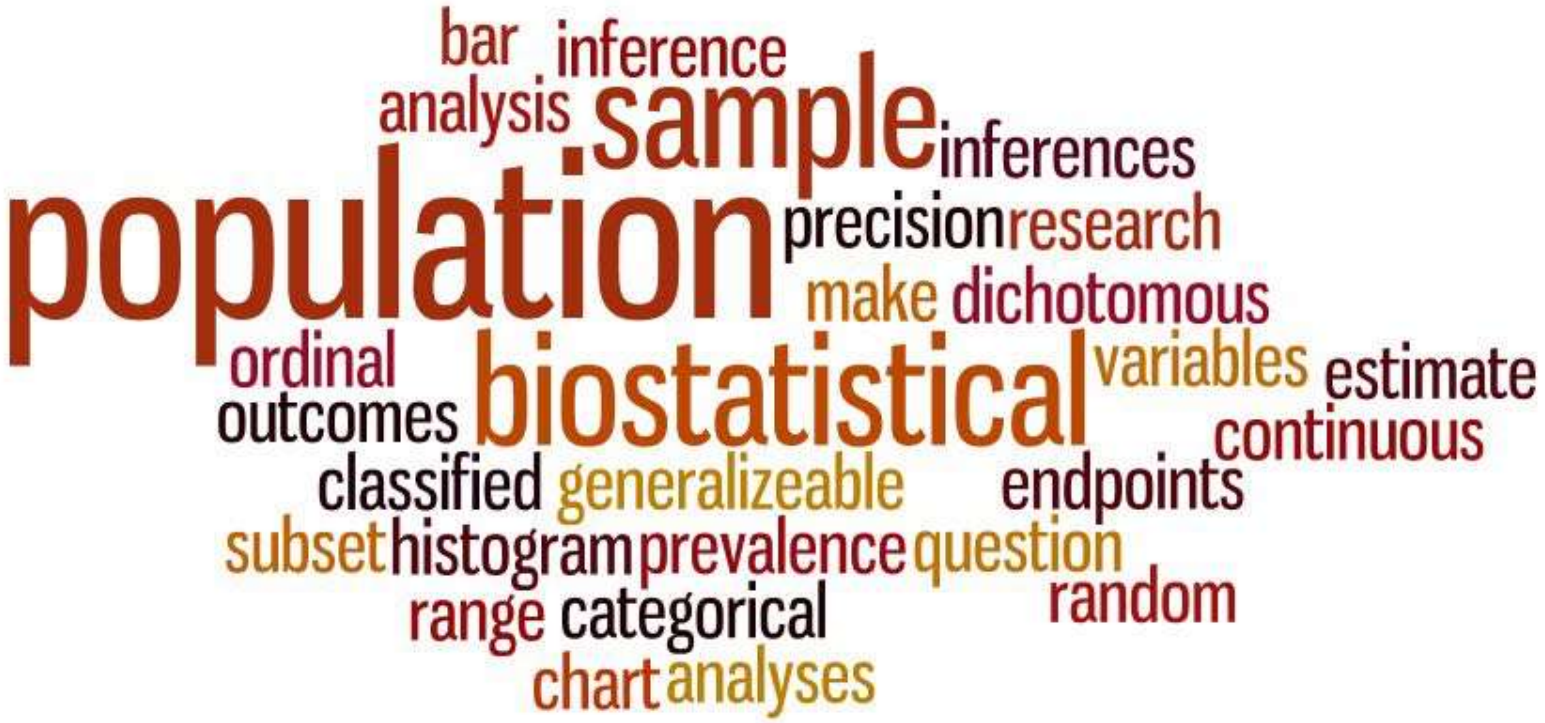


Biostatistics

Yu Zhang, 18110700014@fudan.edu.cn Yuhao Feng, 18210700100@fudan.edu.cn

Oct 25, 2018

Outline

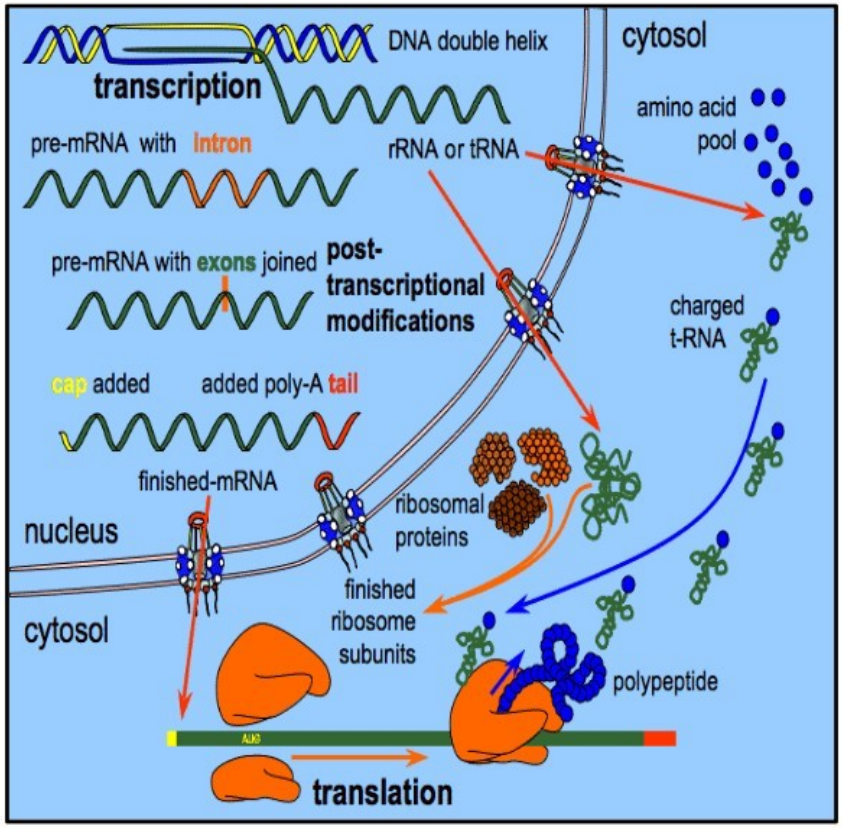
- Introduction to RNA-seq
- Pre-analysis
- Core-analysis
- Advanced-analysis

Introduction to RNA-seq

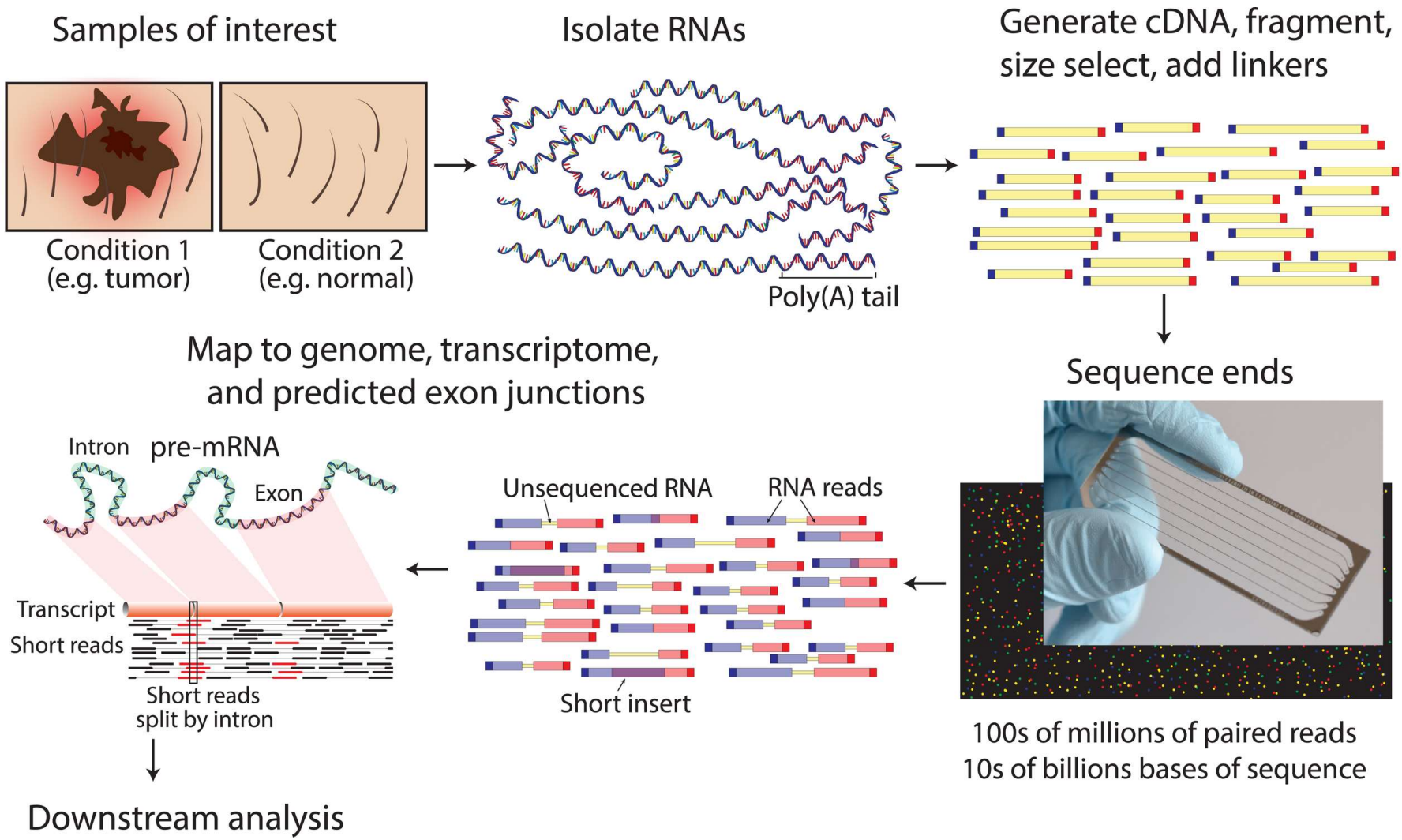
RNA-seq uses NGS to reveal the presence and quantity of RNA in a biological sample at a given moment in time

It has wide application.....

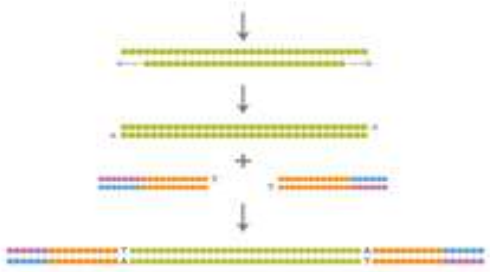
- Differential gene analysis
  - Healthy vs. diseased
  - Time course experiments
  - Different genotypes
- Transcriptional profiling
  - Tissue-specific expression
  - Transcription assembly
- Novel gene identification
- Splice variants identification
- SNP finding
- RNA editing



Overview of RNA-Seq



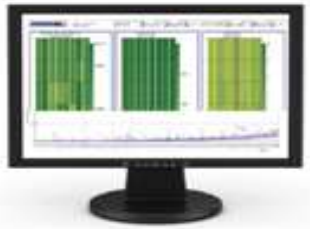
Illumina platform



Library Preparation
~2 h [15 min hands-on (Nextera)]
< 6 h [< 3 h hands-on (TruSeq)]

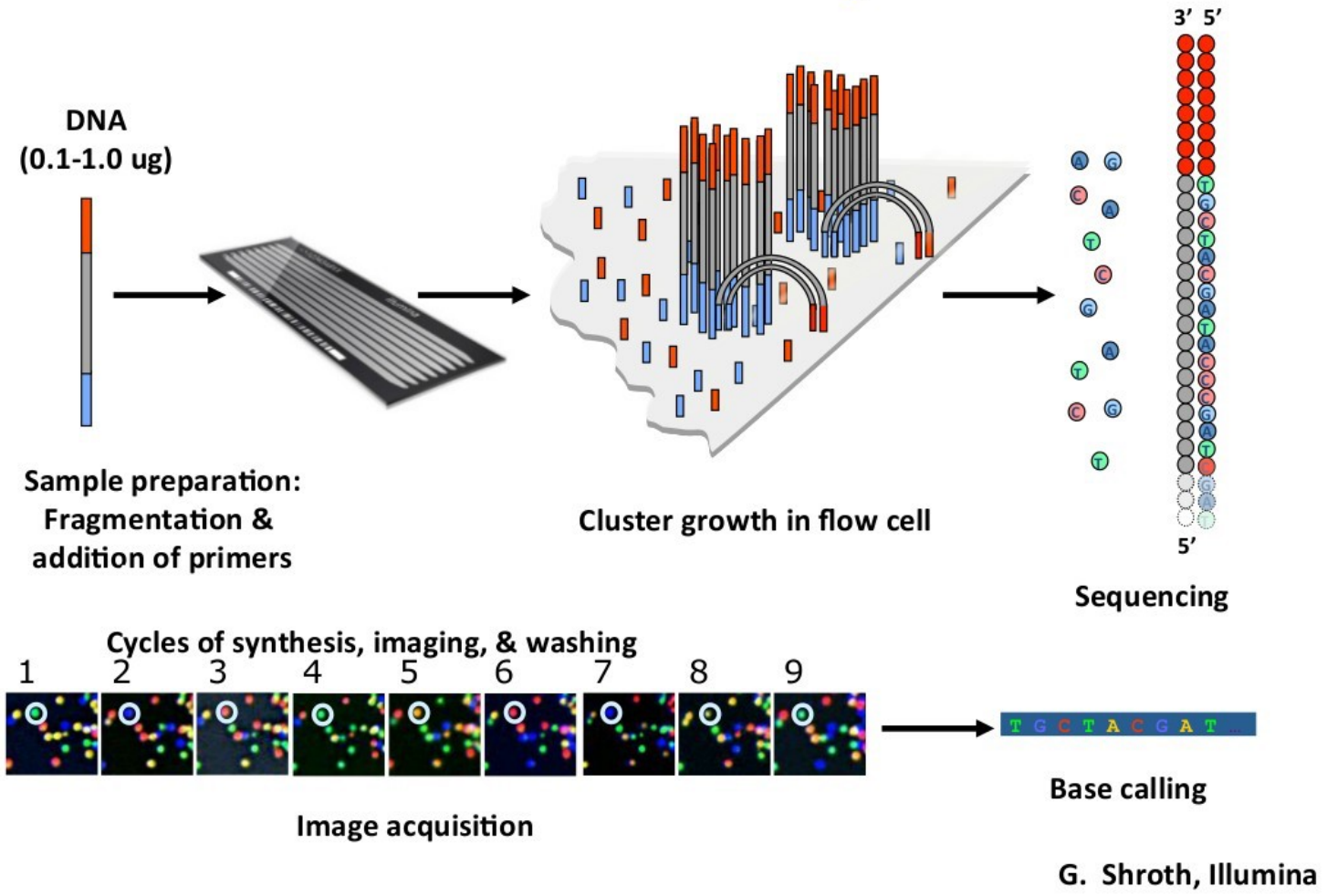
Cluster Generation
~5 h (<10 min hands-on)

