Mcl-1 and phosphorylated ribosomal protein S6. These data demonstrate that Mcl-1 is essential for mammopoiesis and identify EGF as a critical trigger of Mcl-1 translation to ensure survival of milk-producing alveolar cells.

NCB 15 EGF-mediated induction of Mcl-1 at the switch to lactation is essential for alveolar cell survival

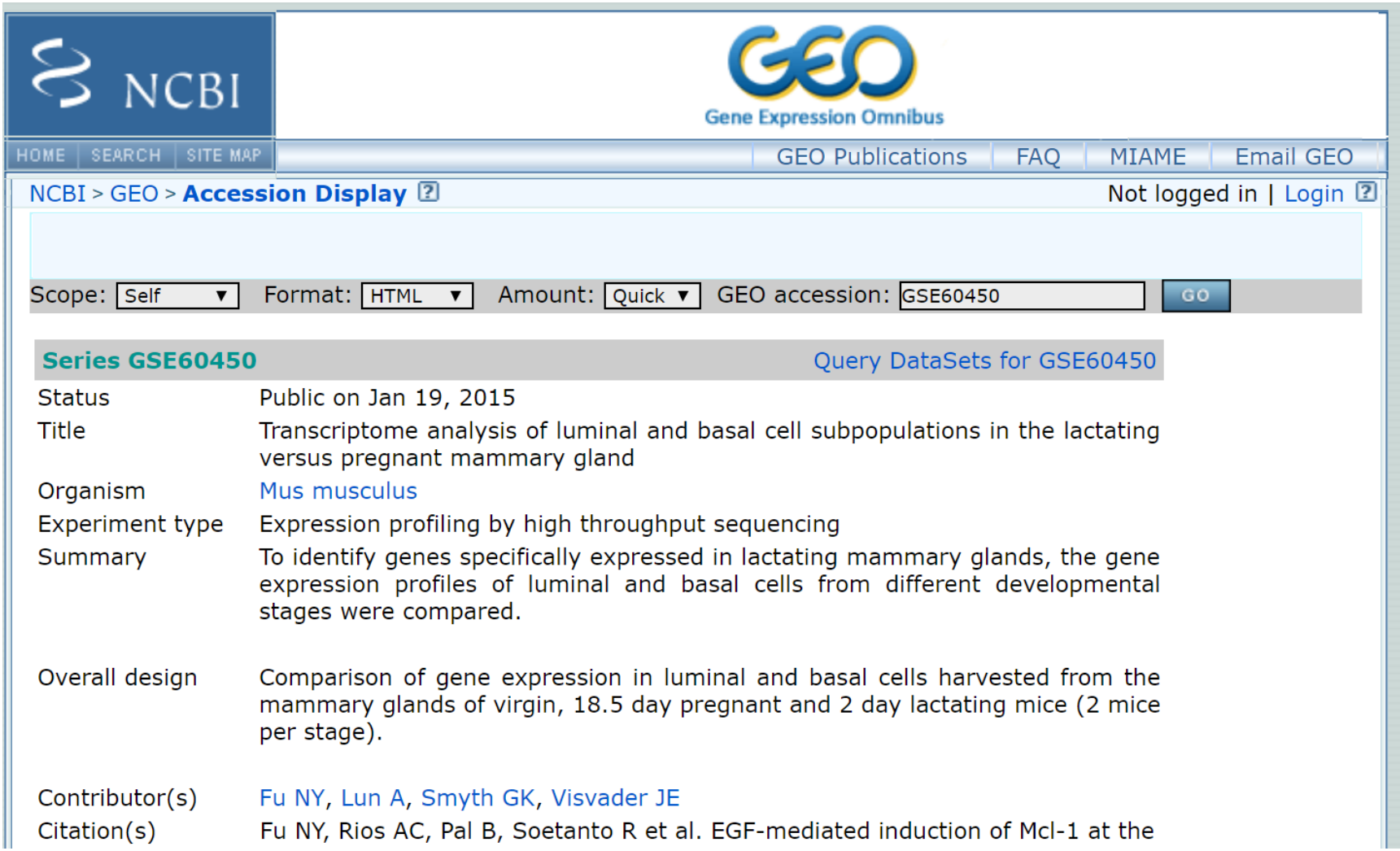
Data example

p.7: We next interrogated potential molecular regulators that could mediate the marked induction of Mcl-1 expression that occurs on lactogenesis.



NCB 15 EGF-mediated induction of Mcl-1 at the switch to lactation is essential for alveolar cell survival

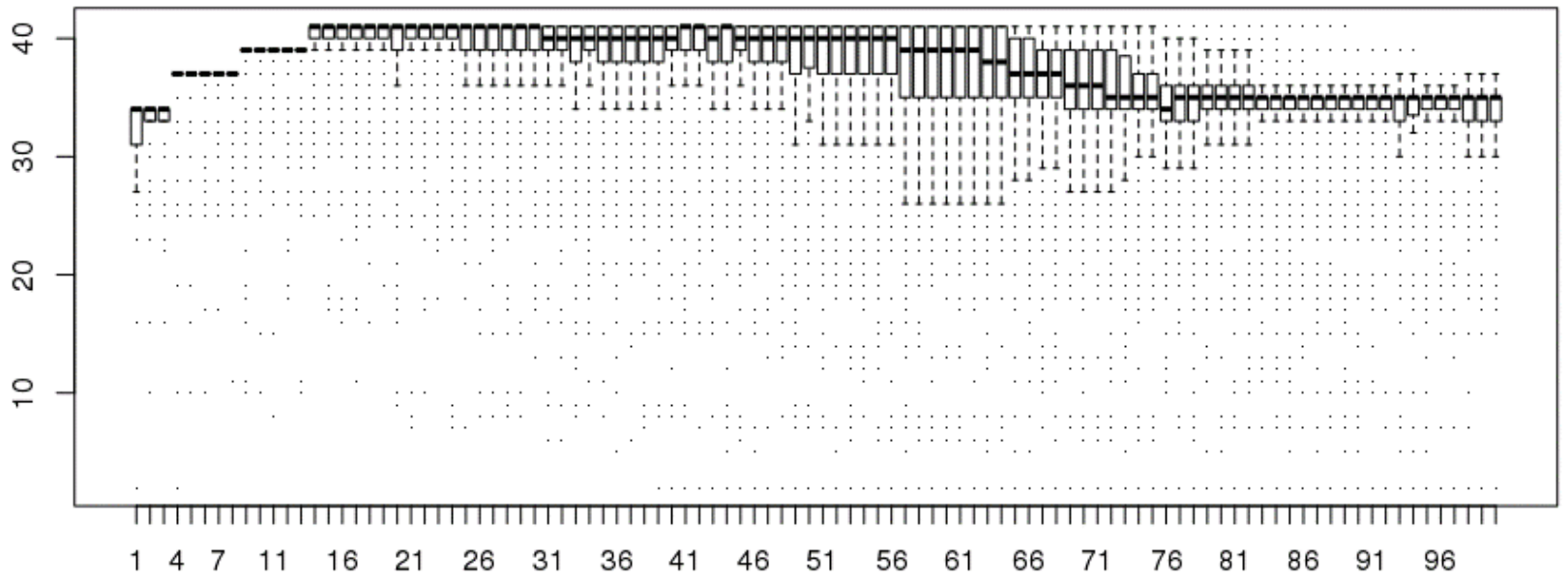
GEO & SRA database



https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE60450

(可找到测序数据的SRA文件)

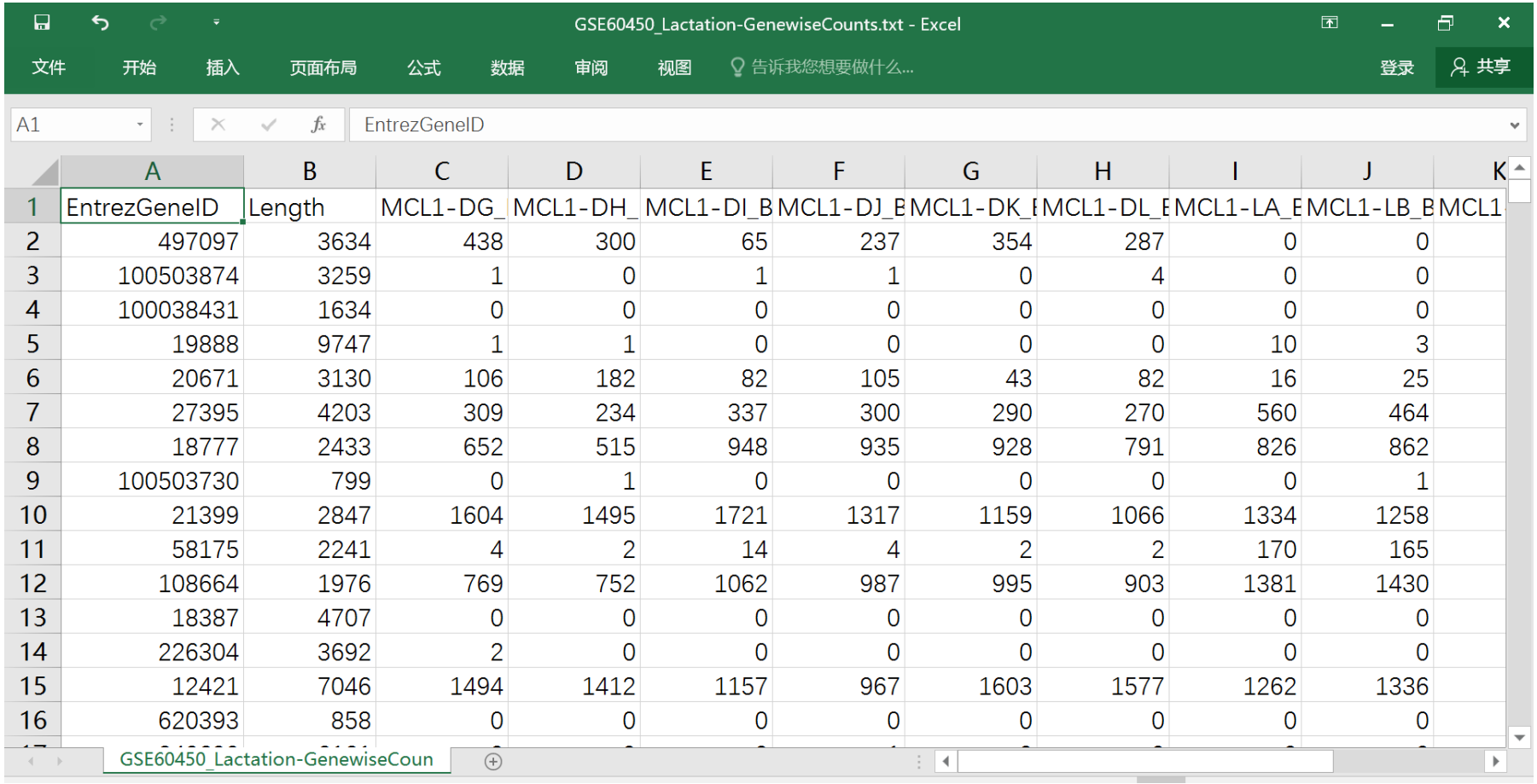
RNA-seq analysis with R: basepair quality



X轴：碱基位置；Y轴：碱基质量值分布

http://combine-australia.github.io/RNAseq-R/07-rnaseq-day2.html
课件：RNA-seq mapping.R

Alignment and read counts per gene


GSE60450_Lactation-GenewiseCounts.txt - Excel

文件 开始 插入 页面布局 公式 数据 审阅 视图 告诉我您想要做什么... 登录 共享

A1 EntrezGeneID

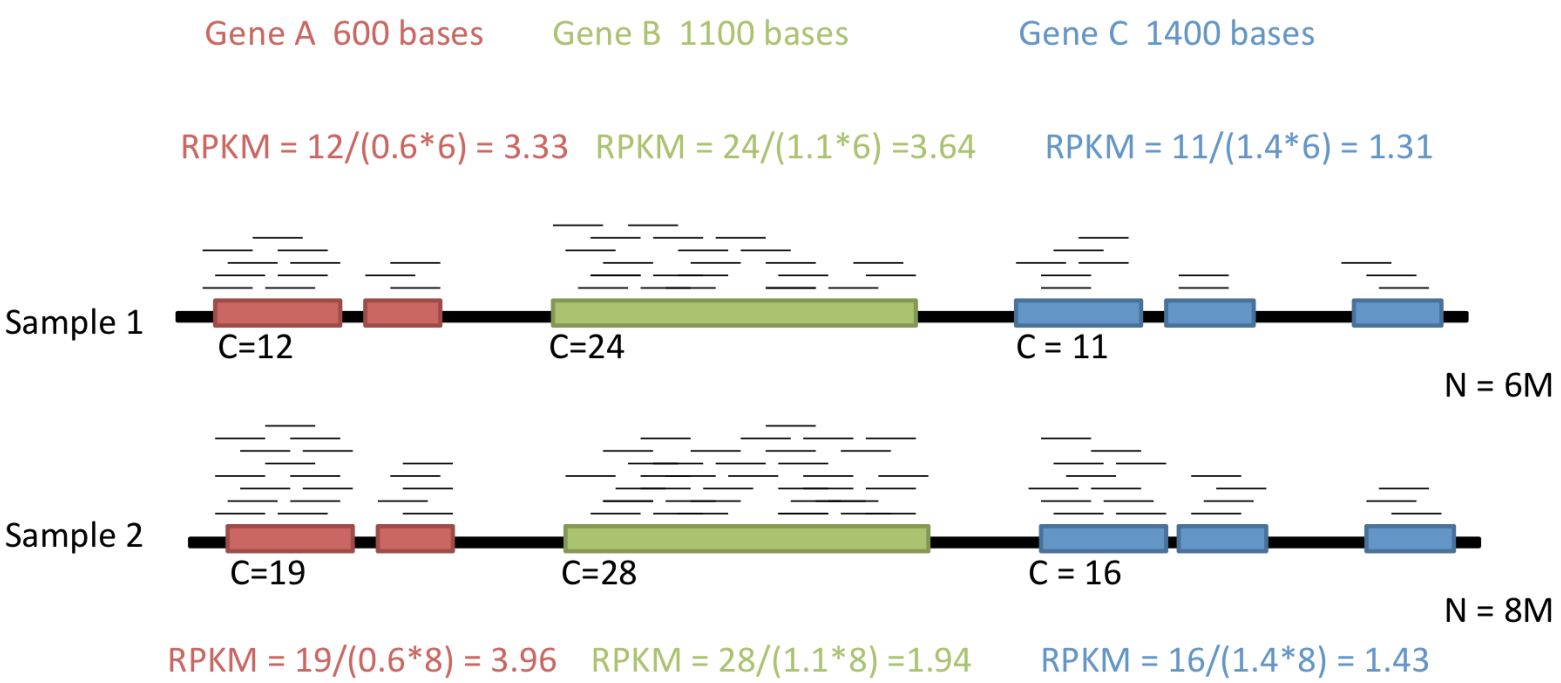
	A	B	C	D	E	F	G	H	I	J	K
1	EntrezGeneID	Length	MCL1-DG_	MCL1-DH_	MCL1-DI_B	MCL1-DJ_B	MCL1-DK_B	MCL1-DL_B	MCL1-LA_B	MCL1-LB_B	MCL1
2	497097	3634	438	300	65	237	354	287	0	0	
3	100503874	3259	1	0	1	1	0	4	0	0	
4	100038431	1634	0	0	0	0	0	0	0	0	
5	19888	9747	1	1	0	0	0	0	10	3	
6	20671	3130	106	182	82	105	43	82	16	25	
7	27395	4203	309	234	337	300	290	270	560	464	
8	18777	2433	652	515	948	935	928	791	826	862	
9	100503730	799	0	1	0	0	0	0	0	1	
10	21399	2847	1604	1495	1721	1317	1159	1066	1334	1258	
11	58175	2241	4	2	14	4	2	2	170	165	
12	108664	1976	769	752	1062	987	995	903	1381	1430	
13	18387	4707	0	0	0	0	0	0	0	0	
14	226304	3692	2	0	0	0	0	0	0	0	
15	12421	7046	1494	1412	1157	967	1603	1577	1262	1336	
16	620393	858	0	0	0	0	0	0	0	0	

