
第二课 转录组计算方法与数据分析

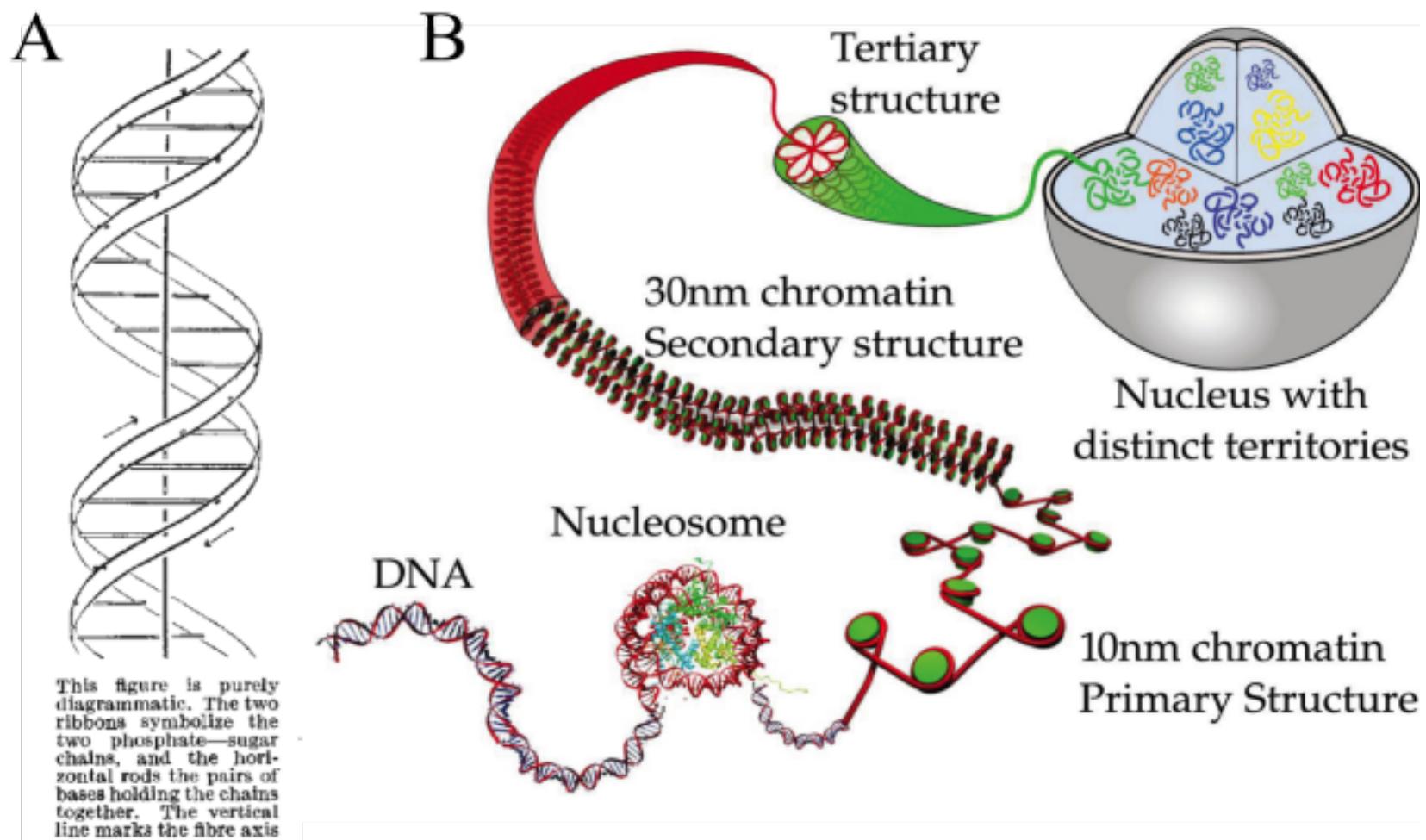
中国医学科学院基础医学研究所

陈阳

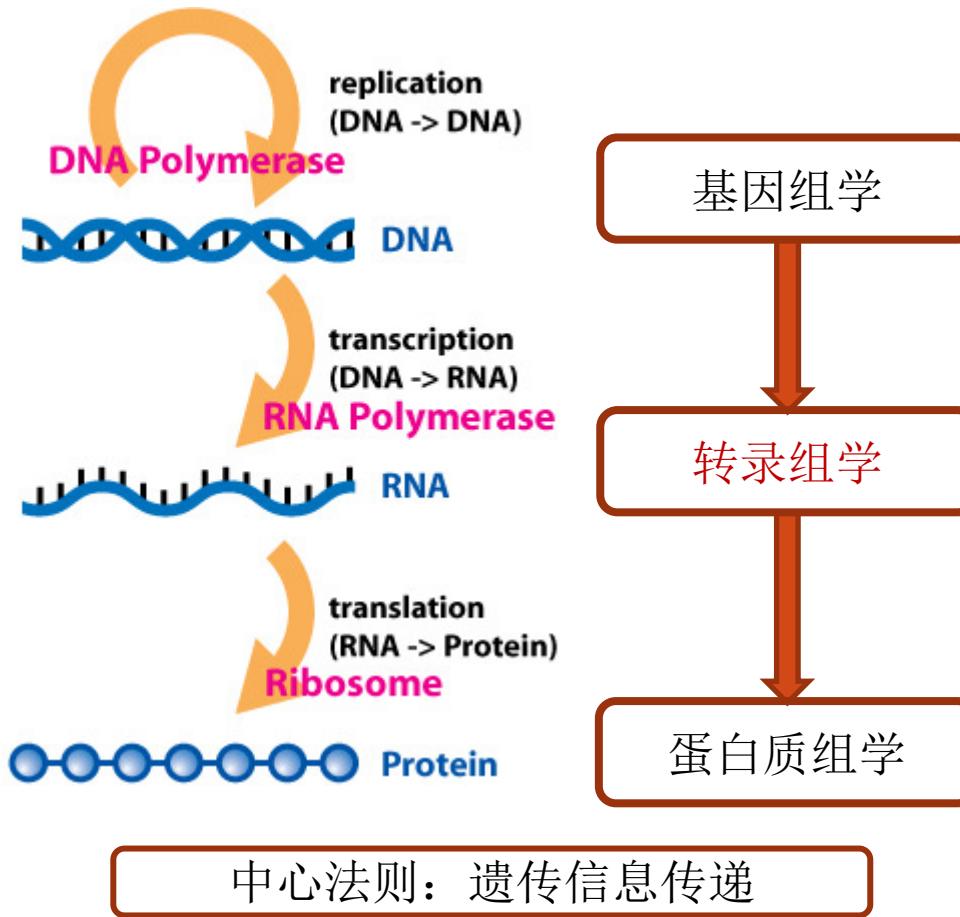
yc@ibms.pumc.edu.cn

20250327

基因组是生命遗传信息的载体



生物中心法则



生物中心法则

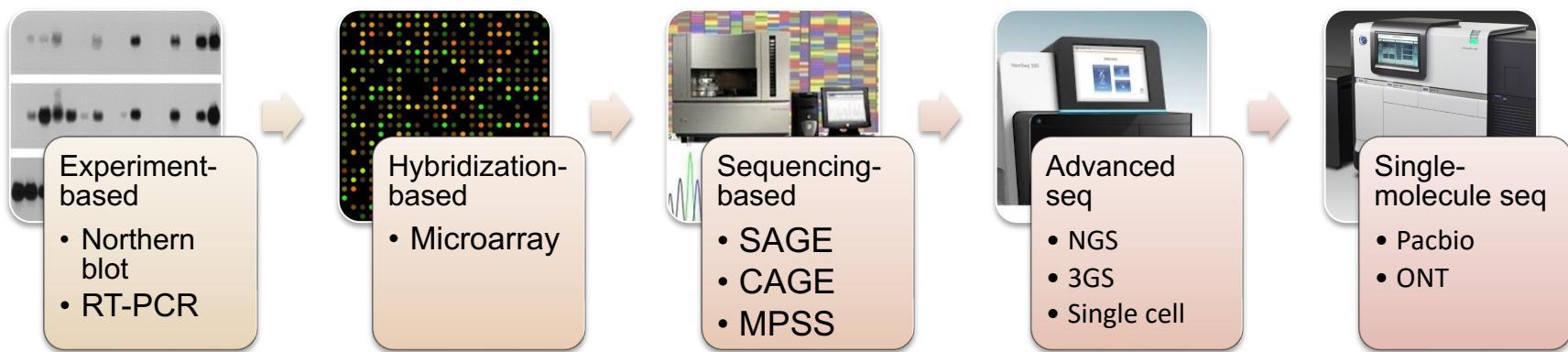
DNA
Learning
Center



Cold
Spring
Harbor
Laboratory

© Copyright, Cold Spring Harbor Laboratory. All rights reserved.

转录物测定技术发展历程

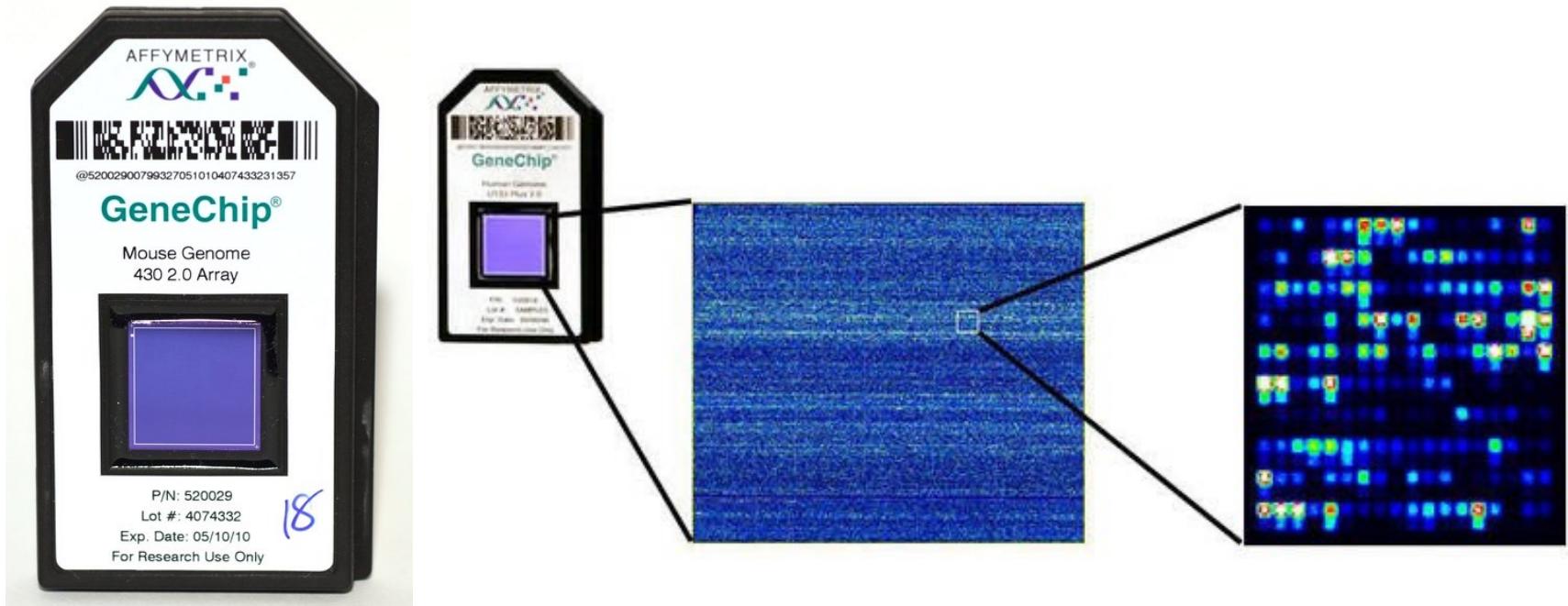


转录组学研究技术革新

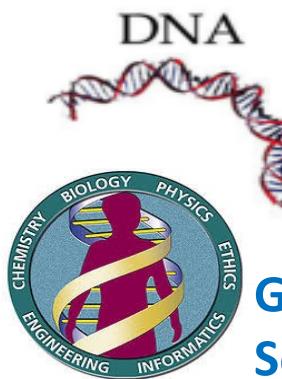
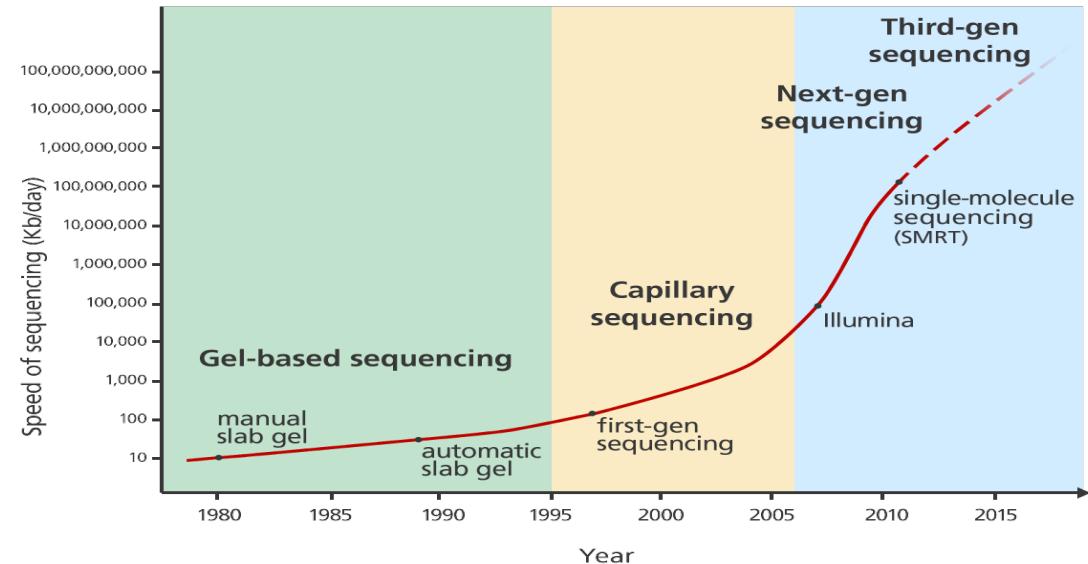
基于芯片的转录组技术

■ 发展历史

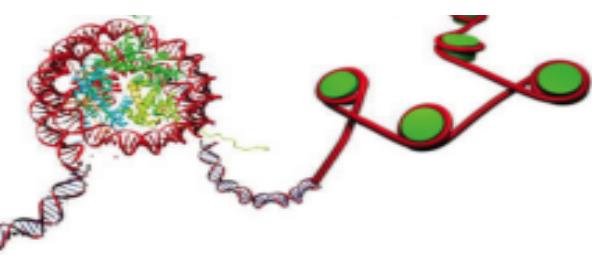
1990s~2006 芯片数据



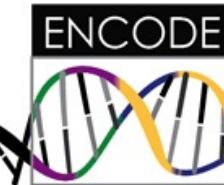
高通量测序技术加快了后基因组时代的研究



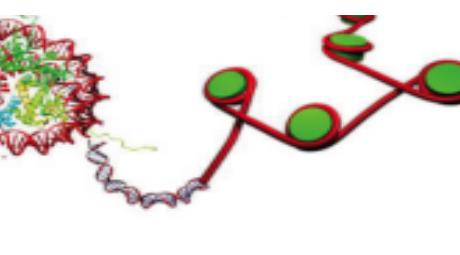
Genome 1990-2003
Scale: DNA molecule & sequence



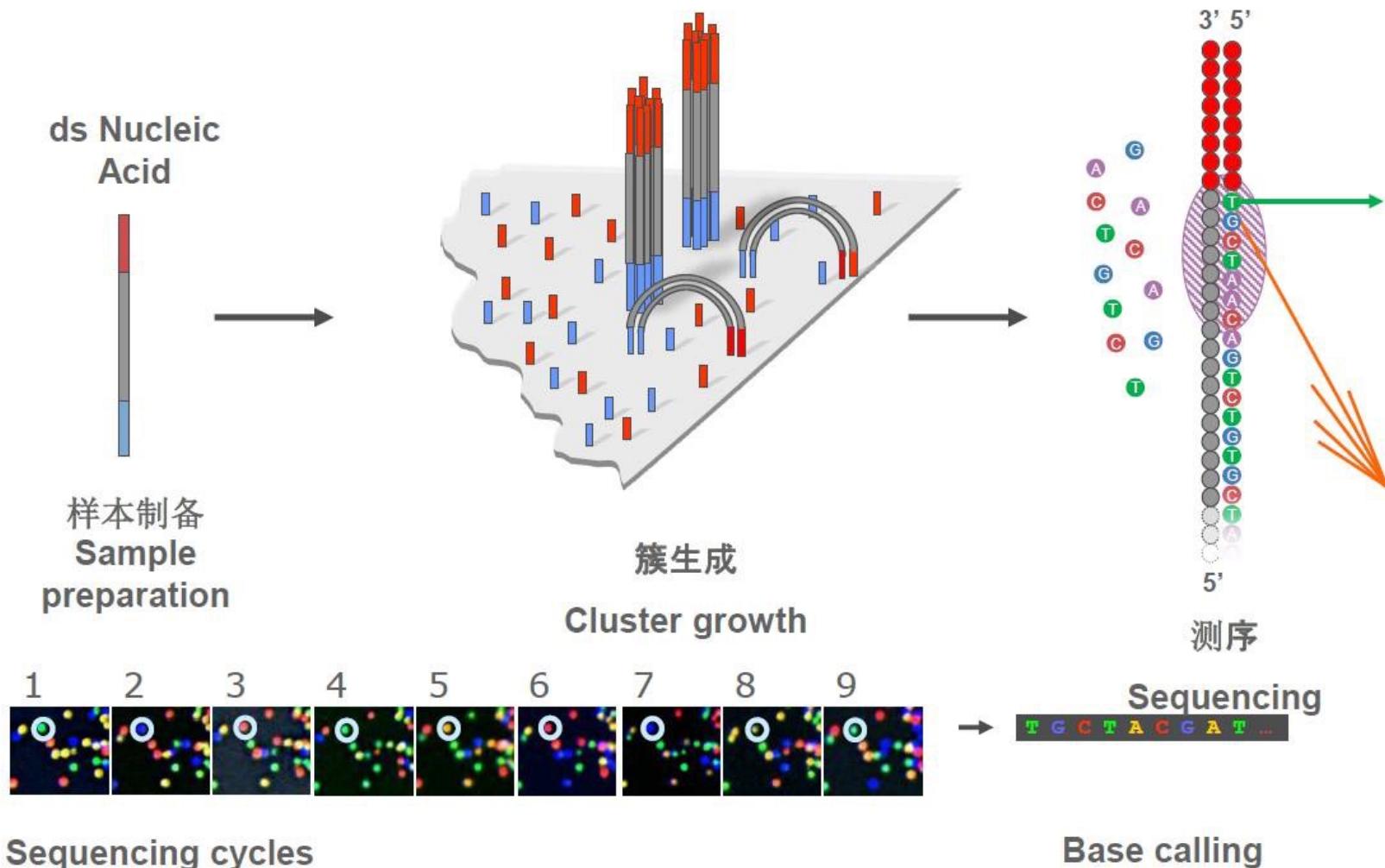
10nm chromatin Primary Structure



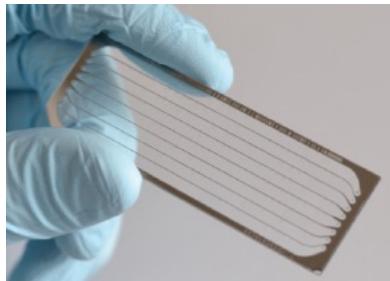
Epigenome 2005-2020
Scale: nucleosome & TF binding



高通量测序基本原理

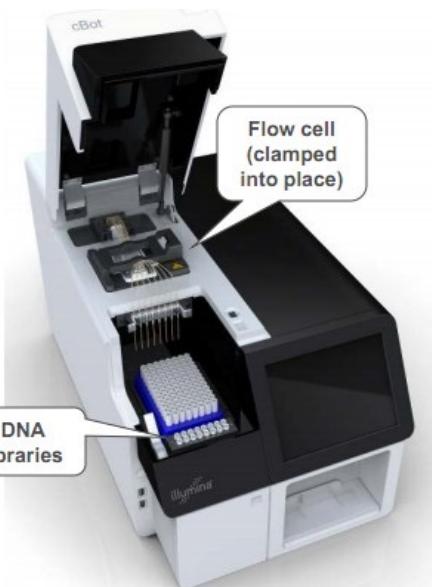
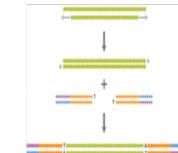