Sequencing by Synthesis
~1.5 to 11 days

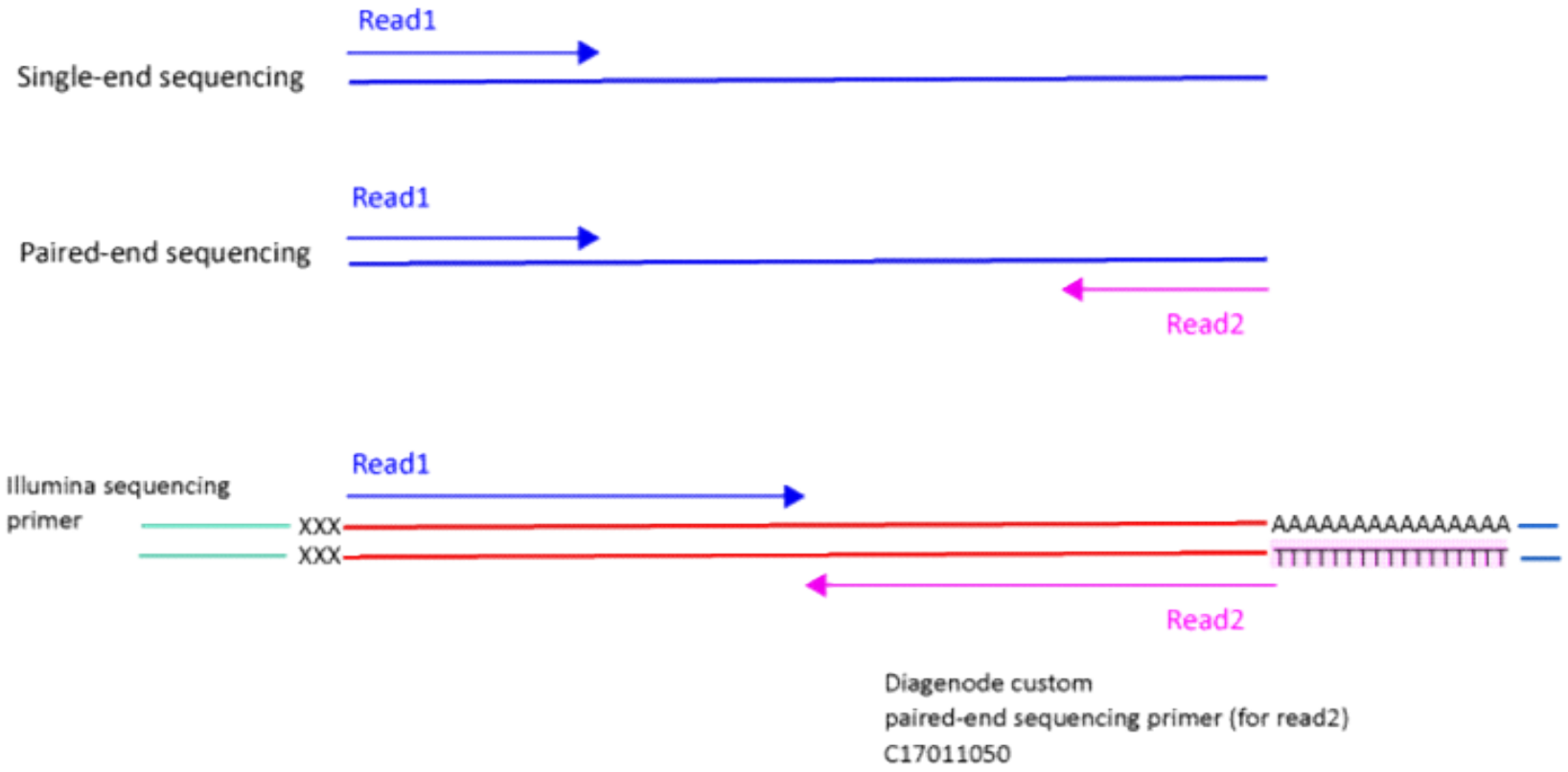


CASAVA
2 days (30 min hands-on)

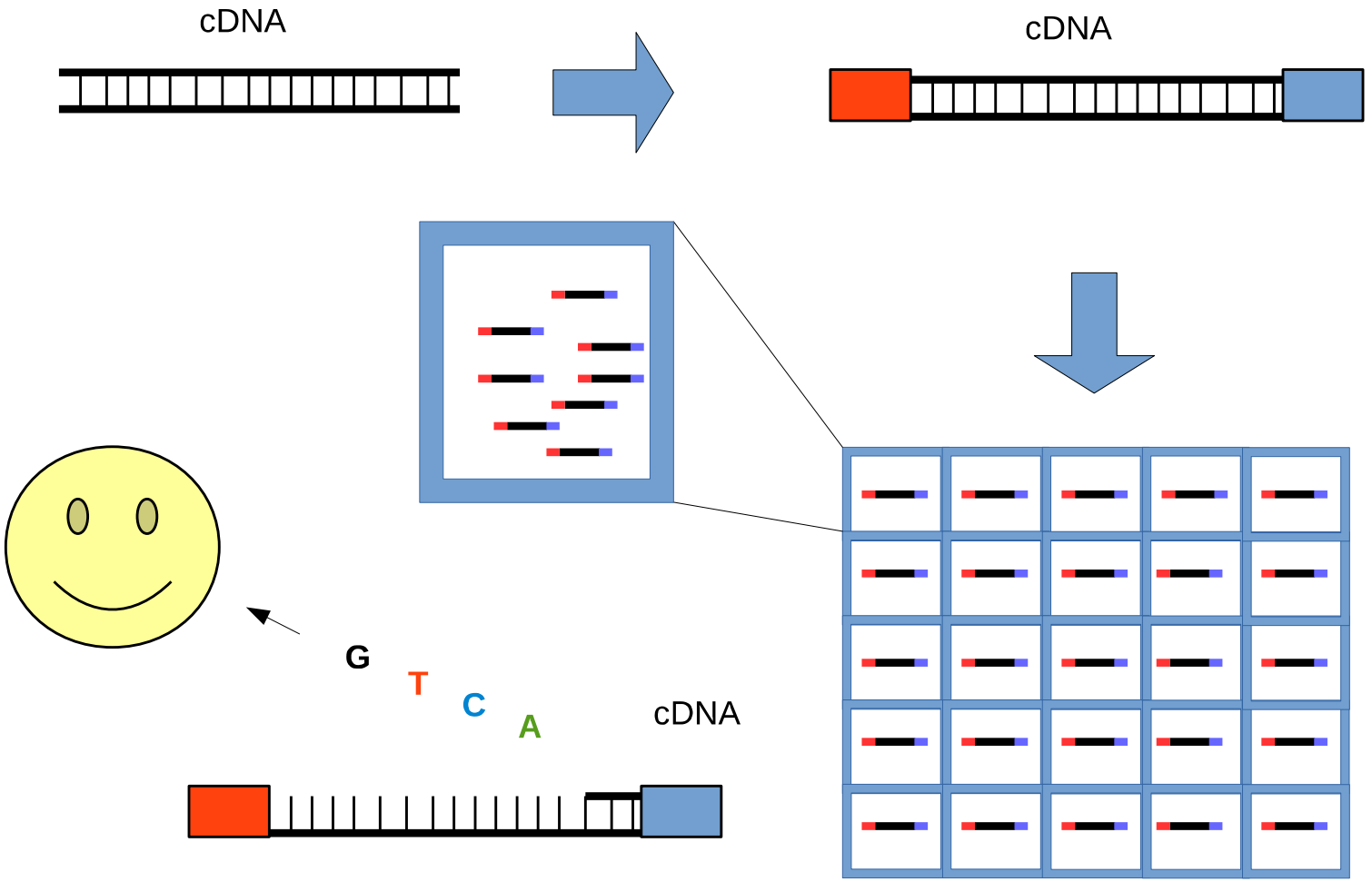
Illumina platform



Single-end sequencing and paired-end sequencing



From cDNA to FASTQ



FASTQ file

```
@ST-E00126:128:HJFLHCCXX:2:1101:7405:1133
TTGCAAAAAATTTCTCTCATTCTGTAGGTTGCCTGTTCACTCTGATGATAGTTTGTTTTGG
+
FFKKKFKKFKF<KK<F,AFKKKKK7FFK77<FKK,<F7K,,7AF<FF7FKK7AA,7<FA,,
```

← 测序信息
← 序列信息
← 附加信息
一般为空

质量信息
Phred值表示
和碱基一一对应

```
@ST-E00126:128:HJFLHCCXX:2:1101:7405:1133

@         开始的标记符号
ST-E00126:128:HJFLHCCXX 测序仪唯一的设备名称
2         lane的编号
1101      tail的坐标
7405      在tail中的X坐标
1133      在tail中的Y坐标
```

Phred quality scores

测序仪自动根据荧光信号的强弱给出参考的测序错误概率：P值

P = 0.01 # 1 base, 4 characters, from 1 to 0.0001

$Q = -10 * \log_{10} P = 20$ # from 0 to 40

P = 10 ^ -(Q / 10)

Phred = Q + 33 = 53 # usually from 33 to 73

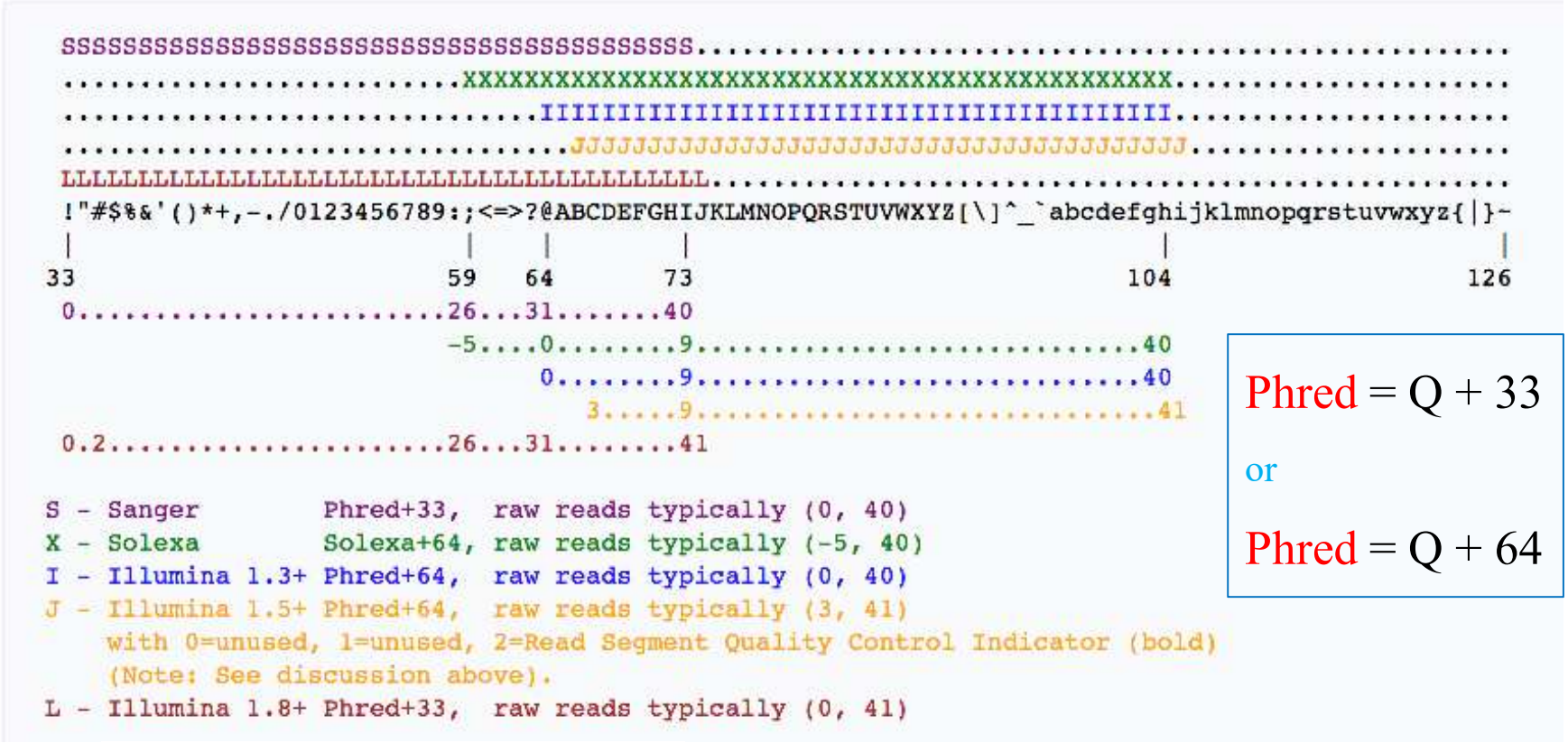
=> ASCII: " 5 " # 1 base, 1 character, from ! to I

ASCII table

Dec	Hx	Oct	Char	Dec	Hx	Oct	Html	Chr	Dec	Hx	Oct	Html	Chr	Dec	Hx	Oct	Html	Chr
0	0	000	NUL (null)	32	20	040		Space	64	40	100	@	@	96	60	140	`	`
1	1	001	SOH (start of heading)	33	21	041	!	!	65	41	101	A	A	97	61	141	a	a
2	2	002	STX (start of text)	34	22	042	"	"	66	42	102	B	B	98	62	142	b	b
3	3	003	ETX (end of text)	35	23	043	#	#	67	43	103	C	C	99	63	143	c	c
4	4	004	EOT (end of transmission)	36	24	044	$	$	68	44	104	D	D	100	64	144	d	d
5	5	005	ENQ (enquiry)	37	25	045	%	%	69	45	105	E	E	101	65	145	e	e
6	6	006	ACK (acknowledge)	38	26	046	&	&	70	46	106	F	F	102	66	146	f	f
7	7	007	BEL (bell)	39	27	047	'	'	71	47	107	G	G	103	67	147	g	g
8	8	010	BS (backspace)	40	28	050	(	(	72	48	110	H	H	104	68	150	h	h
9	9	011	TAB (horizontal tab)	41	29	051	)	)	73	49	111	I	I	105	69	151	i	i
10	A	012	LF (NL line feed, new line)	42	2A	052	*	*	74	4A	112	J	J	106	6A	152	j	j
11	B	013	VT (vertical tab)	43	2B	053	+	+	75	4B	113	K	K	107	6B	153	k	k
12	C	014	FF (NP form feed, new page)	44	2C	054	,	,	76	4C	114	L	L	108	6C	154	l	l
13	D	015	CR (carriage return)	45	2D	055	-	-	77	4D	115	M	M	109	6D	155	m	m
14	E	016	SO (shift out)	46	2E	056	.	.	78	4E	116	N	N	110	6E	156	n	n
15	F	017	SI (shift in)	47	2F	057	/	/	79	4F	117	O	O	111	6F	157	o	o
16	10	020	DLE (data link escape)	48	30	060	0	0	80	50	120	P	P	112	70	160	p	p
17	11	021	DC1 (device control 1)	49	31	061	1	1	81	51	121	Q	Q	113	71	161	q	q
18	12	022	DC2 (device control 2)	50	32	062	2	2	82	52	122	R	R	114	72	162	r	r
19	13	023	DC3 (device control 3)	51	33	063	3	3	83	53	123	S	S	115	73	163	s	s
20	14	024	DC4 (device control 4)	52	34	064	4	4	84	54	124	T	T	116	74	164	t	t
21	15	025	NAK (negative acknowledge)	53	35	065	5	5	85	55	125	U	U	117	75	165	u	u
22	16	026	SYN (synchronous idle)	54	36	066	6	6	86	56	126	V	V	118	76	166	v	v
23	17	027	ETB (end of trans. block)	55	37	067	7	7	87	57	127	W	W	119	77	167	w	w
24	18	030	CAN (cancel)	56	38	070	8	8	88	58	130	X	X	120	78	170	x	x
25	19	031	EM (end of medium)	57	39	071	9	9	89	59	131	Y	Y	121	79	171	y	y
26	1A	032	SUB (substitute)	58	3A	072	:	:	90	5A	132	Z	Z	122	7A	172	z	z
27	1B	033	ESC (escape)	59	3B	073	;	;	91	5B	133	[	[	123	7B	173	{	{
28	1C	034	FS (file separator)	60	3C	074	<	<	92	5C	134	\	\	124	7C	174			|
29	1D	035	GS (group separator)	61	3D	075	=	=	93	5D	135	]	]	125	7D	175	}	}
30	1E	036	RS (record separator)	62	3E	076	>	>	94	5E	136	^	^	126	7E	176	~	~
31	1F	037	US (unit separator)	63	3F	077	?	?	95	5F	137	_	_	127	7F	177	&#127;	DEL

Source: www.LookupTables.com

Quality scores



Phred = Q + 33

or

Phred = Q + 64

$$Solexa\ Q = -10 * \log_{10} \frac{P}{1-P}$$

$$\text{Illumina } Q = -10 * \log_{10} P$$

GEO (Gene Expression Omnibus)

- GEO is a public functional genomics data repository.
- Array- and sequence-based data are accepted.