GSE60450_Lactation-GenewiseCoun

Expression values: RPKM/FPKM

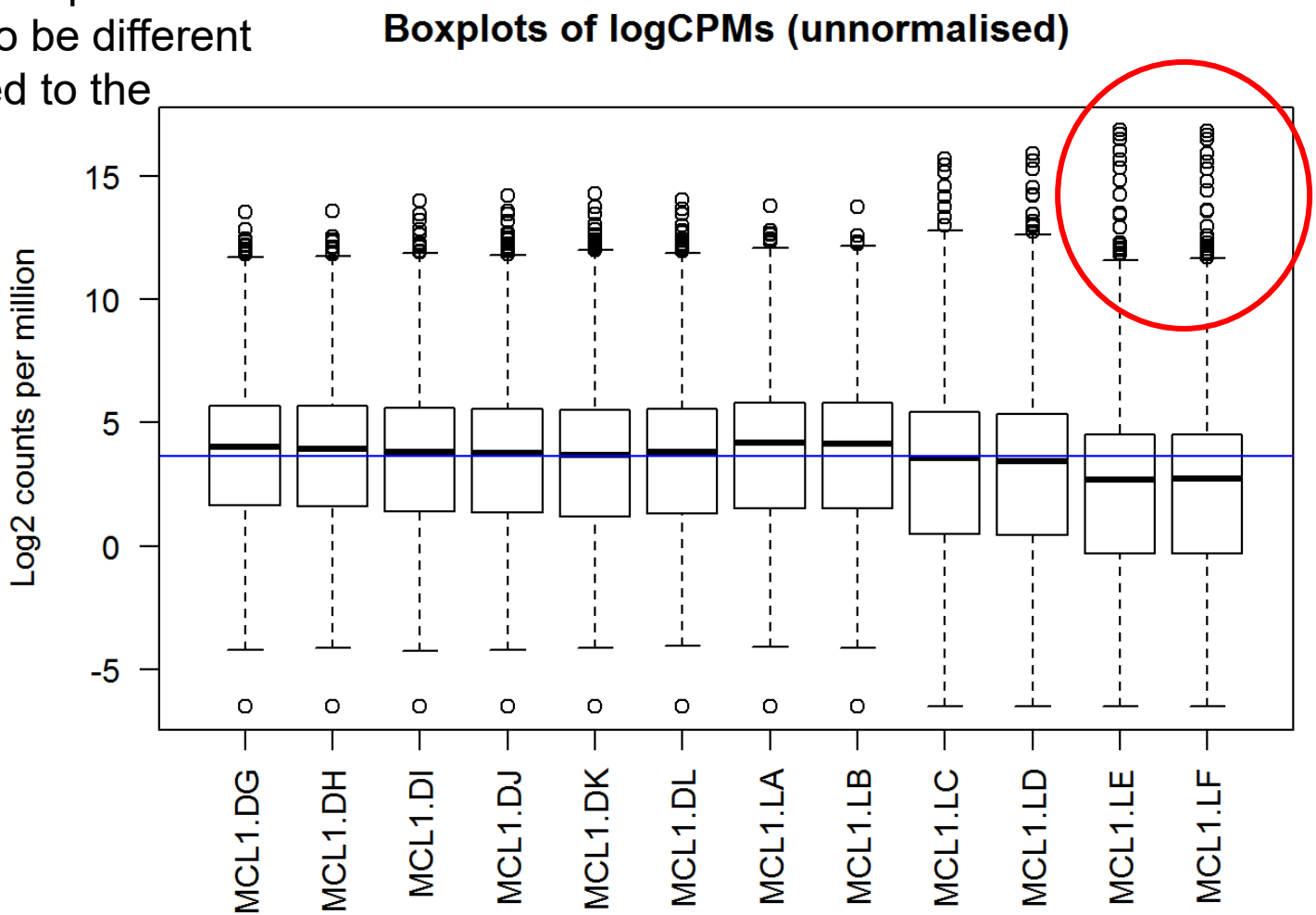
- RPKM (FPKM): **reads (fragments) per kb per million mapped reads**

$$\mathrm{RPKM} = \frac{total\ exon\ reads}{mapped\ reads\ (millions) * exon\ length\ (KB)}$$



RNA-seq analysis with R: count distribution

Do any samples appear to be different compared to the others?

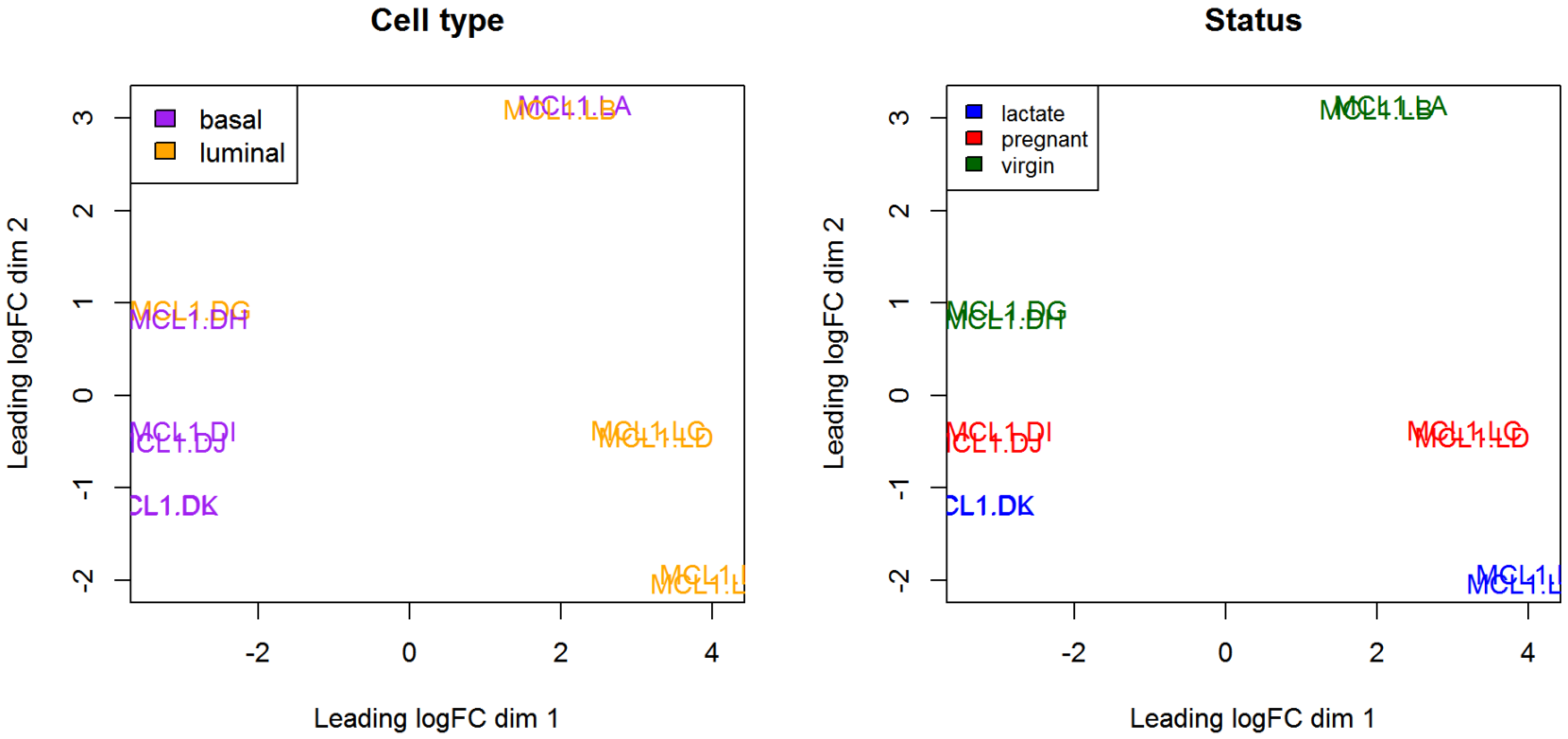


A few genes with large counts may bias the values of other genes

http://combine-australia.github.io/RNAseq-R/06-rnaseq-day1.html

课件：RNA-seq diff_expr.R

RNA-seq analysis with R: MDS (multi-dimensiaonl scaling) clustering

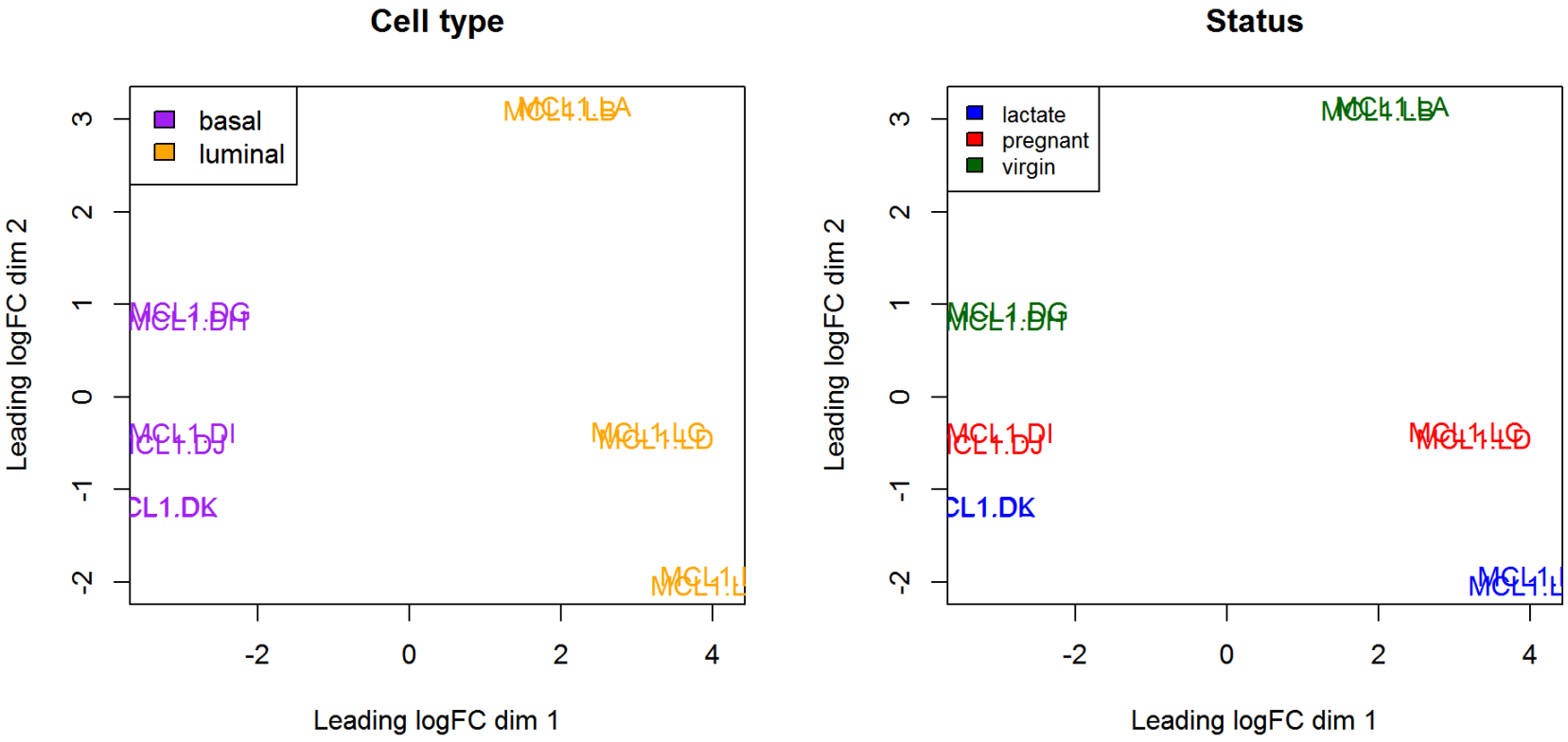


探索性数据分析

Is there something strange going on with the samples? Identify the two samples that don't appear to be in the right place.

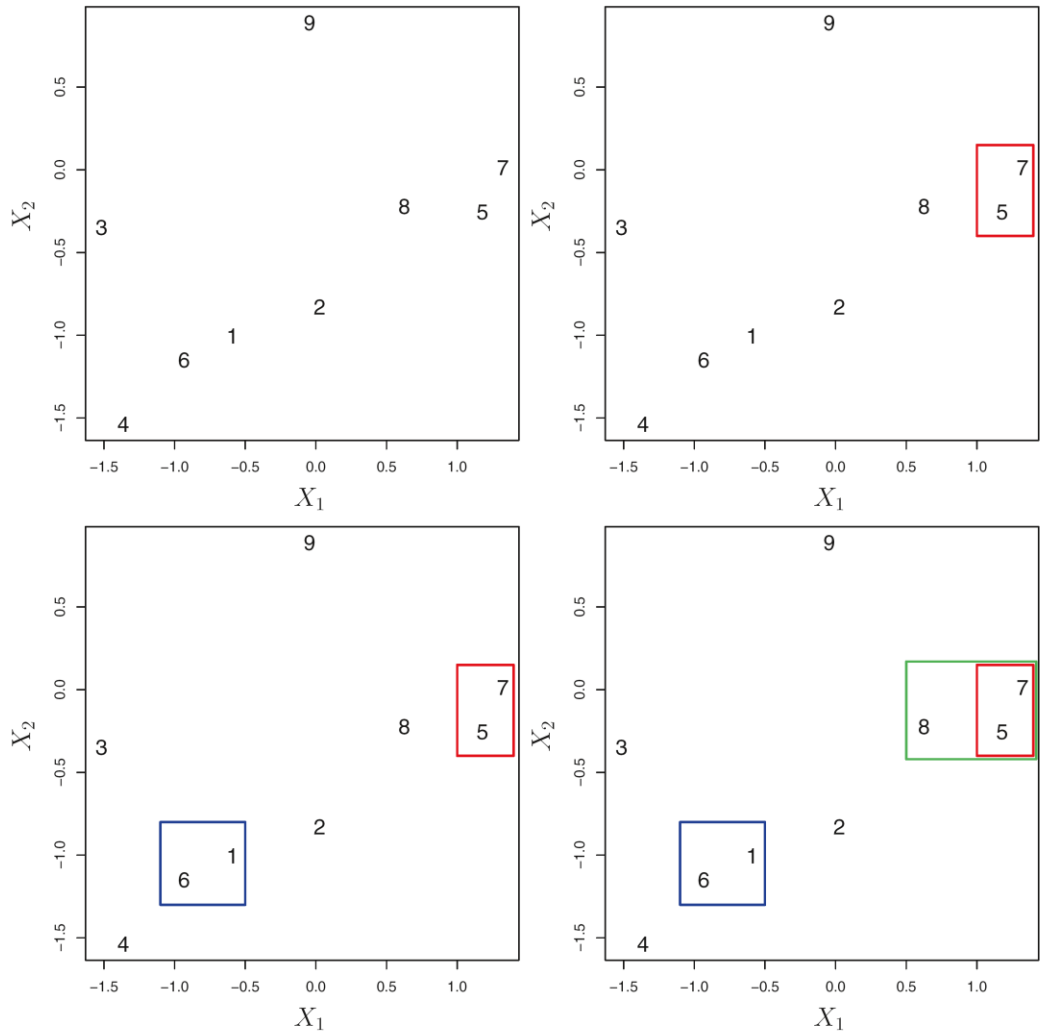
RNA-seq analysis with R: MDS clustering

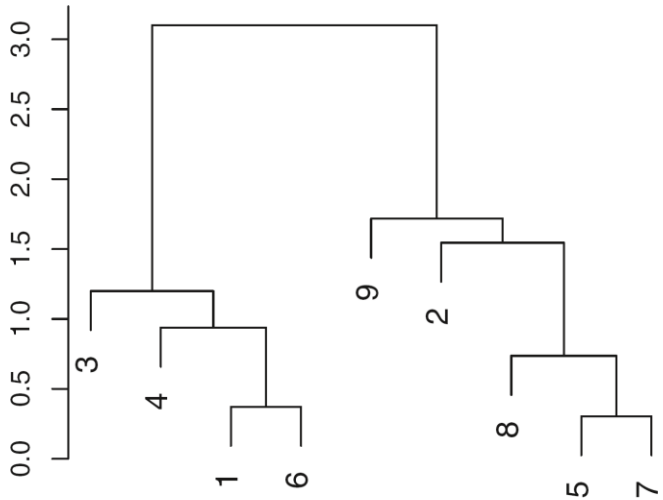
Sample information corrected



What is the greatest source of variation in the data (i.e. what does dimension 1 represent)? What is the second greatest source of variation in the data?