高通量测序基本原理



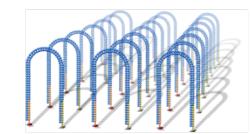
1

文库构建



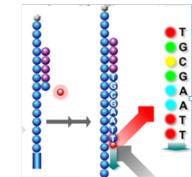
2

Cluster簇生成



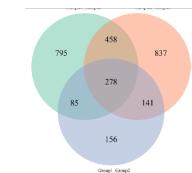
3

SBS测序

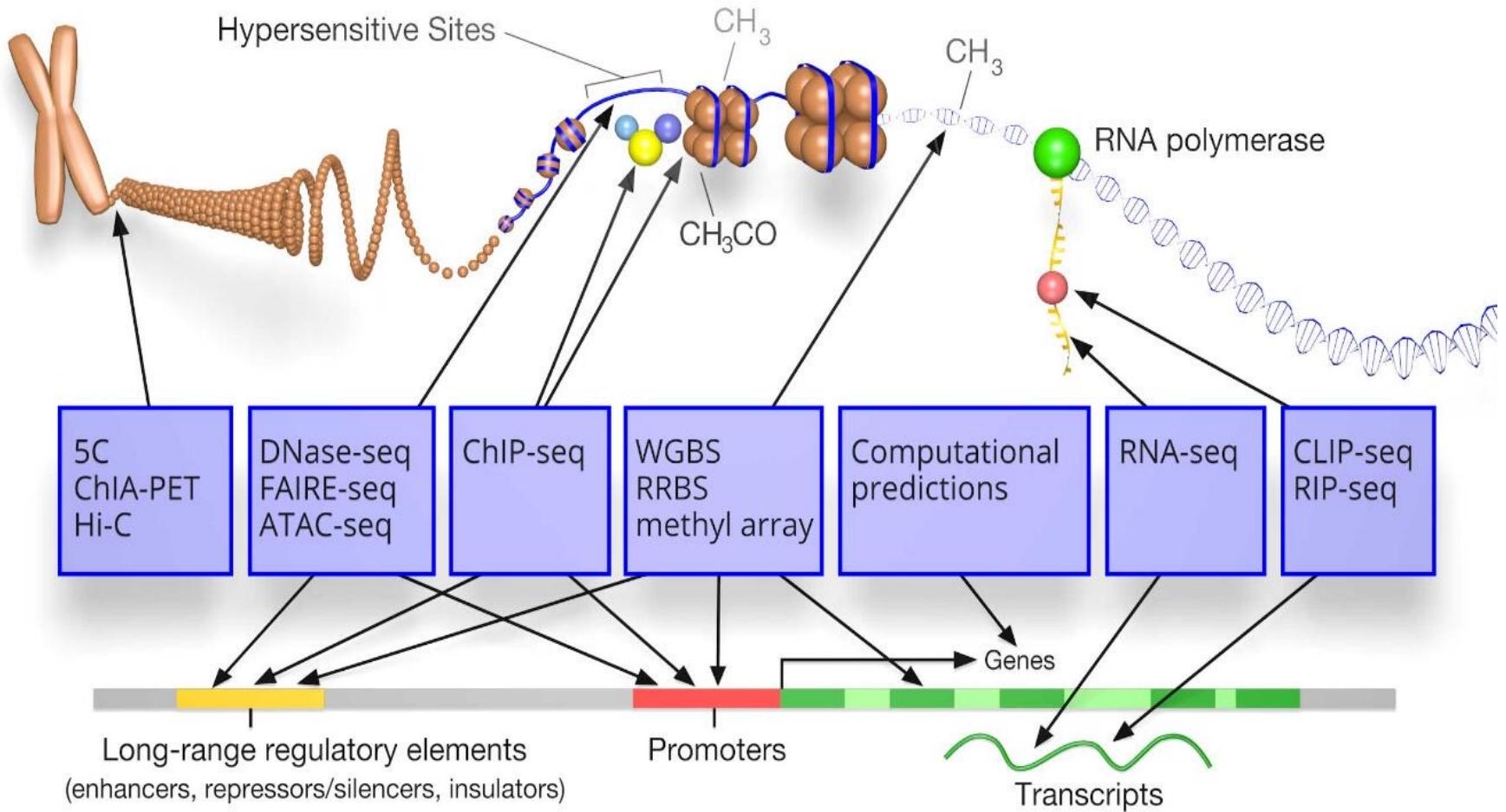


4

数据分析



基于高通量测序的组学技术

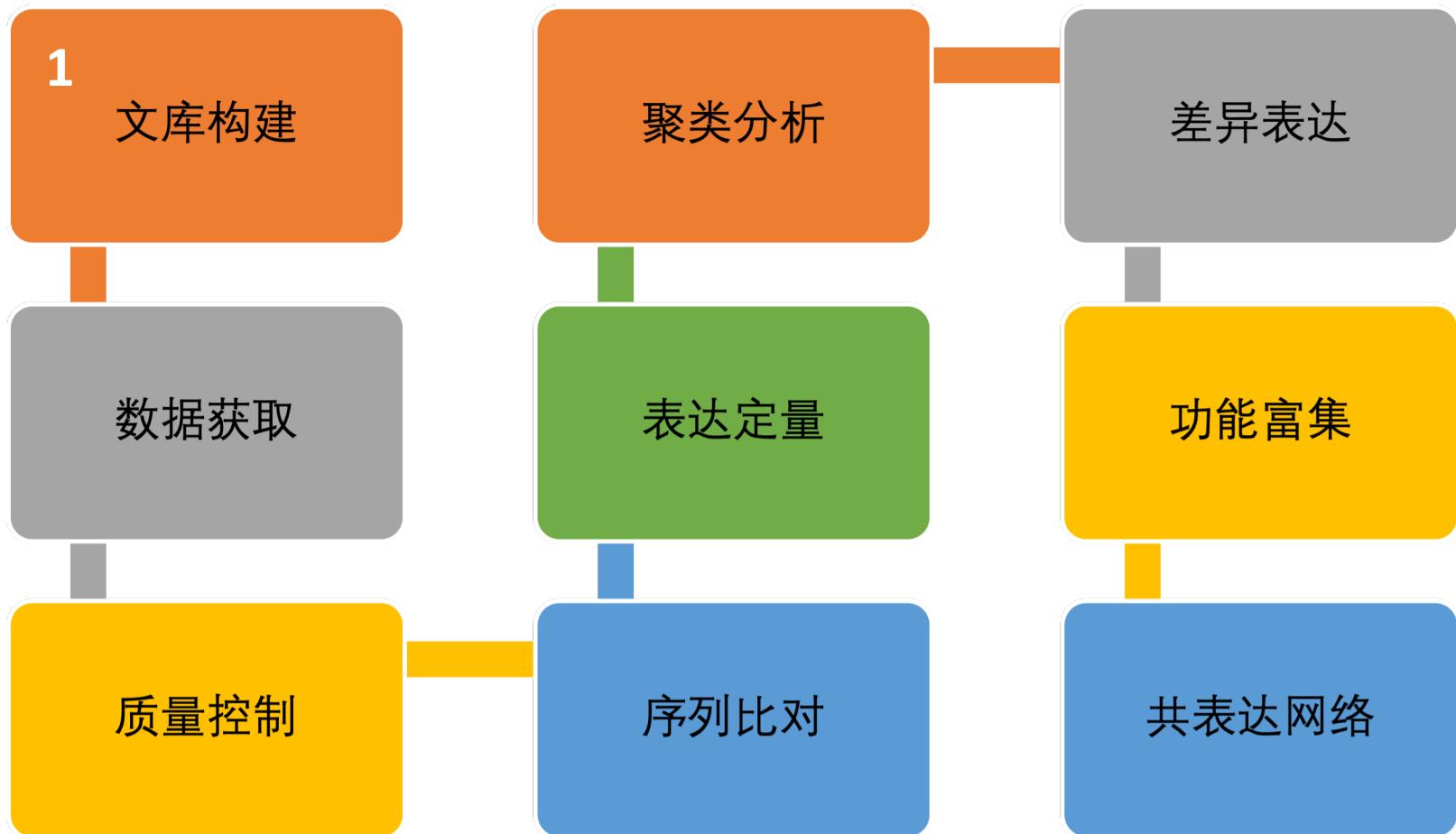


不同的RNA类型

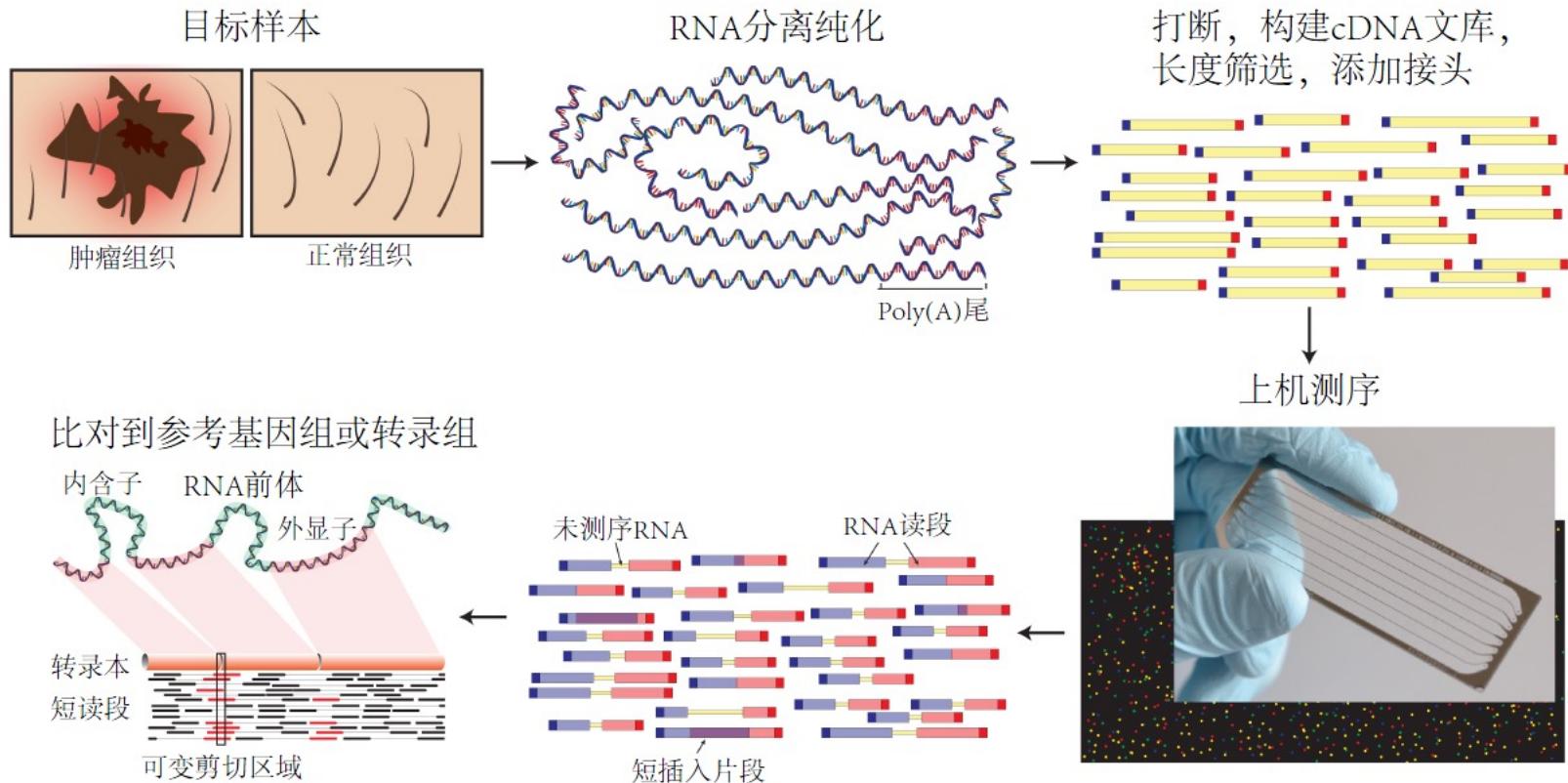
RNA Family	分类	名称	生物学作用
	Coding RNA	Messenger RNA , mRNA	编码蛋白质
		Long noncoding RNA , lncRNA	调控功能
		Circular RNAs , circRNAs	染色质结构
	Noncoding RNA	Small RNA	DNA甲基化
		microRNA , miRNA	基因转录
		Small interference RNA , siRNA	mRNA翻译
		Piwi-protein interacting RNA , piRNA	RNA降解
		Ribosomal RNA , rRNA	形成核糖体

Total RNA [rRNA 80%
tRNA 15%
mRNA 1 ~ 5%
microRNA lncRNA circRNA.....]

转录组测序技术RNA-seq常规流程



RNA-seq文库构建流程



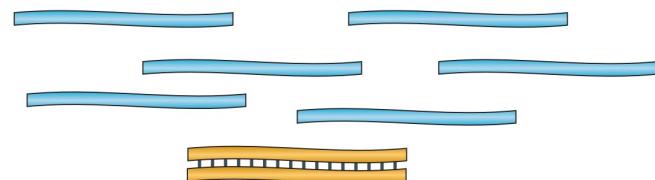
RNA-seq文库构建流程

■ 建库方法

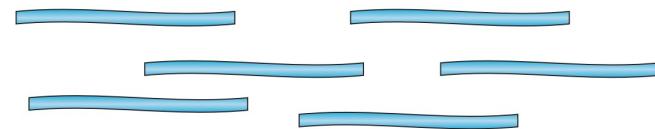
➤ 步骤

a Data generation

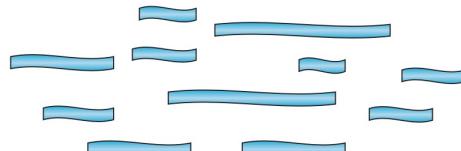
① mRNA or total RNA



② Remove contaminant DNA

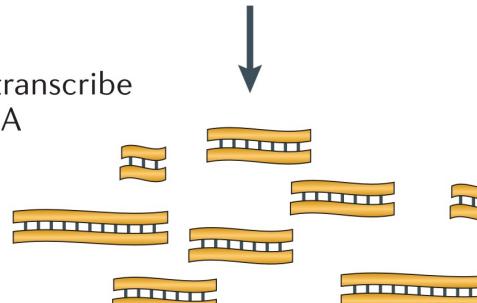


③ Fragment RNA

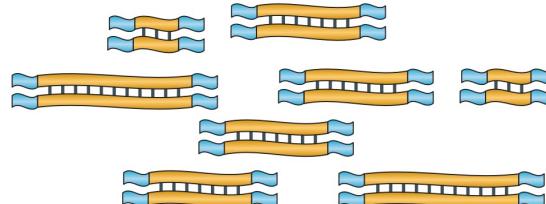


Remove rRNA?
Select mRNA?

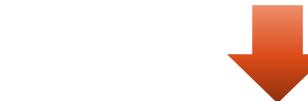
④ Reverse transcribe
into cDNA



⑤ Ligate sequence adaptors



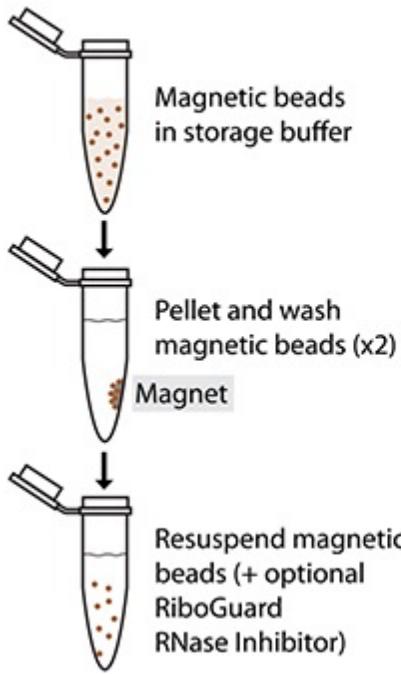
Strand-specific RNA-seq?



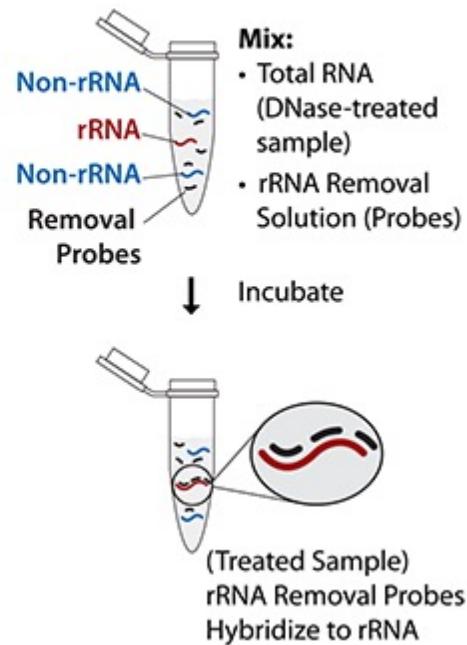
Computational challenge:
From reads to transcriptome

文库构建--rRNA去除

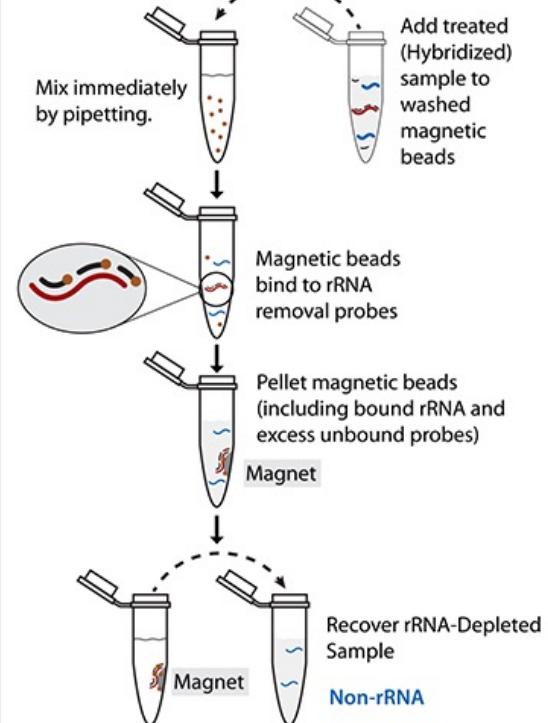
Step 1. Wash the magnetic beads.



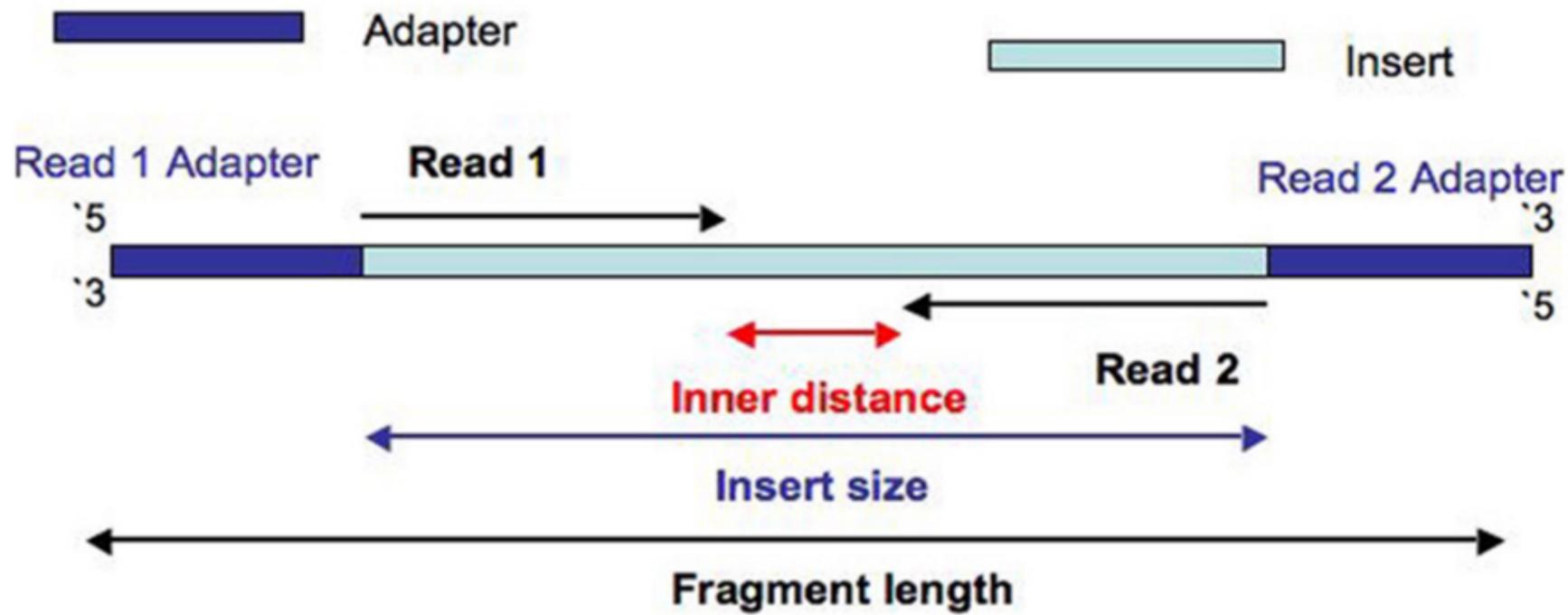
Step 2. Treat sample with rRNA Removal Solution.



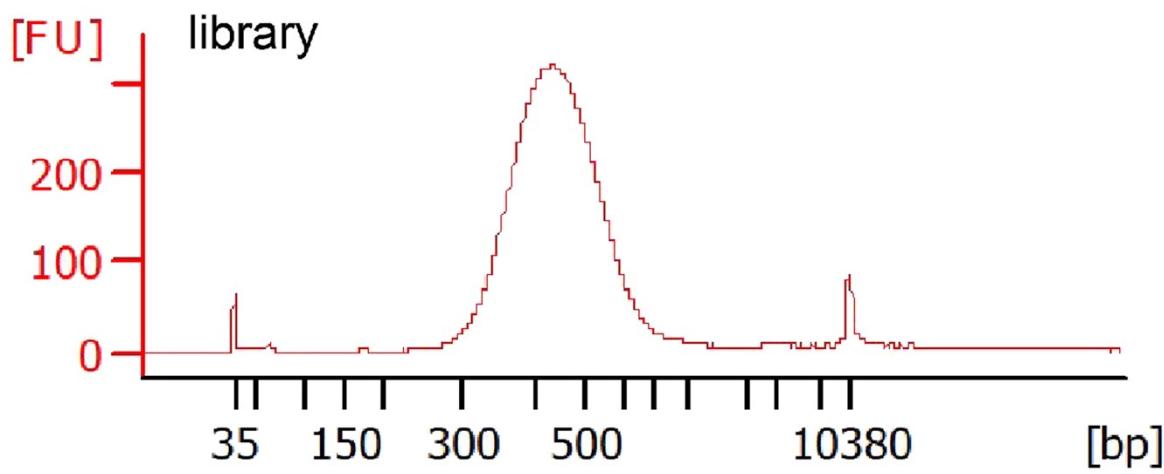
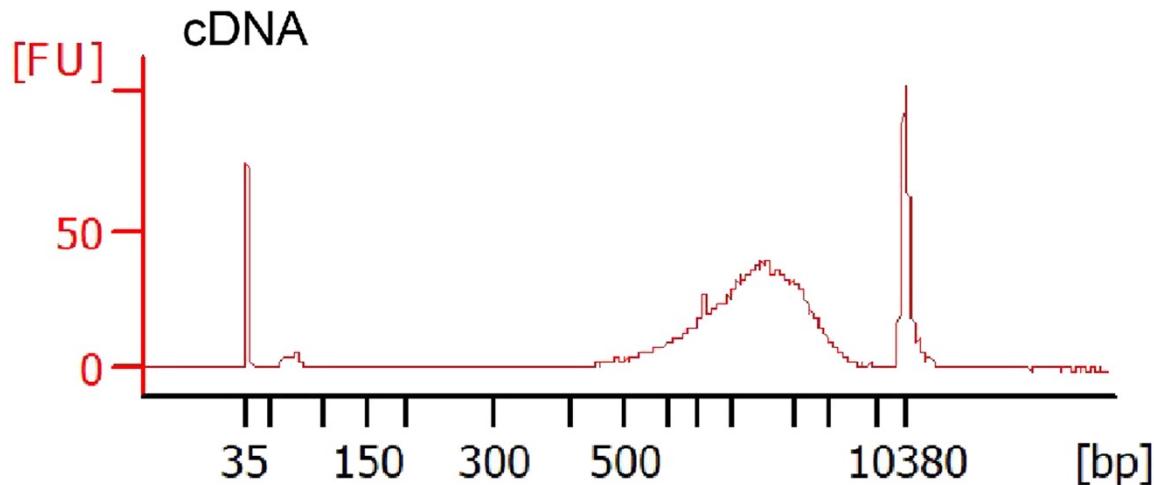
Step 3. Remove rRNA.