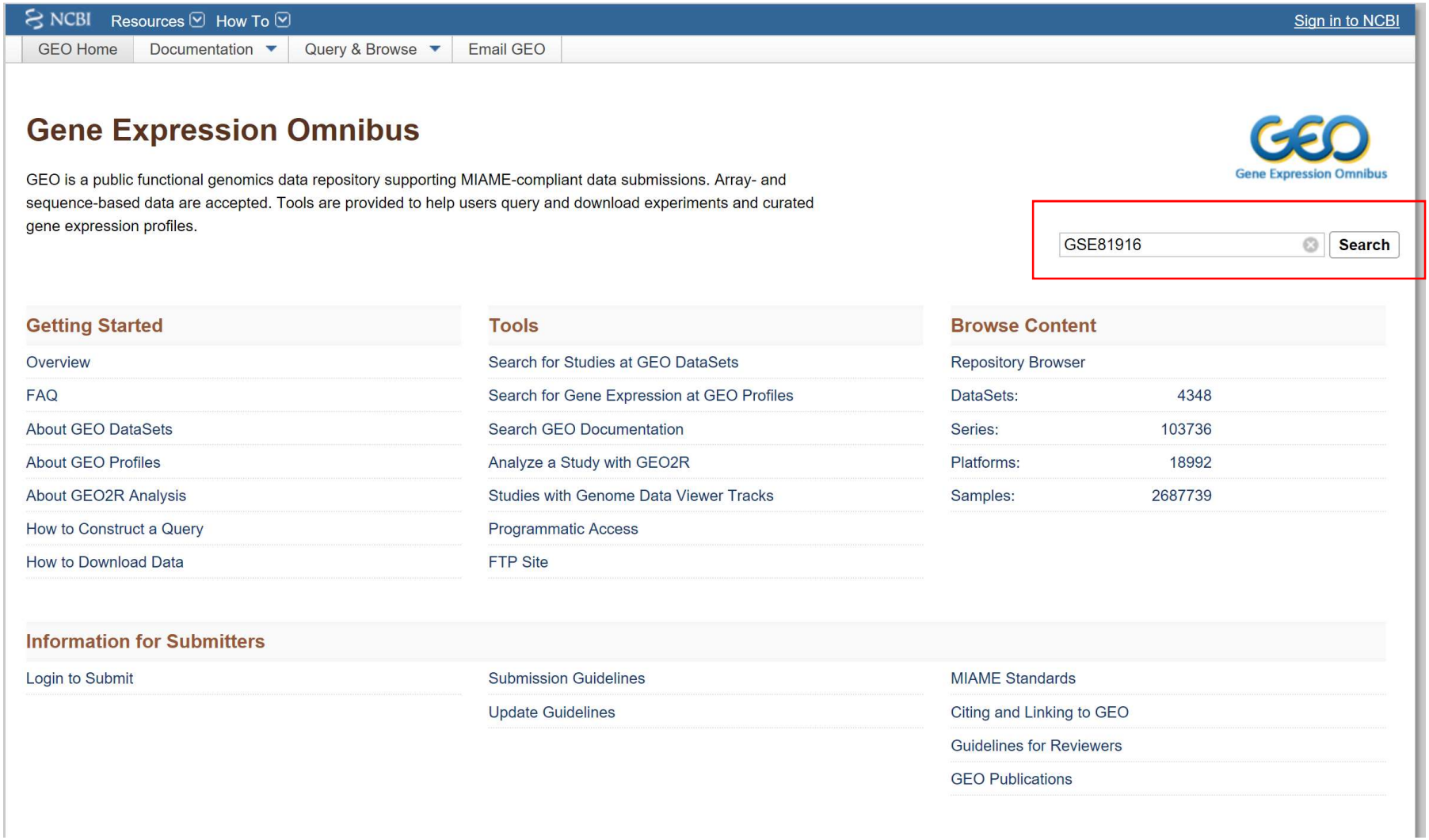
GSM	GEO Sample	SRP	for studies (ERP)
GDS	GEO Dataset	SRS	for samples
GSE	GEO Series	SRX	for experiments
GPL	GEO Platform	SRR	for runs

Get FASTQ file from GEO

Jing H, Khodadadijamayran A, Mao M, et al. **AKAP95 regulates splicing through scaffolding RNAs and RNA processing factors**[J]. Nature Communications, 2016, 7:13347.

Data availability. The RIP-seq an RNA-seq data have been deposited in the Gene Expression Omnibus database, with accession code GSE81916. All other data is available from the author upon reasonable request.

Get FASTQ file from GEO



Series GSE81916	Query DataSets for GSE81916
Status	Public on Sep 08, 2016
Title	AKAP95 regulates splicing through scaffolding RNAs and RNA processing factors
Organisms	Homo sapiens; Mus musculus
Experiment type	Expression profiling by high throughput sequencing
Summary	Alternative splicing of pre-mRNAs significantly contributes to the complexity of gene expression in higher organisms, but the regulation of the splice site selection remains incompletely understood. We have previously demonstrated that a chromatin-associated protein, AKAP95 (AKAP8), has a remarkable activity in enhancing chromatin transcription. In this study, we have shown that AKAP95 physically interacts with many factors involved in transcription and RNA processing, and functionally regulates pre-mRNA splicing. AKAP95 directly promotes splicing in vitro and the inclusion of a specific exon of an endogenous gene FAM126A. The N-terminal YG-rich domain of AKAP95 is important for its binding to RNA processing factors including selective groups of hnRNP proteins, and its zinc finger domains are critical for pre-mRNA binding. Genome-wide binding assays revealed that AKAP95 bound preferentially to proximal intronic regions on a large number of pre-mRNAs in human transcriptome, and AKAP95 depletion predominantly resulted in reduced inclusion of many exons. AKAP95 also selectively coordinates with hnRNP H/F and U proteins in regulating alternative splicing events. We have further shown that AKAP95 directly interacts with itself. Taken together, our results establish AKAP95 as a novel and mostly positive regulator of pre-mRNA splicing and a possible integrator of transcription and splicing regulation, and support a model that AKAP95 regulates pre-mRNA splicing via through scaffolding RNAs and RNA processing factors and facilitating the splice site communication.
Overall design	Samples 1-8 are RNA-immunoprecipitation (RIP)-seq to determine AKAP95 binding to the transcriptome. Samples 9-15 are mRNA-seq to determine effect of AKAP95 knockdown in human 293 cells (9-11) or mouse ES cells (12-15).
Contributor(s)	Khodadadi-Jamayran A, Hu J, Jiang H
Citation(s)	Hu J, Khodadadi-Jamayran A, Mao M, Shah K et al. AKAP95 regulates splicing through scaffolding RNAs and RNA processing factors. *Nat Commun* 2016 Nov 8;7:13347. PMID: 27824034

Get FASTQ file from GEO

Samples (15)
Less...

GSM2177715	Control 293 cell FLAG RIP
GSM2177716	FH-AKAP95 293 cell FLAG RIP
GSM2177717	FH-(101-692) 293 cell FLAG RIP
GSM2177718	FH-ZFc-s 293 cell FLAG RIP
GSM2177719	Control 293 cell AKAP95 RIP
GSM2177720	AKAP95 KD 293 cell AKAP95 RIP
GSM2177721	Control 293 cell RNA Input
GSM2177722	AKAP95 KD 293 cell RNA Input
GSM2177723	Control 293 cell
GSM2177724	AKAP95 KD (miR#8) 293 cell
GSM2177725	AKAP95 KD (miR#12) 293 cell
GSM2177726	E14 cells control shRNA rep1
GSM2177727	E14 cells Akap95 shRNA rep1
GSM2177728	E14 cells control shRNA rep2
GSM2177729	E14 cells Akap95 shRNA rep2

Relations

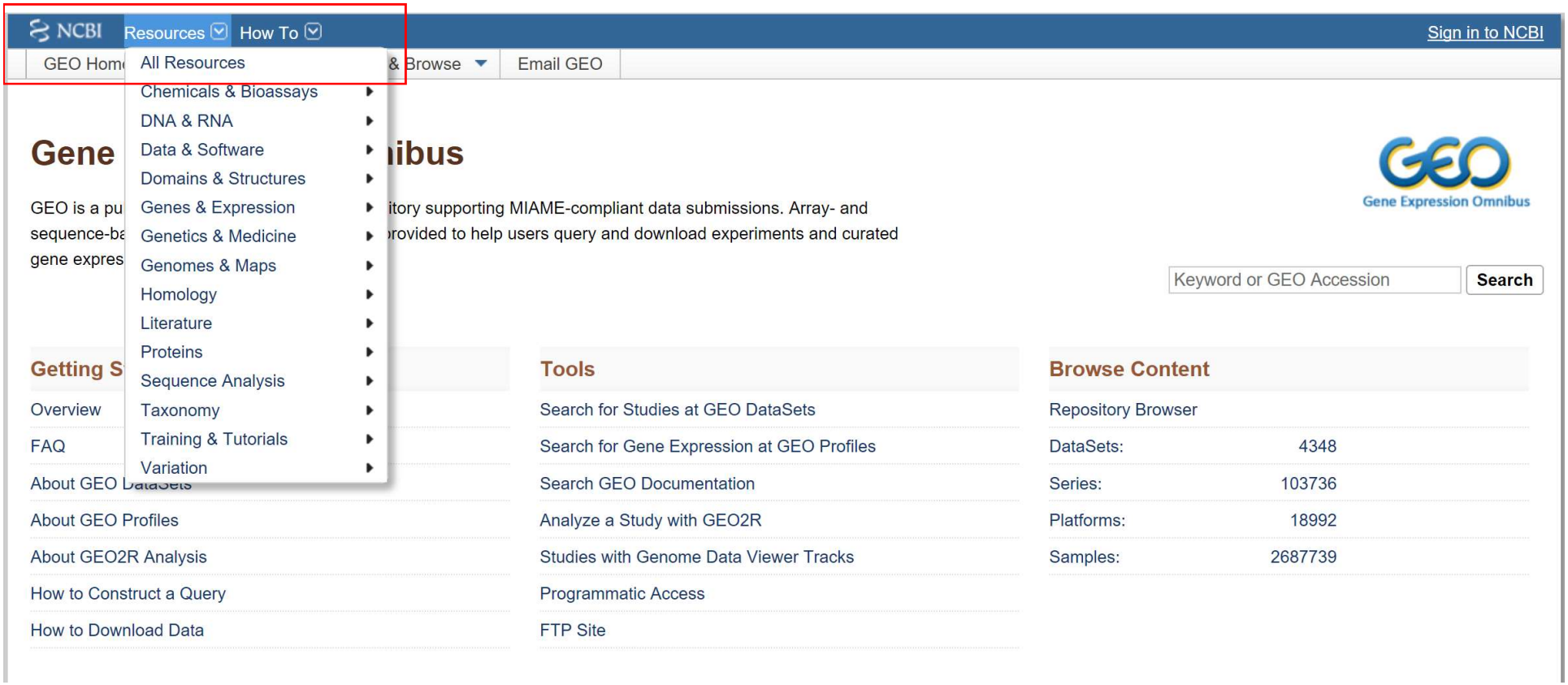
BioSample	SAMN05178628
SRA	SRX1801301

Links:

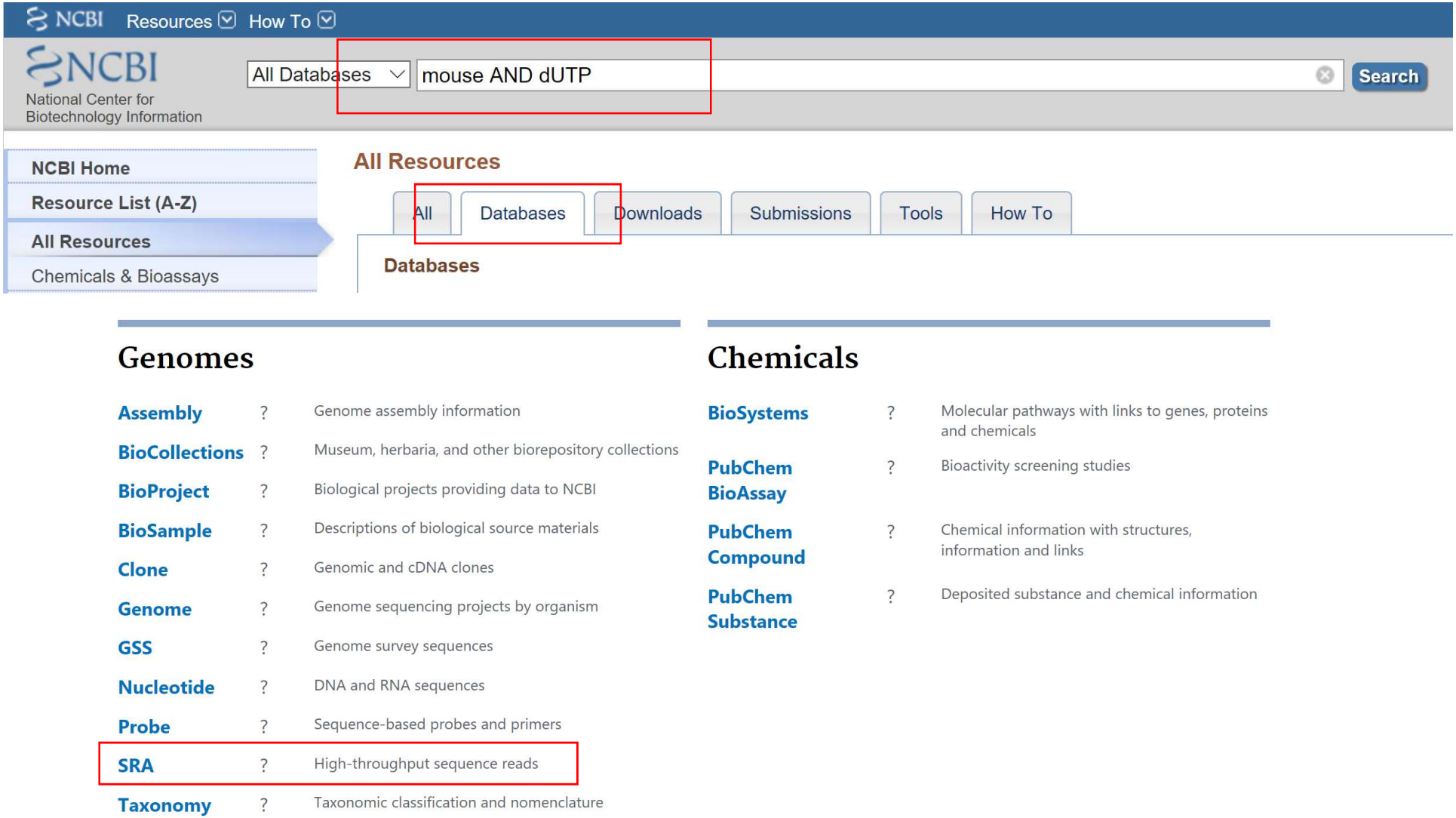
Runs: 1 run, 29.7M spots, 3G bases, 1.3Gb