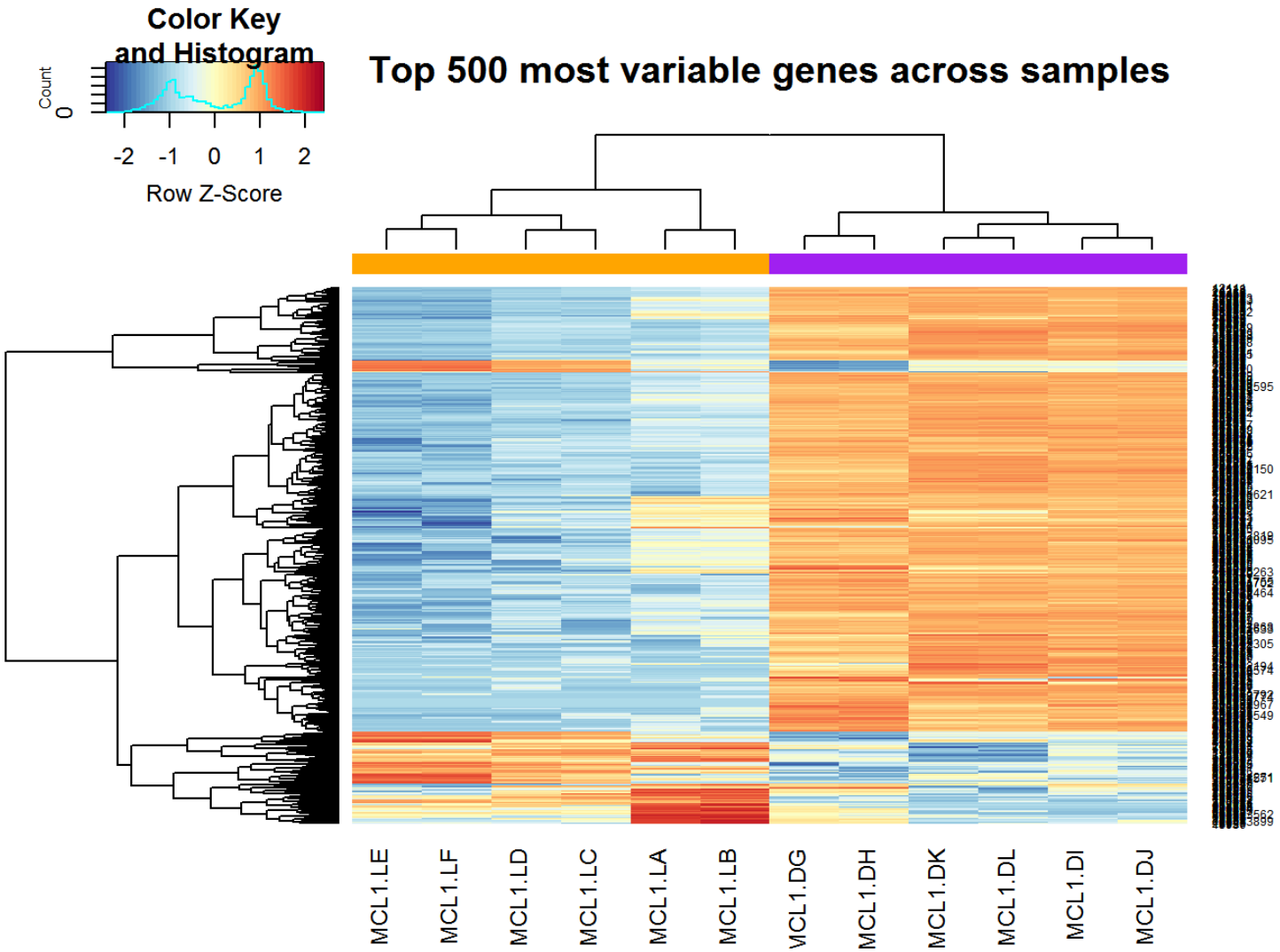
Hierarchical clustering （层次聚类）算法



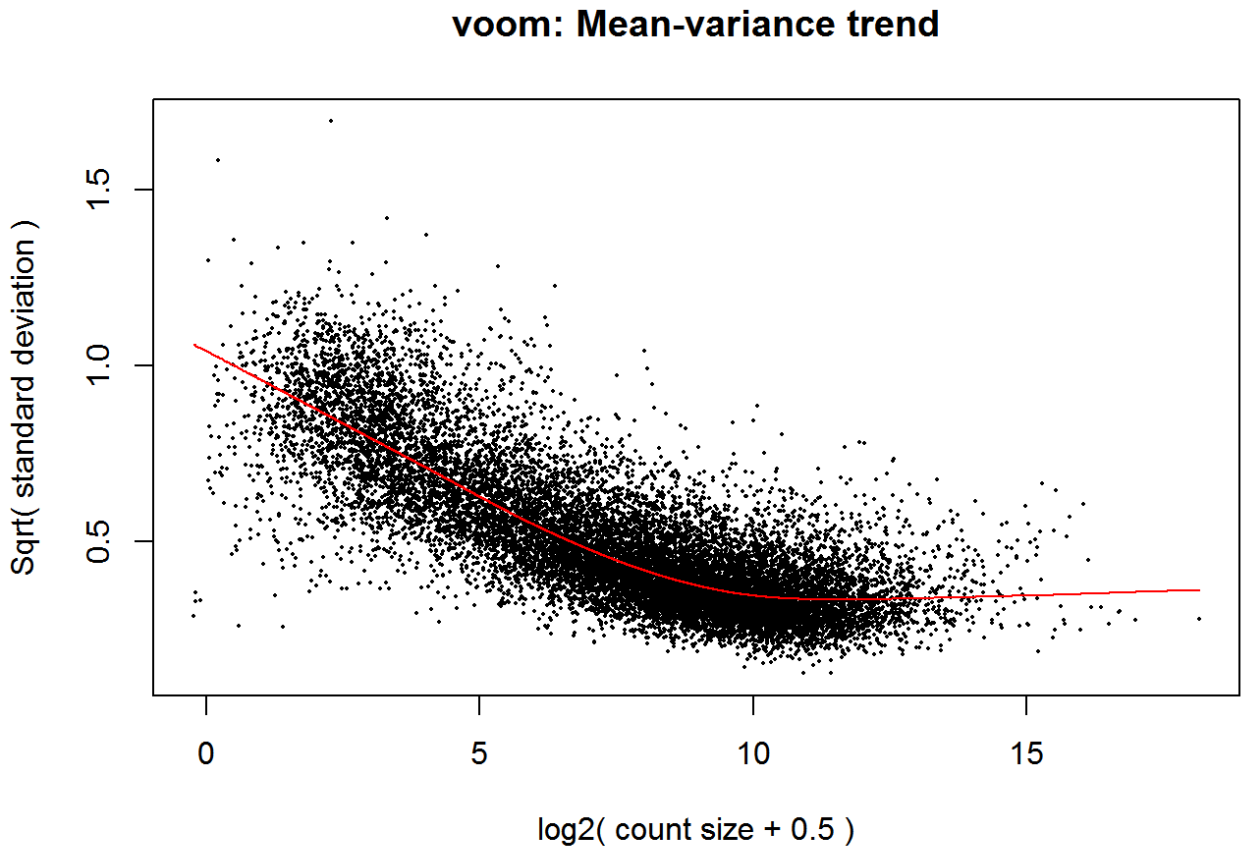


《统计学习导论》，图10-11

RNA-seq analysis with R: Hierarchical clustering with heatmaps



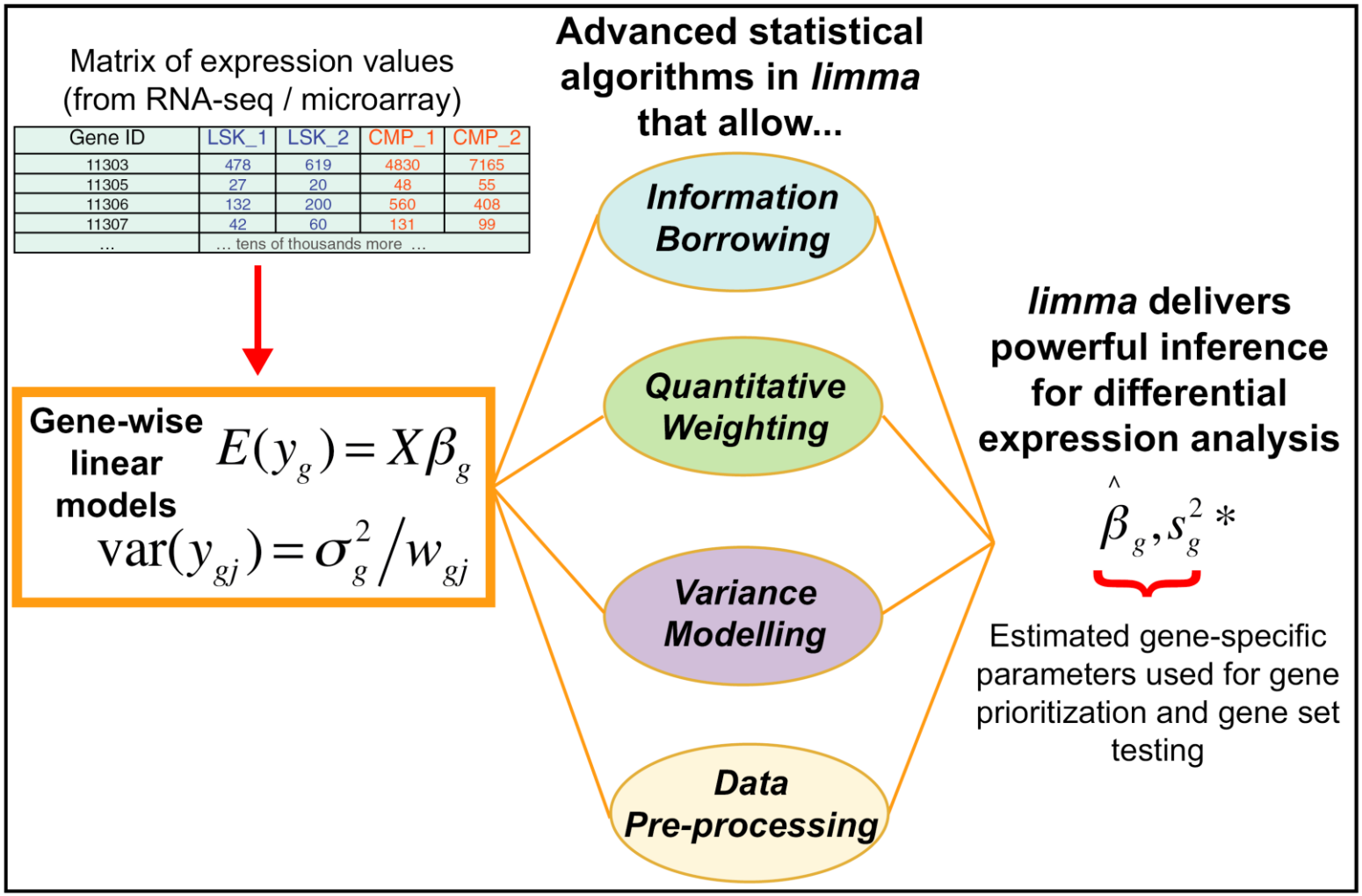
RNA-seq analysis with R: gene variance vs. mean



Problem: 1 vs. 1 or 3 vs. 3 comparison: gene-wise t-test is not robust when using small sample size to estimate the variance of the mean

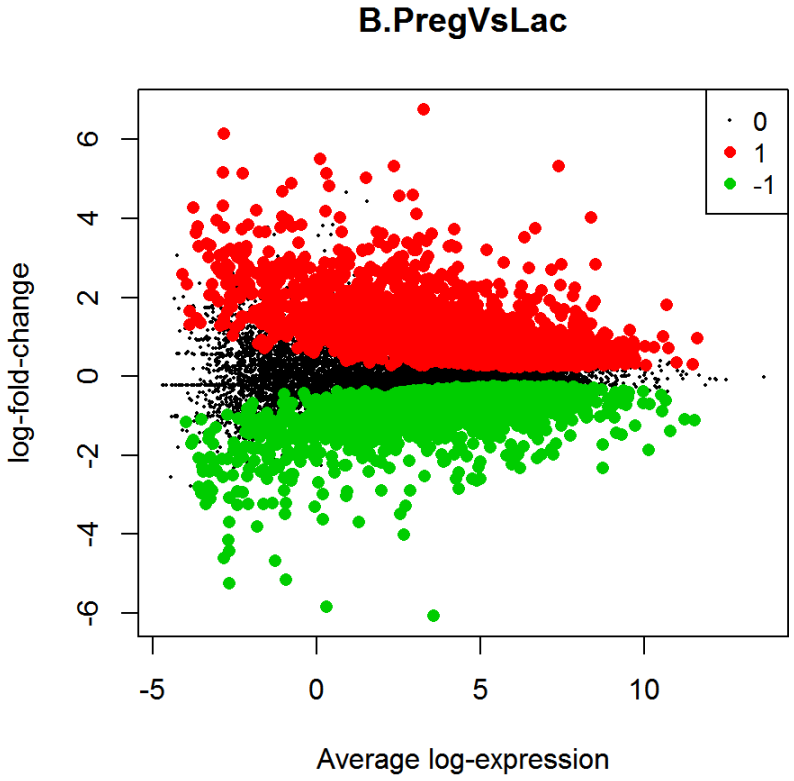
Bayesian methods: borrow variance information acrsoss genes