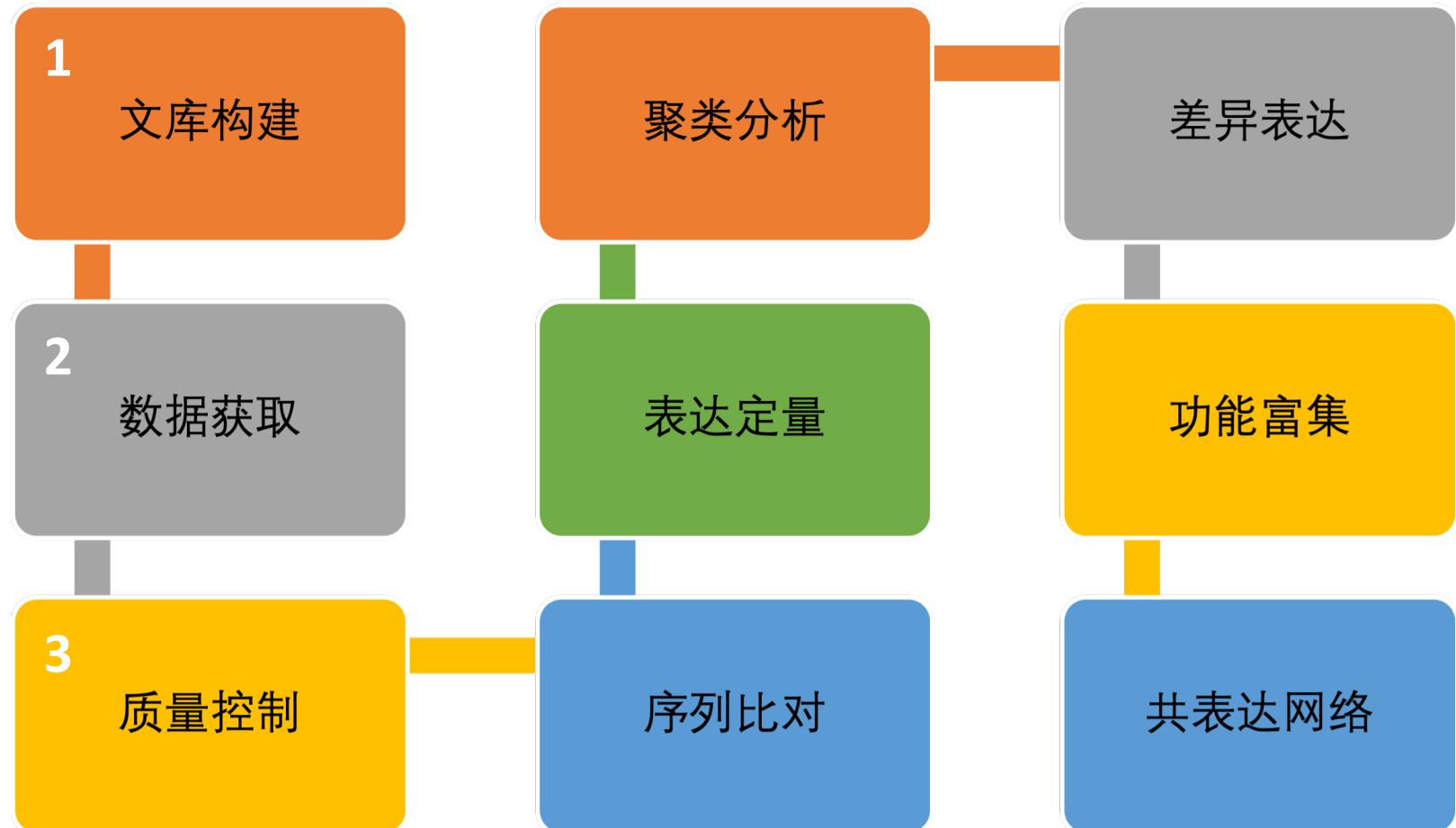
文库构建：插入片段长度



文库构建：质量控制



转录组测序技术RNA-seq常规流程



FastQ文件格式

■ FASTQ文件中，一个序列通常由四行组成：

- 第一行以@开头，之后为序列的标识符以及描述信息（与FASTA格式的描述行类似）
- 第二行为序列信息
- 第三行以+开头，之后可以再次加上序列的标识及描述信息（可选）
- 第四行为质量得分信息，与第二行的序列相对应，长度必须与第二行相同

```
@SEQ_ID
```

```
GATTGGGGTTCAAAGCAGTATCGATCAAATAGTAAATCCATTGTTCAACTCACAGTT
```

```
+
```

```
!''*((((****+))%%%++)(%%%%).1***-+*'')**55CCF>>>>CCCCCCCC65
```

字符与对应的ASCII码

十进制	图形										
33	!	49	1	65	A	81	Q	97	a	113	q
34	"	50	2	66	B	82	R	98	b	114	r
35	#	51	3	67	C	83	S	99	c	115	s
36	\$	52	4	68	D	84	T	100	d	116	t
37	%	53	5	69	E	85	U	101	e	117	u
38	&	54	6	70	F	86	V	102	f	118	v
39	'	55	7	71	G	87	W	103	g	119	w
40	(56	8	72	H	88	X	104	h	120	x
41)	57	9	73	I	89	Y	105	i	121	y
42	*	58	:	74	J	90	Z	106	j	122	z
43	+	59	;	75	K	91	[107	k	123	{
44	,	60	<	76	L	92	\	108	l	124	
45	-	61	=	77	M	93]	109	m	125	}
46	.	62	>	78	N	94	^	110	n	126	~
47	/	63	?	79	O	95	_	111	o		
48	0	64	@	80	P	96	`	112	p		

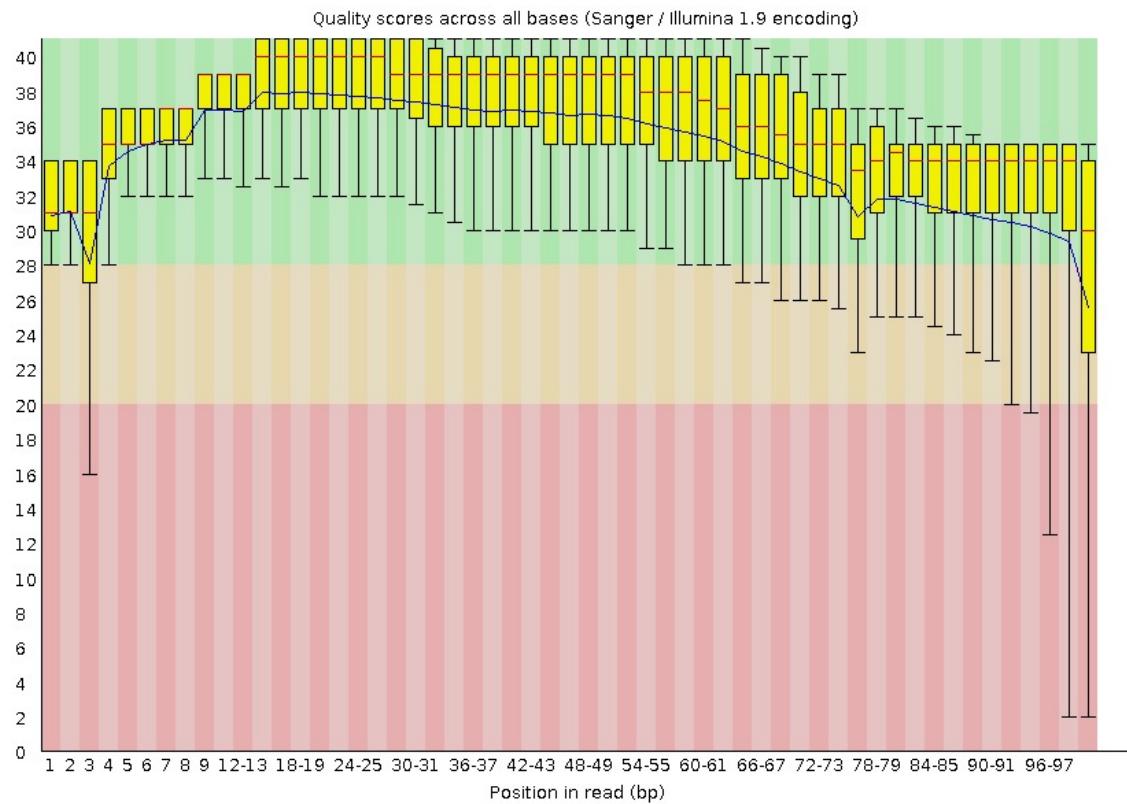
测序质量控制评估

FastQC Report

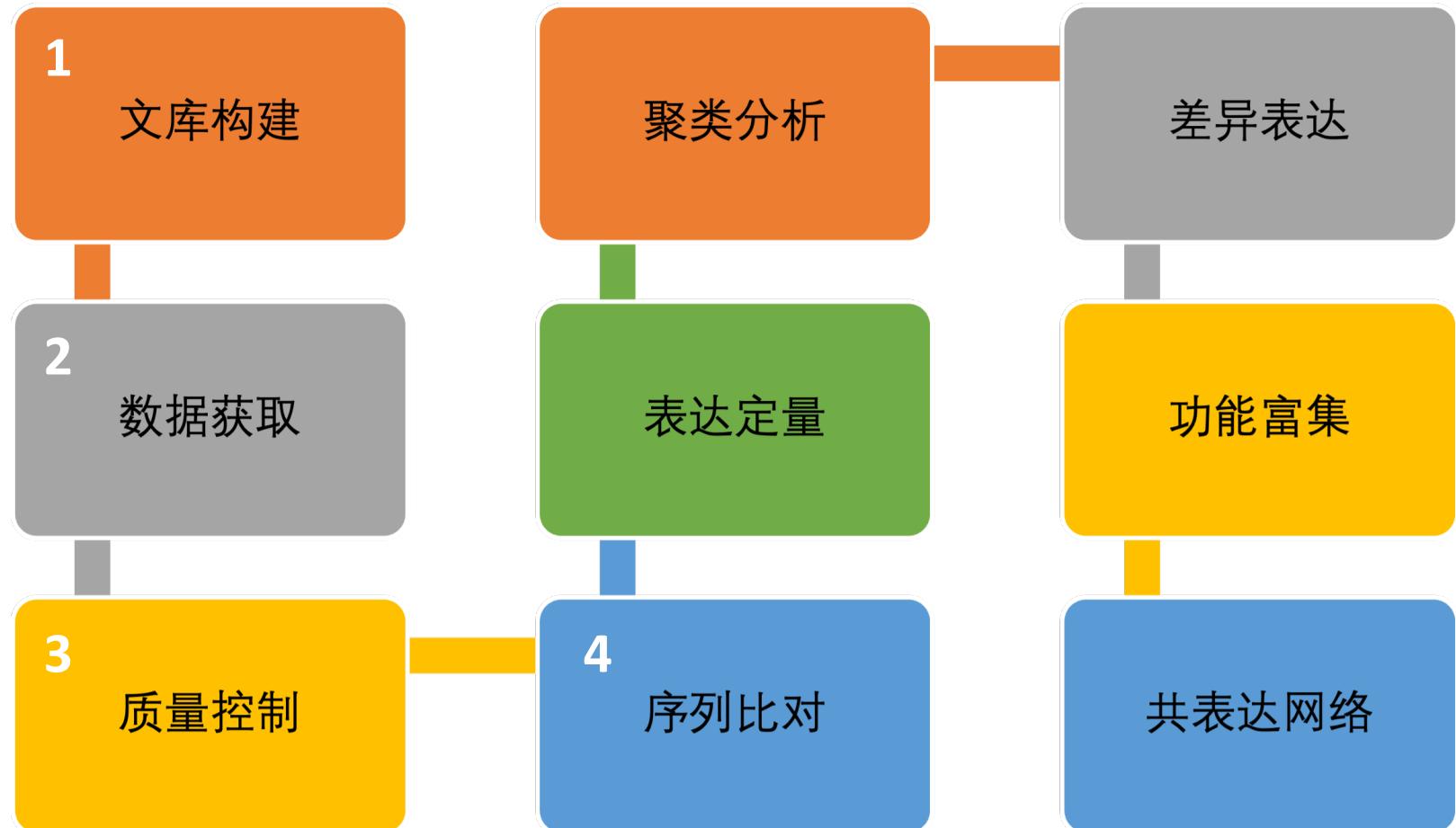
Summary

- ✓ Basic Statistics
- ✓ Per base sequence quality
- ! Per tile sequence quality
- ✓ Per sequence quality scores
- ✗ Per base sequence content
- ✓ Per sequence GC content
- ✓ Per base N content
- ✓ Sequence Length Distribution
- ✗ Sequence Duplication Levels
- ! Overrepresented sequences
- ! Adapter Content
- ✗ Kmer Content

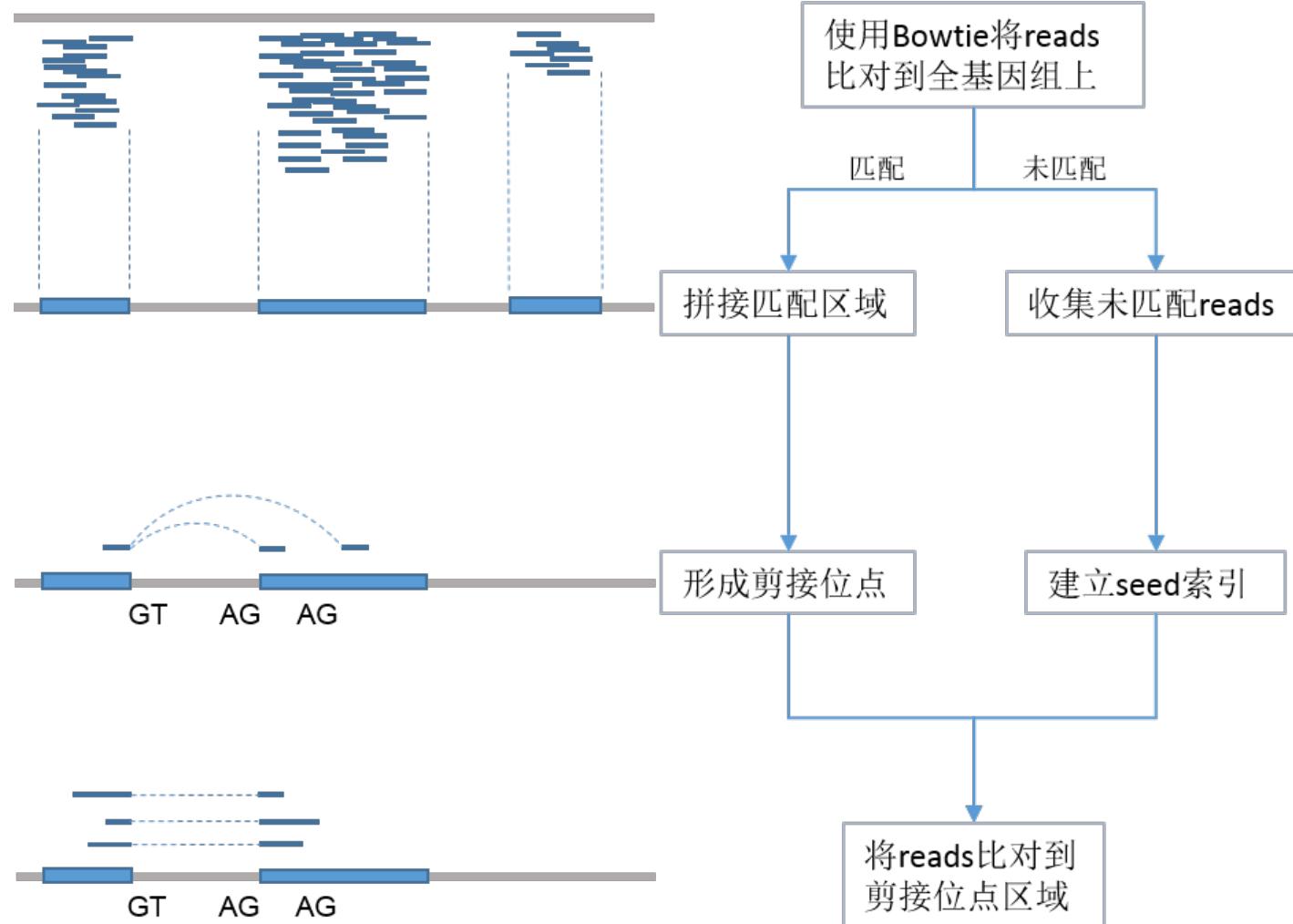
✓ Per base sequence quality



转录组测序技术RNA-seq常规流程



序列比对 (reads mapping)



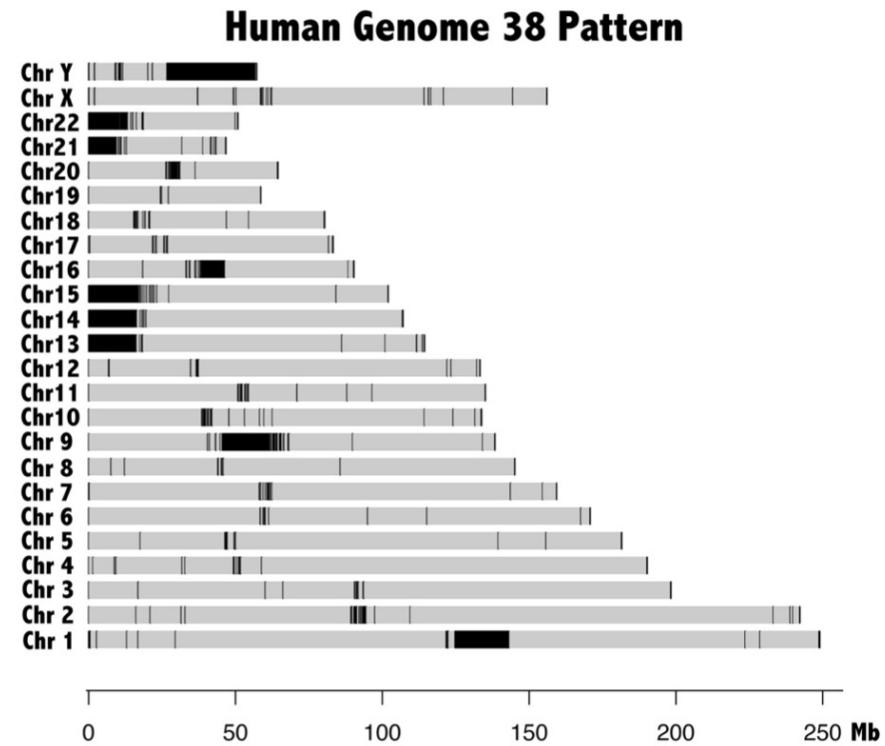
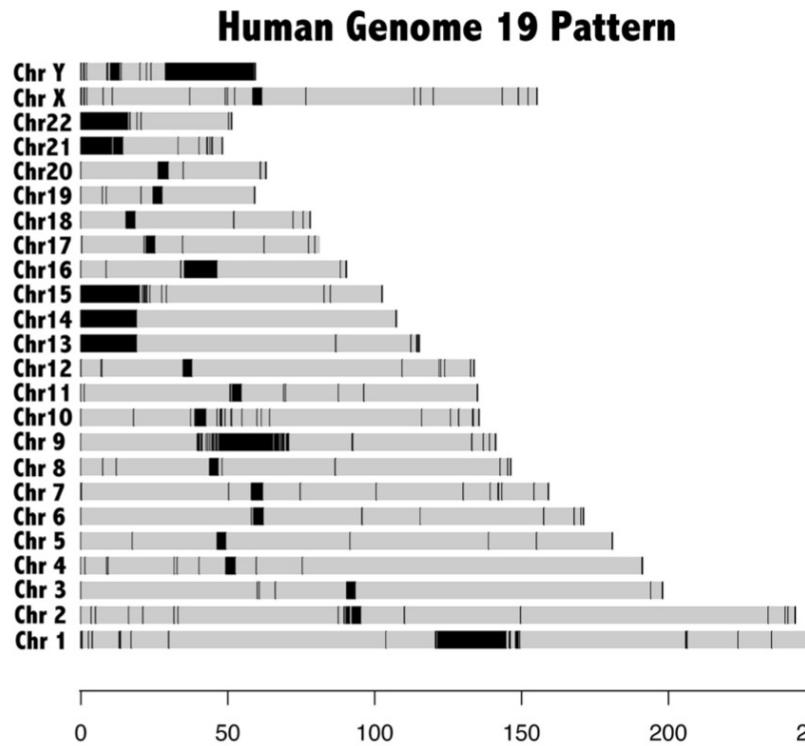
参考基因组

The screenshot shows the Ensembl website. At the top, there is a dark blue header bar with the Ensembl logo on the left and navigation links for BLAST/BLAT, BioMart, Tools, Downloads, Help & Documentation, Blog, and Mirrors. Below the header is a search bar with dropdown menus for 'Search:' (set to 'All species') and 'for' (empty), and a 'Go' button. A placeholder text 'e.g. BRCA2 or rat 5:62797383-63627669 or rs699 or coronary heart disease' is shown below the search bar. The main content area is divided into several sections:

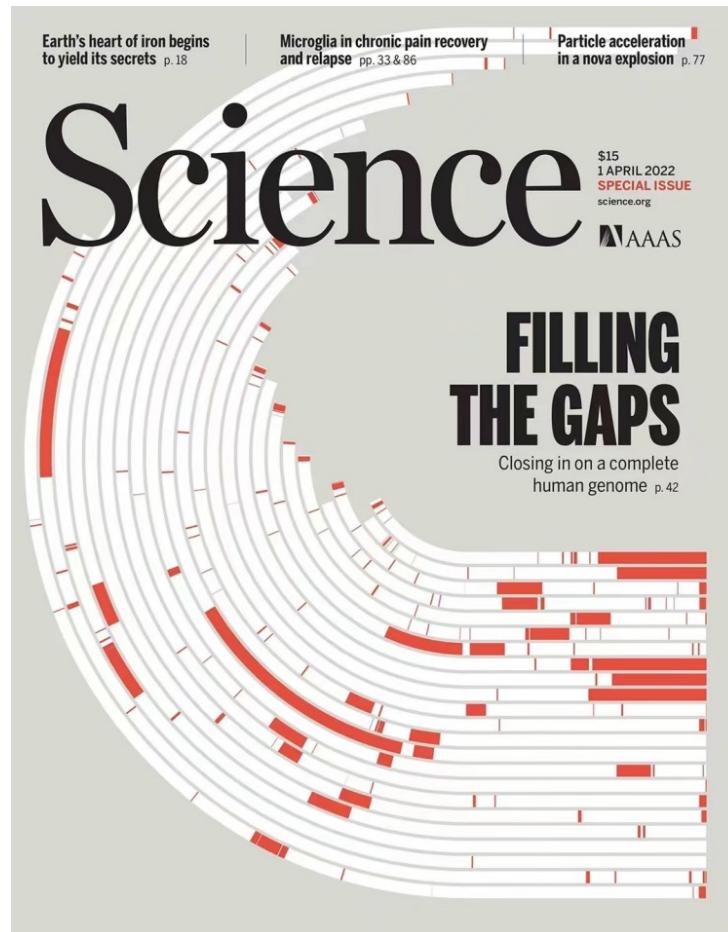
- Browse a Genome**: A text block explaining Ensembl's purpose as a genome browser for vertebrate genomes, mentioning comparative genomics, evolution, sequence variation, and transcriptional regulation. It also lists Ensembl tools like BLAST, BLAT, BioMart, and VEP.
- Favourite genomes**: A section showing icons for Human (GRCh38.p10), Mouse (GRCm38.p5), and Zebrafish (GRCz10). There is also a link to 'Edit favourites'.
- Find a Data Display**: A section with four blue directional arrows pointing right, labeled TABLE, HEATMAP, SEQUENCE, and PIE CHART. To the right, text explains the 'Find a Data Display' page, which allows users to choose a gene, region or variant and browse relevant visualisations. A 'Try it now!' button is located at the bottom right of this section.
- Variant Effect Predictor**: A section featuring the 'Ve!P' logo and a small thumbnail image of a tissue sample.
- Gene expression in different tissues**: A section showing a small thumbnail image of a tissue sample.

- Ensembl数据库

参考基因组hg19与hg38



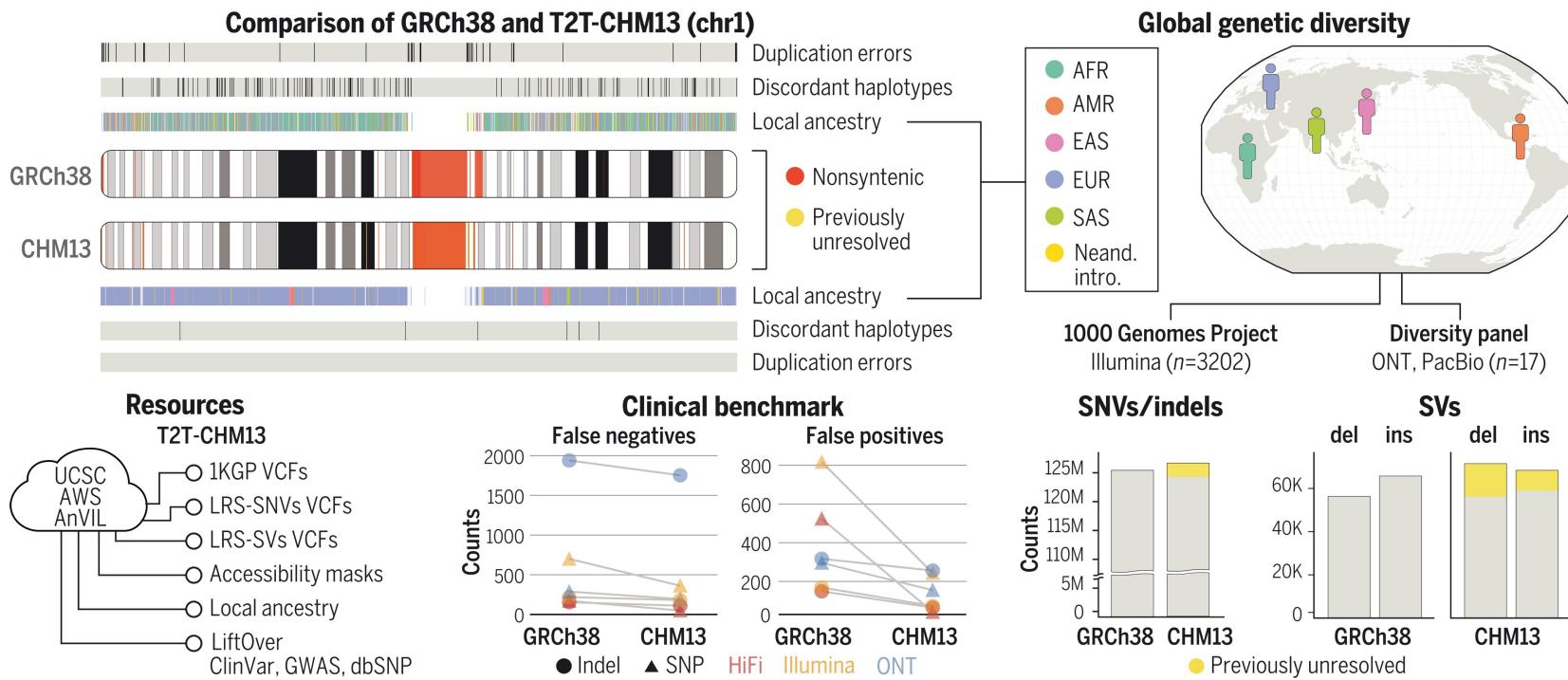
参考基因组T2T-CHM13



The complete sequence of a human genome. 2022 April

参考基因组T2T-CHM13

A complete reference genome improves analysis of human genetic variation



常用比对软件-tophat2

- 调用Bowtie/ Bowtie2，将reads比对到参考基因组上
- 寻找出外显子之间的结合位点

TopHat
A spliced read mapper for RNA-Seq

JOHNS HOPKINS UNIVERSITY
CENTER FOR COMPUTATIONAL BIOLOGY
CCB

Getting started

- [Install quick-start](#)
- [Test the installation](#)
- [Preparing your reference](#)
- [Preparing your reads](#)
- [Running TopHat](#)
- [Examining your results](#)

» **Install quick-start**

Download and extract the latest Bowtie 2 (or Bowtie) releases.