Run	# of Spots	# of Bases	Size	Published
SRR3589957	29,720,636	3G	1.3Gb	2016-09-09

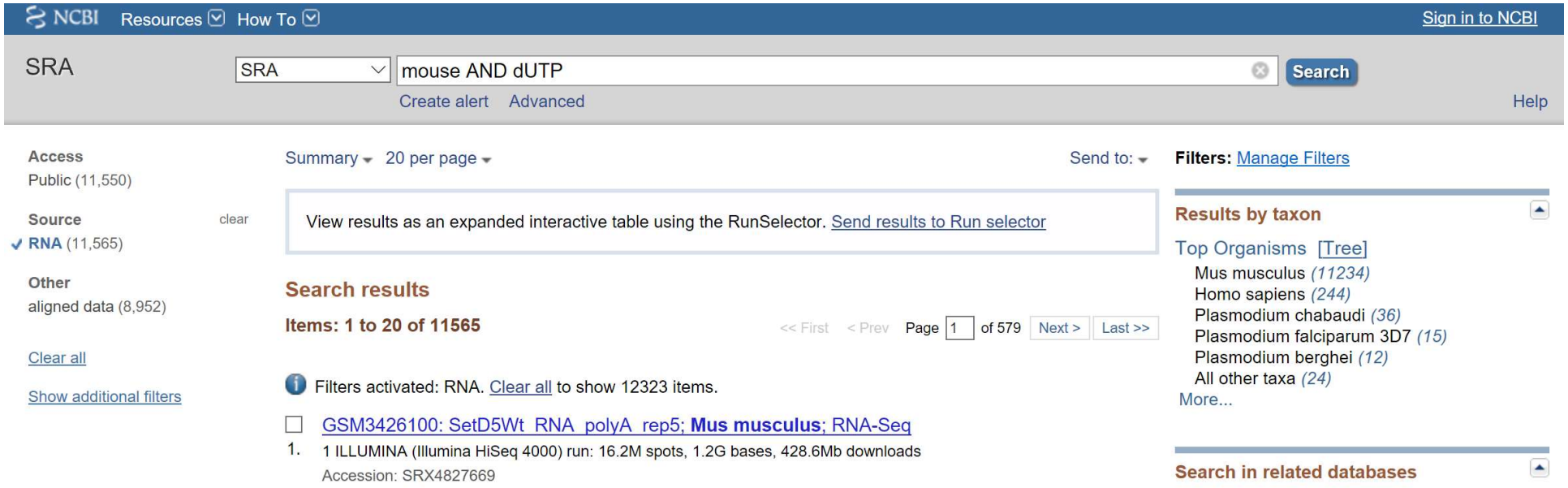
Get FASTQ file from GEO



Get FASTQ file from GEO



Get FASTQ file from GEO



Runs: 1 run, 16.2M spots, 1.2G bases, 428.6Mb

Run	# of Spots	# of Bases	Size	Published
SRR7996772	16,228,032	1.2G	428.6Mb	2018-10-13

Sratoolkit: Convert *.sra into *.fastq

https://trace.ncbi.nlm.nih.gov/Traces/sra/sra.cgi?view=software

SRA Toolkit

Compiled binaries of July 24, 2018, version 2.9.2 release:

- CentOS Linux 64 bit architecture
- Ubuntu Linux 64 bit architecture
- MacOS 64 bit architecture
- MS Windows 64 bit architecture

fastq-dump: Convert SRA data into fastq format

SraRunInfo.csv

	A	B	C	D	E	F	G	
1	Run	ReleaseDa	LoadDate	spots	bases	spots_with	avgLength	size_
2	SRR7468998	#######	#######	14839066	1.11E+09	14839066	74	
3	SRR7468999	#######	#######	15580297	1.16E+09	15580297	74	
4	SRR7469000	#######	#######	13662342	1.02E+09	13662342	74	
5	SRR7469001	#######	#######	19591119	1.46E+09	19591119	74	
6	SRR7469002	#######	#######	16568799	1.24E+09	16568799	74	
7	SRR7469003	#######	#######	13796004	1.03E+09	13796004	74	
8	SRR7469004	#######	#######	15816756	1.18E+09	15816756	74	
9	SRR7469005	#######	#######	13328326	9.94E+08	13328326	74	
10	SRR7469006	#######	#######	15133921	1.13E+09	15133921	74	
11	SRR7469007	#######	#######	16745191	1.25E+09	16745191	74	
12	SRR7469008	#######	#######	16043263	1.2E+09	16043263	74	
13	SRR7469009	#######	#######	14039711	1.05E+09	14039711	74	
14	SRR7469010	#######	#######	14055240	1.05E+09	14055240	74	
15	SRR7469011	#######	#######	6644126	4.95E+08	6644126	74	
16	SRR7469012	#######	#######	14749490	1.1E+09	14749490	74	
17	SRR7469013	#######	#######	13462757	1E+09	13462757	74	
18	SRR7469014	#######	#######	12437095	9.27E+08	12437095	74	
19	SRR7469015	#######	#######	19817417	1.48E+09	19817417	74	

Usage:

```
fastq-dump [options] <path/file> [<path/file> ...]
fastq-dump [options] <accession>
```

fastq-dump --split-3 -O ./ ← 1 单端测序
2 双端测序
3 计算机自己识别

SRR3589956_1.fastq

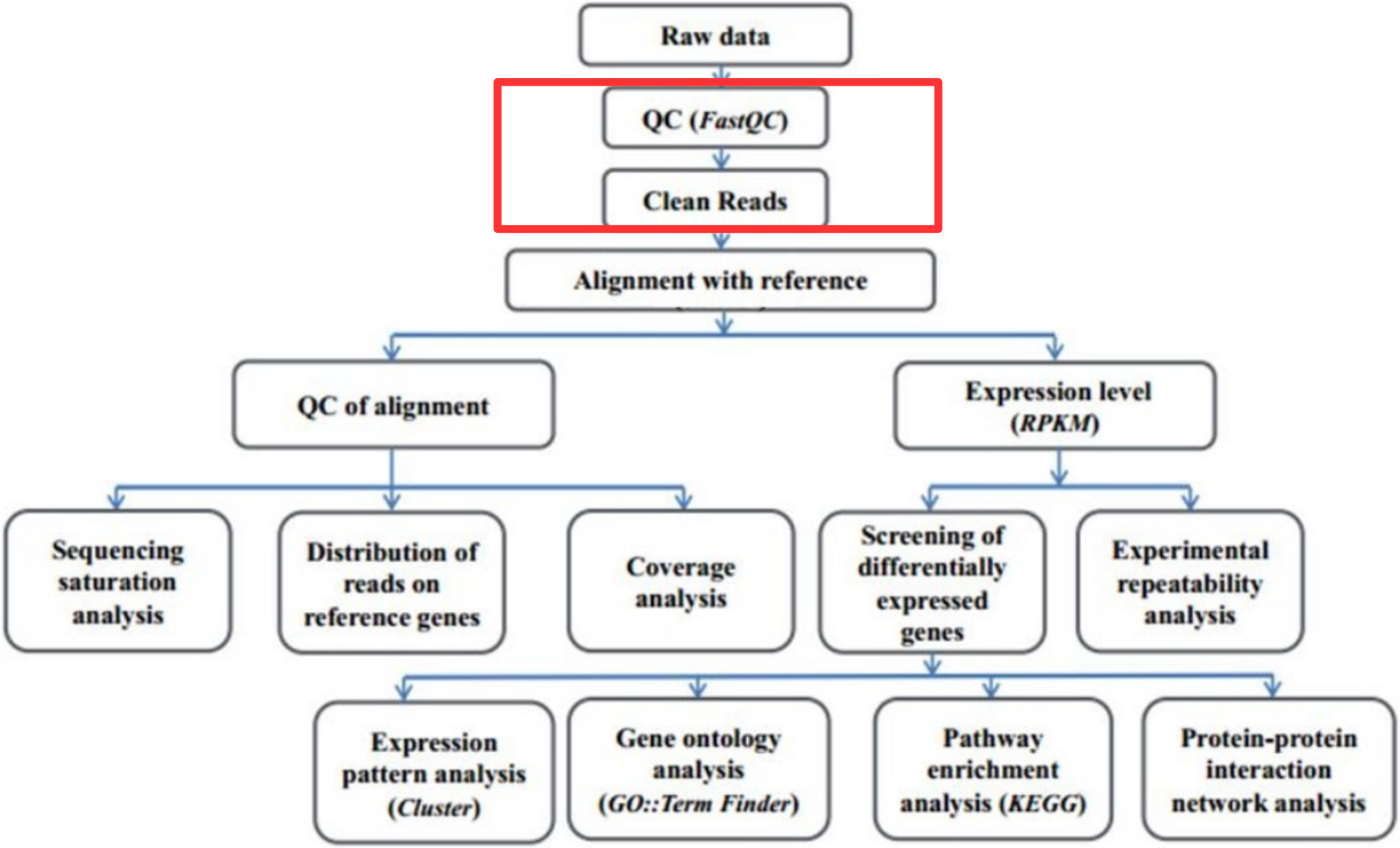
↑ 得到双端测序的序列

```python
#!/usr/bin/env python
# _*_ coding: utf-8 _*_
# @author: Drizzle_Zhang
# @file: download_sra.py
# @time: 2018/10/22 11:37

import subprocess
from time import time
import sys


def download(ls_sra, path_out):
    with open(ls_sra, 'r') as fi:
        downlsit = []
        for line in fi:
            tmp_line = line.strip().split(',')
            if tmp_line[0] != 'Run':
                downlsit.append(tmp_line[0])

    subprocesses = []
    for one in downlsit:
        subprocesses.append(subprocess.Popen('fastq-dump --split-3 -O ' + path_out + ' ' + one, shell=True))

    for sub_process in subprocesses:
        sub_process.wait()

    return


if __name__ == '__main__':
    start = time()
    download(sys.argv[1], sys.argv[2])
    end = time()
    print(end - start)
```

```
[zy@rna GSM3243797]$ python36 download_sra.py SraRunInfo.csv ./ > output
2018-10-22T04:49:43 fastq-dump.2.9.2 err: param empty while validating argument list - expected accession
```

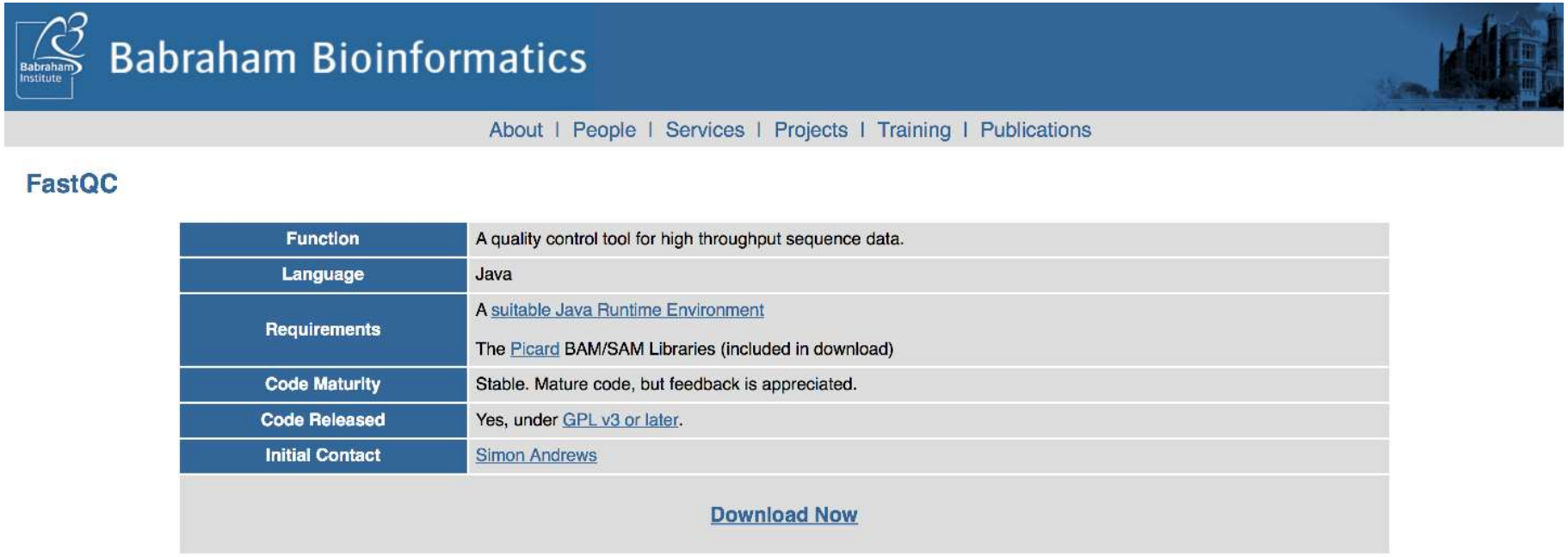
Pre-analysis

Quantitative Analysis Pipeline



FastQC: quality control

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



FastQC: quality control

fastqc [-o output dir] [--(no)extract] [-f fastq|bam|sam] [-c contaminant file] [-t threads] seqfile1 .. seqfileN

- -o 用来指定输出文件的目录，需要注意的是，FastQC 不会自动创建新目录，故指定的目录必须存在；
- FastQC 输出结果为 .zip 文件，默认参数为 --extract (自动解压缩)，执行时加上 --noextract 则不解压缩；
- -f 用来指定输入文件格式，如果不指定则自动检测；
- -c 用来指定一个文件，这个文件里面存放可能存在的污染序列，FastQC 会在这个文件里面搜索 reads 中的 overrepresented sequences；
- -t 用来指定同时处理的文件个数；
- seqfile1 等是需要处理的文件名称；

FastQC: quality control

```
[zy@rna course_rna_seq]$ fastqc -o ./result_fastQC/ --noextract -t 10 SRR2084236.fastq
```

```
Started analysis of SRR2084236.fastq
Approx 5% complete for SRR2084236.fastq
Approx 10% complete for SRR2084236.fastq
Approx 15% complete for SRR2084236.fastq
Approx 20% complete for SRR2084236.fastq
Approx 25% complete for SRR2084236.fastq
Approx 30% complete for SRR2084236.fastq
Approx 35% complete for SRR2084236.fastq
Approx 40% complete for SRR2084236.fastq
Approx 45% complete for SRR2084236.fastq
Approx 50% complete for SRR2084236.fastq
Approx 55% complete for SRR2084236.fastq
Approx 60% complete for SRR2084236.fastq
Approx 65% complete for SRR2084236.fastq
Approx 70% complete for SRR2084236.fastq
Approx 75% complete for SRR2084236.fastq
Approx 80% complete for SRR2084236.fastq
Approx 85% complete for SRR2084236.fastq
Approx 90% complete for SRR2084236.fastq
Approx 95% complete for SRR2084236.fastq
Analysis complete for SRR2084236.fastq
```

```
[zy@rna course_rna_seq]$ cd result_fastQC/
[zy@rna result_fastQC]$ ls
SRR2084236_fastqc.html  SRR2084236_fastqc.zip
```

FastQC Report

Sat 20 Oct 2018
SRR2084236.fastq

Summary

- Basic Statistics
- Per base 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