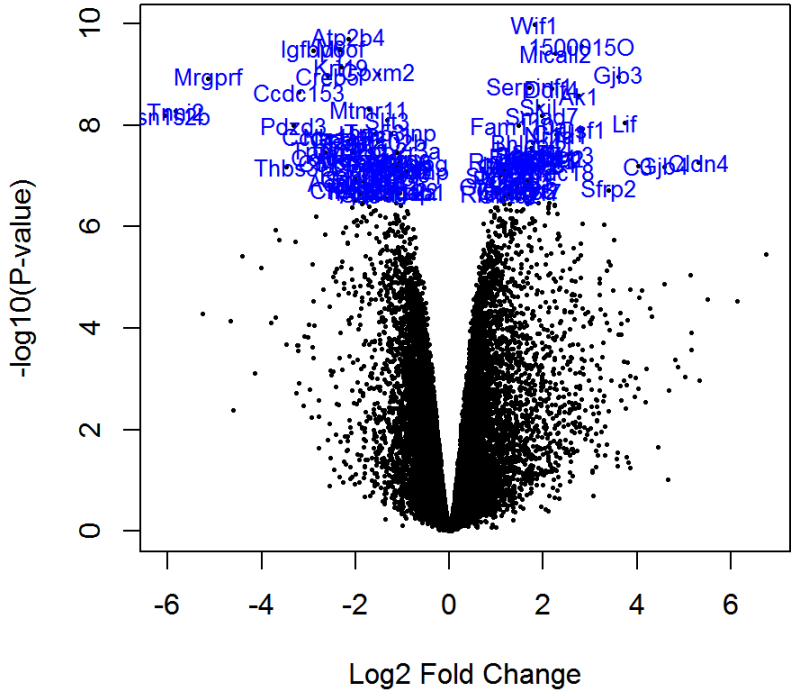
Limma analysis



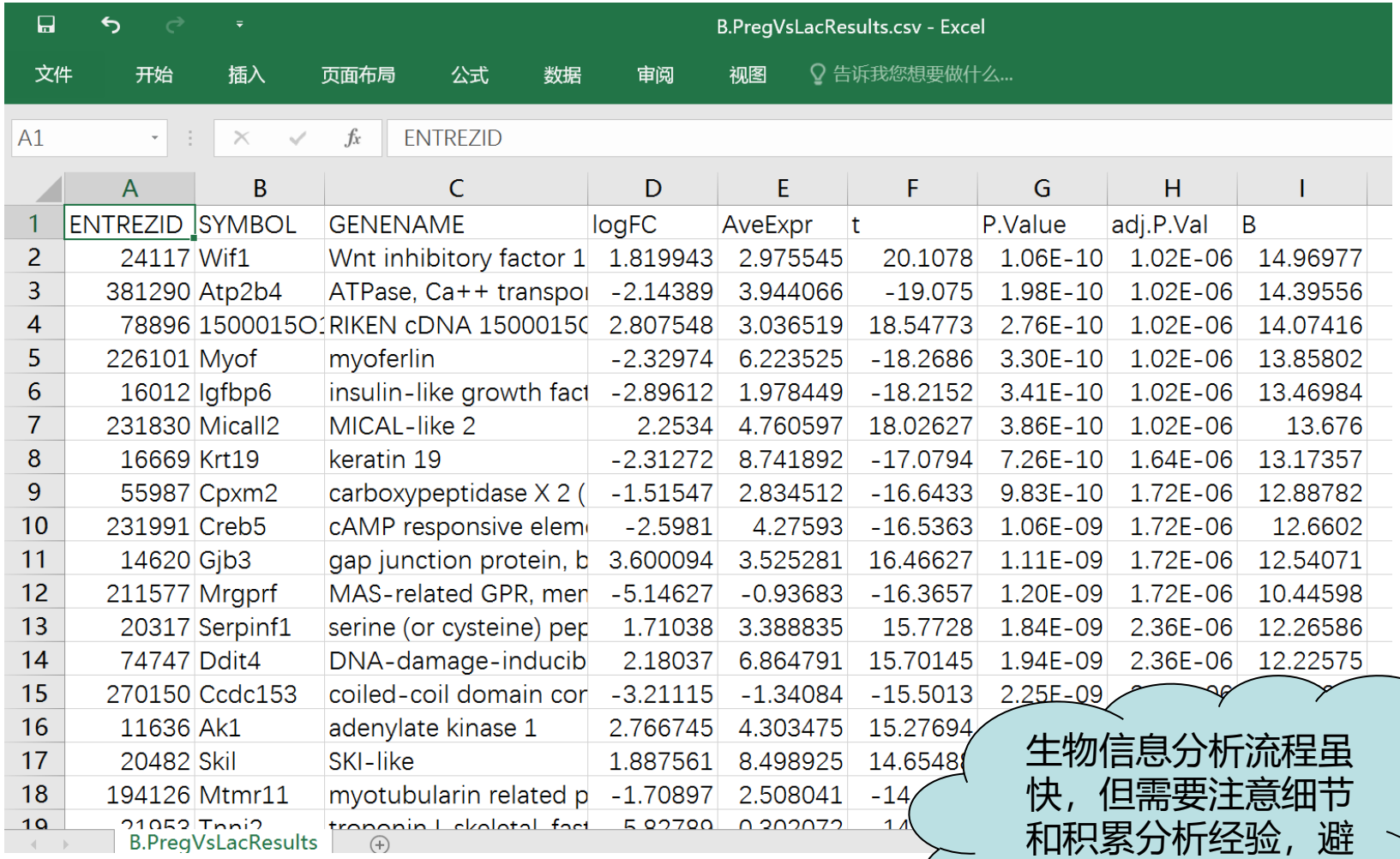
NAR 15 limma powers differential expression analyses for RNA-sequencing and microarray studies

RNA-seq analysis with R: Differential expression





分析结果



	A	B	C	D	E	F	G	H	I
1	ENTREZID	SYMBOL	GENENAME	logFC	AveExpr	t	P.Value	adj.P.Val	B
2	24117	Wif1	Wnt inhibitory factor 1	1.819943	2.975545	20.1078	1.06E-10	1.02E-06	14.96977
3	381290	Atp2b4	ATPase, Ca++ transpo	-2.14389	3.944066	-19.075	1.98E-10	1.02E-06	14.39556
4	78896	1500015O	RIKEN cDNA 1500015(	2.807548	3.036519	18.54773	2.76E-10	1.02E-06	14.07416
5	226101	Myof	myoferlin	-2.32974	6.223525	-18.2686	3.30E-10	1.02E-06	13.85802
6	16012	Igfbp6	insulin-like growth fact	-2.89612	1.978449	-18.2152	3.41E-10	1.02E-06	13.46984
7	231830	Micall2	MICAL-like 2	2.2534	4.760597	18.02627	3.86E-10	1.02E-06	13.676
8	16669	Krt19	keratin 19	-2.31272	8.741892	-17.0794	7.26E-10	1.64E-06	13.17357
9	55987	Cpxm2	carboxypeptidase X 2 (	-1.51547	2.834512	-16.6433	9.83E-10	1.72E-06	12.88782
10	231991	Creb5	cAMP responsive elem	-2.5981	4.27593	-16.5363	1.06E-09	1.72E-06	12.6602
11	14620	Gjb3	gap junction protein, b	3.600094	3.525281	16.46627	1.11E-09	1.72E-06	12.54071
12	211577	Mrgprf	MAS-related GPR, mem	-5.14627	-0.93683	-16.3657	1.20E-09	1.72E-06	10.44598
13	20317	Serpinf1	serine (or cysteine) pep	1.71038	3.388835	15.7728	1.84E-09	2.36E-06	12.26586
14	74747	Ddit4	DNA-damage-inducib	2.18037	6.864791	15.70145	1.94E-09	2.36E-06	12.22575
15	270150	Ccdc153	coiled-coil domain cor	-3.21115	-1.34084	-15.5013	2.25E-09		
16	11636	Ak1	adenylate kinase 1	2.766745	4.303475	15.27694			
17	20482	Skil	SKI-like	1.887561	8.498925	14.6548			
18	194126	Mtmr11	myotubularin related p	-1.70897	2.508041	-14			
19	21953	Tnni2	troponin I, skeletal, fast	5.82789	0.302072	14			

B.PregVsLacResults

生物信息分析流程虽快，但需要注意细节和积累分析经验，避免错误结果和解释

B.PregVsLacResults.csv

RNA-seq analysis with R: Gene set enrichment

```
# Rerun goana with gene length information
          go_length <- goana(fit.cont,coef="B.PregVsLac",species="Mm",covariate=gene_length)
          topGO(go_length, n=10)
```

```
                                          Term Ont    N  Up Down
GO:0003723                                   RNA binding  MF 1454 476  160
GO:1990904                  ribonucleoprotein complex  CC  742 293   63
GO:0030529 intracellular ribonucleoprotein complex  CC  741 292   63
GO:0042254                        ribosome biogenesis  BP  244 139    7
GO:0022613    ribonucleoprotein complex biogenesis  BP  361 178   24
GO:0022626                         cytosolic ribosome  CC  107  80    1
GO:0006364                            rRNA processing  BP  171 102    2
GO:0016072                     rRNA metabolic process  BP  197 111    8
GO:0005840                                  ribosome  CC  212 107    5
GO:0022625       cytosolic large ribosomal subunit  CC   58  49    0
                    P.Up P.Down
GO:0003723 8.494584e-55      1
GO:1990904 8.278850e-54      1
GO:0030529 2.003347e-53      1
GO:0042254 1.651659e-48      1
GO:0022613 5.780765e-48      1
GO:0022626 4.702353e-44      1
```

The N column represents the total number of genes that are annotated with each GO term. The Up and Down columns represent the number of differentially expressed genes that overlap with the genes in the GO term. The P.Up and P.Down columns contain the p-values for over-representation of the GO term across the set of up- and down-regulated genes, respectively. The output table is sorted by the minimum of P.Up and P.Down by default.

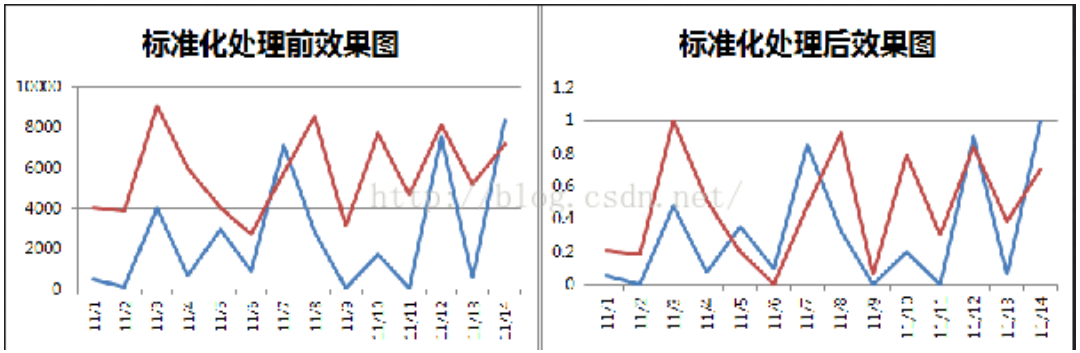
本节内容

- 高通量测序技术
- 测序原始数据
- RNA-seq分析流程
- 统计**模型**：数据归一化
- **演示：**Linux、云计算、分析资源

归一化（Normalization）

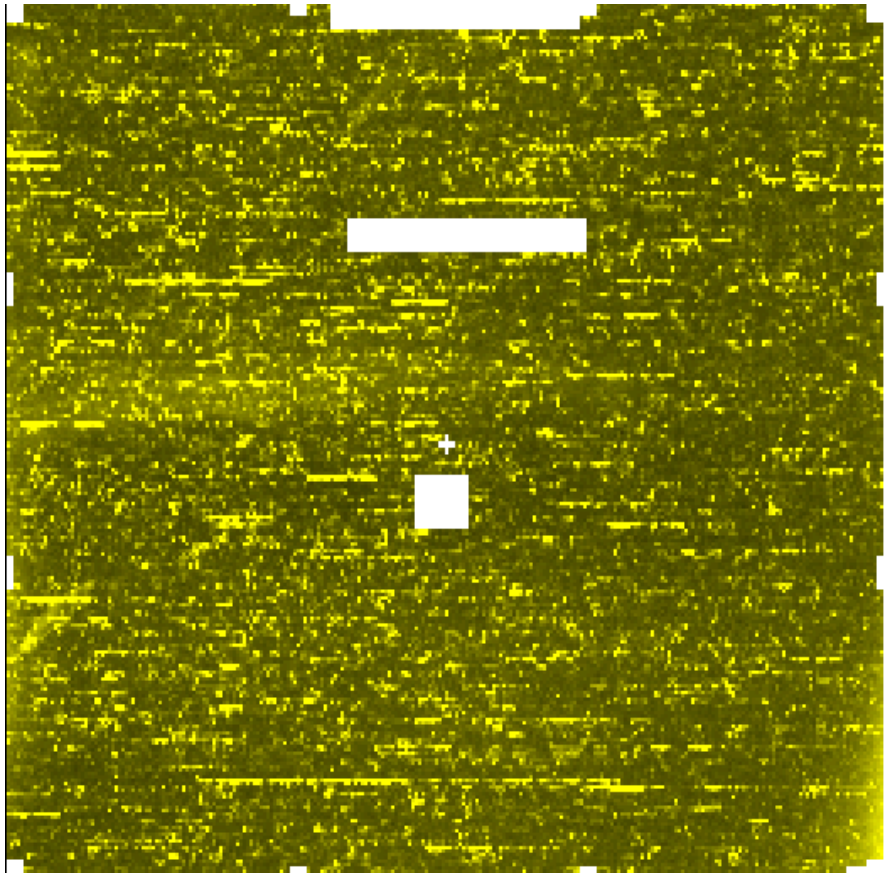
- 把数据变为（0，1）之间的小数。主要是为了方便数据处理，因为将数据映射到0～1范围之内，可以使处理过程更加便捷、快速。

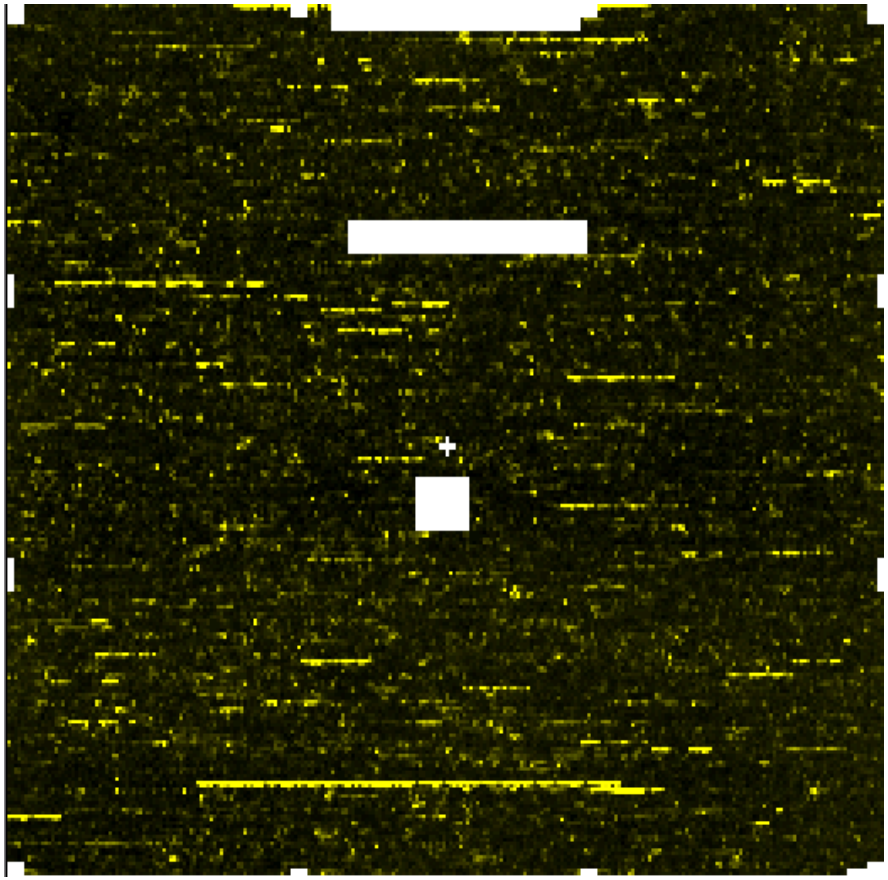
- 把有量纲表达式变换为无量纲表达式，成为纯量。经过归一化处理的数据，处于同一数量级，可以消除指标之间的量纲和量纲单位的影响，提高不同数据指标之间的可比性。
- **常用算法**：
  - 线性转换，即min-max归一化： `y=(x-min)/(max-min)`
  - 对数函数转换： `y=log10(x)`