Note that you can use either Bowtie 2 (the default) or Bowtie (--bowtie1) and you will need the following Bowtie 2 (or Bowtie) programs in your PATH:

- bowtie2 (or bowtie)
- bowtie2-build (or bowtie-build)
- bowtie2-inspect (or bowtie-inspect)

Installing a pre-compiled binary release

In order to make it easy to install TopHat we provide a few binary packages to save users from the occasionally frustrating process of building TopHat themselves, which requires a certain development environment and the [Boost](#) libraries installed. To use the binary packages, simply download the appropriate one for your platform, unpack it, and make sure the [TopHat](#) binaries are in a directory in your PATH environment variable (or create a symbolic link to the included tophat2 script somewhere in your PATH, see below)

Note: if you want to be able to install and run this new version without overwriting a previous TopHat version already installed on your system, make sure you unpack the new version into a different directory from the old version, then instead of copying the new programs in a directory in your PATH just create a symbolic link from the tophat2 wrapper script in this new directory to a directory in your shell's PATH. For example, assuming the ~/bin directory is in your PATH and you unpack tophat-2.0.0.Linux_x86_64.tar.gz under your home directory:

```
cd  
tar xvfz tophat-2.0.0.Linux_x86_64.tar.gz  
cd ~/bin
```

Site Map

- [Home](#)
- [Getting started](#)
- [Manual](#)
- [Index and annotation downloads](#)
- [FAQ](#)
- [Protocol](#)

News and updates

New releases and related tools will be announced through the Bowtie [mailing list](#).

Getting Help

Questions and comments about TopHat can be posted on the [Tuxedo Tools Users Google Group](#). Please use tophat.cufflinks@gmail.com for private communications only. Please do not email technical questions to TopHat contributors directly.

Releases

version 2.1.1	2/23/2016
Source code	
Linux x86_64 binary	
Mac OS X x86_64 binary	

Related Tools

常用比对软件-STAR

- STAR能够发现非典型拼接
- 嵌合（融合）转录本，并能够比对全长RNA序列

Index of /shares/gingeraslab/www-data/dobin/STAR

<u>Name</u>	<u>Last modified</u>	<u>Size</u>	<u>Description</u>
 Parent Directory		-	
 Old/	2020-04-09 17:03	-	
 STARgenomes/	2020-04-10 20:13	-	

<https://github.com/alexdobin/STAR>

<https://github.com/alexdobin/STAR/blob/master/doc/STARmanual.pdf>

<https://labshare.cshl.edu/shares/gingeraslab/www-data/dobin/STAR/>

常用比对软件-HISAT2

- HISAT2找到junction正确率最高
- 总数上却比TopHat和STAR少



HISAT2 is a fast and sensitive alignment program for mapping next-generation sequencing reads (both DNA and RNA) to a population of human genomes as well as to a single reference genome. Based on an extension of BWT for graphs (Sirén et al. 2014), we designed and implemented a graph FM index (GFM), an original approach and its first implementation. In addition to using one global GFM index that represents a population of human genomes, **HISAT2** uses a large set of small GFM indexes that collectively cover the whole genome. These small indexes (called local indexes), combined with several alignment strategies, enable rapid and accurate alignment of sequencing reads. This new indexing scheme is called a Hierarchical Graph FM index (HGFM).

The [HISAT-3N paper](#) published at *Genome Research*. 7/1/2021

HISAT-3N beta release 12/14/2020

HISAT-3N is a software system for analyzing nucleotide conversion sequencing reads. See the [HISAT-3N](#) for more details.

Index files are moved to the AWS Public Dataset Program. 9/3/2020

We have moved HISAT2 index files to the AWS Public Dataset Program. See the [link](#) for more details.

HISAT 2.2.1 release 7/24/2020

Search

Main

About

Manual

HISAT-3N

Download

HowTo

Links

Funding

This work was supported in part by the National Human Genome Research Institute under grants R01-HG006102 and R01-HG006677, and NIH grants R01-LM06845 and R01-GM083873 and NSF grant CCF-0347992 to Steven L. Salzberg

文件格式

FASTQ文件--> SAM文件--> BAM文件

- SAM 文件是Sequence Alignment/Map Format的简写
- BAM文件是SAM文件的二进制格式
- 使用SAMtools 可以查看二进制的bam文件。

BIOINFORMATICS APPLICATIONS NOTE

Vol. 25 no. 16 2009, pages 2078–2079
doi:10.1093/bioinformatics/btp352

Sequence analysis

The Sequence Alignment/Map format and SAMtools

Heng Li^{1,†}, Bob Handsaker^{2,†}, Alec Wysoker², Tim Fennell², Jue Ruan³, Nils Homer⁴, Gabor Marth⁵, Goncalo Abecasis⁶, Richard Durbin^{1,*} and 1000 Genome Project Data Processing Subgroup⁷

¹Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Cambridge, CB10 1SA, UK, ²Broad Institute of MIT and Harvard, Cambridge, MA 02141, USA, ³Beijing Institute of Genomics, Chinese Academy of Science, Beijing 100029, China, ⁴Department of Computer Science, University of California Los Angeles, Los Angeles, CA 90095,

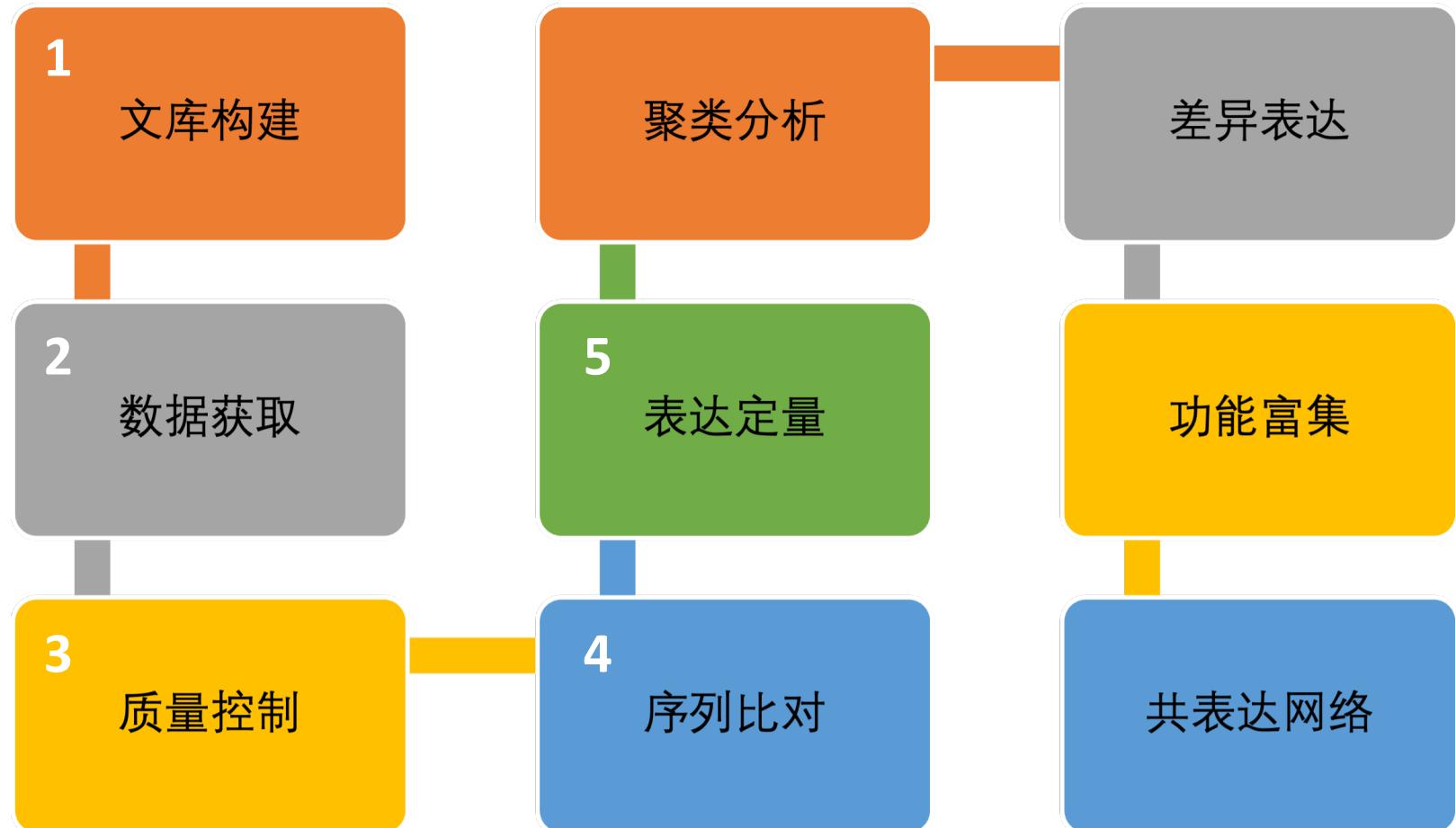
⁵Department of Biology, Boston College, Chestnut Hill, MA 02467, ⁶Center for Statistical Genetics, Department of Biostatistics, University of Michigan, Ann Arbor, MI 48109, USA and ⁷<http://1000genomes.org>

Received on April 28, 2009; revised on May 28, 2009; accepted on May 30, 2009

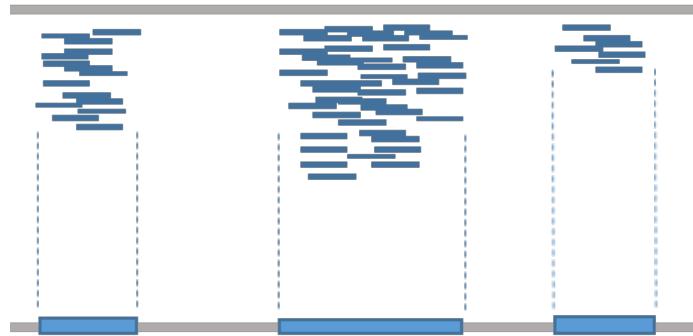
Advance Access publication June 8, 2009

Associate Editor: Alfonso Valencia

转录组测序技术RNA-seq常规流程

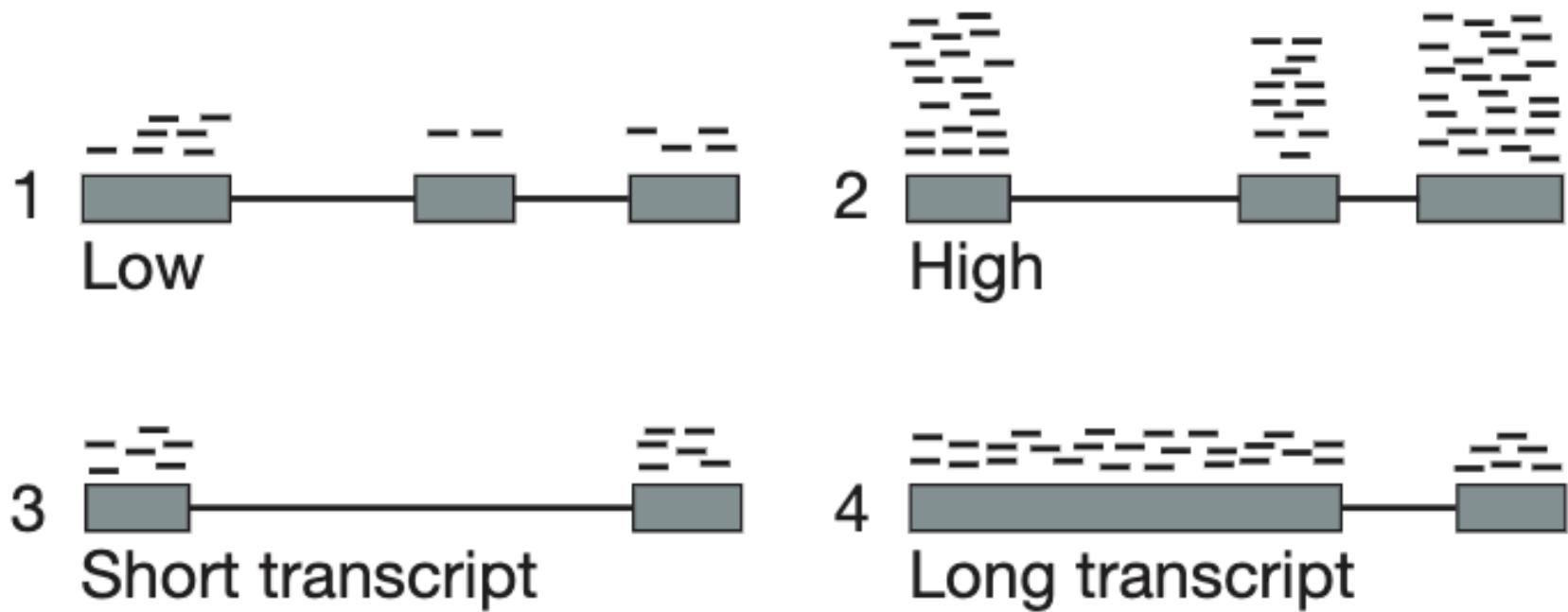


从reads到count



Count: 能够比对到基因上的reads的总数

为什么要数据归一化?



基因长度与测序深度

1. 基因长度：目标基因有多少个碱基对(不含内含子)

- a. 基因最长的转录本长度
- b. 多个转录本长度的平均值
- c. 非重叠外显子长度之和 ($L_1+L_2+L_3+L_4$)
- d. 非重叠cds序列(即编码序列)长度之和



2. 测序深度：测序得到的碱基总量(bp)与基因组(或转录组、测序目标区域)大小的比值

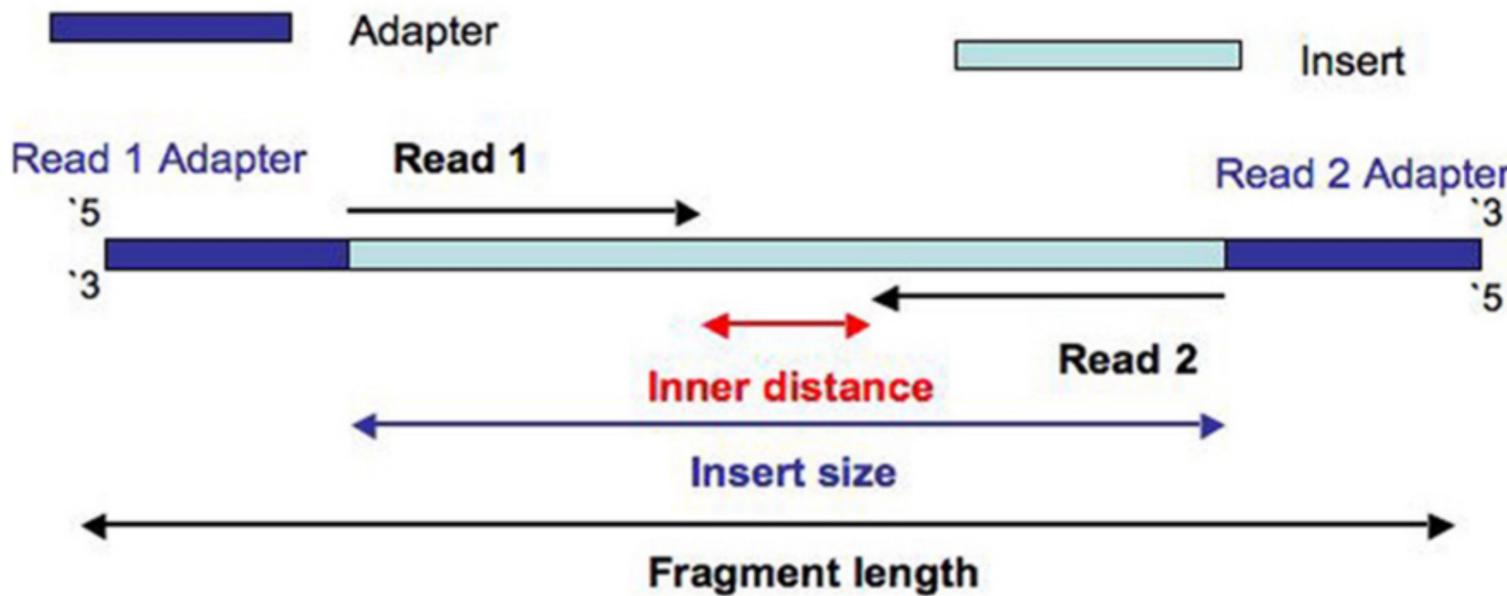
PCR扩增后，DNA片段越多，测序的时候得到的碱基总量也就越多，测序深度越大

RPKM

$$RPKM = \frac{ExonMappedReads * 10^9}{TotalMappedReads * ExonLength}$$



FPKM



$$FPKM = \frac{ExonMappedFragments * 10^9}{TotalMappedFragments * ExonLength}$$

TPM

$$TPM = \frac{Ni/Li * 10^6}{sum(N1/L1 + N2/L2 + \dots + Nn/Ln)}$$



RPKM vs TPM

RPKM

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

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

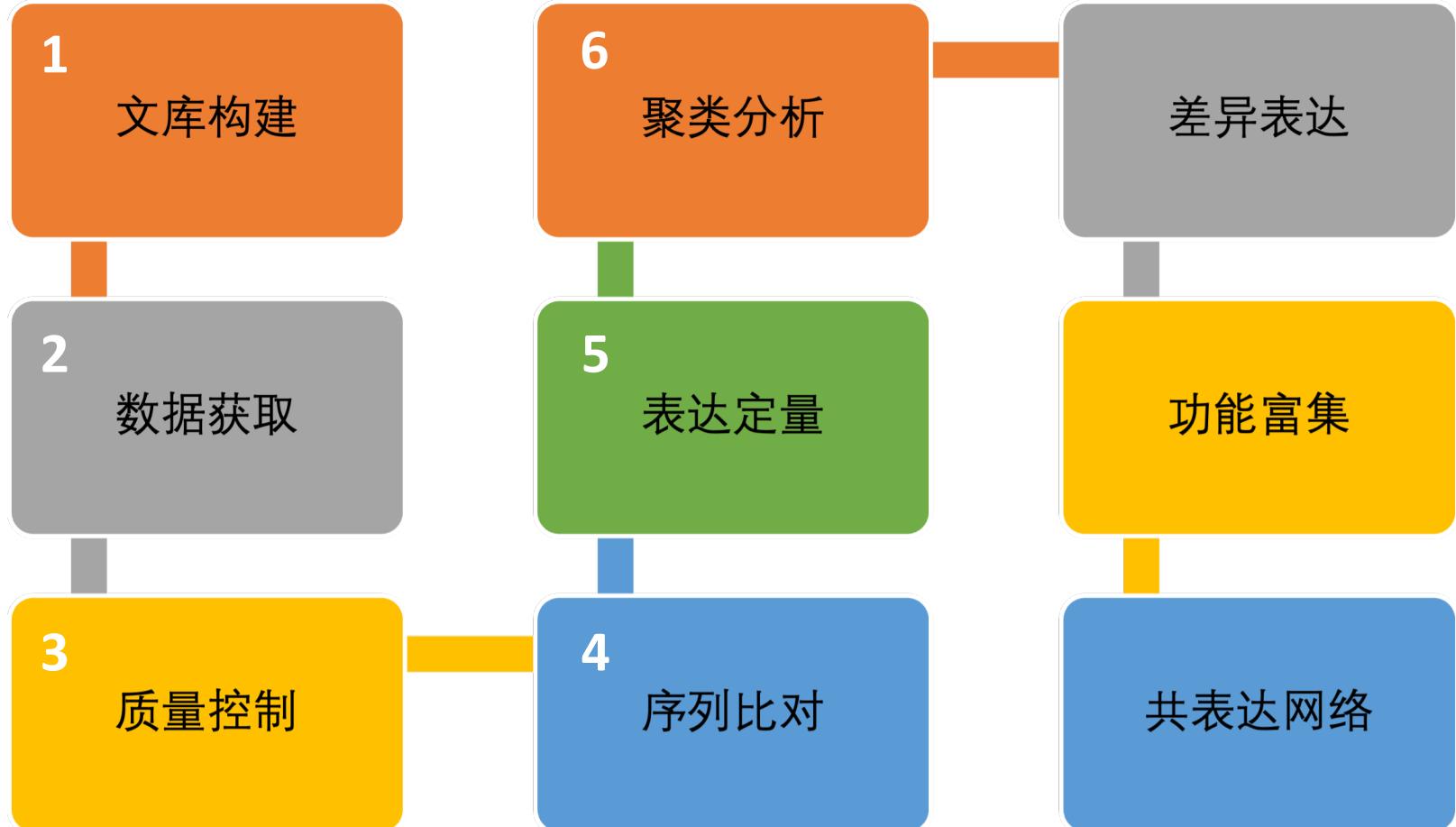
Total: 4.29 4.5 4.25

TPM

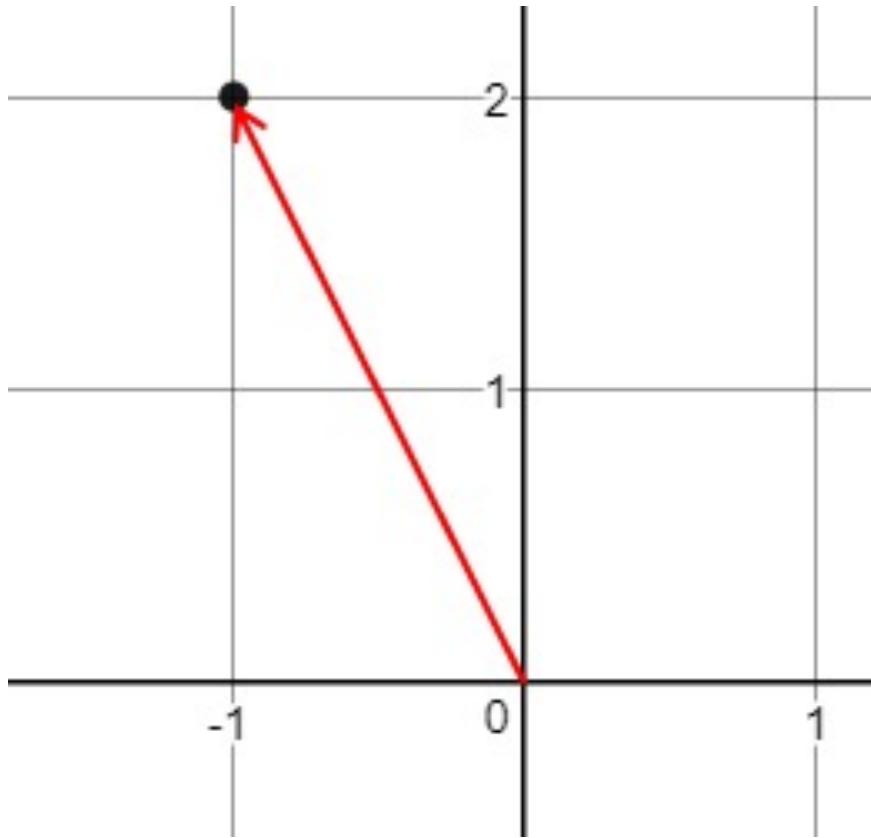
Gene Name	Rep1 TPM	Rep2 TPM	Rep3 TPM
A (2kb)	3.33	2.96	3.326
B (4kb)	3.33	3.09	3.326
C (1kb)	3.33	3.95	3.326
D (10kb)	0	0	0.02