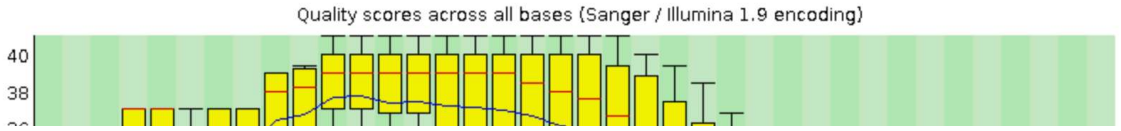
合格是对勾
警告是叹号
不合格红叉

Basic Statistics

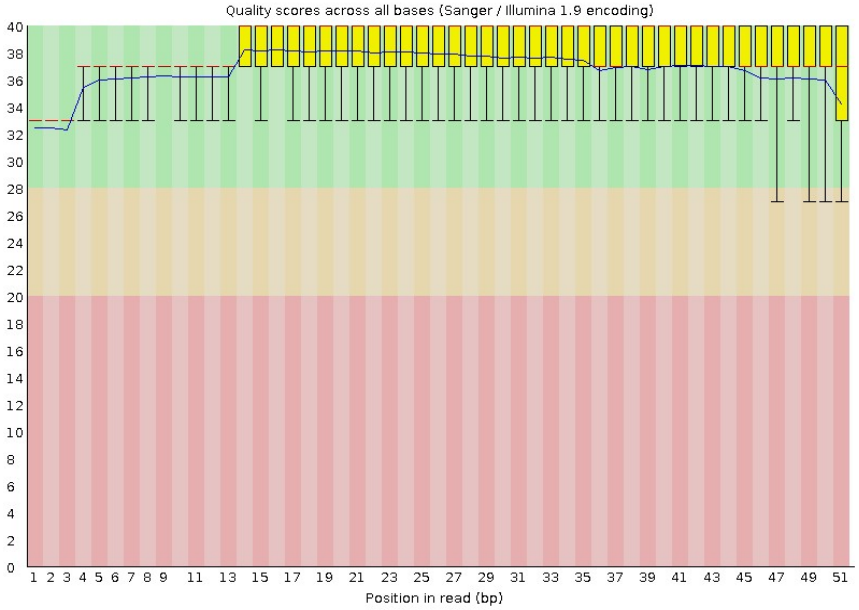
Measure	Value
Filename	SRR2084236.fastq
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	926266
Sequences flagged as poor quality	0
Sequence length	20-150
%GC	43

简要的统计信息
包括文件名、文件类型、
测序平台的信息等

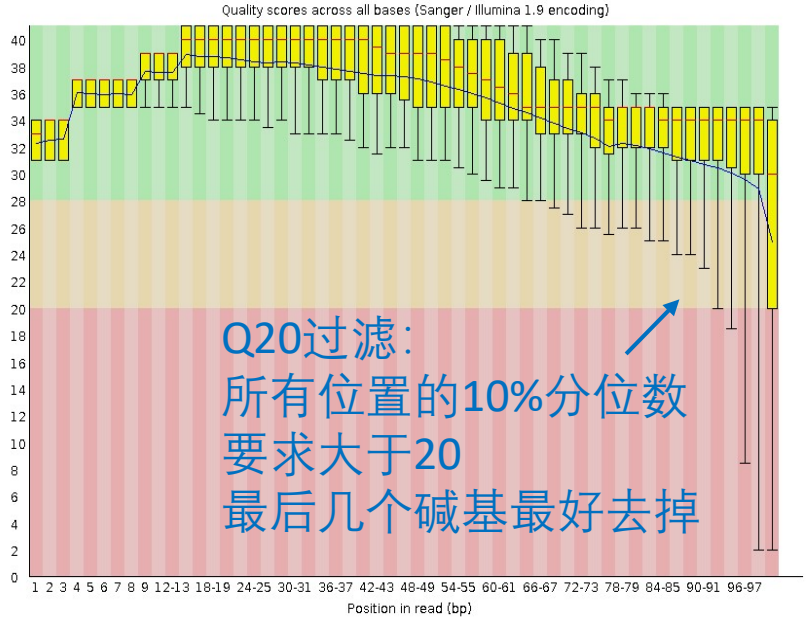
Per base sequence quality



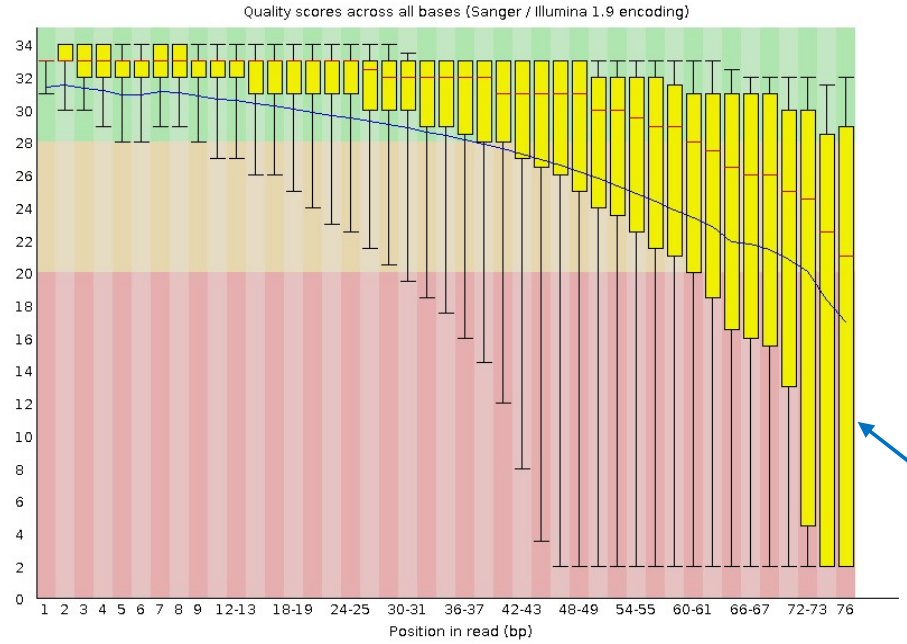
Per base sequence quality



Per base sequence quality



Per base sequence quality

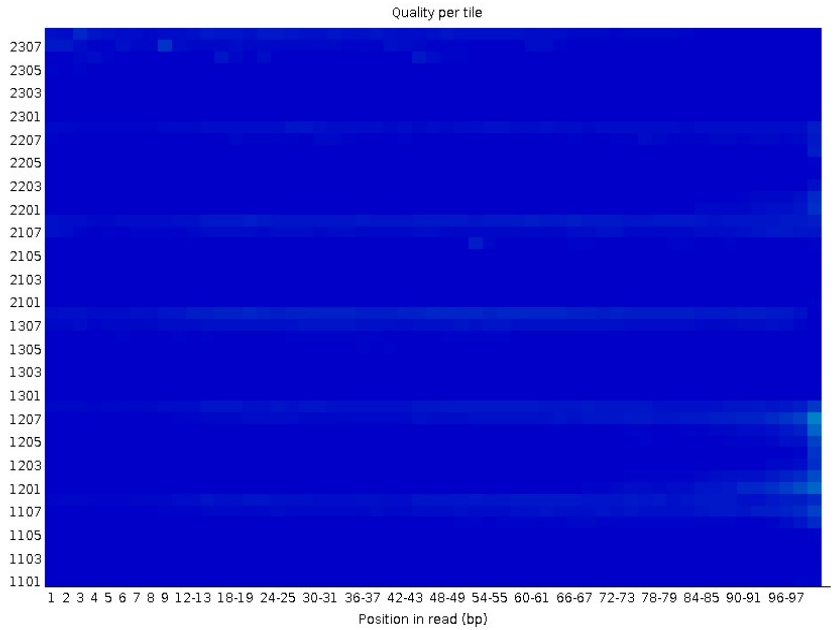


拖尾严重，数据质量差，
需要换数据分析

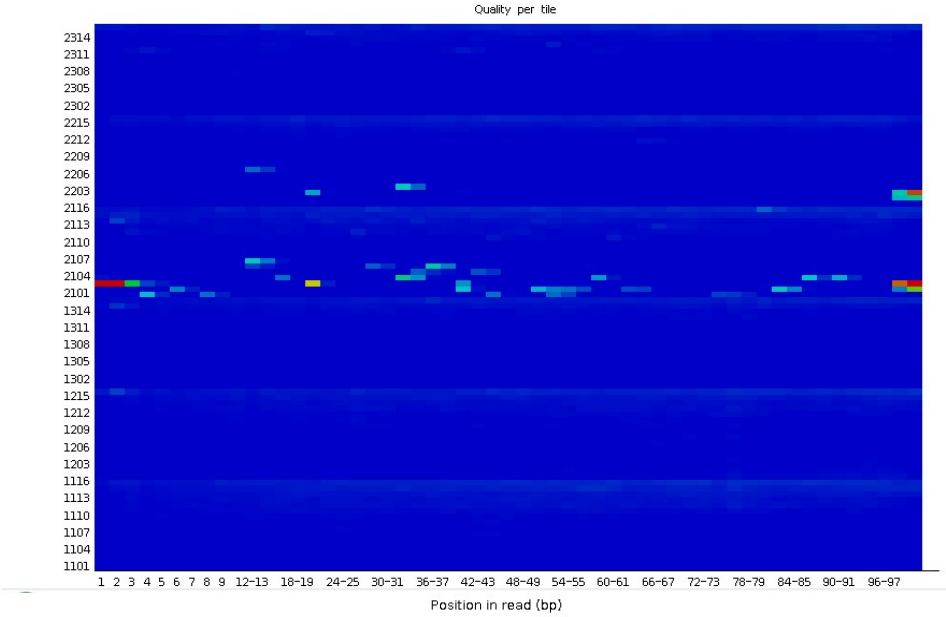
X：Position in read (bp)

Y：Base quality

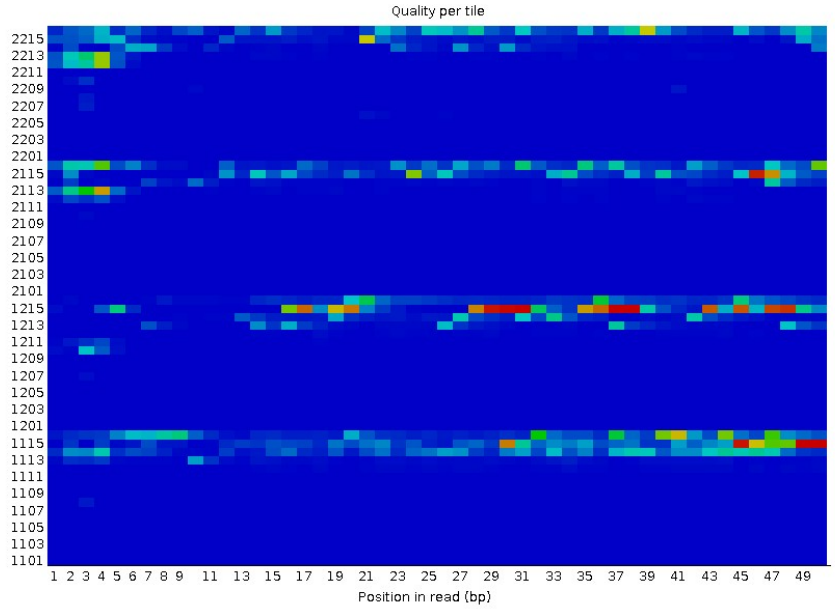
Per tile sequence quality



Per tile sequence quality



Per tile sequence quality

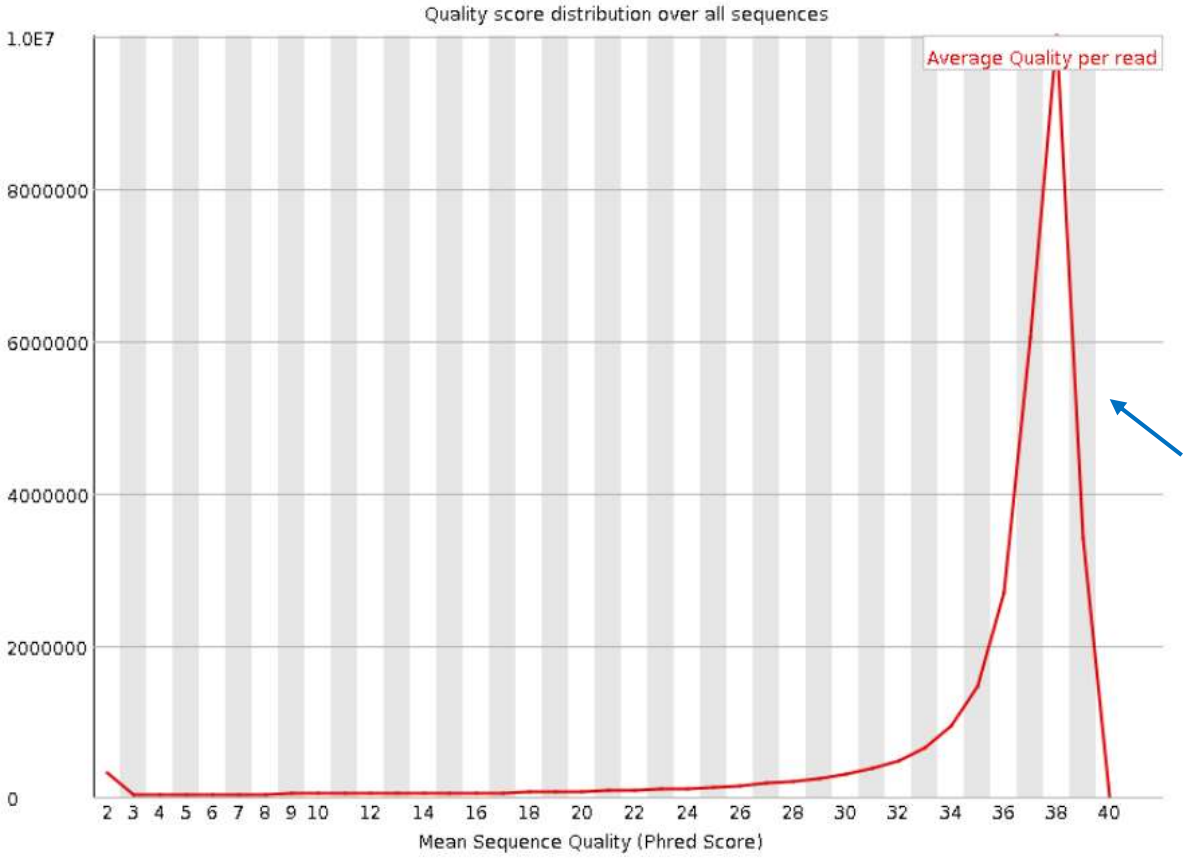


X：Position in read (bp)

Y：Tile (According to read name)