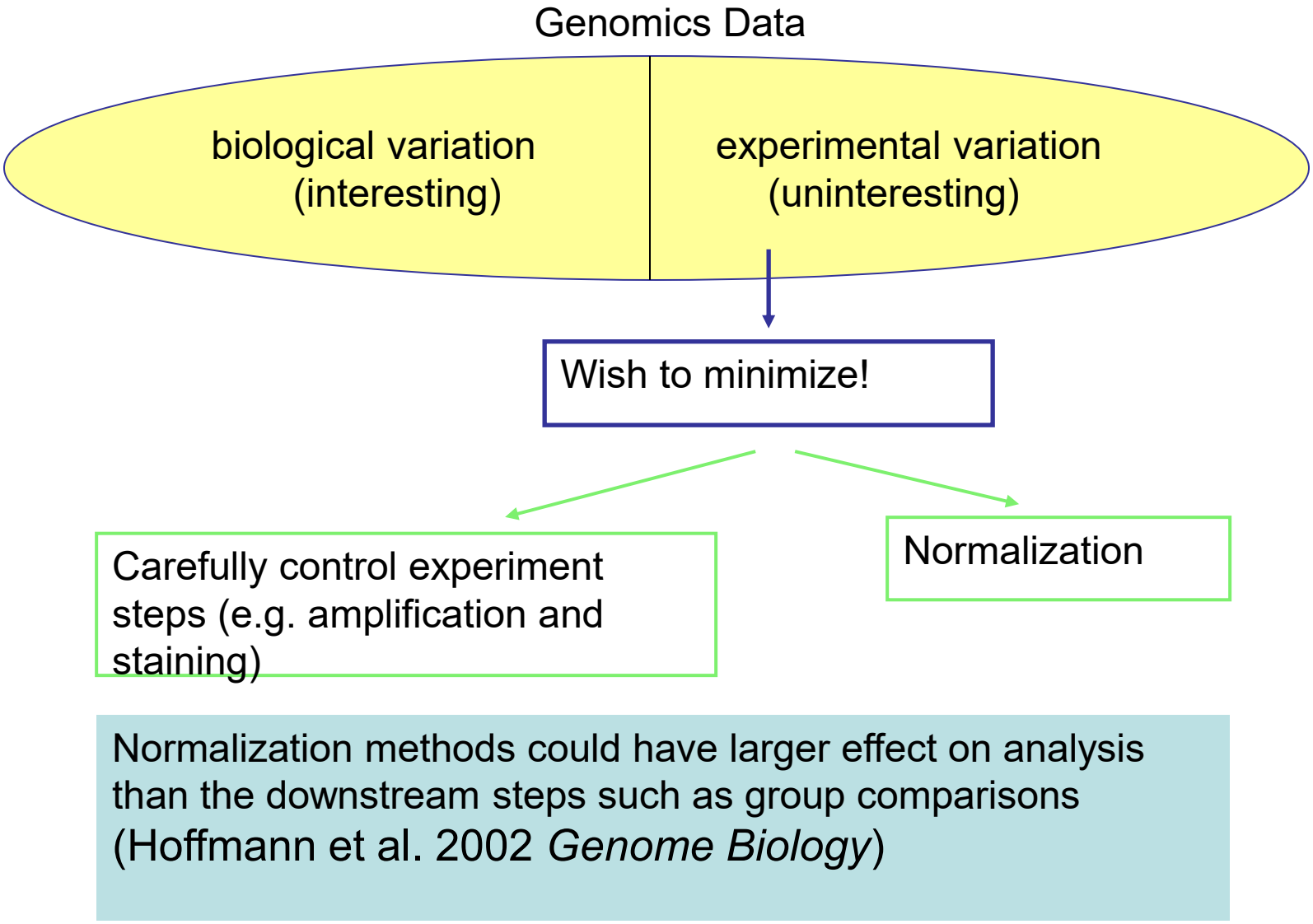
https://blog.csdn.net/zyf89531/article/details/45922151
https://blog.csdn.net/pipisorry/article/details/52247379

Normalization: Arrays may have different brightness





Normalization: Motivation



Normalization methods could have larger effect on analysis than the downstream steps such as group comparisons (Hoffmann et al. 2002 *Genome Biology*)

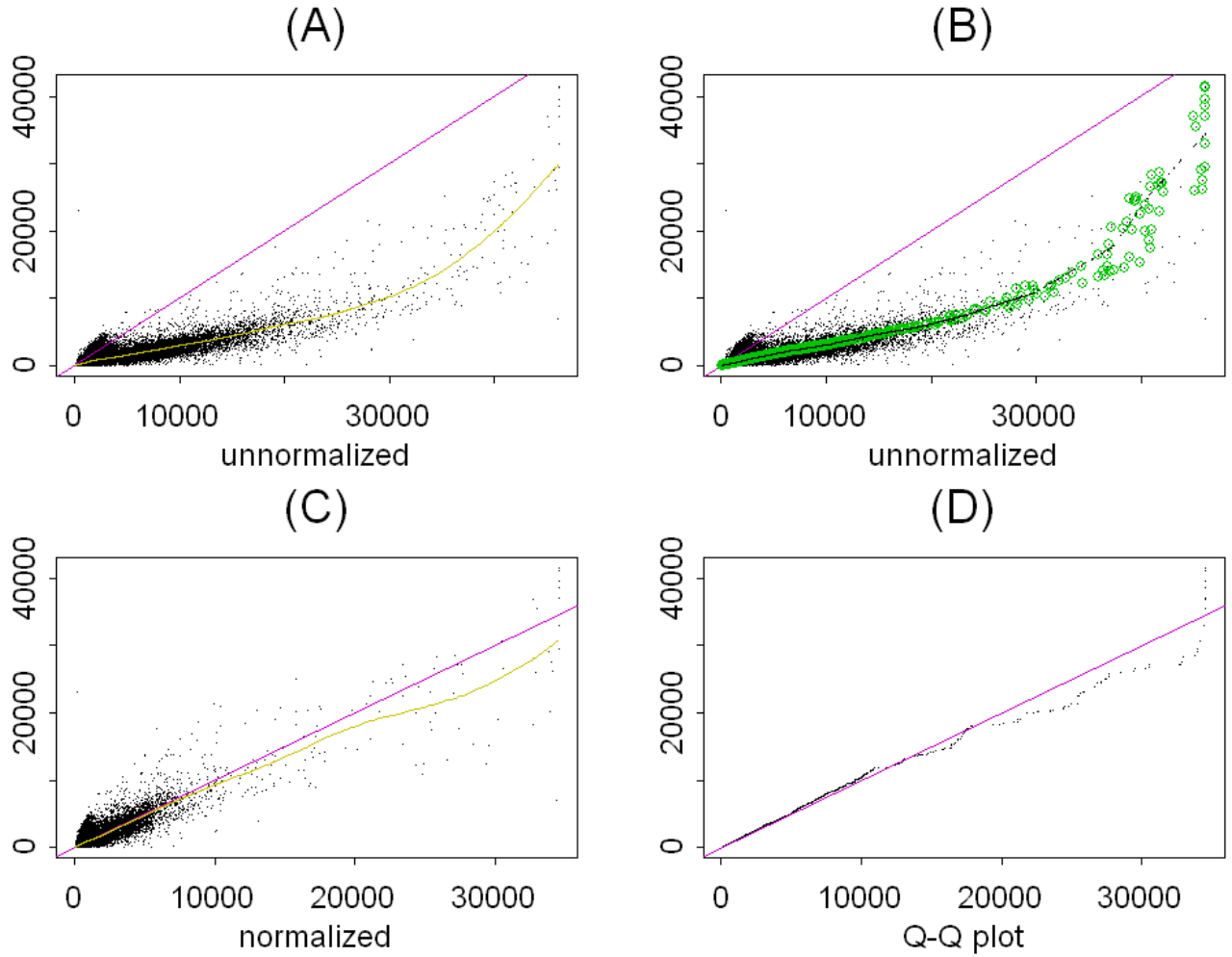
Normalization: Nonlinear methods, “Invariant Set”

Li and Wong 2001b, Tseng et al. 2001,

Schadt et al. 2001, Stuart et al. 2001

- It’s not desired that normalization eliminates the real changes of differentially expressed genes
- Select a subset of probes or genes that are presumably non-differentially expressed as the basis for normalization
- Use absolute rank differences as iterative heuristics.
- Use nonlinear curve fitting to determine the normalization relation.

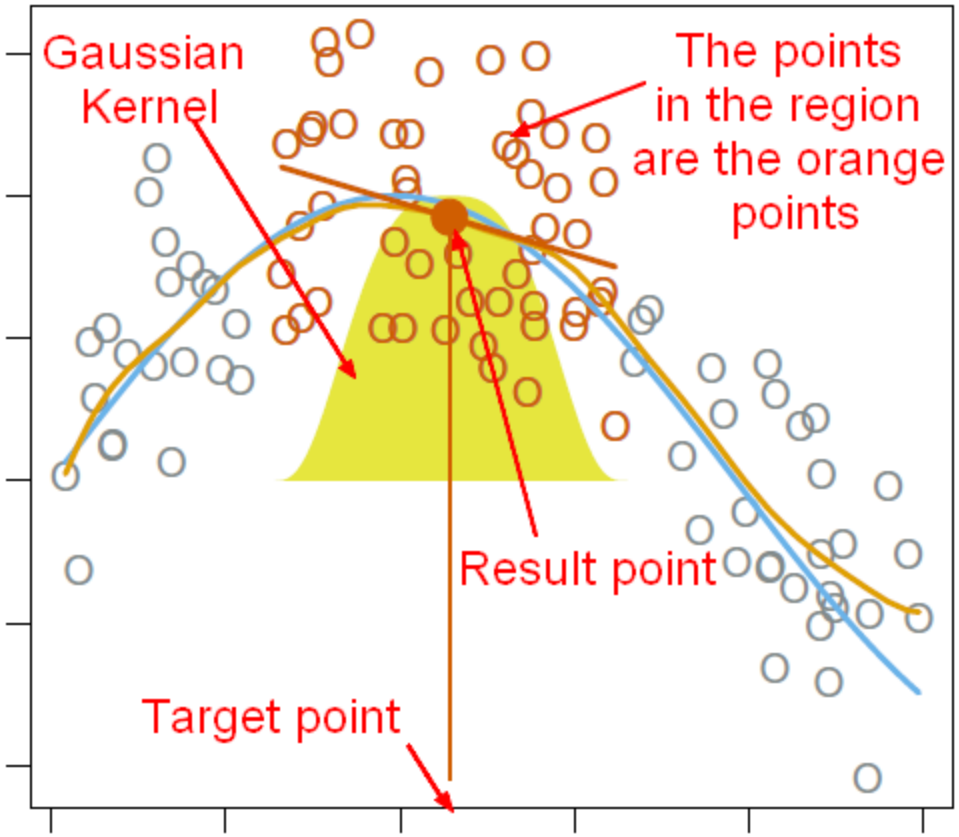
Normalization: Nonlinear methods, “Invariant Set”



Two different samples. The smoothing spline in (A) is affected by several points at the lower-right corner, which might belong to differentially expressed genes. Whereas the “invariant set” does not include these points when determining normalization curve, leading to a different normalization relationship at the high end.

Genome Biology 2001 Model-based analysis of oligonucleotide arrays model validation design issues and standard error application

LOWESS (locally weighted scatterplot smoothing)



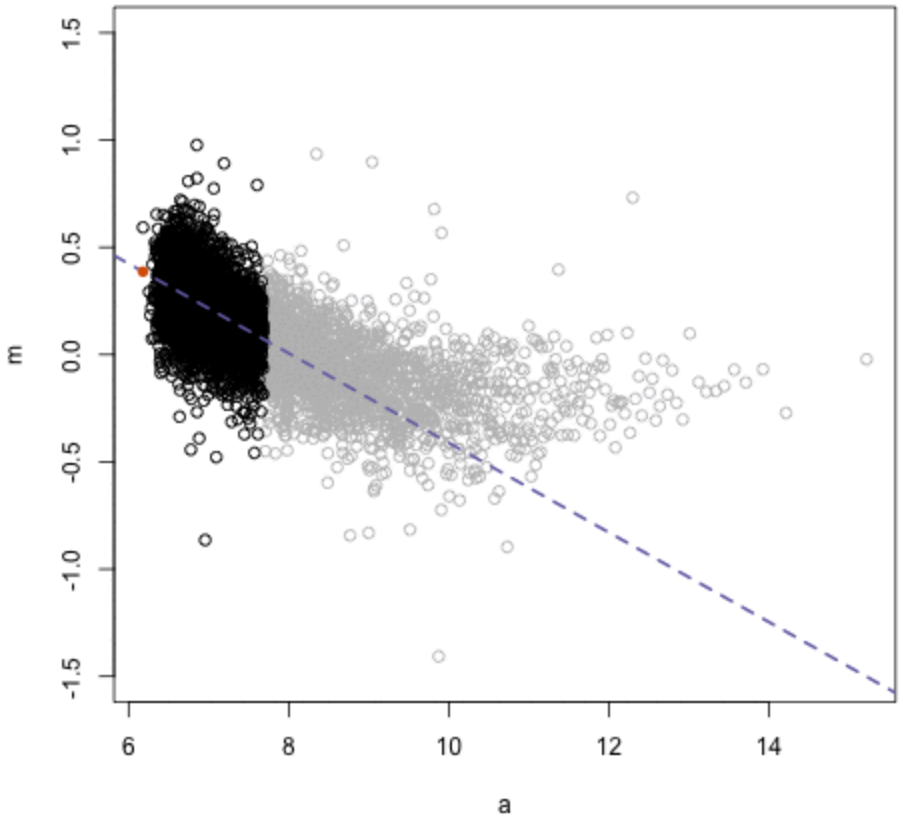
A linear function is fitted only on a local set of point delimited by a region. The polynomial is fitted using weighted least squares. The weights are given by the heights of the kernel (the weighting function).

We obtain then a fitted polynomial model but retains only the point of the model at the target point (x). The target point then moves away on the x axis and the procedure repeats and that traces out the orange curve.

https://gerardnico.com/data_mining/local_regression

R function **lowess()**: https://www.rdocumentation.org/packages/stats/versions/3.6.1/topics/lowess

Loess explained in a GIF

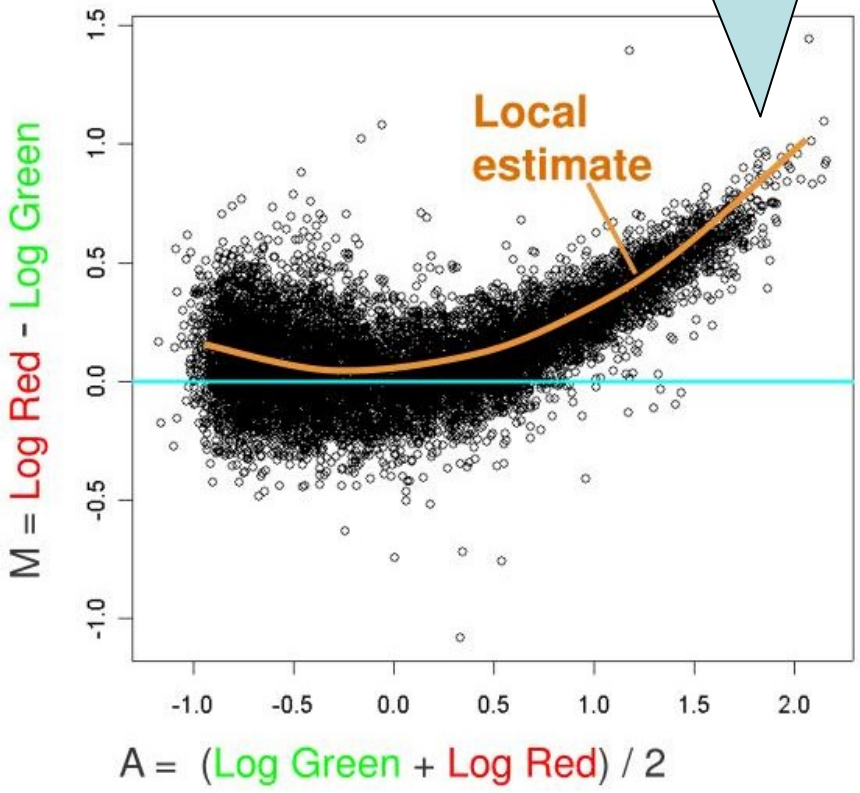


Rafael Irizarry: “Local regression (loess) is one of the statistical procedures I most use.”