Total: 10 10 10

转录组测序技术RNA-seq常规流程



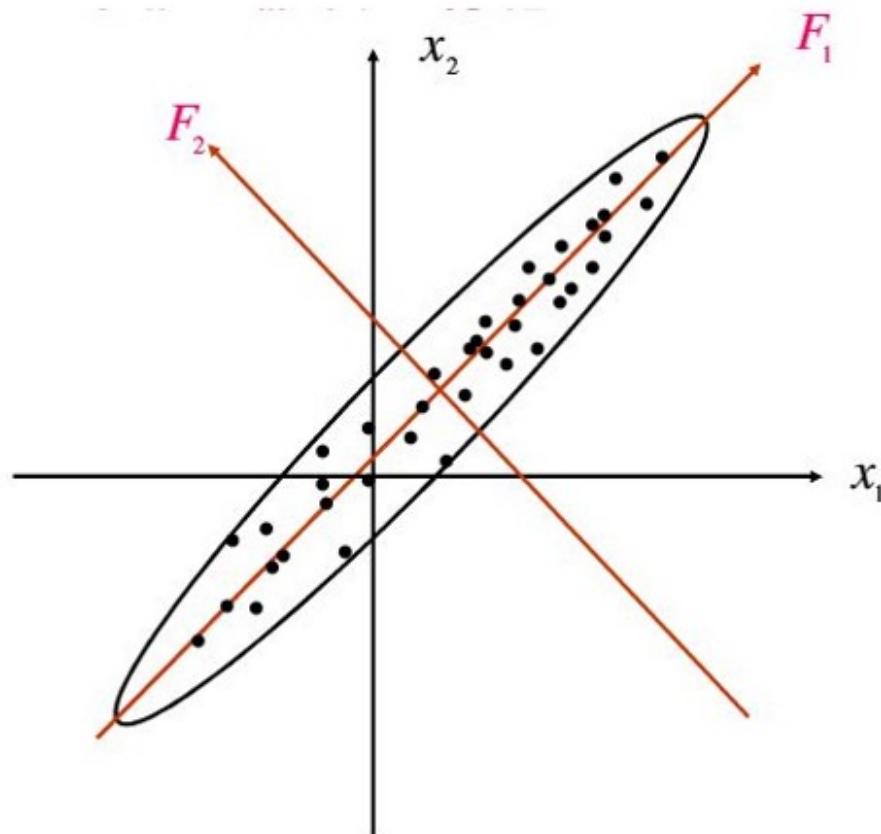
向量 (vector)



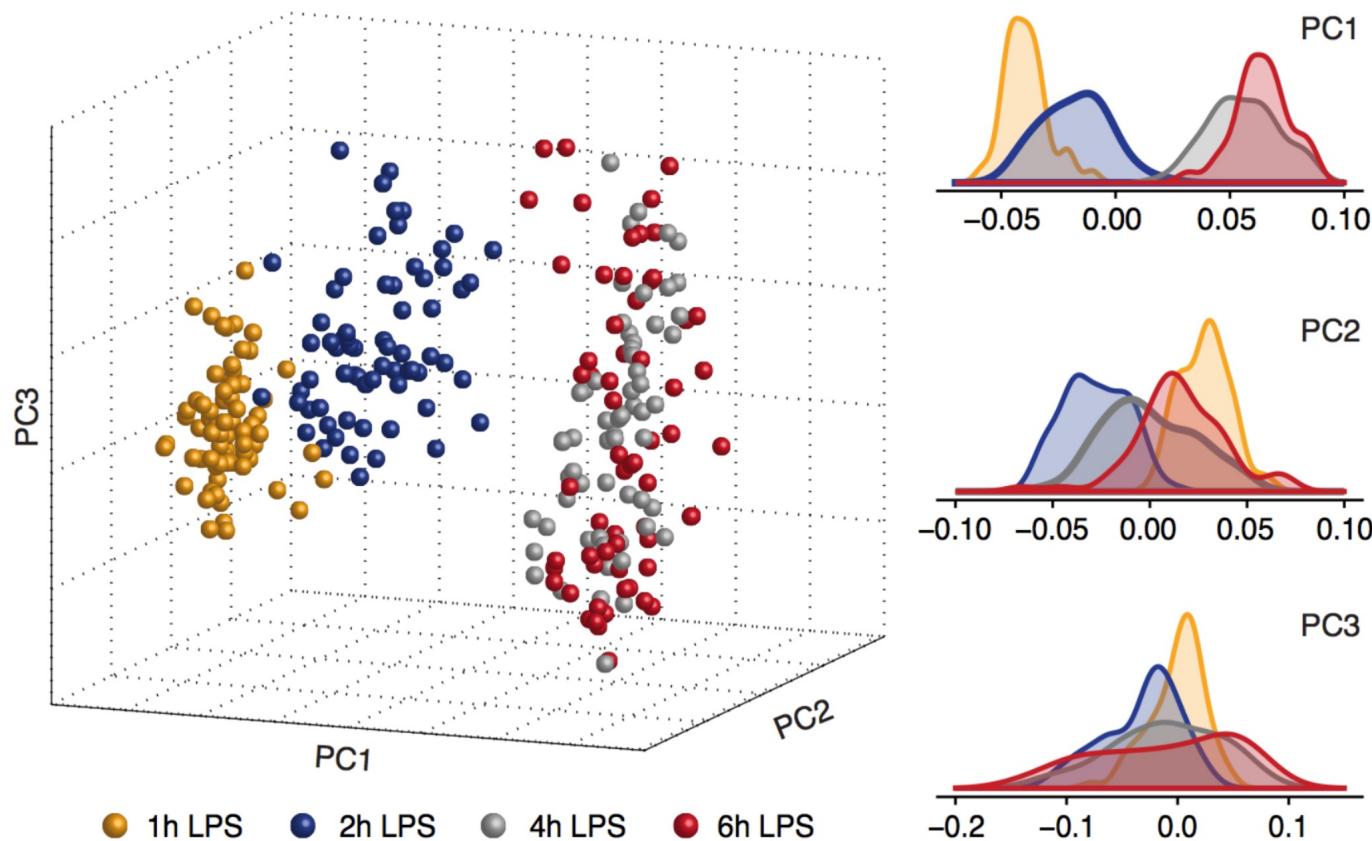
一个高维向量可以用一个数组来表示

$$\mathbf{V} = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \end{bmatrix}$$

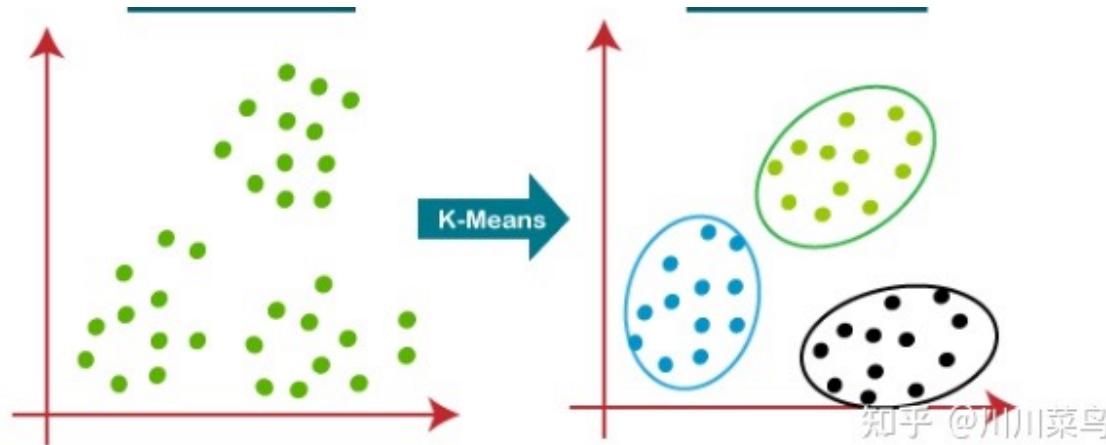
主成分分析 (Principal Component Analysis, PCA)



在特定维度投影后的特征分布

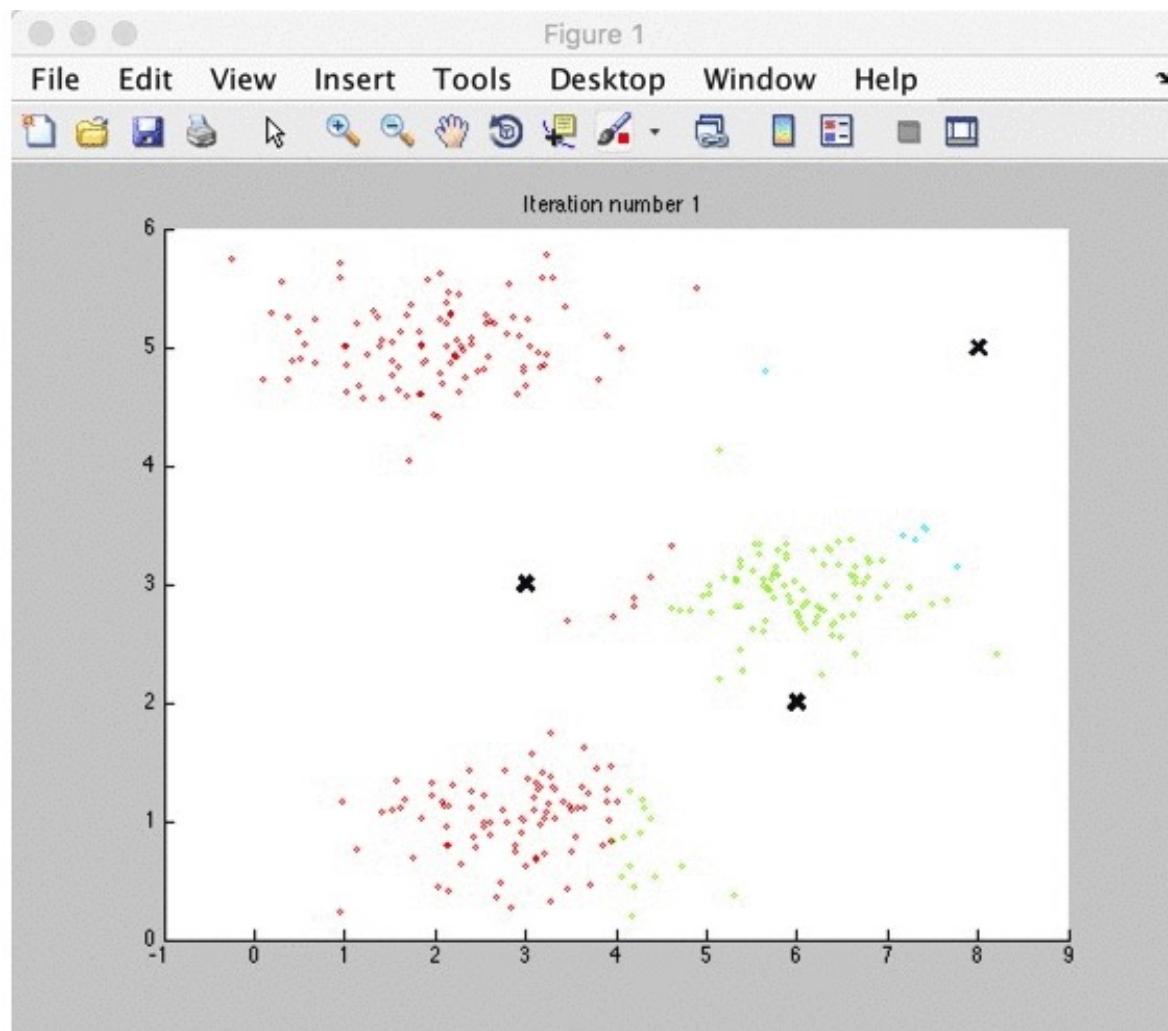


K-均值聚类

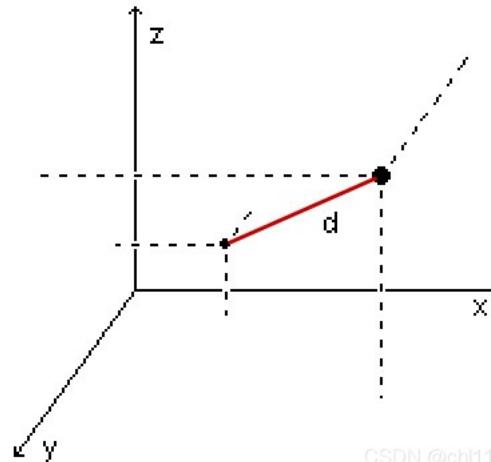
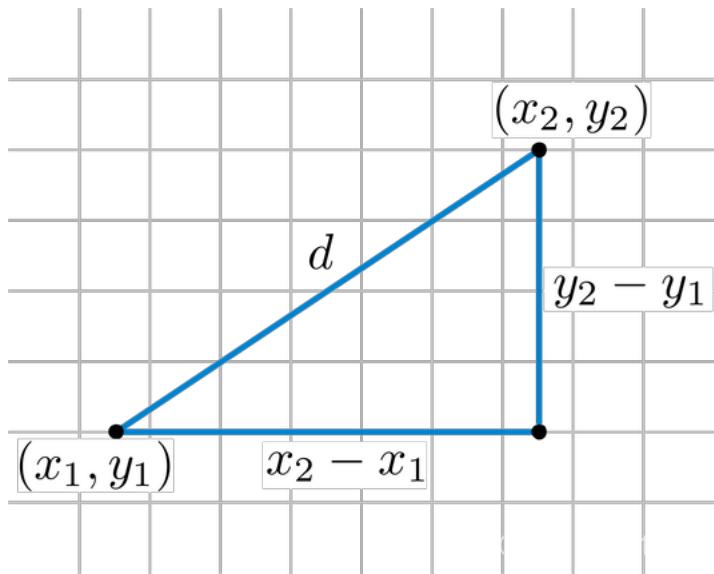


1. 选择聚类的个数k (k-means算法传递超参数的时候，只需设置最大的K值)
2. 任意产生k个聚类，然后确定聚类中心，或者直接生成k个中心。
3. 对每个点确定其聚类中心点。
4. 再计算其聚类新中心。
5. 重复以上步骤直到满足收敛要求。 (通常就是确定的中心点不再改变。)

K-均值聚类



欧式距离

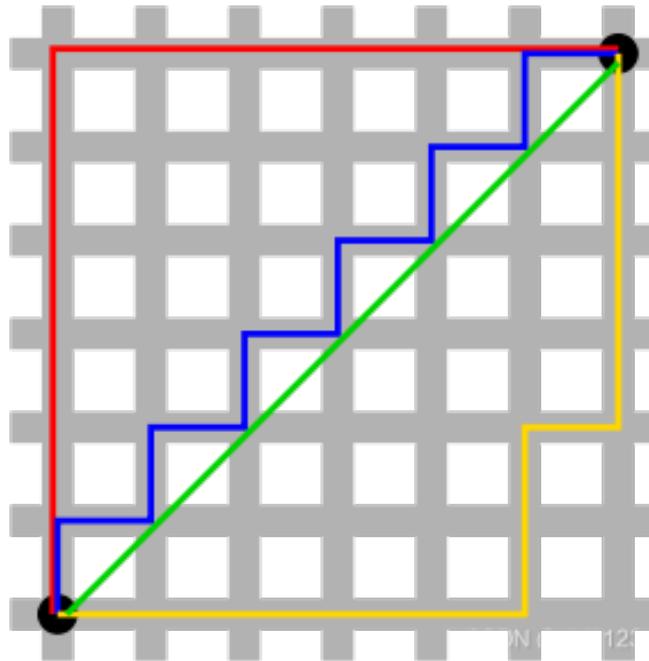
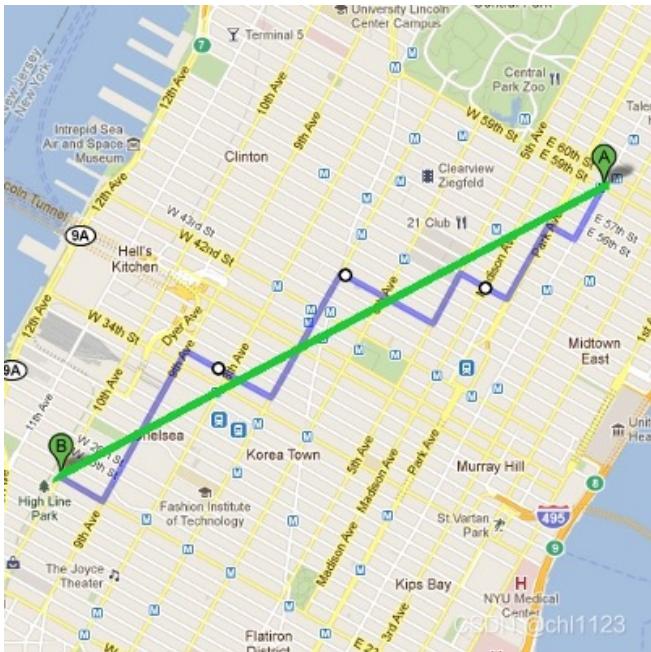


CSDN @chl1123

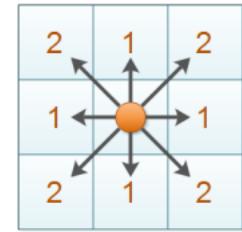
$$2D: \text{distance} = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$$

$$3D: \text{distance} = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2 + (z_2 - z_1)^2}$$

曼哈顿距离

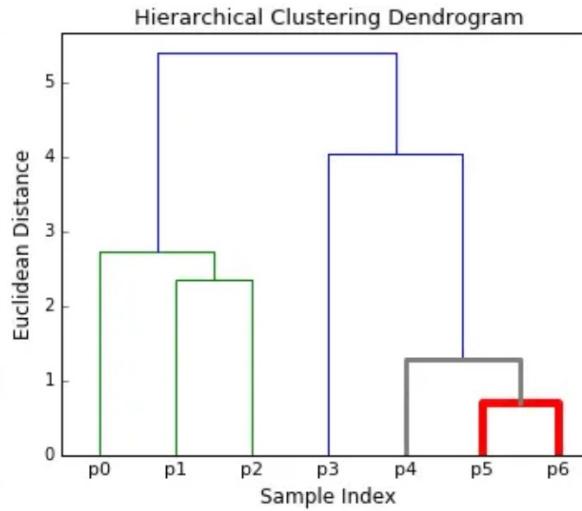
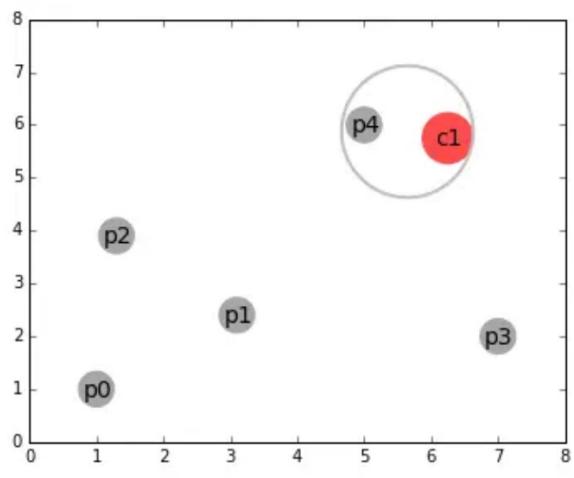
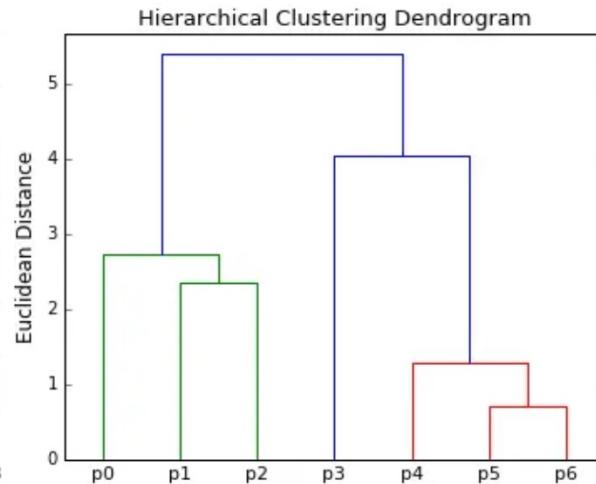
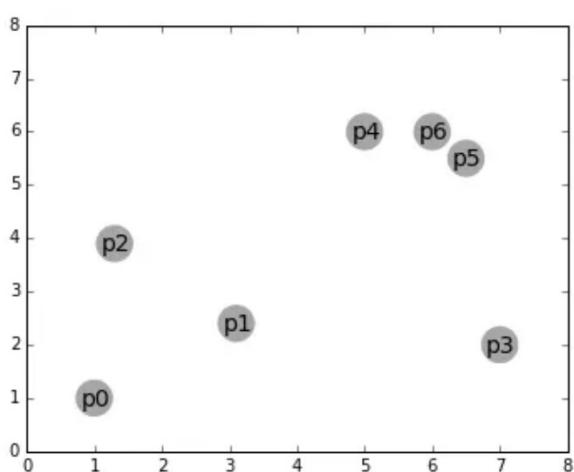