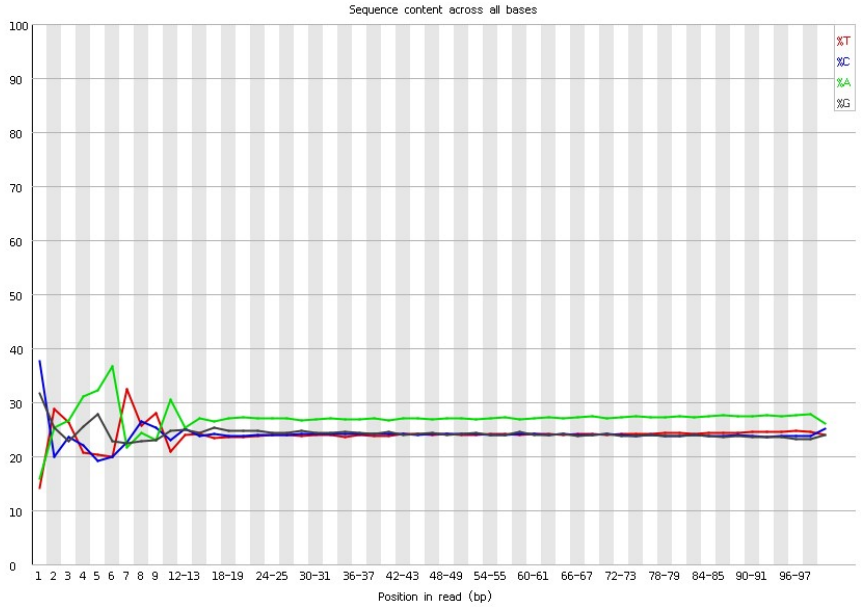
蓝色代表测序质量很高
暖色代表测序质量不高
各平台有区别
传统分析不做质控

Per sequence quality scores

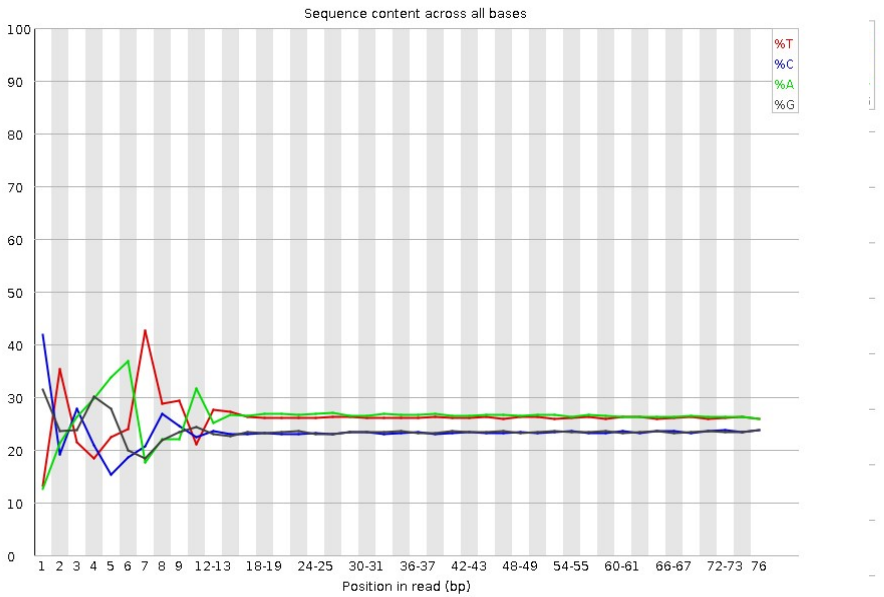


测序结果主要集中在高分
测序质量良好

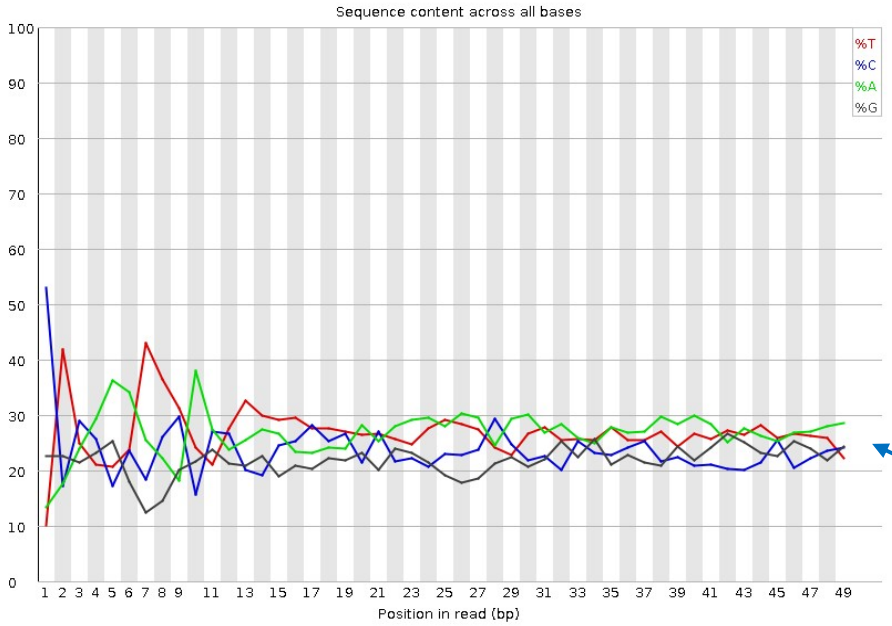
Per base sequence content



Per base sequence content



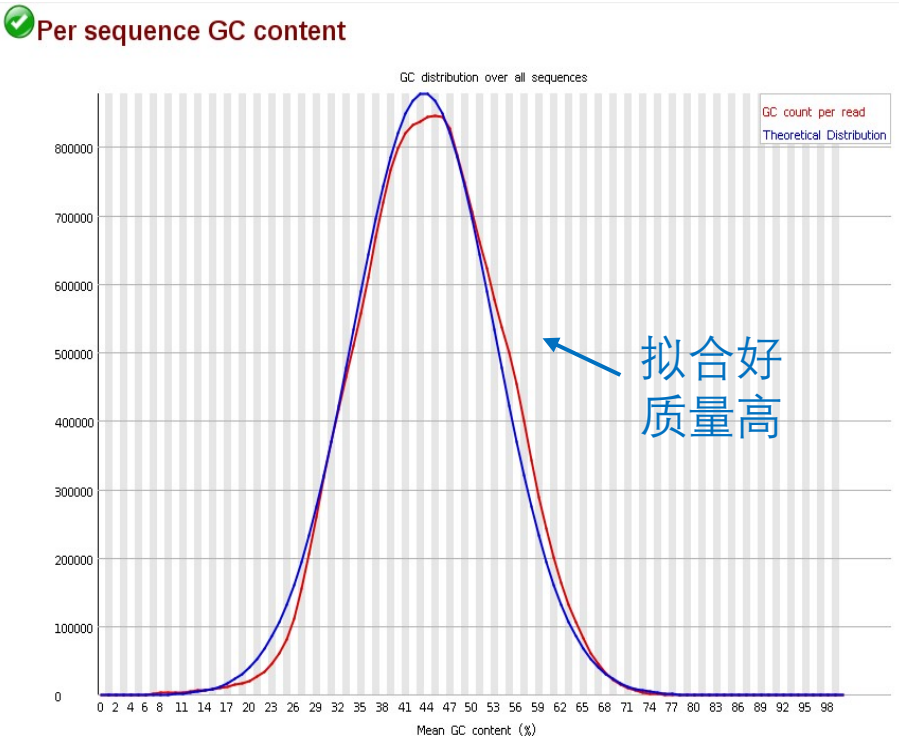
Per base sequence content

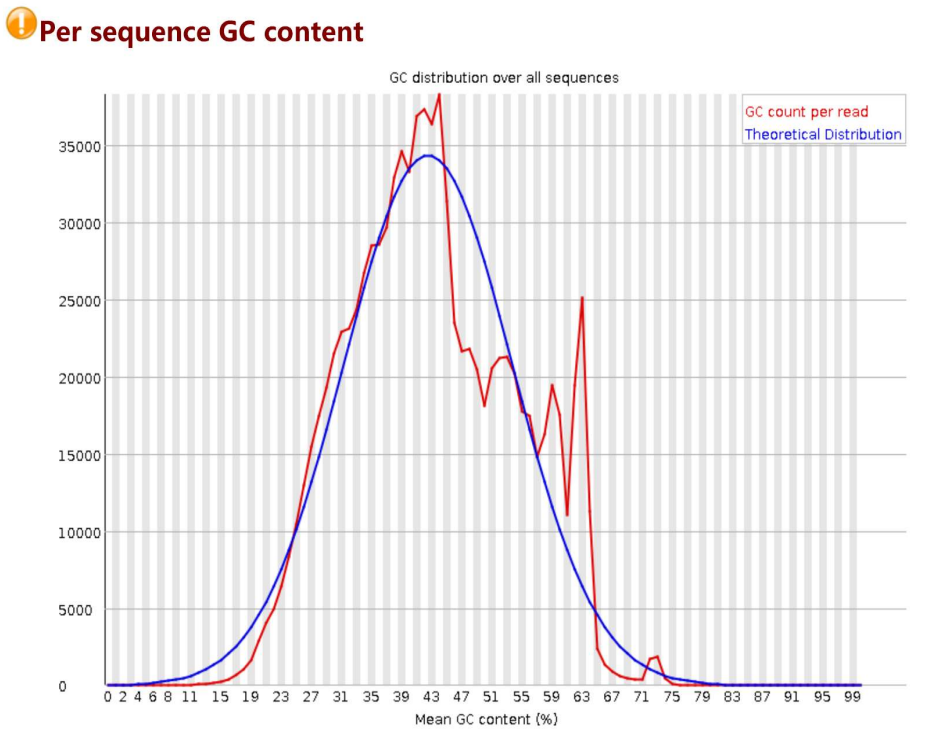


X:Position in read (bp)

Y:Content(A,T,G,C)

理论上AT应该相等，GC应该相等
本身随机性不够的序列，跑出来就应该是乱的





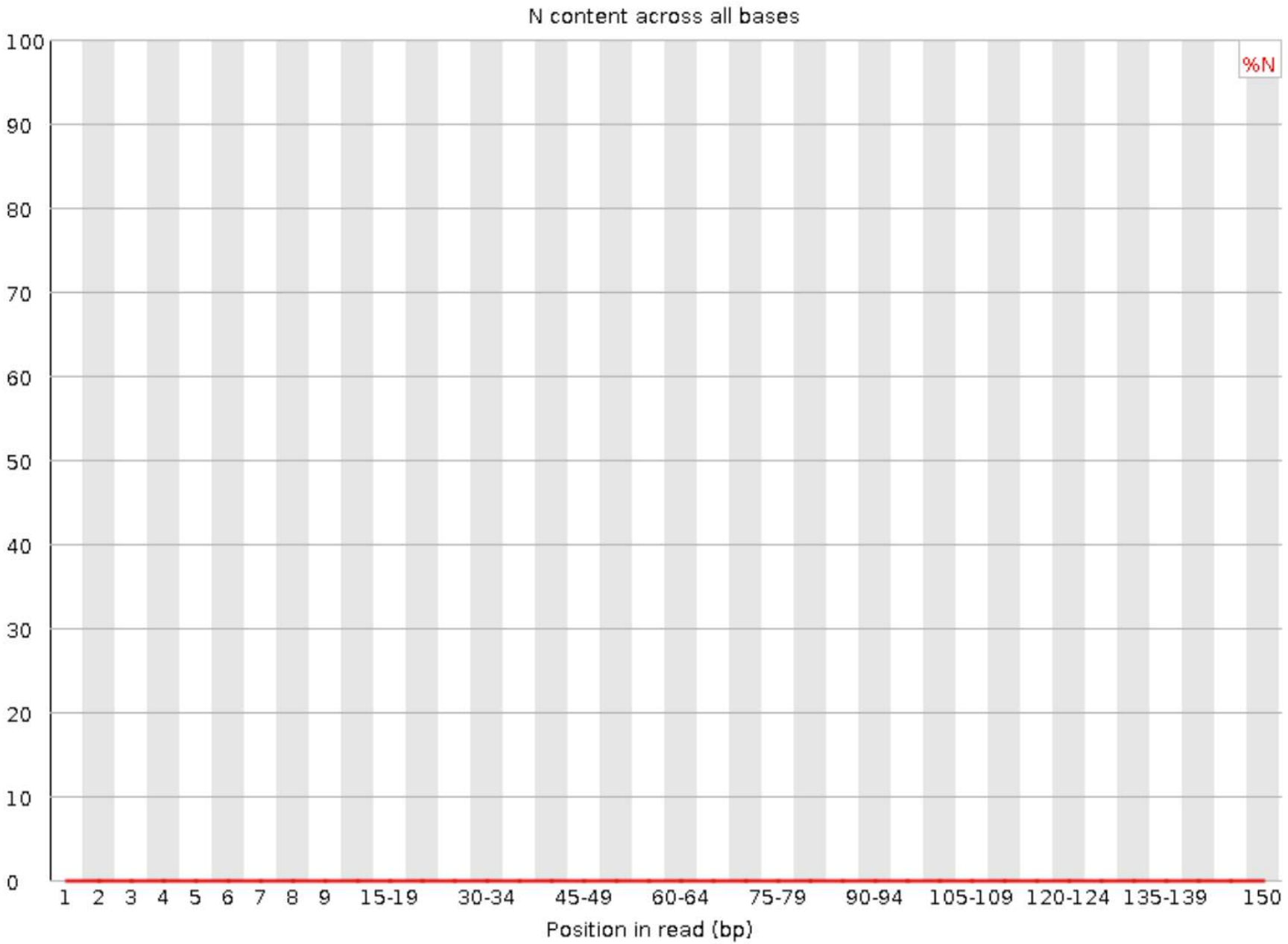
X:Mean GC content(%)