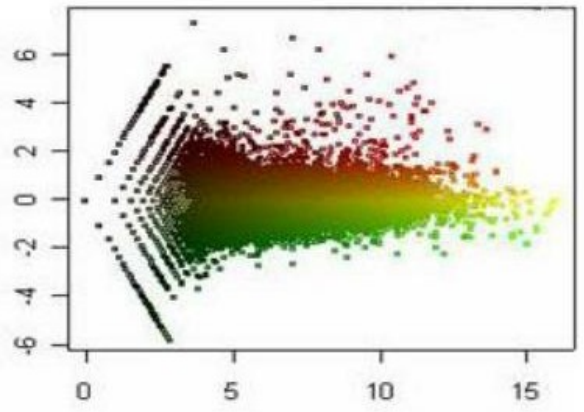
https://simplystatistics.org/2014/02/13/loess-explained-in-a-gif/

Lowess normalization (M-A plot)

系统误差：表达量高的基因大都上调?

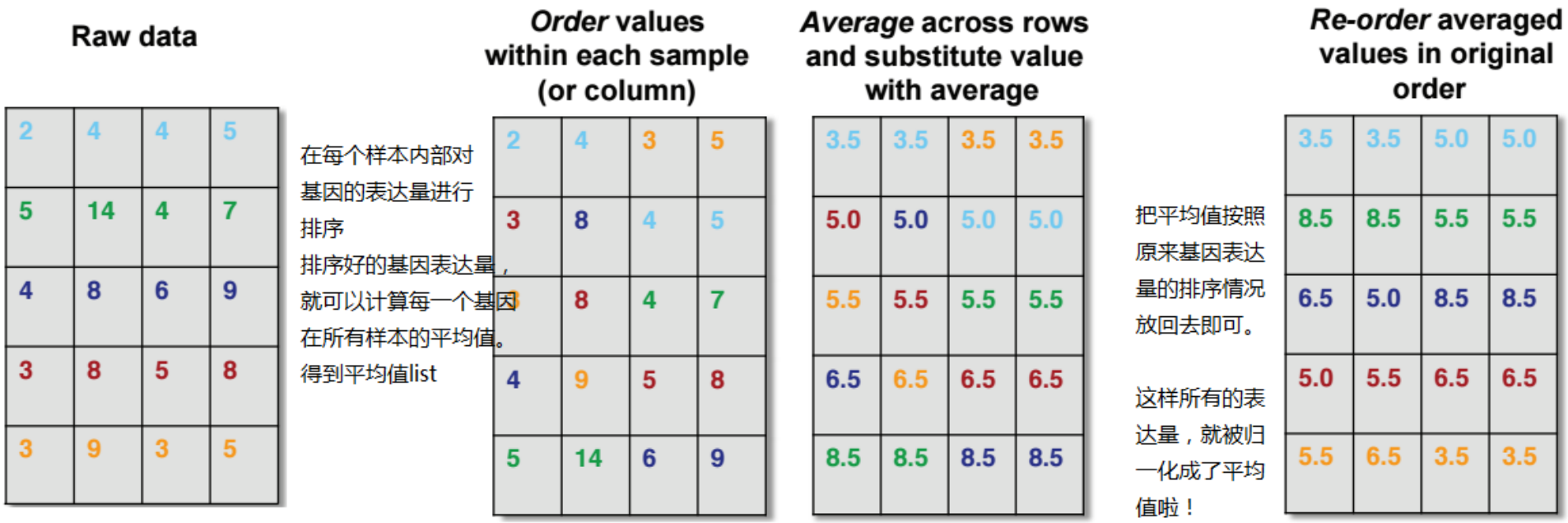


Use the estimate to bend **the banana straight**



https://www.slideserve.com/briana/microarray-analysis

Quantile normalization

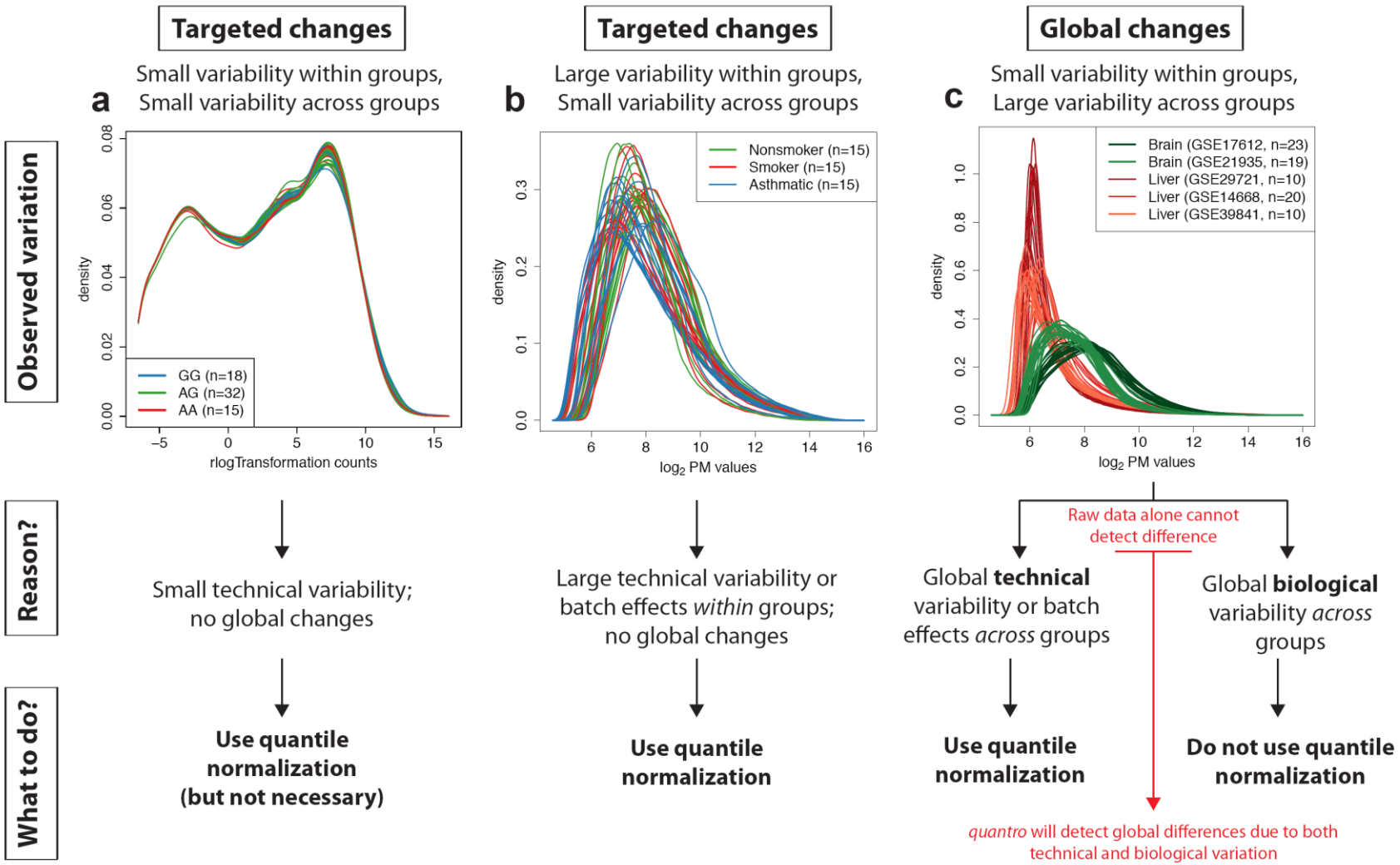


表达量矩阵的每一列是一个样本，每一行是一个基因（不同颜色），值是表达量

假设：每个样本中所有基因的表达量分布大致相同

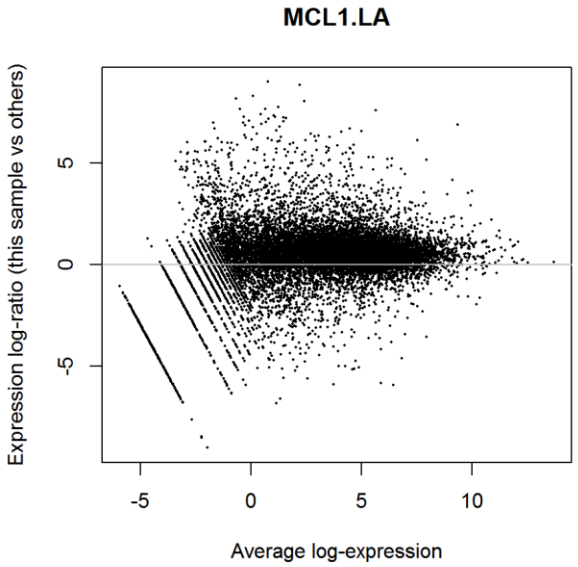
https://vip.biotrainee.com/d/81-quantile-normalization （生信技能树）

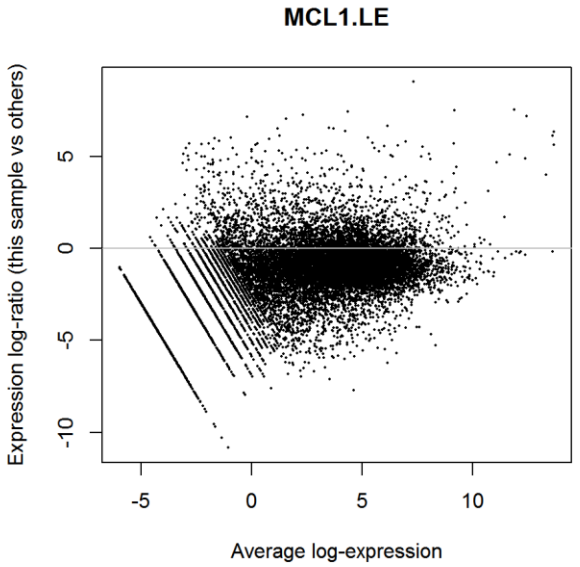
When to use Quantile Normalization



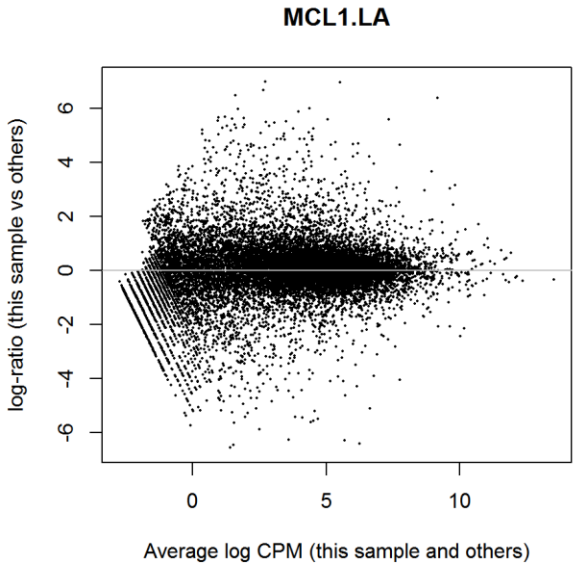
RNA-seq analysis with R: Normalization

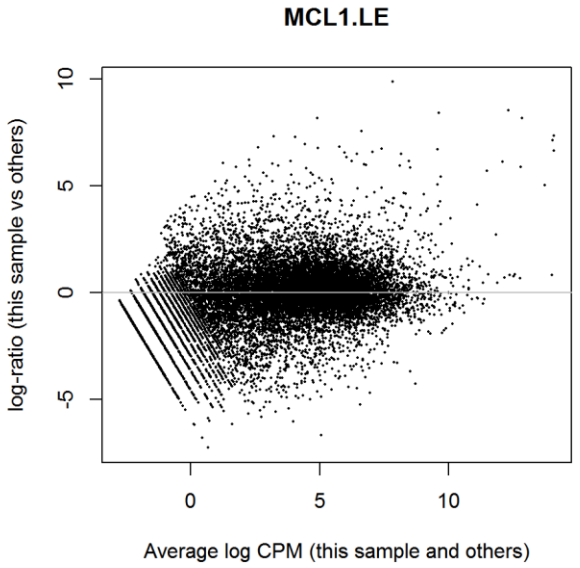
Before normaliation



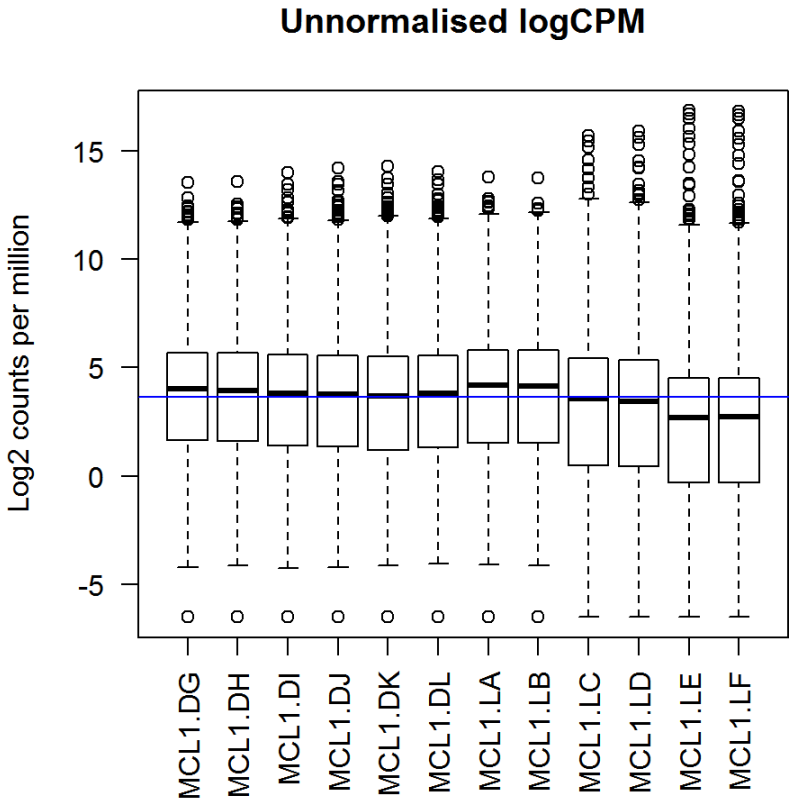


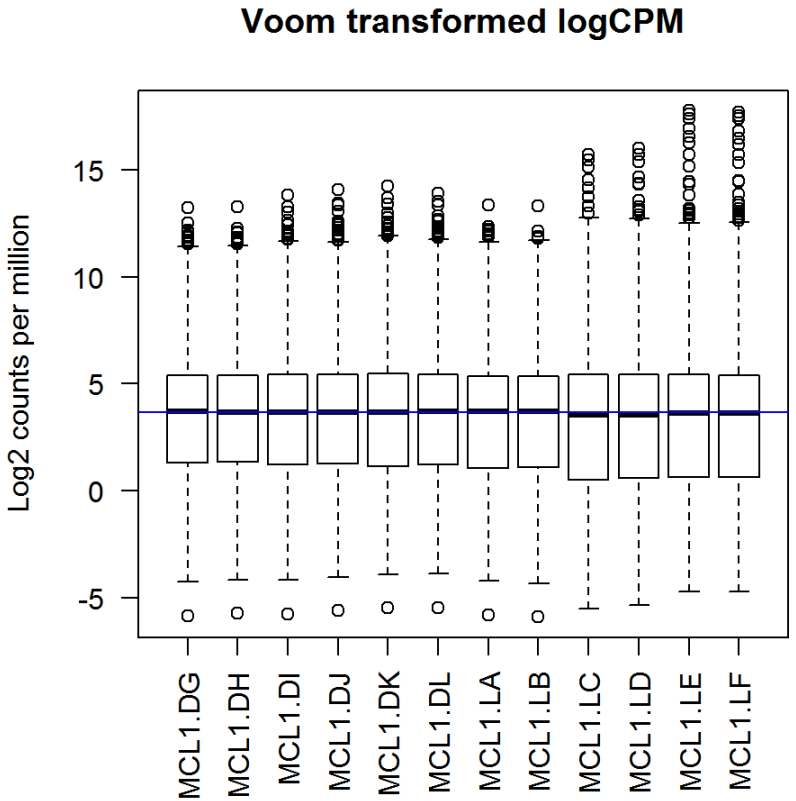
After normaliation





RNA-seq analysis with R: Normalization





本节内容

- 高通量测序技术
- 测序原始数据
- RNA-seq分析流程
- 统计**模型**：数据归一化
- **演示：**Linux、云计算、分析资源

Linux 常用命令

1、ls命令

就是 list 的缩写，通过 ls 命令不仅可以查看 linux 文件夹包含的文件，而且可以查看文件权限(包括目录、文件夹、文件权限)□查看目录信息等等。

常用参数搭配：

```
ls -a 列出目录所有文件，包含以.开始的隐藏文件
ls -A 列出除.及..的其它文件
ls -r 反序排列
ls -t 以文件修改时间排序
ls -S 以文件大小排序
ls -h 以易读大小显示
ls -l 除了文件名之外，还将文件的权限、所有者、文件大小等信息详细列出来
```

https://www.runoob.com/w3cnote/linux-common-command-2.html

Linux中修改环境变量

- **直接运行export命令定义变量【只对当前shell（BASH）有效（临时的）】**

- 在shell的命令行下直接使用[export 变量名=变量值]定义变量，该变量只在当前的shell（BASH）或其子shell（BASH）下是有效的，**shell关闭了，变量也就失效了，再打开新shell时就没有这个变量，需要使用的话还需要重新定义。**