Manhattan Distance

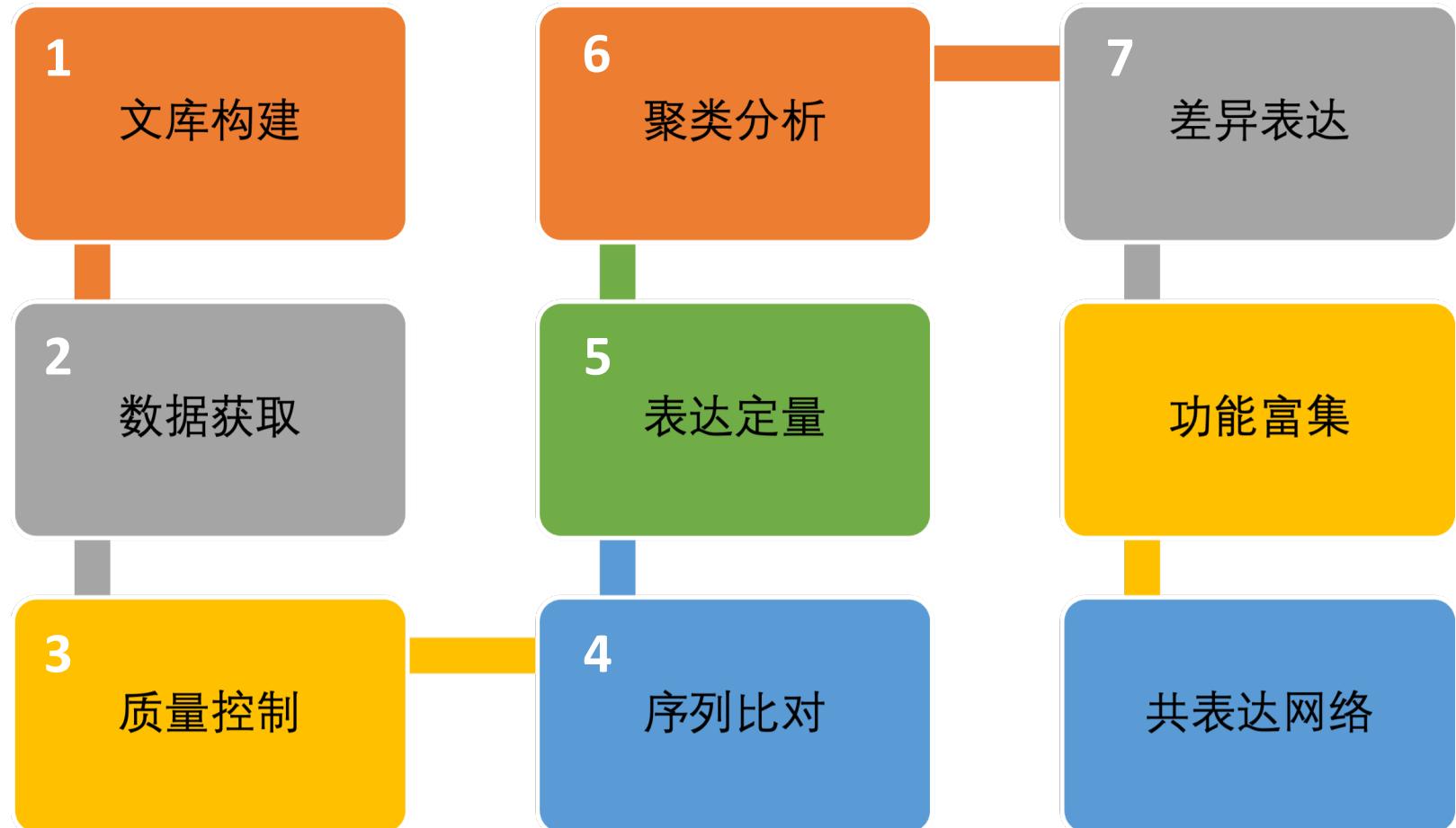


$$|x_1 - x_2| + |y_1 - y_2|$$

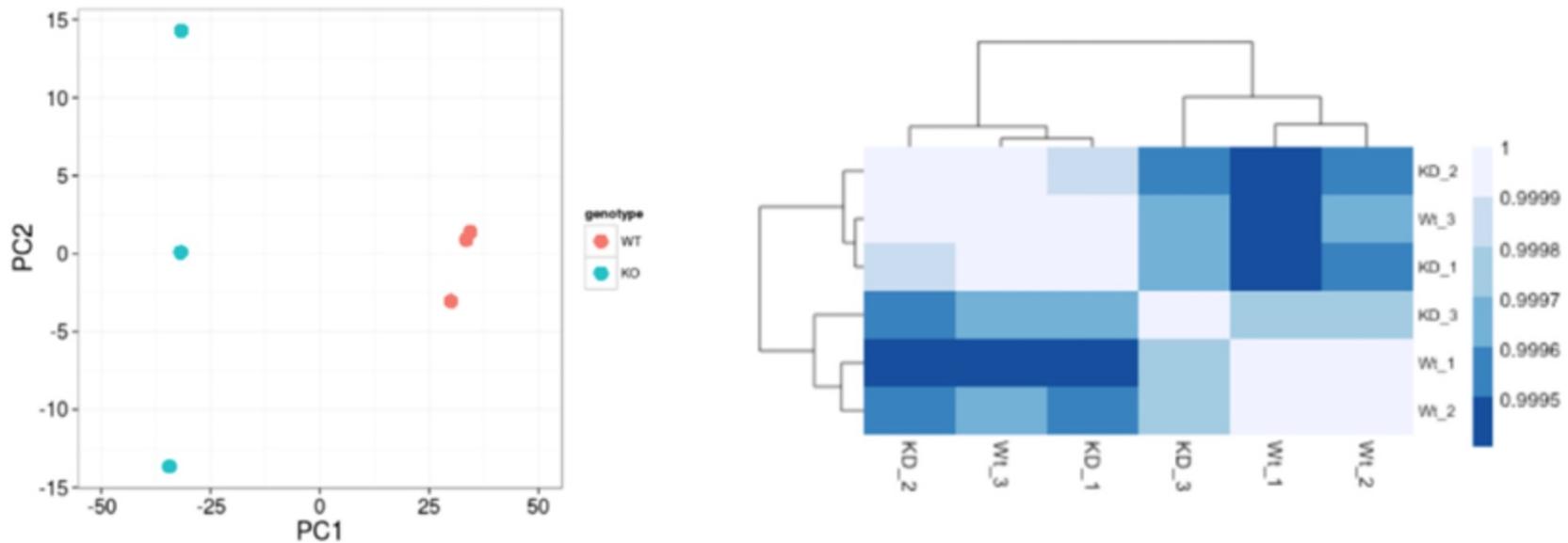
层次聚类



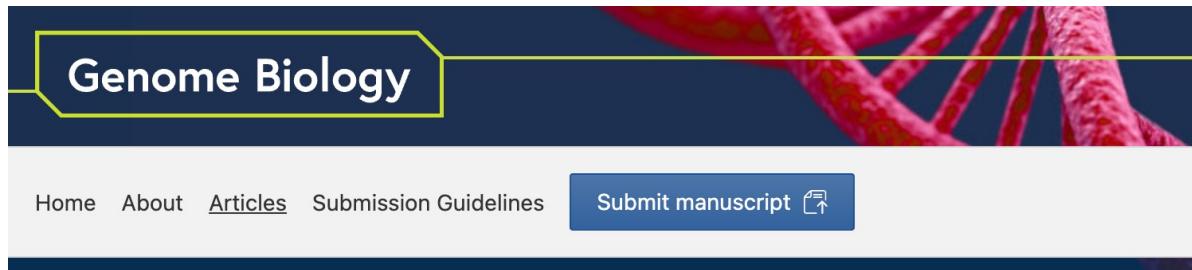
转录组测序技术RNA-seq常规流程



在差异分析之前完成样本的聚类分析



DESeq2



Genome Biology

<https://genomebiology.biomedcentral.com/articles/10.1186/s13059-014-0550-8> ::

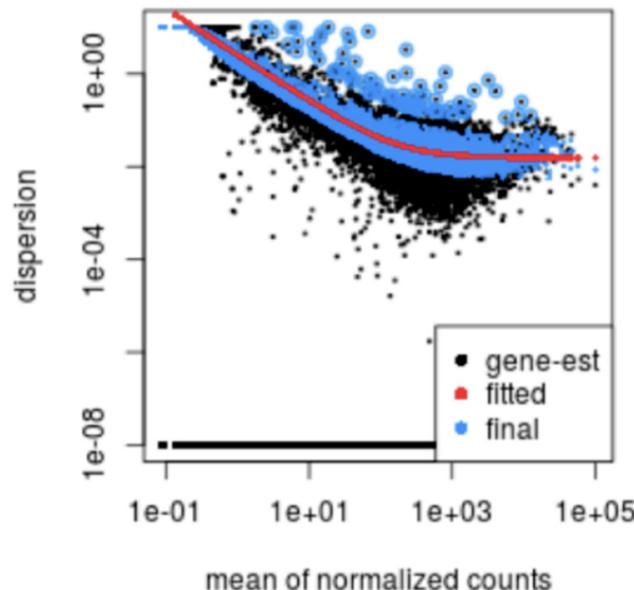
Moderated estimation of fold change and ... - Genome Biology

by MI Love · 2014 · Cited by 62469 — We present **DESeq2**, a method for differential analysis of count data, using shrinkage estimation for dispersions and fold changes to improve ...

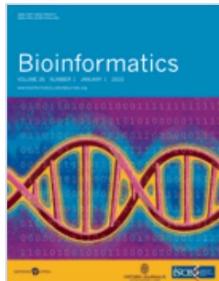
<https://genomebiology.biomedcentral.com/articles/10.1186/s13059-014-0550-8>

DESeq2

- 基于负二项分布模型
- 推荐3个以上重复样本
- 输入数据类型为counts矩阵
- 离散度 (dispersion) 与表达量成负相关，与方差成正相关



Bioinformatics

[Issues](#)[Advance articles](#)[Submit ▾](#)[Alerts](#)[About ▾](#)[Bioinformatics](#)

Volume 26, Issue 1
January 2010

JOURNAL ARTICLE

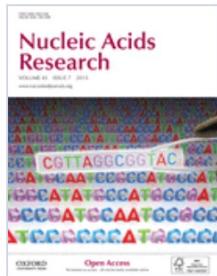
edgeR: a Bioconductor package for differential expression analysis of digital gene expression data

Mark D. Robinson , Davis J. McCarthy, Gordon K. Smyth [Author Notes](#)

Bioinformatics, Volume 26, Issue 1, January 2010, Pages 139–140,
<https://doi.org/10.1093/bioinformatics/btp616>

Published: 11 November 2009 [Article history ▾](#)

- 有样本重复的Bulk RNA-seq，选择quasi-likelihood(QL) F-test
- ScRNA-seq或没有重复样本的数据，选择likelihood ratio test



Volume 43, Issue 7

20 April 2015

JOURNAL ARTICLE

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

Matthew E. Ritchie, Belinda Phipson, Di Wu, Yifang Hu, Charity W. Law, Wei Shi, Gordon K. Smyth 

Nucleic Acids Research, Volume 43, Issue 7, 20 April 2015, Page e47,

<https://doi.org/10.1093/nar/gkv007>

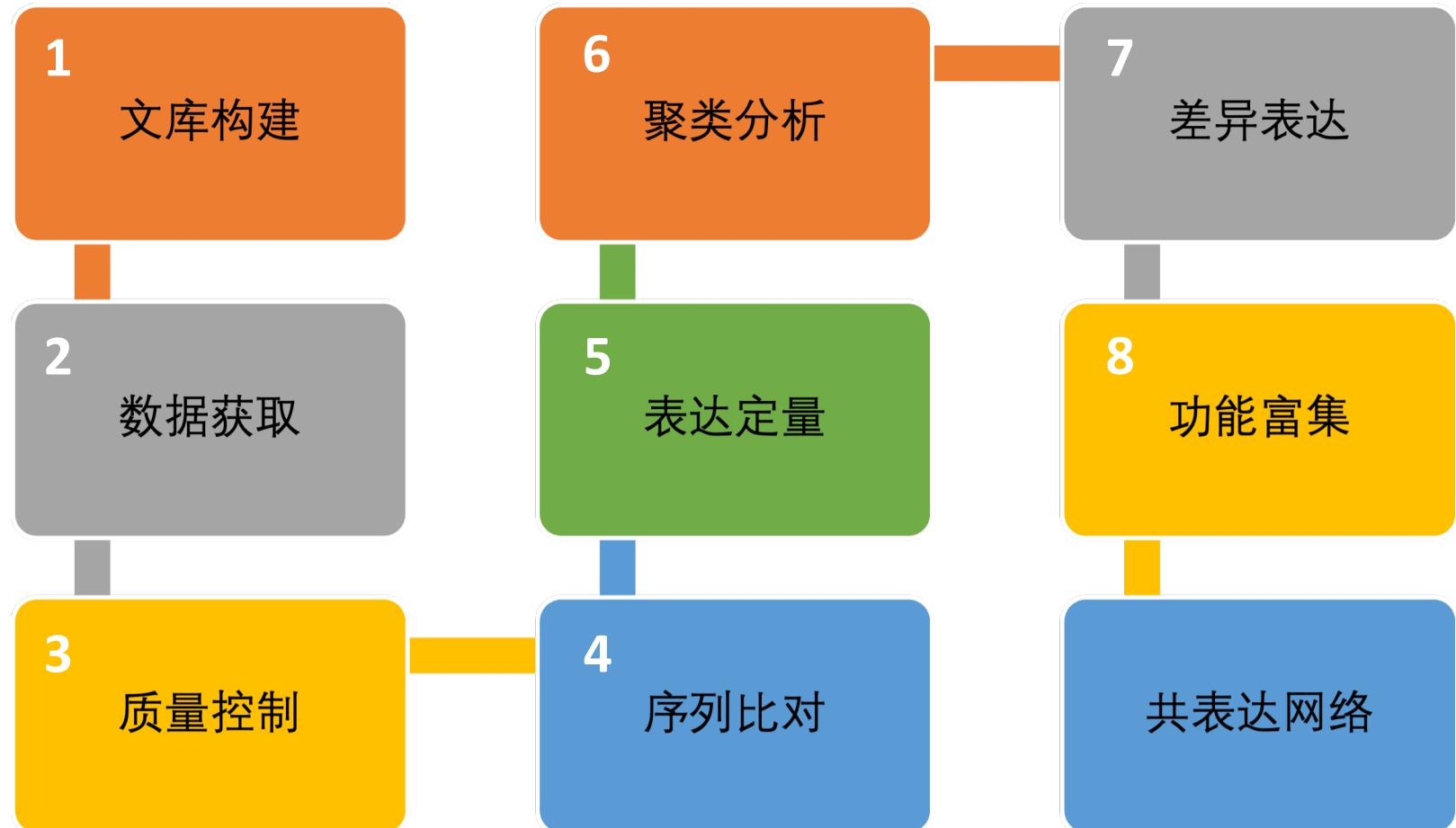
Published: 20 January 2015 Article history ▾

- 有适合芯片和测序数据的差异分析
- 适合小样本的差异分析

三种方法的R包官方说明书

- Analyzing RNA-seq data with DESeq2 (bioconductor.org)
 - edgeR: differential analysis of sequence read count data User's Guide (bioconductor.org)
 - limma usersguide.pdf (bioconductor.org)
-

转录组测序技术RNA-seq常规流程

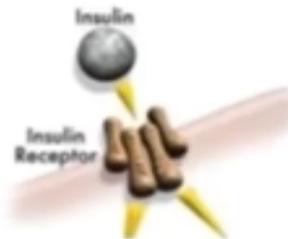


Gene Ontology, GO, 基因功能国际标准分类体系

MF

1. Molecular Function

An elemental activity or task or job

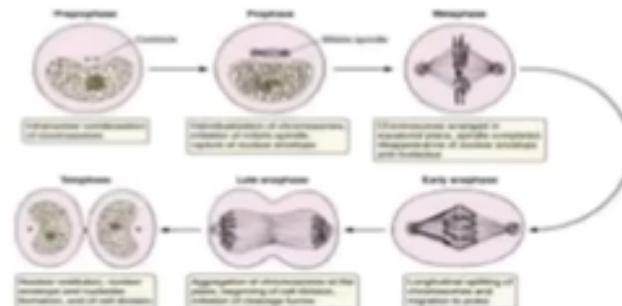


- protein kinase activity
- insulin receptor activity

BP

2. Biological Process

A commonly recognized series of events



- cell division

CC

3. Cellular Component

Where a gene product is located



- mitochondrion
- mitochondrial matrix
- mitochondrial inner membrane

DAVID使用

DAVID Bioinformatics Resources Laboratory of Human Retrovirology and Immunoinformatics (LHRI)

Home Start Analysis Shortcut to DAVID Tools Technical Center Downloads & APIs Term of Service About DAVID About LHRI

点击开始分析

Overview

The Database for Annotation, Visualization and Integrated Discovery (**DAVID**) provides a comprehensive set of functional annotation tools for investigators to understand the biological meaning behind large lists of genes. These tools are powered by the comprehensive **DAVID Knowledgebase** built upon the DAVID Gene concept which pulls together multiple sources of functional annotations. For any given gene list, DAVID tools are able to:

- Identify enriched biological themes, particularly GO terms
- Discover enriched functional-related gene groups
- Cluster redundant annotation terms
- Visualize genes on BioCarta & KEGG pathway maps
- Display related many-genes-to-many-terms on 2-D view.
- Search for other functionally related genes not in the list
- List interacting proteins
- Explore gene names in batch
- Link gene-disease associations
- Highlight protein functional domains and motifs
- Redirect to related literatures
- Convert gene identifiers from one type to another.
- And more

Hot Links

Multiple positions available in LHRI

The Laboratory of Human Retrovirology and Immunoinformatics (LHRI) has collaborated with the National Institute of Allergy and Infectious Diseases (NIAID) and supported NIAID clinical trials for patients infected with HIV mutants resisting anti-retroviral therapy. LHRI has isolated the multiple-class drug-resistant (MDR) variants from patients and characterized each variant's drug sensitivity and infectivity. The study aims to define salvage therapy and develop novel therapy (chemotherapy and immunotherapy). During the investigation, LHRI has characterized the emergence of novel mutations on drug susceptibility and viral replication. LHRI is a pioneer in researching the anti-viral cytokine, Interleukin-27, DNA-repair protein (Ku70)-mediated innate immune response against HIV and other virus co-infection, and novel subsets of immune cells. LHRI maintains the Database for Annotation, Visualization and Integrated Discovery ([DAVID](#)).

(1) [Scientist I - Virology position](#) available to perform the defective proviral study in our [Basic Research Section](#).

(2) [Scientist-Cytokines and HIV](#) available in our [Basic Research Section](#). We are looking for a cytokine immunologist who is interested in virus (HIV/ HSV/KHSV) pathogenesis in myeloid immune cell types (macrophages, dendritic cells and microglia cells).

(3) [Postdoctoral Fellow](#) available in our [Basic Research Section](#). This position is an excellent opportunity for a young Ph.D. who has no experience in virus research and seeks a career in a new research field. You will learn how to handle infectious RNA viruses and investigate the mechanism of virus infection and the interaction of host cell proteins using HIV (lentivirus) variants under an SOP following the NIAID guideline in the BSL2* laboratory. LHRI does not consider virus experience in the past, instead seeks a highly motivated researcher. The knowledge learned in the lab can apply to future COIVD or other virus studies.

(4) [Computational Scientist - Bioinformatics position](#) available to develop bioinformatics analysis pipelines and manage independent research projects in our [Bioinformatics Section](#).

DAVID使用

Analysis Wizard

Tell us how you like the tool
Contact us for questions

Step 1. Submit your gene list through left panel.

第一步：导入基因列表/文件

An example:

Copy/paste IDs to "box A" -> Select Identifier as "Affy_ID" -> List Type as "Gene List" -> Click "Submit" button

第二步：选择ID类型

第三步：基因列表/背景

第四步：提交运行

DAVID Analysis Wizard Step-by-Step Guide

- Step 1: Upload Gene List (Left Panel)
 - A: Paste a list (Text area)
 - B: Choose From a File (File browser)
 - Multi-List File (checkbox)
- Step 2: Select Identifier (Dropdown menu)
 - OFFICIAL_GENE_SYMBOL
 - MGI_ID
 - MIRBASE_ID
 - MRNA_GI_ACCESSION
 - NASONIABASE_ID
 - PROTEIN_GI_ACCESSION
 - PSEUDOCAP_ID
 - REFSEQ_MRNA
 - REFSEQ_PROTEIN
 - RGD_ID
 - SGD_ID
 - TAIR_ID
 - UNIGENE
 - UNIPROT_ACCESSION
 - UNIPROT_ID
 - VECTORBASE_ID
 - WORMBASE_GENE_ID
 - XENBASE_ID
 - ZFIN_ID
 - Not Sure
- Step 3: List Type (Radio buttons)
 - Gene List (selected)
 - Background
- Step 4: Submit List (Button)

geneList.txt - 记事本

文件(F) 编辑(E) 格式(O) 查看(V)
帮助(H)

名称

- DEA.R
- enrichment.R
- geneList.txt (highlighted)
- shell.txt
- WGCNA.R

geneList.txt 修改日期: 2018/1/9 10:56
文本文档 大小: 6.33 KB

已选择 1 项

AT5G59320
AT5G09530
AT5G52300
AT2G19900
AT1G52690
AT1G04560
AT2G46680
AT2G42560
AT5G66400
AT3G02480
AT1G43160
AT3G57020
AT1G29090
AT5G53870
AT5G06760
AT4G33550
AT2G47770
AT5G43840
AT5G17460
AT2G37870

DAVID使用

DAVID BIOINFORMATICS DATABASE

Home Start Analysis Shortcut to DAVID Tools

Upload List Background

Gene List Manager

Select to limit annotations by one or more species [Help](#)

- Use All Species -
Arabidopsis thaliana(579)

Select Species

List Manager Help

tair

Select List to:
Use Rename
Remove Combine
Show Gene List

Annotation Summary Results

Current Gene List: tair
Current Background: Arabidopsis thaliana

Functional Categories (0 selected)
Gene_Ontology (1 selected)
General_Annotations (0 selected)
Literature (0 selected)
Main_Accessions (0 selected)
Pathways (1 selected)
Protein_Domains (0 selected)
Protein_Interactions (0 selected)
Tissue_Expression (0 selected)

Gene_Ontology (1 selected)
GOTERM_BP_1 72.4% 419
GOTERM_BP_2 72.4% 419
GOTERM_BP_3 71.5% 414
GOTERM_BP_4 66.3% 384
GOTERM_BP_5 61.3% 355
GOTERM_BP_ALL 72.4% 419
GOTERM_BP_DIRECT 72.4% 419

Pathways (1 selected)
EC_NUMBER 28.3% 164
KEGG_PATHWAY 17.3% 100
REACTOME_PATHWAY 8.6% 50

*** If you are looking for help, click [here](#).

*** Red annotation categories denote DAVID defined defaults ***

Combined View for Selected Annotation
[Functional Annotation Clustering](#)
[Functional Annotation Chart](#)
[Functional Annotation Table](#)

选择功能类型（GO, KEGG）

Step 2. Analyze above gene list with one of DAVID tools

Which DAVID tools to use?

Functional Annotation Tool

Functional Annotation Clustering
Functional Annotation Chart
Functional Annotation Table

Gene Functional Classification Tool
Gene ID Conversion Tool
Gene Name Batch Viewer

功能分析

DAVID使用

Annotation Summary Results

Current Gene List: tair

Current Background: Arabidopsis thaliana

- + Functional_Categories (0 selected)
- + Gene_Ontology (1 selected)
- + General_Annotations (0 selected)
- + Literature (0 selected)
- + Main_Accessions (0 selected)
- + Pathways (1 selected)
- + Protein_Domains (0 selected)
- + Protein_Interactions (0 selected)
- + Tissue_Expression (0 selected)

***Red annotation categories denote DAVID defined

Combined View for Selected Annotation

Functional Annotation Clustering

Functional Annotation Chart

Functional Annotation Table

Functional Annotation Clustering

Current Gene List: tair
Current Background: Arabidopsis thaliana
579 DAVID IDs

Options Classification Stringency Medium ▾
Rerun using options Create Sublist

9 cluster(s)

Annotation Cluster	Enrichment Score	G	RT	Count	P_Value	Benjamini
Annotation Cluster 1	4.82	G	RT	6	1.2E-7	1.0E-5
GOTERM_BP_DIRECT	seed oilbody biogenesis	RT	RT	7	2.5E-5	1.6E-3
GOTERM_BP_DIRECT	lipid storage	RT	RT	5	1.1E-3	3.3E-2
Annotation Cluster 2	3.6	G	RT	14	1.7E-5	5.2E-4
KEGG_PATHWAY	Phenylpropanoid biosynthesis	RT	RT	9	9.1E-4	3.3E-2
GOTERM_BP_DIRECT	hydrogen peroxide catabolic process	RT	RT	17	1.0E-3	3.4E-2
GOTERM_BP_DIRECT	response to oxidative stress	RT	RT	5	1.1E-4	5.4E-3
Annotation Cluster 3	2.83	G	RT	8	1.4E-4	6.2E-3
GOTERM_BP_DIRECT	response to water	RT	RT	4	2.0E-1	7.8E-1
GOTERM_BP_DIRECT	cold acclimation	RT	RT			
GOTERM_BP_DIRECT	response to stress	RT	RT			