Y:GC countperread

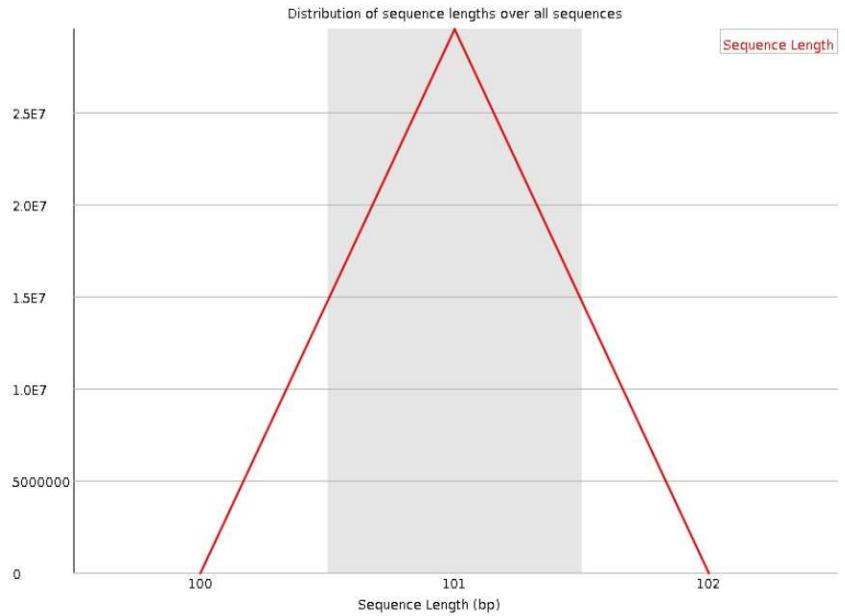
Blue:TheoreticalDistribution

Red:ActualDistribution

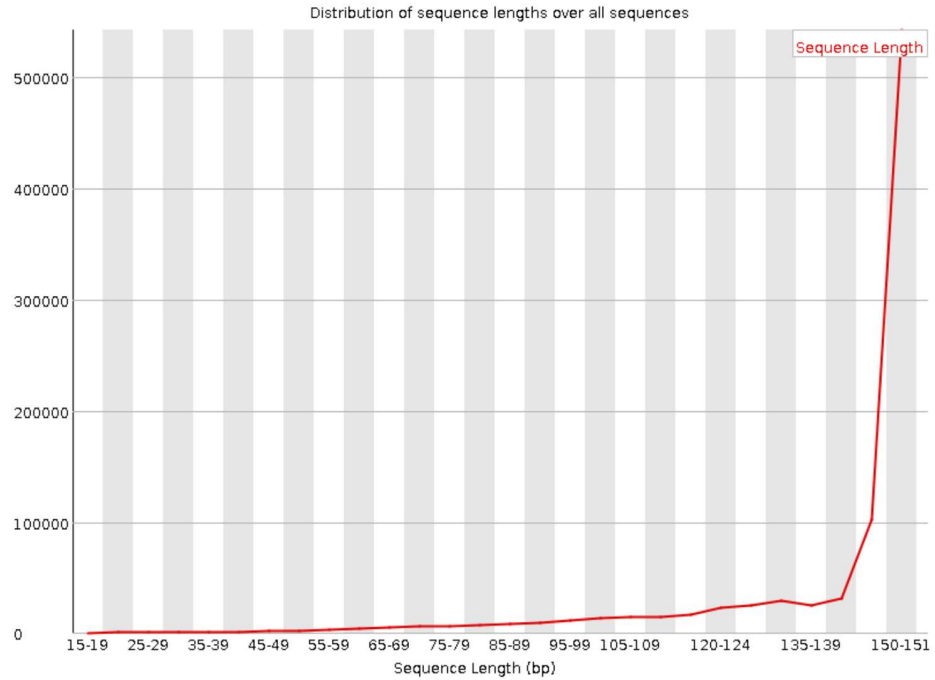
Per base N content



Sequence Length Distribution

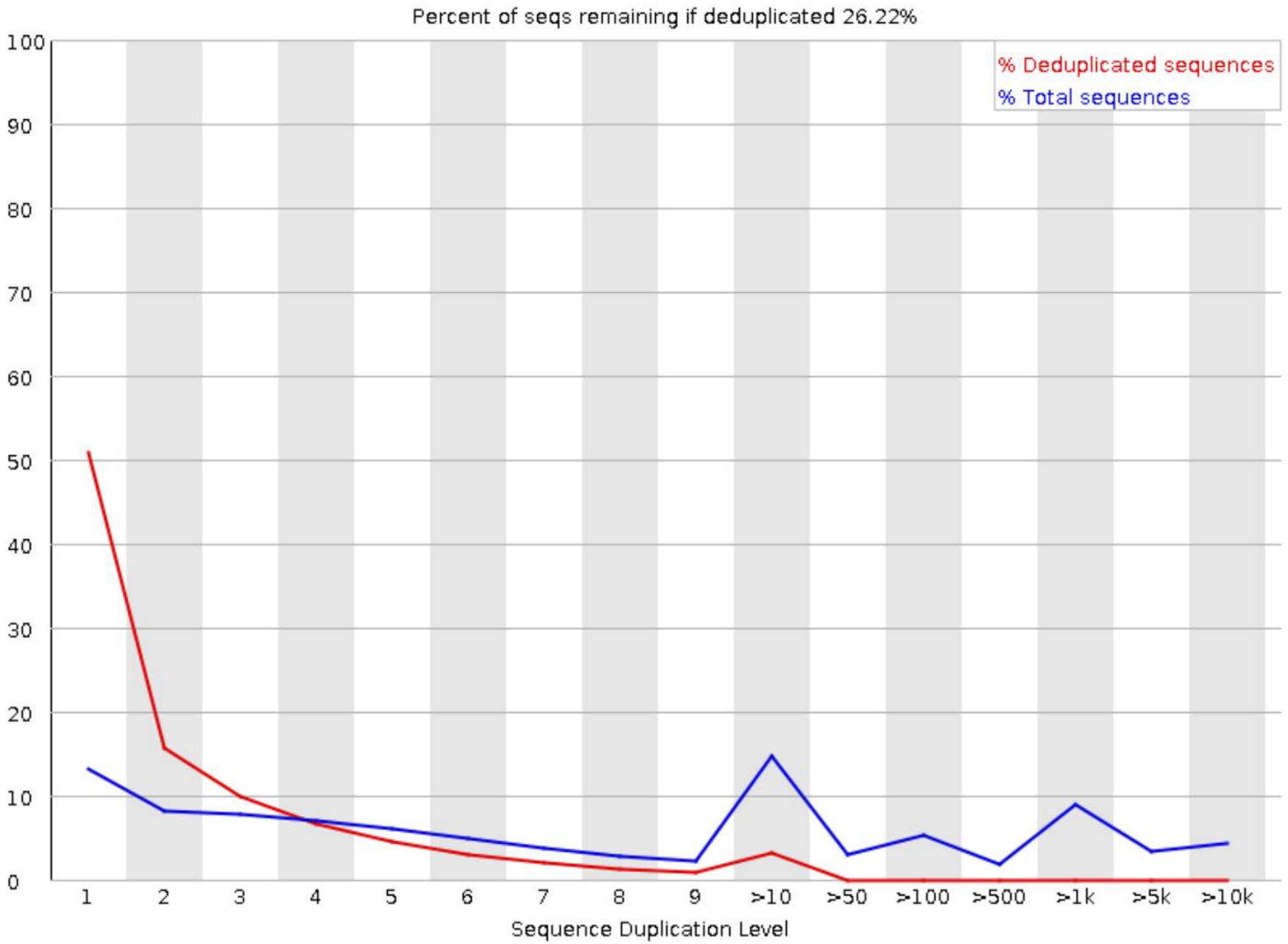


Sequence Length Distribution



每次测序仪测出来的长度理论上完全相等
有少量偏差不影响后续分析

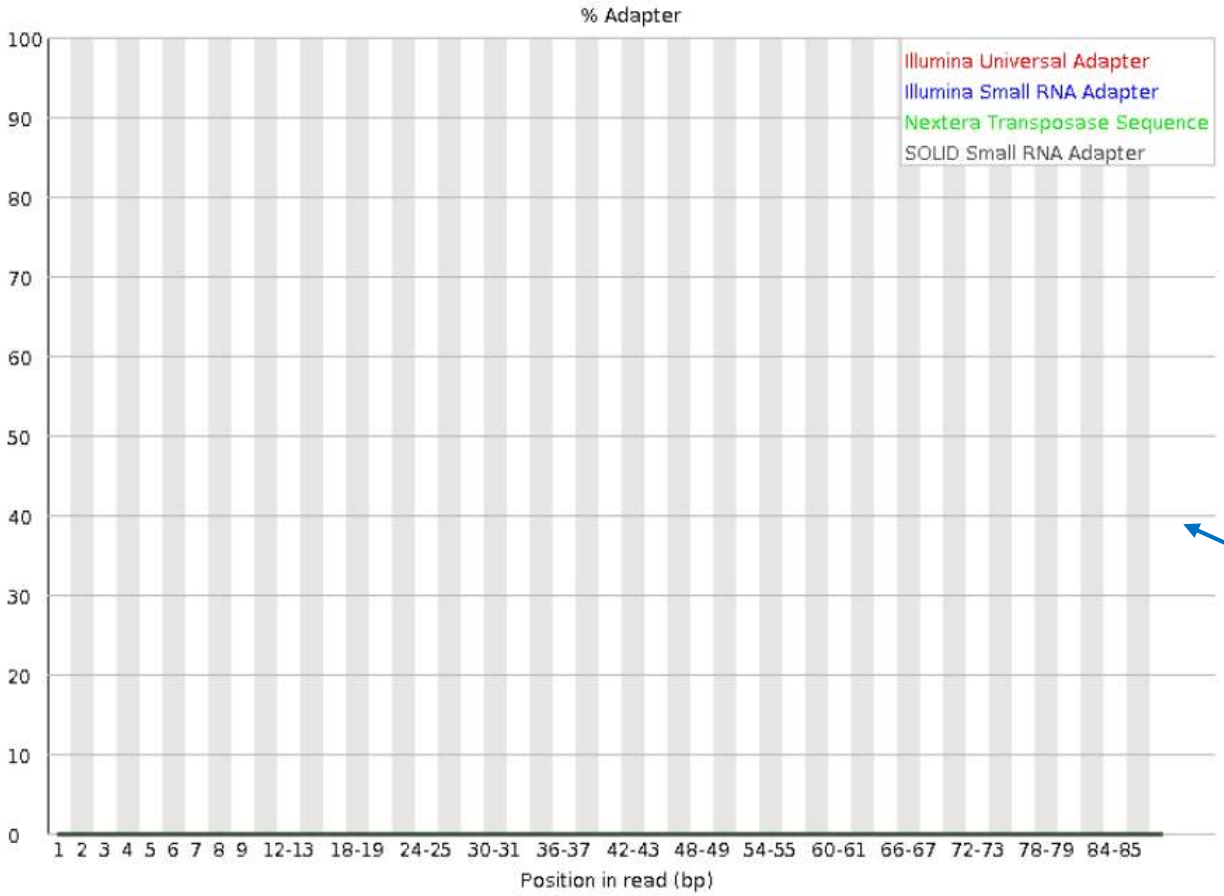
Sequence Duplication Levels



Overrepresented sequences

Sequence	Count	Percentage	Possible Source
AGCTTTGCAACCATACTCCCCCCGGAACCCAAAGACTTTGGTTTCCCGGA	19466	2.1015561404607315	No Hit
CAGCTTTGCAACCATACTCCCCCCGGAACCCAAAGACTTTGGTTTCCCGG	11684	1.2614087098090614	No Hit
GATTAATGAAAACATTCTTGGCAAATGCTTTCGCTCTGGTCCGTCTTGCG	10249	1.106485609965172	No Hit
CAACTTTCCCTTACGGTACTTGTTGACTATCGGTCTCGTGCCGGTATTTA	7893	0.852131029315553	No Hit
GCAACCATACTCCCCCCGGAACCCAAAGACTTTGGTTTCCCGGAAGCTGC	7822	0.844465844584601	No Hit
CCTTTTTCTGGAGAACGGGGTCTCGCTATATTGCCCAGGCAGGTCTCGAA	7029	0.7588532883642496	No Hit
AATGAAAACATTCTTGGCAAATGCTTTCGCTCTGGTCCGTCTTGCGCCGG	5562	0.6004754573740158	No Hit
ACCATTAAAAGTTTATTGTTTTATTTCAATATTCAGGAACAGTGTACACT	5107	0.5513534988869289	No Hit
GGAGAACGGGGTCTCGCTATATTGCCCAGGCAGGTCTCGAACTCCTGGGC	4602	0.49683352298367855	No Hit
GCTTTGCAACCATACTCCCCCCGGAACCCAAAGACTTTGGTTTCCCGGAA	4468	0.48236683630836064	No Hit
ATTCGGTTCAGTCTAATCCTTTTTGTAGTCACTCATAGGCCAGACTTAGG	4197	0.45310958191275513	No Hit
AAACCCTGTTCTTGGGTGGGTGTGGGTATAATACTAAGTTGAGATGATAT	3762	0.40614683039213356	No Hit
AAGAGGAAAACCCGGTAATGATGTCGGGGTTGAGGGATAGGAGGAGAATG	3578	0.3862821263006523	No Hit
AACCTTTCCTTATGAGCATGCCTGTGTTGGGTTGACAGTGAGGGTAATAA	3553	0.3835831175925706	No Hit
ATCTCTTCTCCTCCACCTACACCCTTACTACAACTCTCAAAACCTACCCT	3448	0.3722472810186275	No Hit

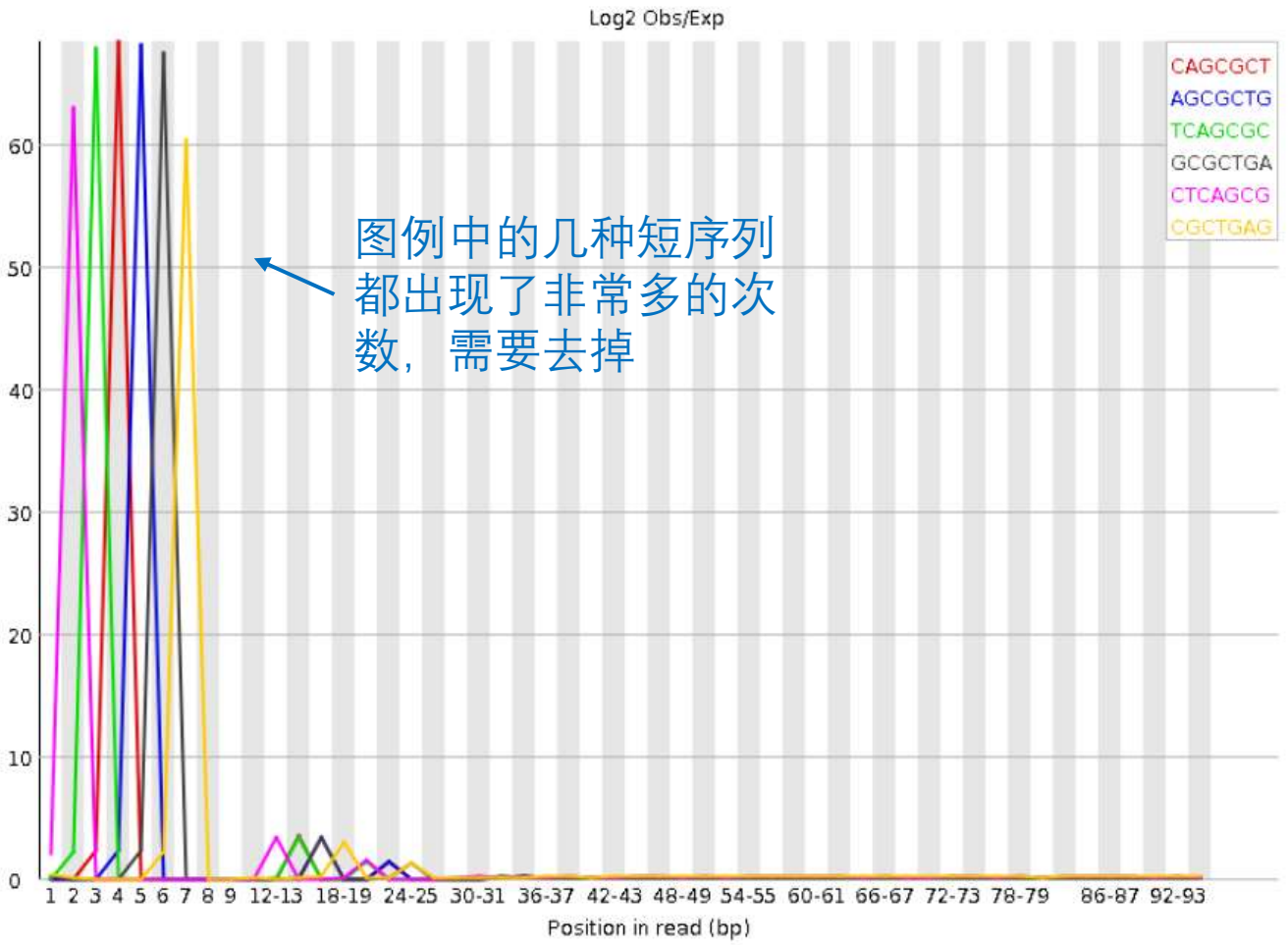
Adapter Content



如果adapter没有去除干净，需要去接头再分析

常见测序片段：adapter--index--insert--adapter
测序时会得到部分adapter

Kmer Content



图例中的几种短序列都出现了非常多的次数，需要去掉

在序列中某些特征的短序列重复出现的次数

Sequence	Count	PValue	Obs/Exp Max	Max Obs/Exp Position
CGTAACG	20	3.3845408E-6	132.1207	3
AACCCGG	845	0.0	95.37708	9
AAACCCG	905	0.0	89.053734	8
CGAGATG	130	0.0	86.38662	1
TCGGTTC	1065	0.0	85.599335	3
CCTTACG	1525	0.0	85.33698	9
ATTCGGT	740	0.0	84.80721	1
TCGTGCA	790	0.0	80.27587	9
TAAACGC	520	0.0	80.03466	6
GTCGTGC	795	0.0	79.770996	8

Data cleaning

Per base sequence content

Per base sequence quality

Adaptor content...