- 例如：export PATH=/usr/local/webserver/php/bin:$PATH

- **环境变量的查看**

  (1)使用echo命令查看单个环境变量。例如:

  ```
  echo $PATH
  ```

  (2)使用env查看所有环境变量。例如:

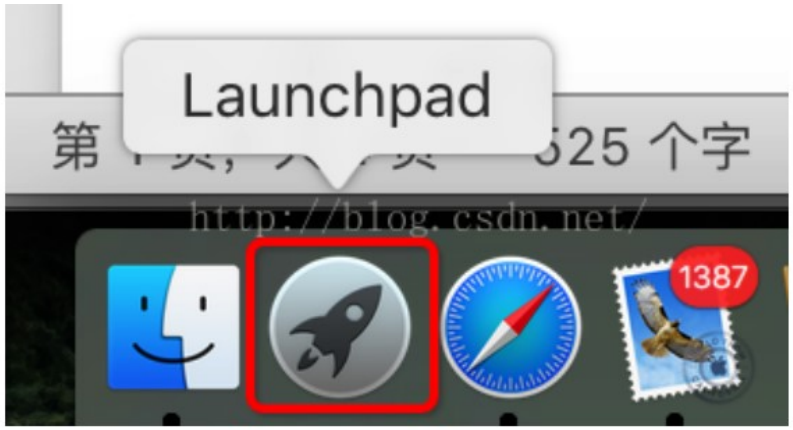
  ```
  env echo
  ```

  (3)使用set查看所有本地定义的环境变量。例如:

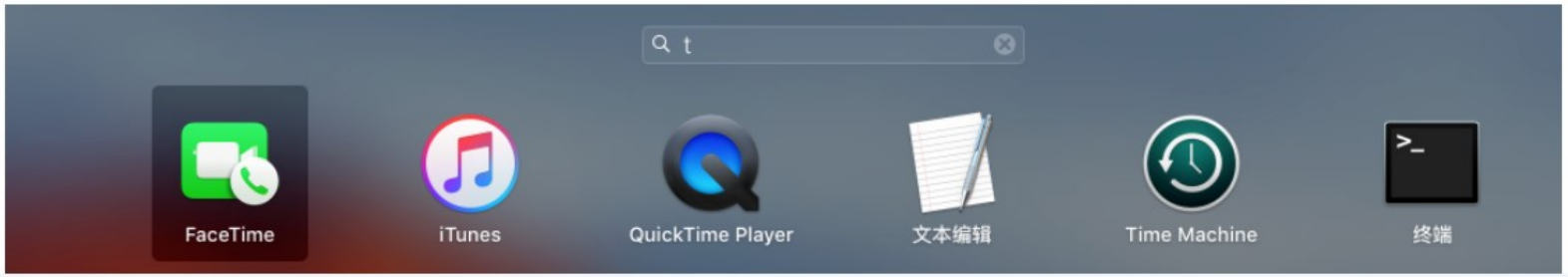
  ```
  set
  ```

https://blog.csdn.net/u011630575/article/details/49839893

Mac下如何调出命令行窗口

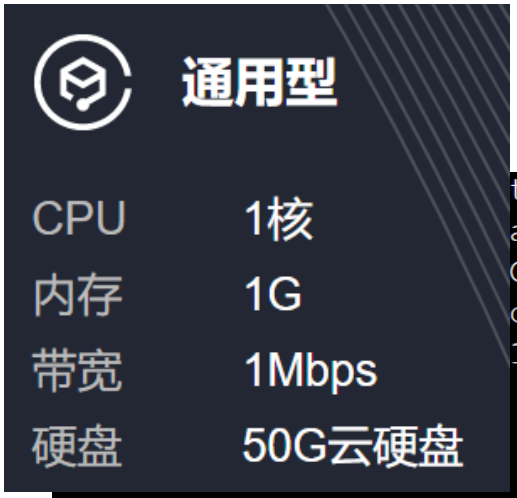


调出的界面搜索terminal



常用命令：https://www.jianshu.com/p/fdddcdf56e06

云计算



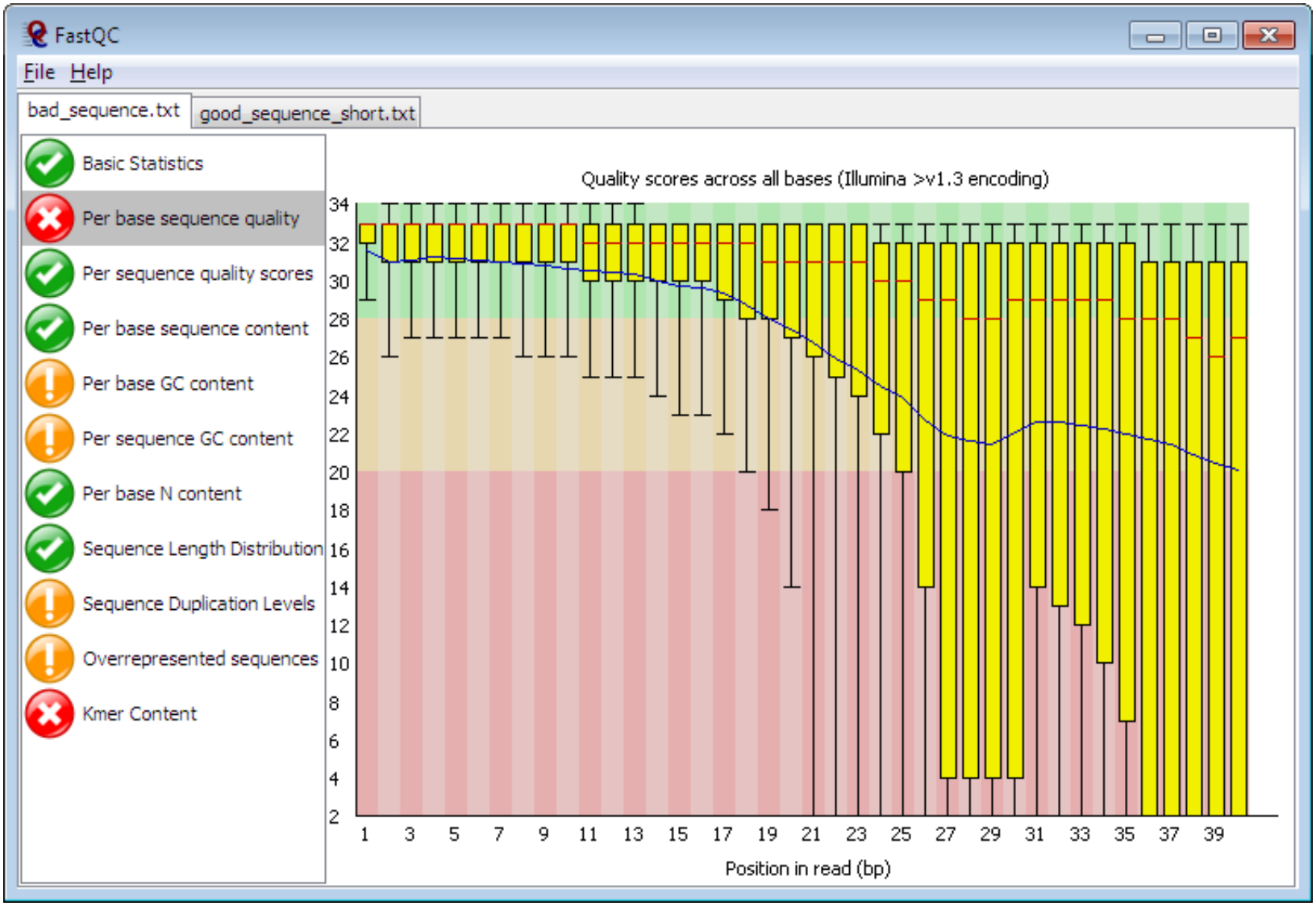

```
tie2-2.4.2-linux-x86_64]# * Connection closed *
ablished *
Oct 20 14:14:00 CST 2020 from 62.219.167.67 on ssh:notty
ogin attempts since the last successful login.
11:49:37 2020 from 111.230.154.177
 ls

 cd cheng
[root@VM-0-7-centos cheng]# ls
bowtie2-2.4.2-linux-x86_64  bowtie2-2.4.2-linux-x86_64.zip
[root@VM-0-7-centos cheng]# cd bowtie2-2.4.2-linux-x86_64/
[root@VM-0-7-centos bowtie2-2.4.2-linux-x86_64]# ls
AUTHORS                 bowtie2-build-s          eg1.sam                 MANUAL
bowtie2                 bowtie2-build-s-debug    example                 MANUAL.markdown
bowtie2-align-l         bowtie2-inspect          lambda_virus.1.bt2      NEWS
bowtie2-align-l-debug   bowtie2-inspect-l        lambda_virus.2.bt2      README.md
bowtie2-align-s         bowtie2-inspect-l-debug  lambda_virus.3.bt2      scripts
bowtie2-align-s-debug   bowtie2-inspect-s        lambda_virus.4.bt2      TUTORIAL
bowtie2-build           bowtie2-inspect-s-debug  lambda_virus.rev.1.bt2
bowtie2-build-l         BOWTIE2_VERSION          lambda_virus.rev.2.bt2
bowtie2-build-l-debug   doc                      LICENSE
[root@VM-0-7-centos bowtie2-2.4.2-linux-x86_64]#
```

腾讯云试用：https://cloud.tencent.com/act/free

腾讯云的“Linux基础入门”在线实验课

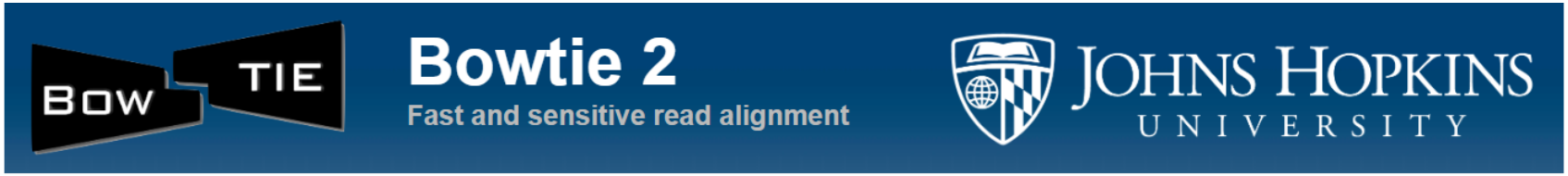
https://cloud.tencent.com/developer/labs/gallery

测序数据质控：FastQC



https://www.bioinformatics.babraham.ac.uk/projects/fastqc/

序列比对软件： Bowtie2



bowtie2-2.4.2-sra-linux-x86_64.zip

bowtie2-2.4.2-linux-aarch64.zip

bowtie2-2.4.2-source.zip

bowtie2-2.4.2-macos-x86_64.zip

bowtie2-2.4.2-linux-x86_64.zip ⟸ uname -a 命令可以直接显示 Linux 系统架构的命令

Site Map

Home
News archive
Manual
Getting started
Frequently Asked Questions
Tools that use Bowtie

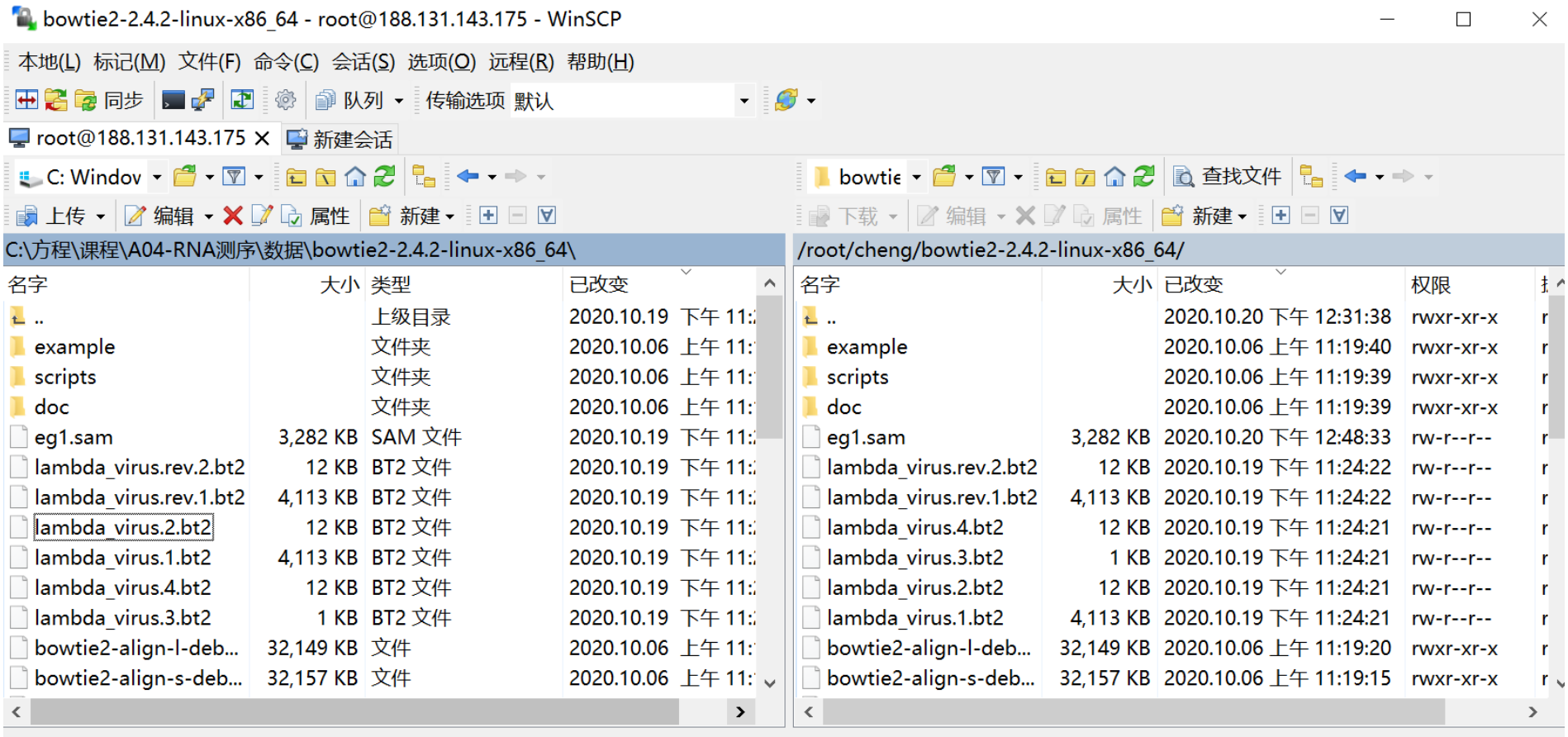
Latest Release

install with bioconda

Bowtie2 v2.4.2 10/05/20

Please cite: Langmead B, Salzberg S. Fast gapped-read alignment with Bowtie 2. *Nature Methods*. 2012, 9:357-359.