Functional Annotation Chart

Current Gene List: tair
Current Background: Arabidopsis thaliana
579 DAVID IDs

Options

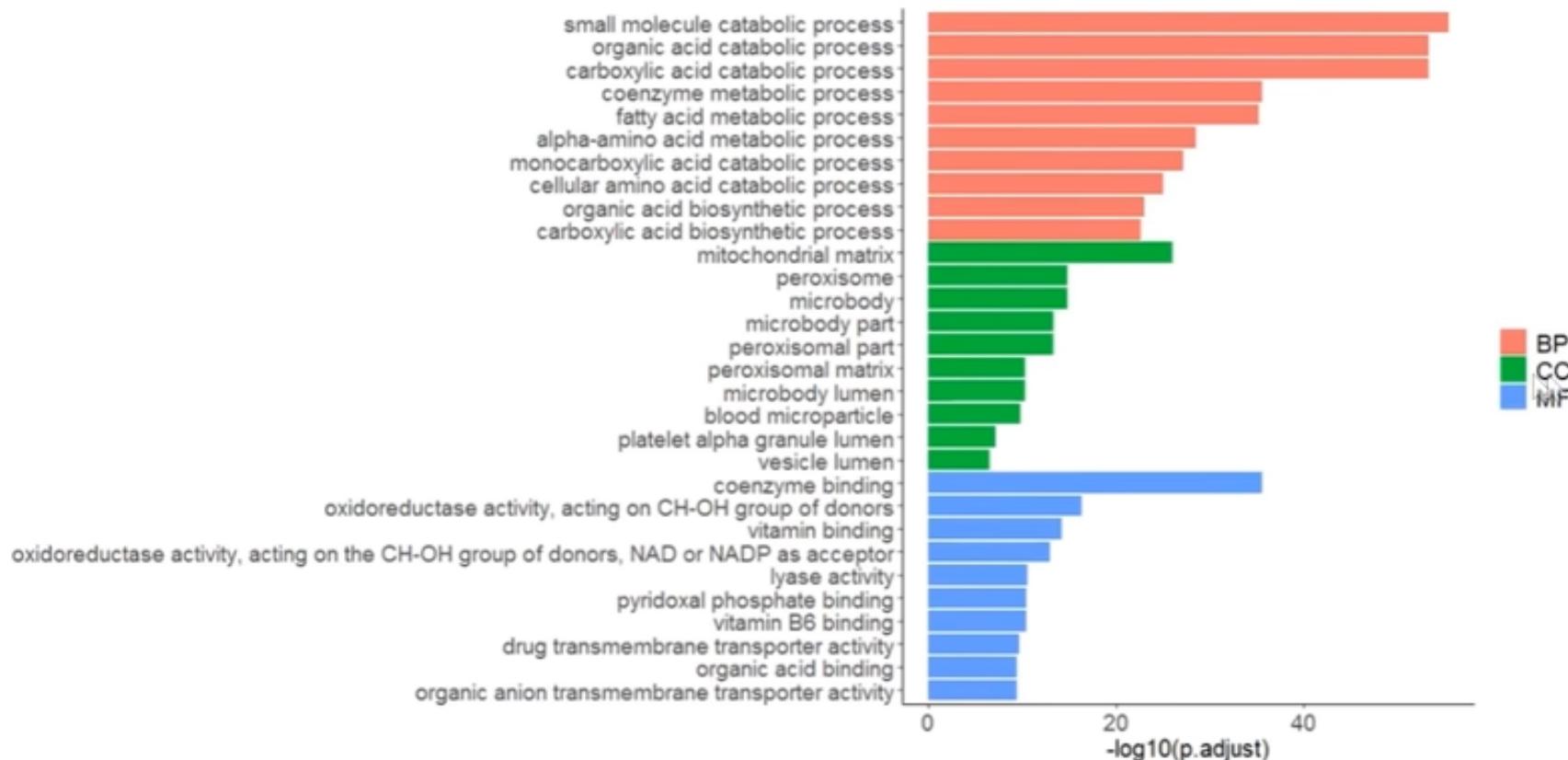
Rerun Using Options Create Sublist

1 chart records

Sublist	Category	Term	RT	Genes	Count	%	P_Value	Benjamini
	GOTERM_BP_DIRECT	response to abscisic acid	RT	50	8.6	1.2E-22	5.5E-20	
	GOTERM_BP_DIRECT	response to water deprivation	RT	39	6.7	3.9E-19	8.7E-17	
	GOTERM_BP_DIRECT	oxidation-reduction process	RT	64	11.1	1.6E-8	2.4E-6	
	KEGG_PATHWAY	Cutin, suberine and wax biosynthesis	RT	9	1.6	4.1E-8	2.5E-6	
	GOTERM_BP_DIRECT	response to osmotic stress	RT	16	2.8	1.1E-7	1.2E-5	
	GOTERM_BP_DIRECT	seed oilbody biogenesis	RT	6	1.0	1.2E-7	1.0E-5	
	GOTERM_BP_DIRECT	response to salt stress	RT	31	5.4	6.6E-7	4.9E-5	

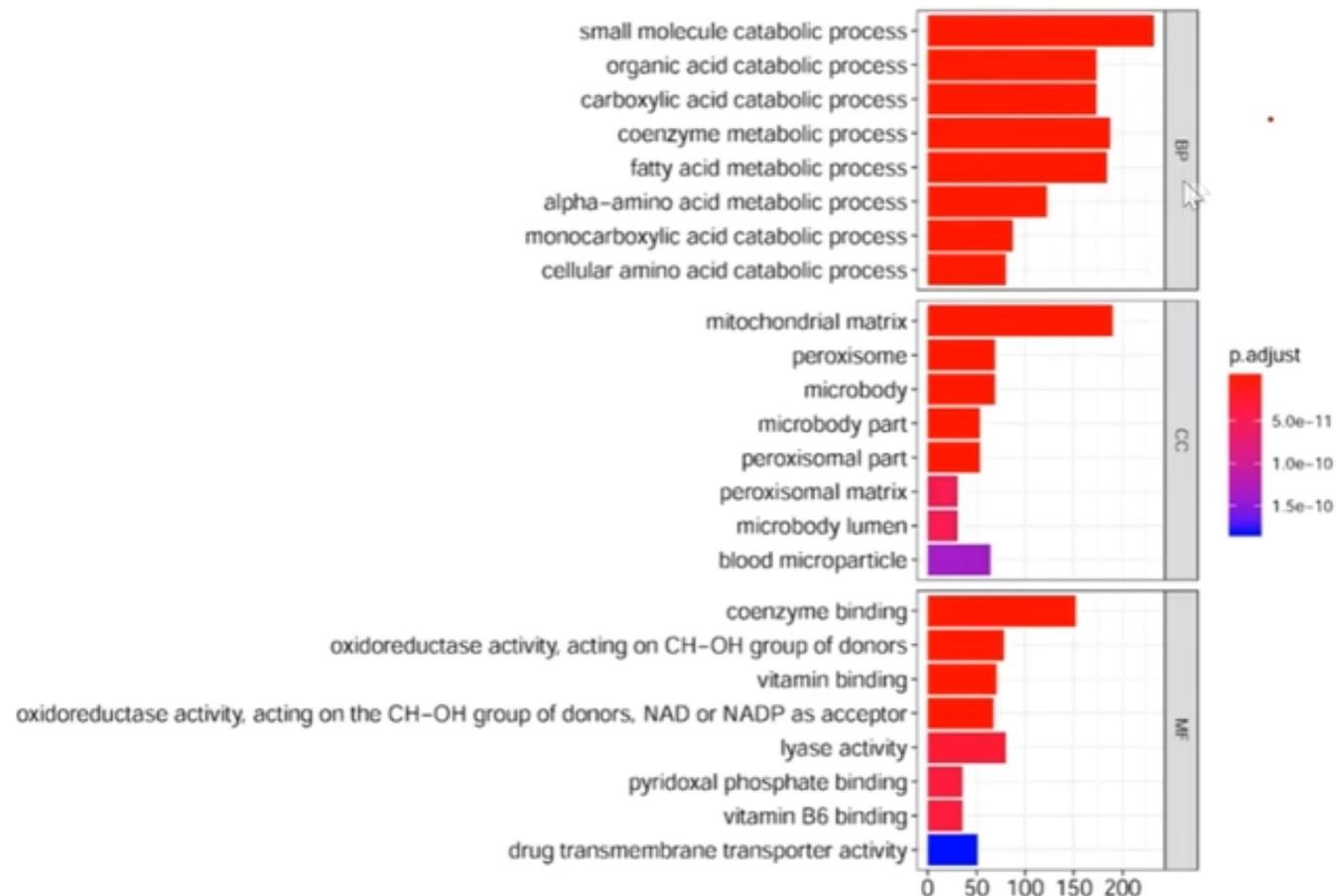
GO分析结果展示

GO富集分析-柱形图1

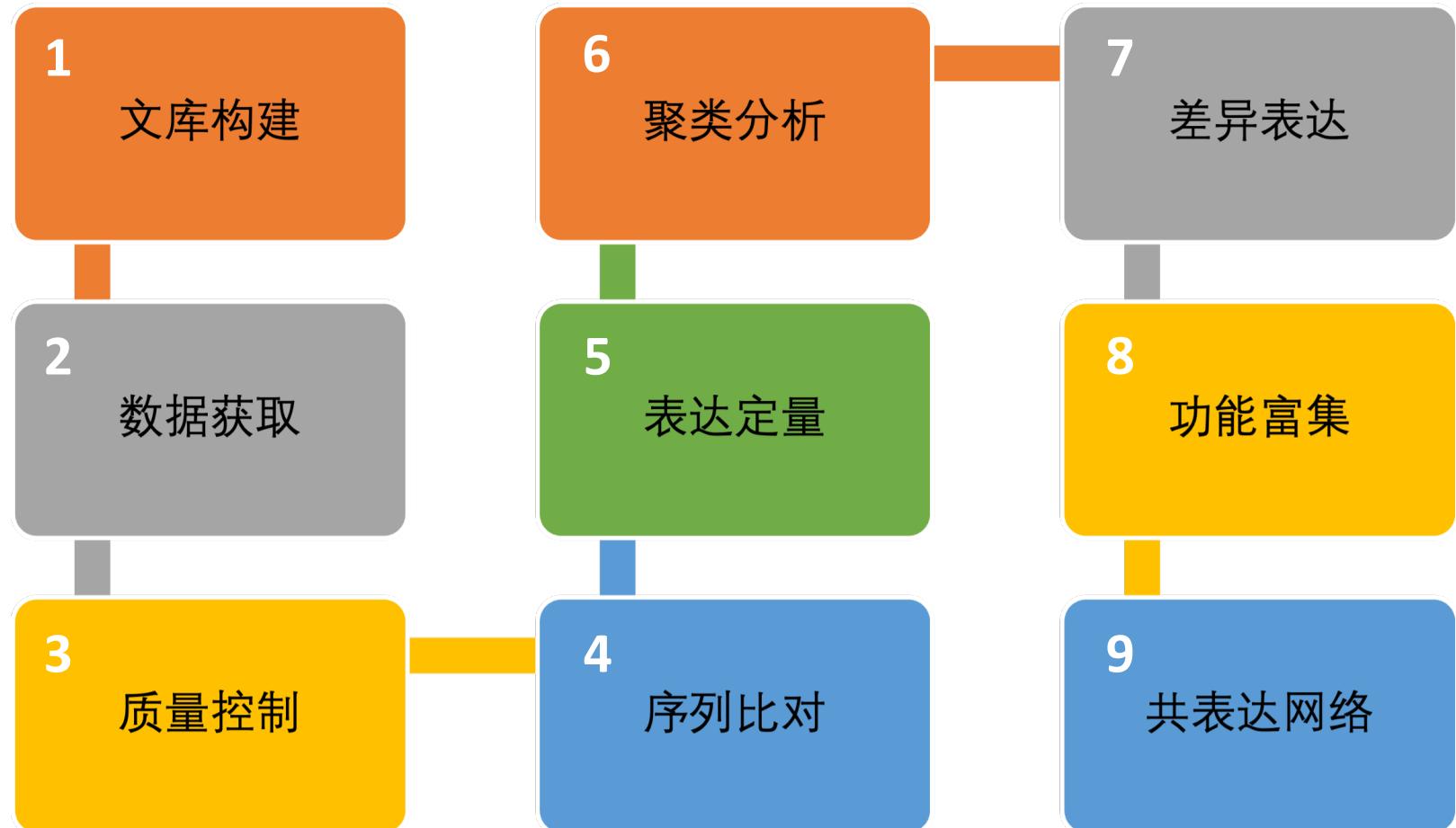


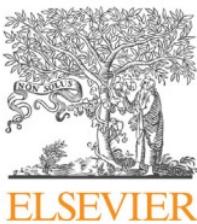
GO分析结果展示

GO富集分析-柱形图2



转录组测序技术RNA-seq常规流程





Biomaterials 91 (2016) 11–22

Contents lists available at ScienceDirect

Biomaterials

journal homepage: www.elsevier.com/locate/biomaterials



Engineering EMT using 3D micro-scaffold to promote hepatic functions for drug hepatotoxicity evaluation



Jingyu Wang ^a, Fengling Chen ^b, Longwei Liu ^c, Chunxiao Qi ^a, Bingjie Wang ^c,
Xiaojun Yan ^a, Chenyu Huang ^d, Wei Hou ^e, Michael Q. Zhang ^b, Yang Chen ^{b, **},
Yanan Du ^{a, f, *}

^a Department of Biomedical Engineering, School of Medicine, Collaborative Innovation Center for Diagnosis and Treatment of Infectious Diseases, Tsinghua University, Beijing, 100084, PR China

^b MOE Key Laboratory of Bioinformatics, Bioinformatics Division & Center for Synthetic and Systems Biology, Department of Automation, TNLIST, Tsinghua University, Beijing, 100084, PR China

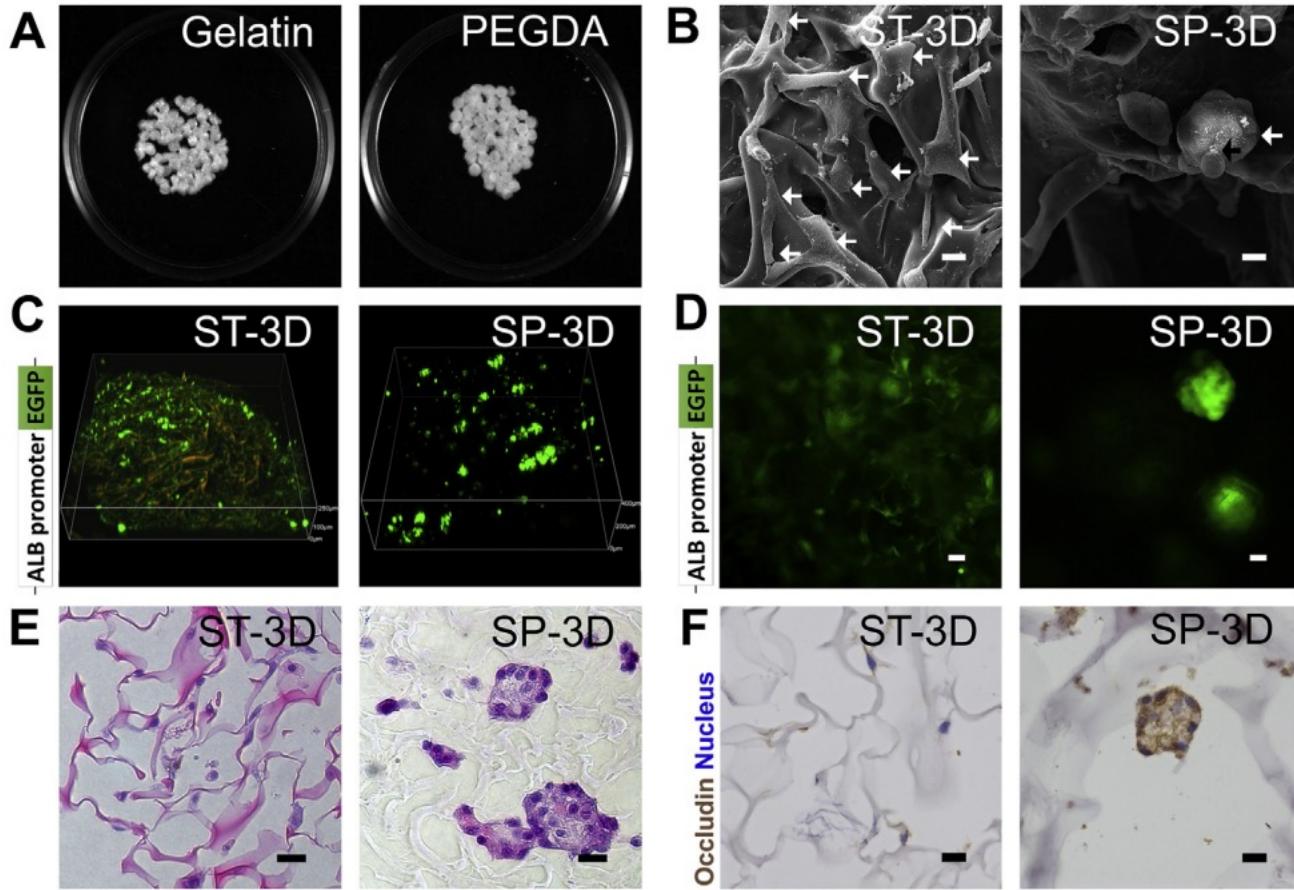
^c The School of Life Science, Tsinghua University, Beijing, 100084, PR China

^d Department of Plastic, Reconstructive and Aesthetic Surgery, Beijing Tsinghua Changgung Hospital, Medical Center, Tsinghua University, Beijing, 102218, PR China

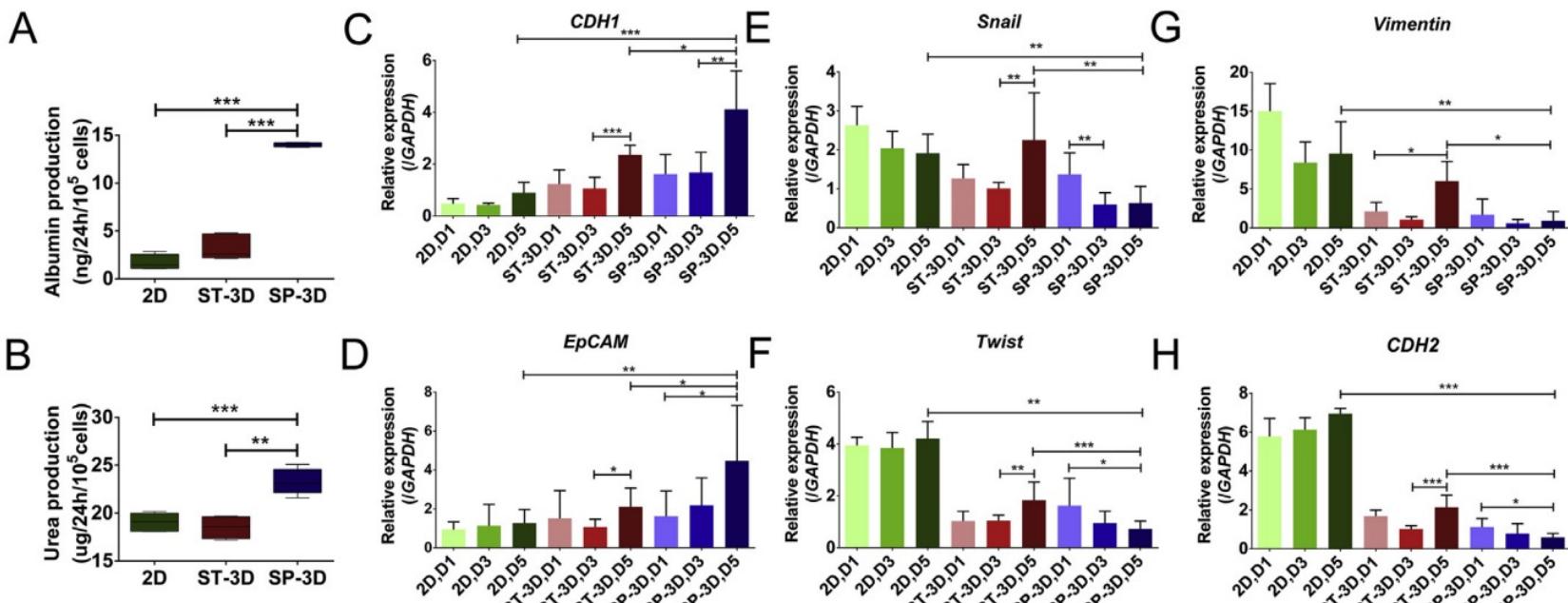
^e Tianjin Second People's Hospital and Tianjin Institute of Hepatology, Tianjin, 300192, PR China

^f China Orthopedic Regenerative Medicine Group (CORMed), Hangzhou, Zhejiang, 310058, PR China