在本地和云服务器之间传输文件



https://cloud.tencent.com/document/product/213/2131

Bowtie 2: Lambda phage example

Indexing a reference genome

To create an index for the Lambda phage reference genome included with Bowtie 2, create a new temporary directory (it doesn't matter where), change into that directory, and run:

```
$BT2_HOME/bowtie2-build $BT2_HOME/example/reference/lambda_virus.fa lambda_virus
```

Aligning example reads

Stay in the directory created in the previous step, which now contains the `lambda_virus` index files. Next, run:

```
$BT2_HOME/bowtie2 -x lambda_virus -U $BT2_HOME/example/reads/reads_1.fq -S eg1.sam
```

Using SAMtools/BCFtools downstream

Use `samtools view` to convert the SAM file into a BAM file. BAM is the binary format corresponding to the SAM text format. Run:

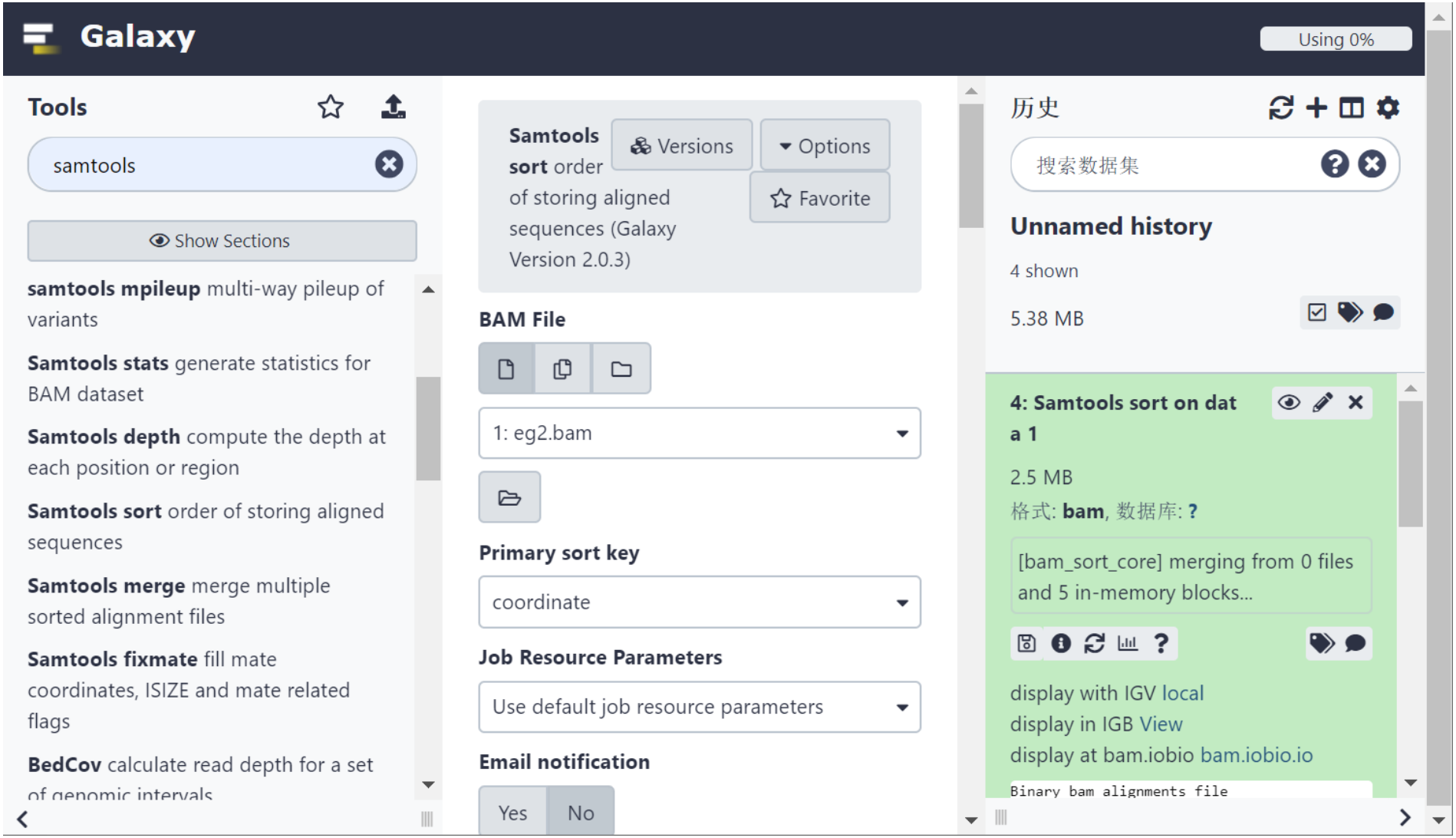
```
samtools view -bS eg2.sam > eg2.bam
```

Use `samtools sort` to convert the BAM file to a sorted BAM file.

```
samtools sort eg2.bam -o eg2.sorted.bam
```

http://bowtie-bio.sourceforge.net/bowtie2/manual.shtml#getting-started-with-bowtie-2-lambda-phage-example

Galaxy生信分析网站



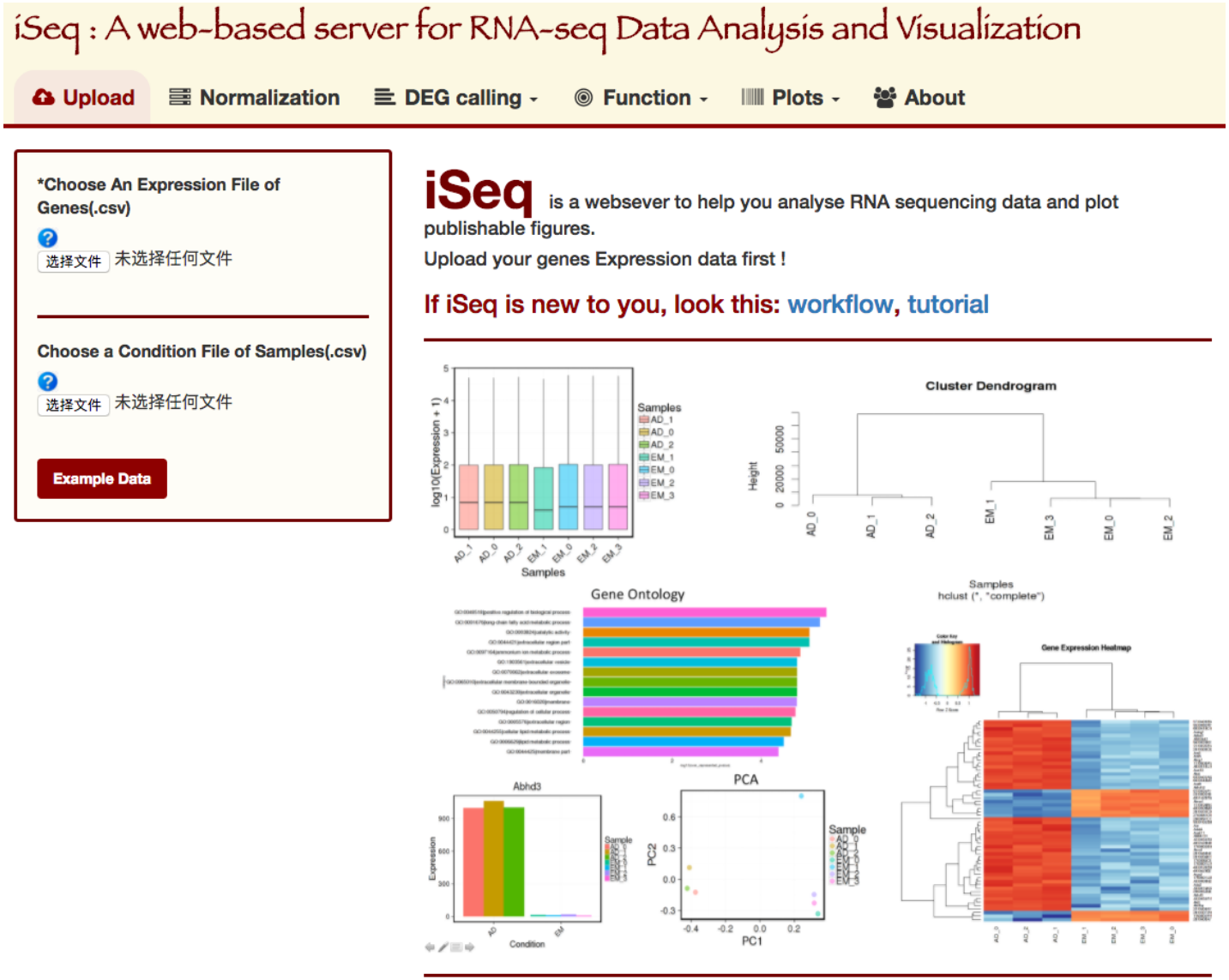
https://usegalaxy.org/

IGV浏览器

http://software.broadinstitute.org/software/igv/

帮助页：https://software.broadinstitute.org/software/igv/AlignmentData

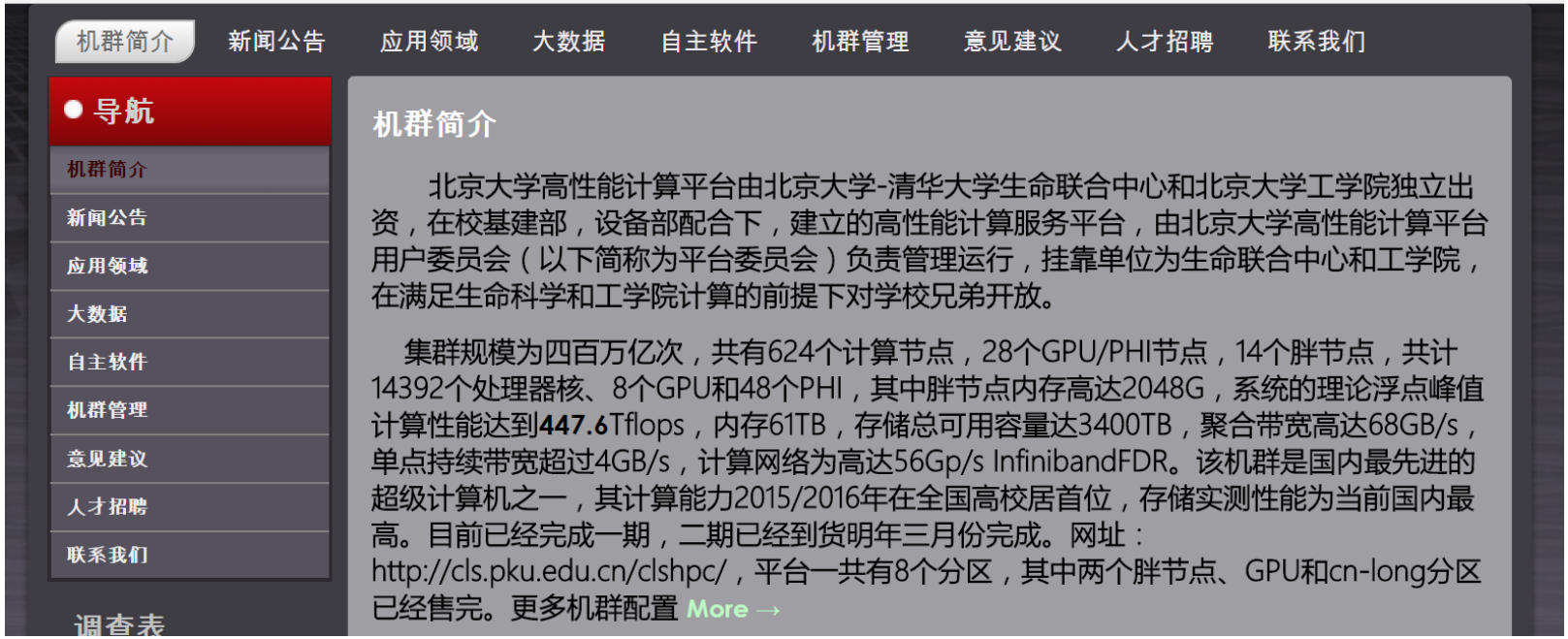
iSeq : A web-based server for RNA-seq Data Analysis and Visualization



http://iseq.cbi.pku.edu.cn

北京大学“北极星”高性能计算平台

http://cls.pku.edu.cn:8080/clshpc/



陈芳进、裴剑锋老师《定量生物学计算与软件使用》课程

- 交叉学院/秋季/3学分，研究生课程号08402070
- 陈芳进老师，20年9月: “定量生物学计算与软件使用”课程是一门理论与实践结合的课程。课程讲授的主要内容为Matlab基础和使用、Linux基础和使用、Python基础和使用、生物网络分析、机器学习、深度学习、以及其他定量生物学计算方法和资源等。其中linux基础和使用包括Linux shell，文件传输，GPU加速，任务并行，北极星集群使用和任务提交。上课地点：老化学楼东配楼101教室，时间为：星期二7-9节课，9月22日开始。需要设备：自带笔记本。注：本学期使用新的教学集群，环境和北极星一致，登录不需要vpn和堡垒机，为密钥登录。今年不再使用虚拟机。

课后练习

1. 阅读文献

   《NCB15 EGF-mediated induction of Mcl-1 at the switch to lactation is essential for alveolar cell survival》

   思考作者为什么做RNA-seq实验（图6），它的数据分析结果对课题的推进和结论有哪些帮助?

2. 数据分析练习

   a) 安装和运行“演示”部分的软件和流程

   遇到安装、运行错误，可以先上网搜索，但如果花费超过半小时，及时问助教或同学

   b) 结合课堂内容和《RNAseq-R教程》（http://combine-australia.github.io/RNAseq-R/），运行课件中的R代码，理解所用函数的功能和输出结果。比较得到的结果和（1）中文献分析结果的异同。

延伸阅读

- 用R做RNA-seq分析的其他教程
  - https://bioinformatics-core-shared-training.github.io/RNAseq-R/
  - http://bioconductor.org/help/workflows/rnaseqGene/
  - http://biocluster.ucr.edu/~rkaundal/workshops/R_mar2016/RNAseq.html

- 课件目录《文献阅读》中选择感兴趣的文献阅读。

RNA-seq初学者的感悟

陈霞（生信平台实习同学）

RNA-seq背景学习

首先，对于学习流程，个人偏向先看懂背景，再进行数据分析，所以在此推荐一篇最新的关于RNA-seq的review，在这里你可以了解到关于RNA-seq的全面的应用的知识，以及在RNA-seq的分析过程中会遇到的一些分析策略选择的建议。链接：http://www.ncbi.nlm.nih.gov/pubmed/26813401（此篇因为讲得范围广，涉及的点多，初学者就会有很多疑惑）

RNA-seq分析练习

在上述提到的综述中，会看到关于Cufflinks和Tophat两个软件在RNA-seq分析中的应用。并且有篇关于使用它们进行RNA-seq的数据分析的指导性文献（链接：http://www.ncbi.nlm.nih.gov/pubmed/22383036）接下来，可以利用这篇文献进行实战。

这时需要有基本的linux的操作知识，在此分享从生物秀上下载的生物信息学的教程 http://wenku.baidu.com/view/266f8ee658fb770bf78a55e6，教程很长先看懂指令可以满足初期的软件的使用，后续的可以慢慢看，了解生物信息学的一些软件及资源。

实战过程中：一定要先理清分析的目的，这样使用软件会不易出错，并且多看这个软件的manuals：http://cole-trapnell-lab.github.io/cufflinks/manual/
对于该文献中涉及的数据可以在GEO网站上下载。

软件处理完原始数据，对于软件生成的数据，我用了一个笨办法就是在R中直接打开看数据的内部到底是什么。虽然比较费时，但是理解软件的生成数据和进行下一步数据可视化是很有效的。若有更好的办法，也希望大家共享。

http://forum.cbi.pku.edu.cn/forum/upload/forum.php?mod=viewthread&tid=166