不同3D培养条件下细胞群生长有差异

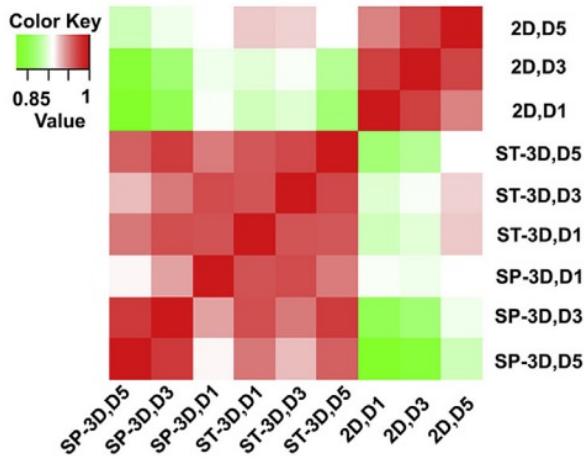


不同3D培养条件下细胞功能有差异

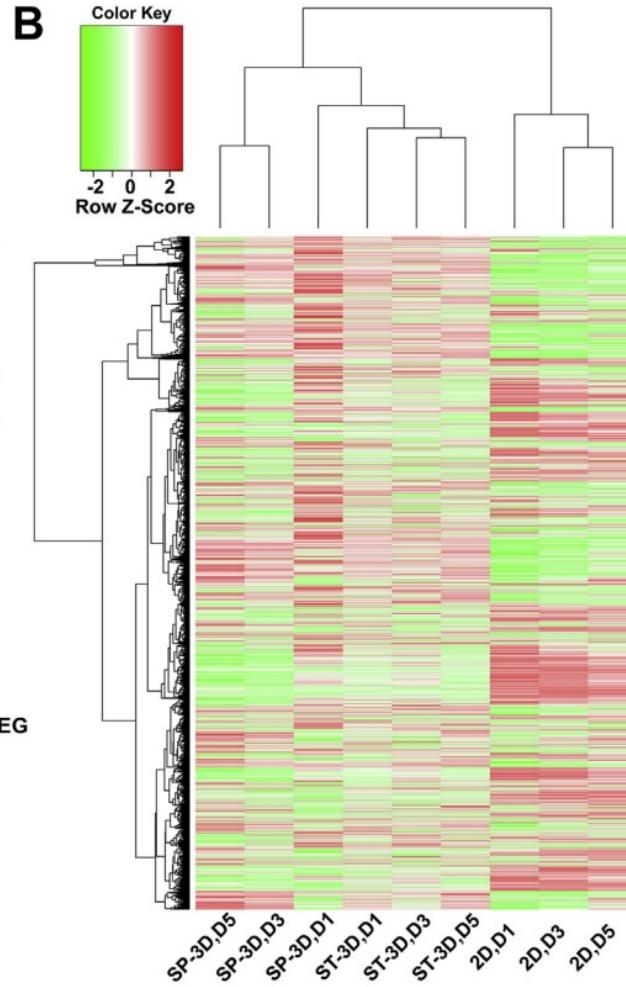


不同3D培养条件下细胞的转录组差异

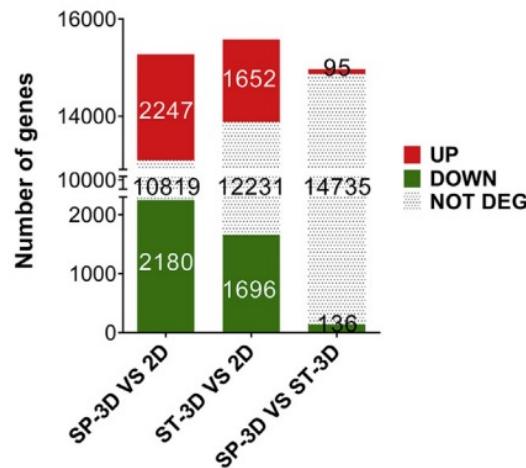
A



B



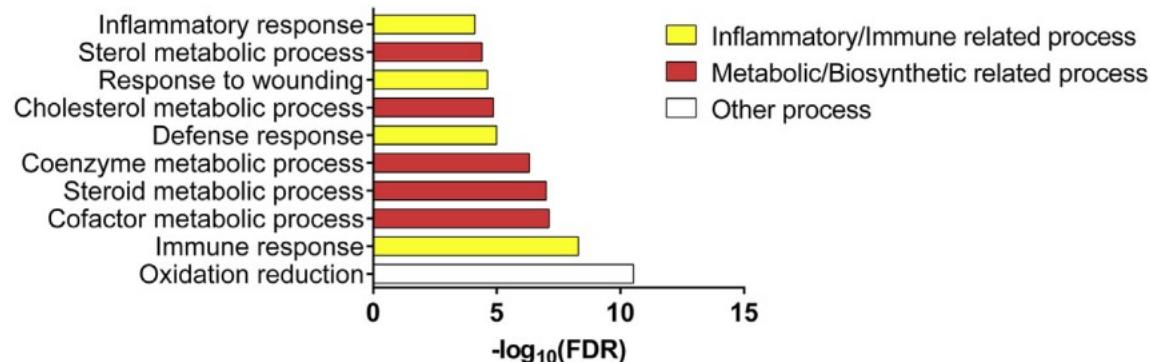
C



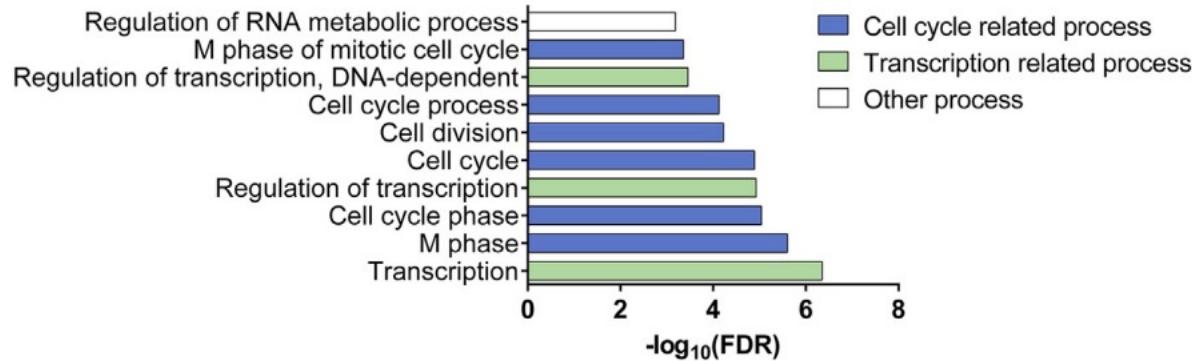
差异表达基因的功能富集分析

D

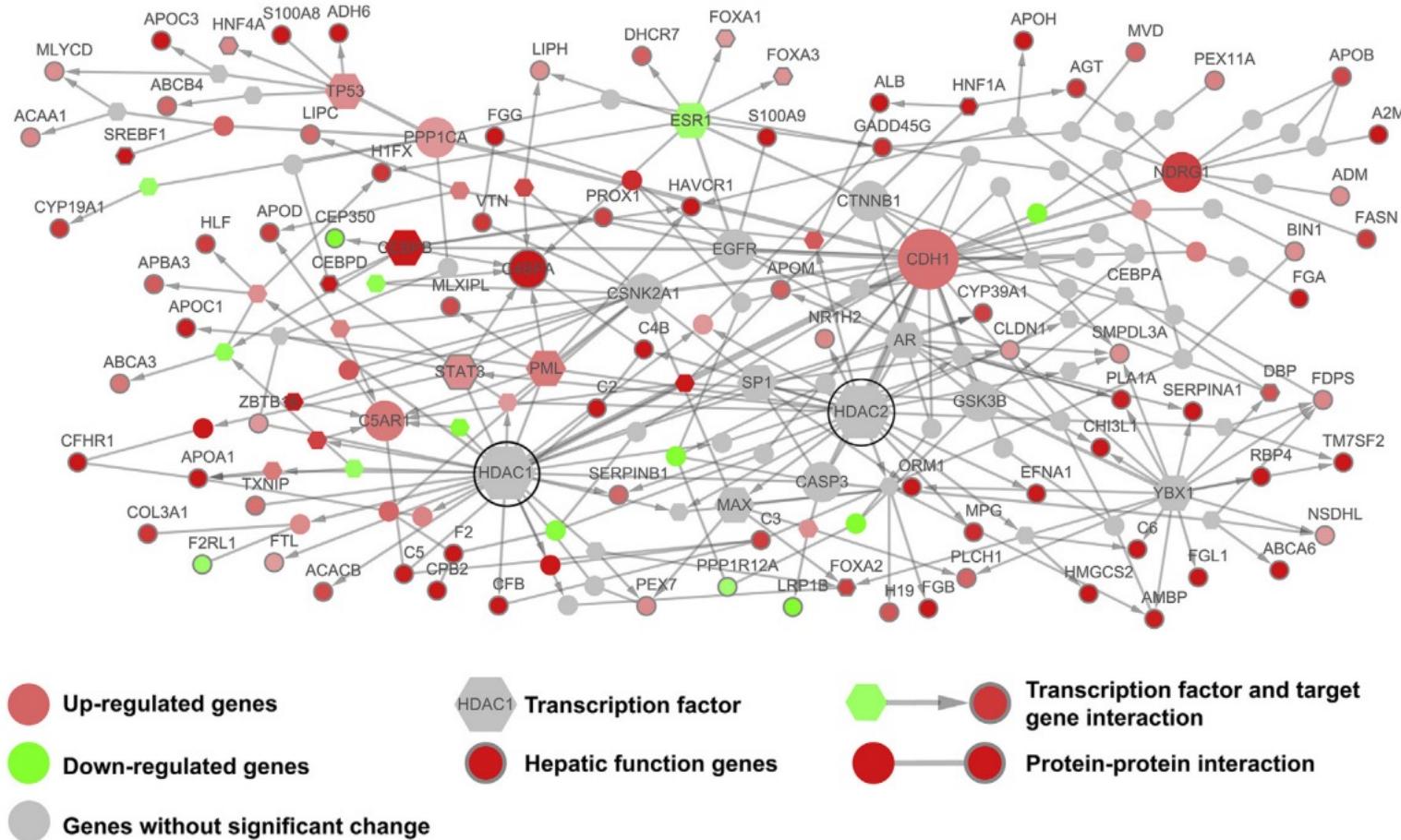
Top 10 GO terms of up-regulated DEGs (SP-3D VS 2D)



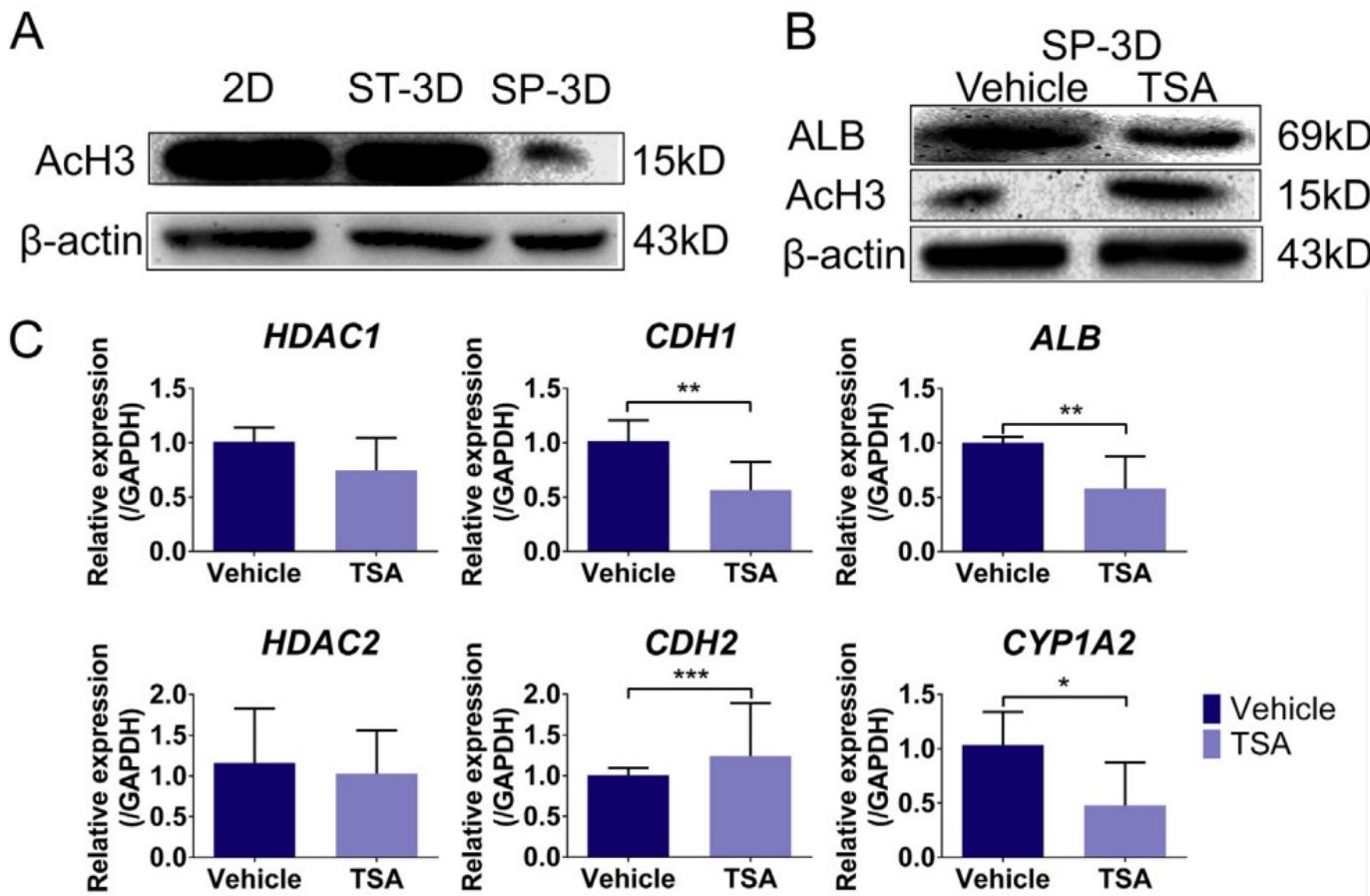
Top 10 GO terms of down-regulated DEGs (SP-3D VS 2D)



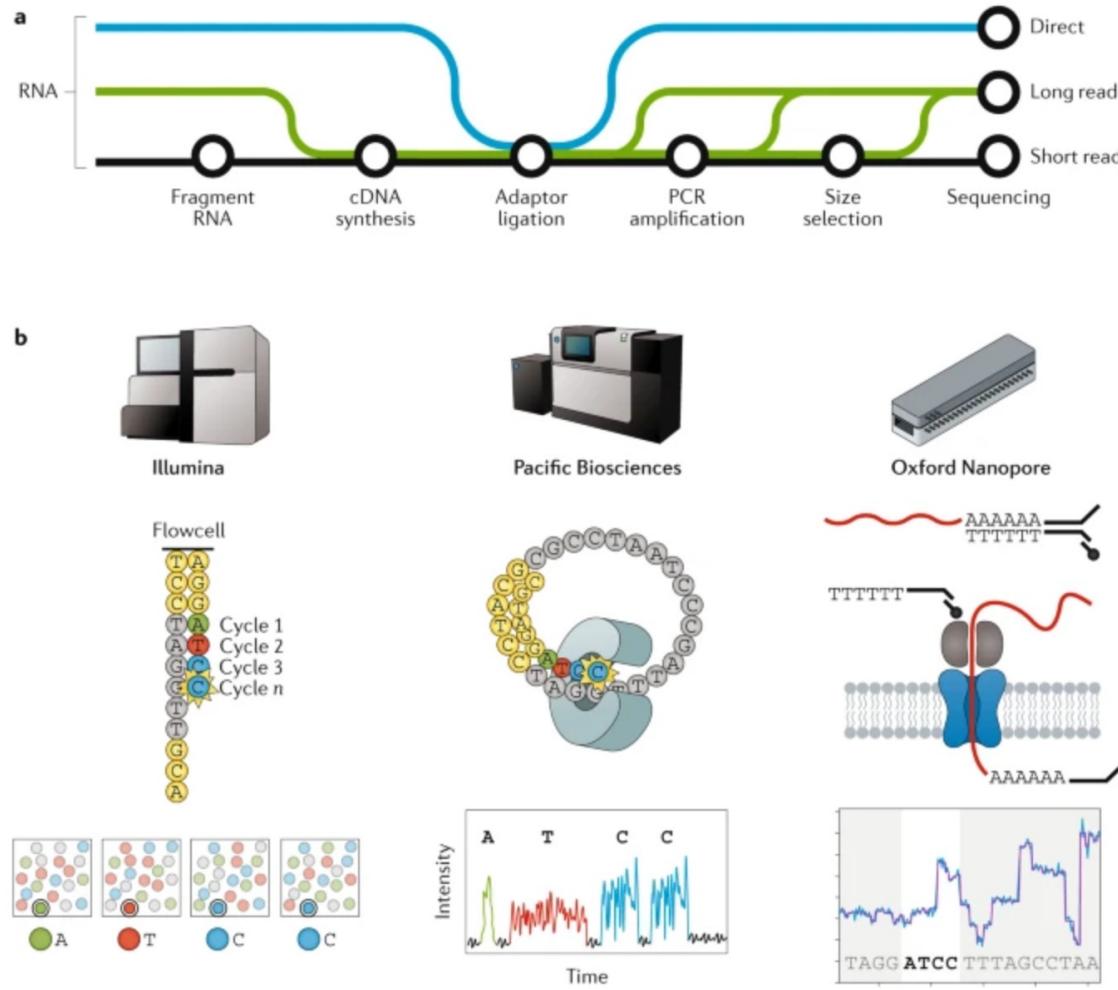
差异表达基因相关的分子相互作用网络



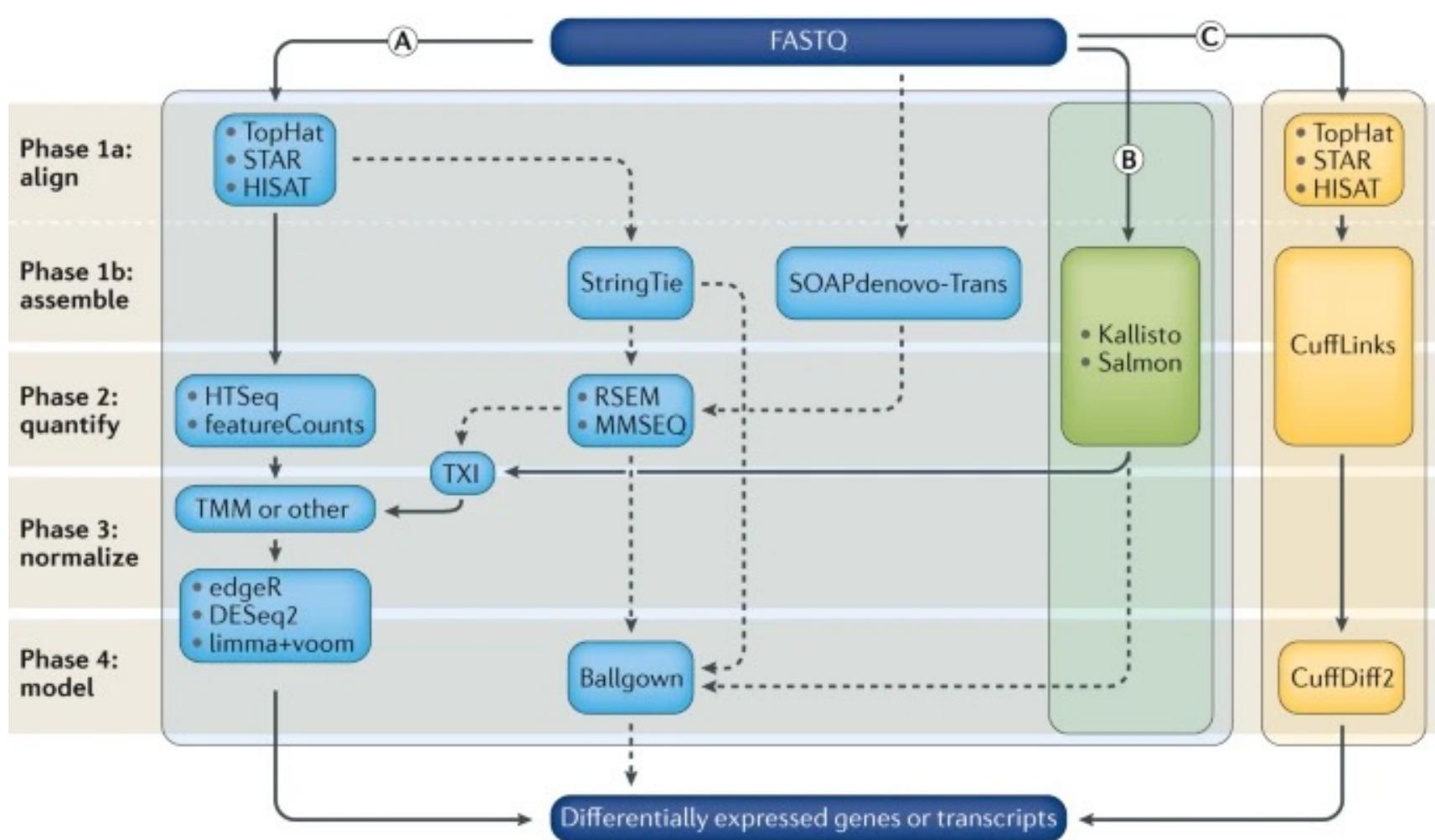
关键调控基因的分子功能验证



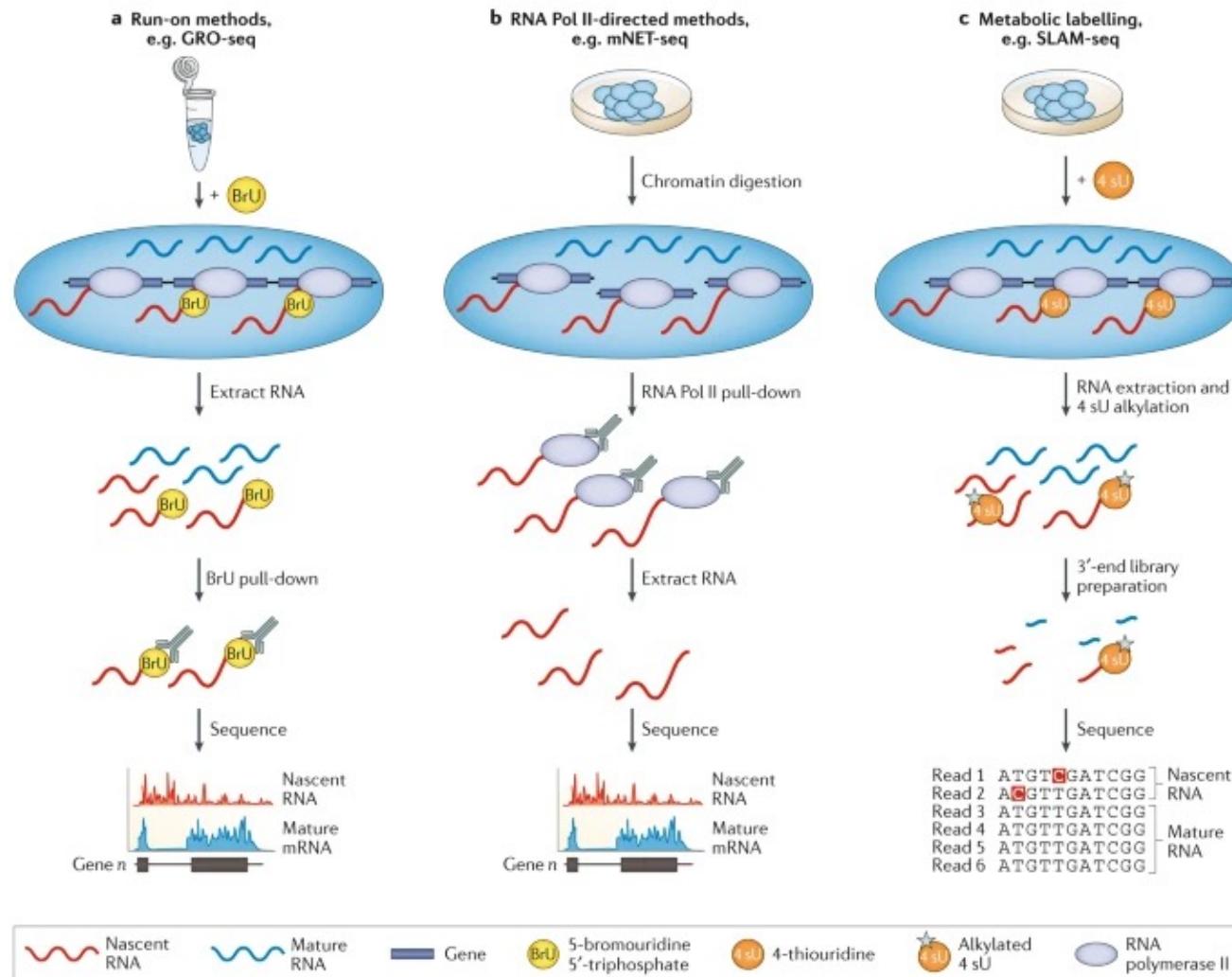
Advances in RNA-seq technologies



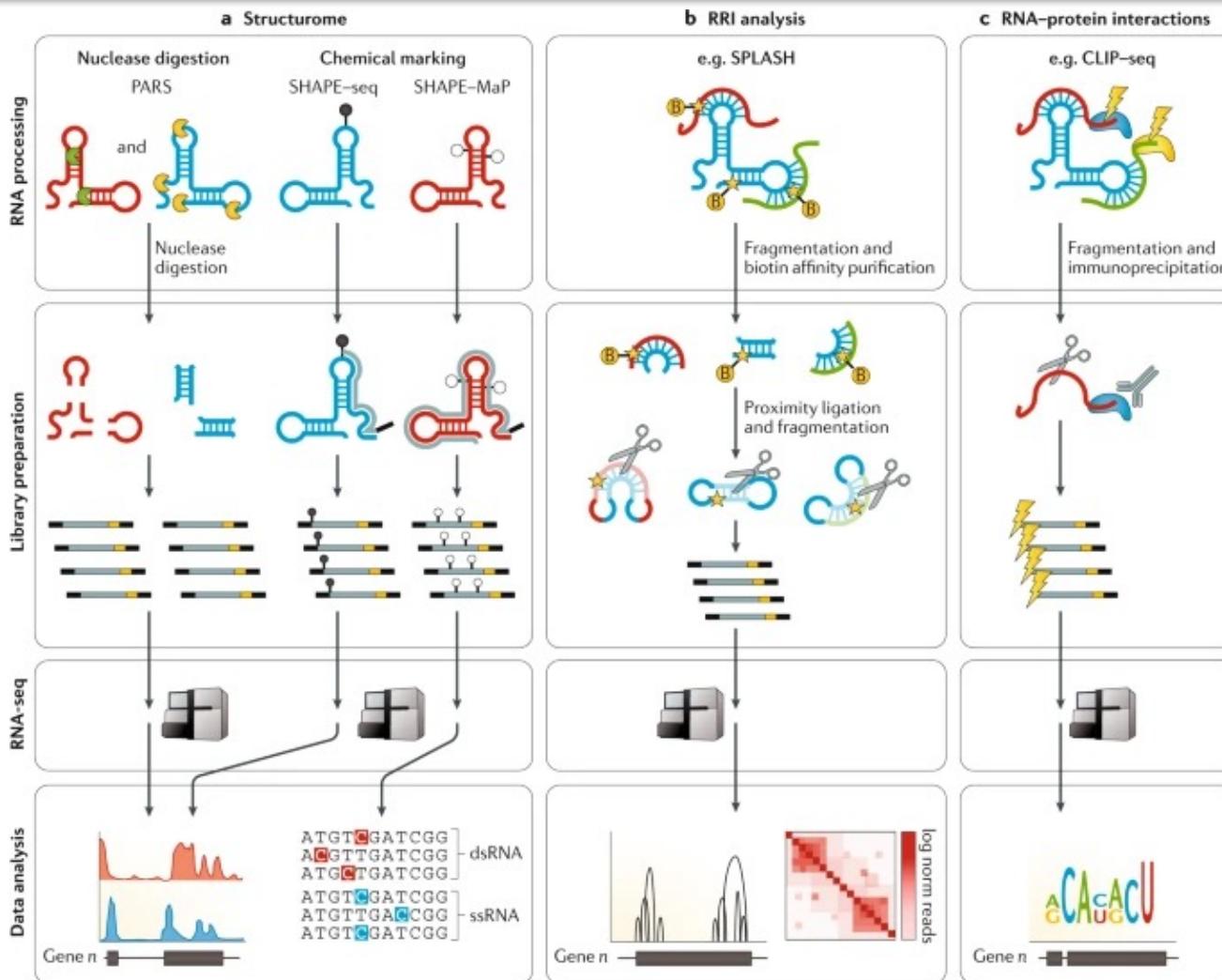
RNA-seq data analysis workflow for differential gene expression.



The key concepts of nascent RNA and translatome analysis



The key concepts of RNA structure and RNA–protein interaction analysis



Thanks for attention!