Cutadapt, FASTX_Toolkit, Trim Galore...

Trim Galore

Trim Galore是对FastQC和Cutadapt的包装。适用于所有高通量测序，包括RRBS(Reduced Representation Bisulfite-Seq), Illumina、Nextera 和smallRNA测序平台的双端和单端数据。主要功能包括两步：
第一步首先去除低质量碱基，然后去除3' 末端的adapter, 如果没有指定具体的adapter，程序会自动检测前1million的序列，然后对比前12-13bp的序列是否符合以下类型的adapter:

- Illumina: AGATCGGAAGAGC
- Small RNA: TGGAATTCTCGG
- Nextera: CTGTCTCTTATA

Trim Galore

Trim Galore是对FastQC和Cutadapt的包装。适用于所有高通量测序，包括RRBS(Reduced Representation Bisulfite-Seq), Illumina、Nextera 和smallRNA测序平台的双端和单端数据。主要功能包括两步：
第一步首先去除低质量碱基，然后去除3' 末端的adapter, 如果没有指定具体的adapter，程序会自动检测前1million的序列，然后对比前12-13bp的序列是否符合以下类型的adapter:

- Illumina: AGATCGGAAGAGC
- Small RNA: TGGAATTCTCGG
- Nextera: CTGTCTCTTATA

Trim Galore

--quality：设定Phred quality score阈值，默认为20。

--phred33：：选择-phred33或者-phred64，表示测序平台使用的Phred quality score。

--adapter：输入adapter序列。也可以不输入，Trim Galore!会自动寻找可能性最高的平台对应的adapter。自动搜选的平台三个，也直接显式输入这三种平台，即--illumina、--nextera和--small_rna。

--stringency：设定可以忍受的前后adapter重叠的碱基数，默认为1（非常苛刻）。可以适度放宽，因为后一个adapter几乎不可能被测序仪读到。

--length：设定输出reads长度阈值，小于设定值会被抛弃。

--paired：对于双端测序结果，一对reads中，如果有一个被剔除，那么另一个会被同样抛弃，而不管是否达到标准。

--retain_unpaired：对于双端测序结果，一对reads中，如果一个read达到标准，但是对应的另一个要被抛弃，达到标准的read会被单独保存为一个文件。

--gzip和--dont_gzip：清洗后的数据zip打包或者不打包。

--output_dir：输入目录。需要提前建立目录，否则运行会报错。

-- trim-n : 移除read一端的reads

Trim Galore

```
[zy@rna course_rna_seq]$ trim_galore --illumina --paired SRR7469011_1.fastq SRR7469011_2.fastq -o
./result_trim_glaore/
```

```
SUMMARISING RUN PARAMETERS
==========================
Input filename: SRR7469011_1.fastq
Trimming mode: paired-end
Trim Galore version: 0.4.4
Cutadapt version: 1.18
Quality Phred score cutoff: 20
Quality encoding type selected: ASCII+33
Adapter sequence: 'AGATCGGAAGAGC' (Illumina TruSeq, Sanger iPCR; user defined)
Maximum trimming error rate: 0.1 (default)
Minimum required adapter overlap (stringency): 1 bp
Minimum required sequence length for both reads before a sequence pair gets removed: 20 bp

Writing final adapter and quality trimmed output to SRR7469011_1_trimmed.fq
```

Trim Galore

```
=== Summary ===

Total reads processed:                  6,644,126
Reads with adapters:                      757,604 (11.4%)
Reads written (passing filters):        6,644,126 (100.0%)

Total basepairs processed:    329,322,716 bp
Quality-trimmed:                1,424,984 bp (0.4%)
Total written (filtered):     326,907,068 bp (99.3%)
```

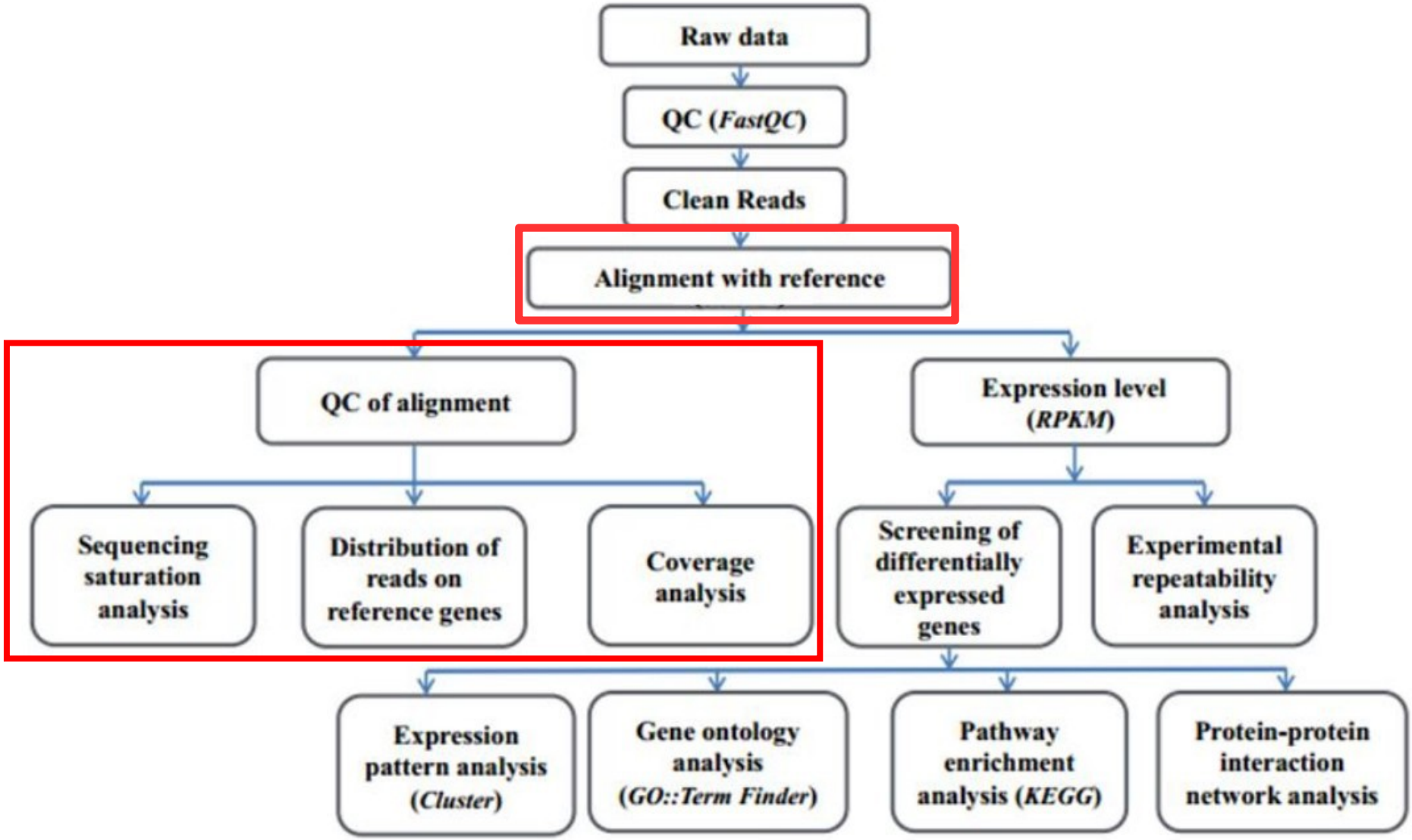
Trim Galore

```
=== Adapter 1 ===

Sequence: AGATCGGAAGAGC; Type: regular 3'; Length: 13; Trimmed: 757604 times.

No. of allowed errors:
0-9 bp: 0; 10-13 bp: 1

Bases preceding removed adapters:
  A: 32.1%
  C: 29.0%
  G: 23.0%
  T: 15.9%
  none/other: 0.0%

Overview of removed sequences
length  count   expect  max.err error counts
1       602974  1661031.5       0       602974
2       97618   415257.9        0       97618
3       46078   103814.5        0       46078
4       7126    25953.6 0       7126
5       3275    6488.4  0       3275
6       214     1622.1  0       214
7       9       405.5   0       9
9       7       25.3    0       0 7
10      38      6.3     1       1 37
11      20      1.6     1       0 20
12      8       0.4     1       0 8
13      2       0.1     1       0 2
14      2       0.1     1       0 2
15      4       0.1     1       0 4
```

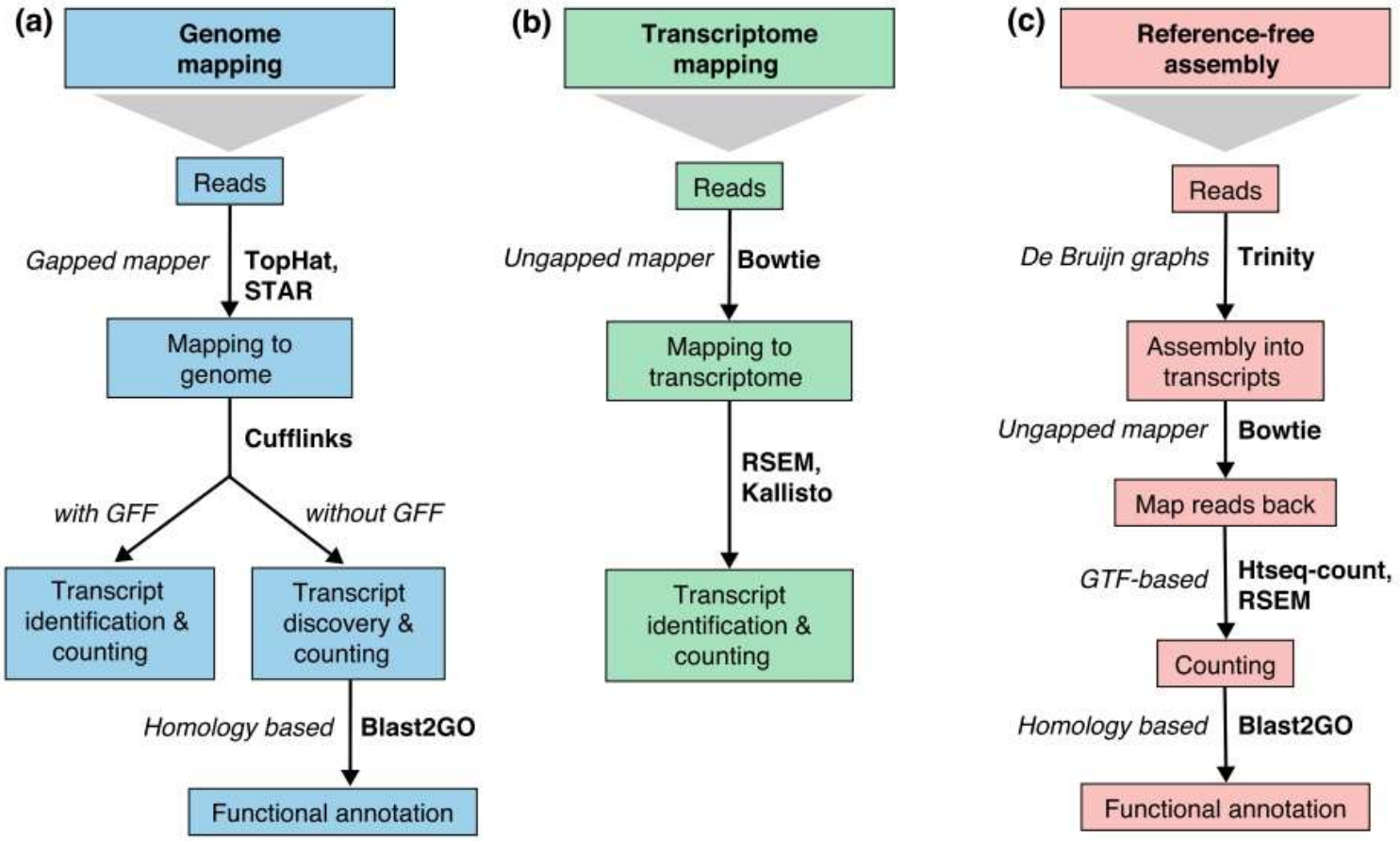
Trim Galore

```
>>>>> Now validing the length of the 2 paired-end infiles: SRR7469011_1_trimmed.fq and SRR7469011_2_trimm
ed.fq <<<<<
Writing validated paired-end read 1 reads to SRR7469011_1_val_1.fq
Writing validated paired-end read 2 reads to SRR7469011_2_val_2.fq

Total number of sequences analysed: 6644126

Number of sequence pairs removed because at least one read was shorter than the length cutoff (20 bp): 72
225 (1.09%)

Deleting both intermediate output files SRR7469011_1_trimmed.fq and SRR7469011_2_trimmed.fq
```

Quantitative Analysis Pipeline



Genome mapping, Transcriptome mapping and Reference-free assembly



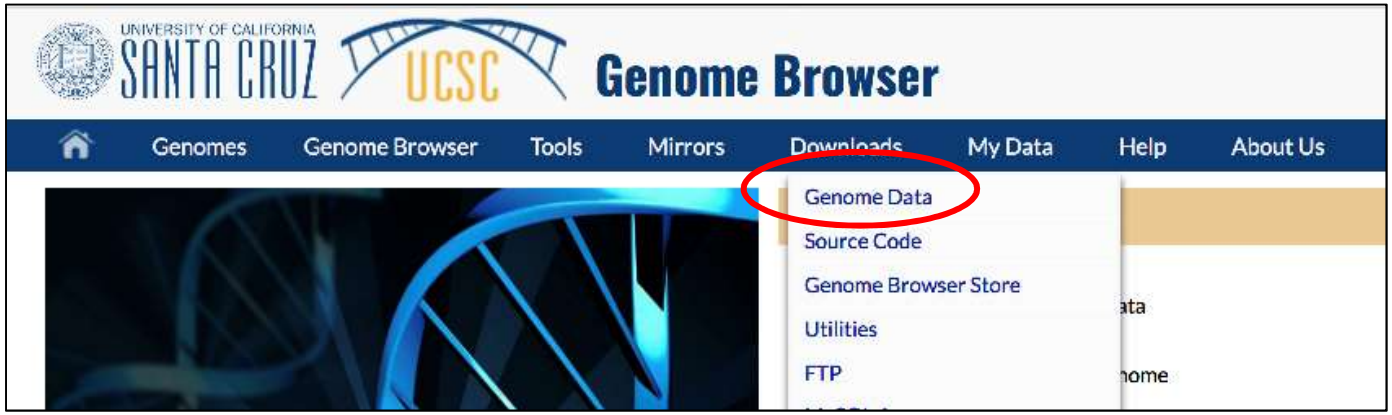
回帖：将reads与参考基因组/转录组的序列进行比较和匹配

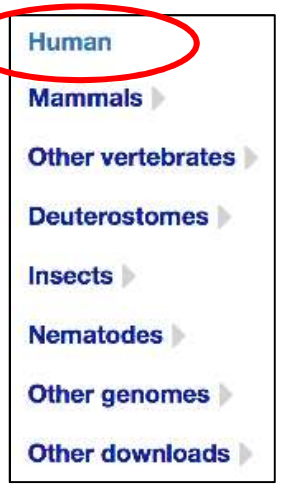
Reference genome

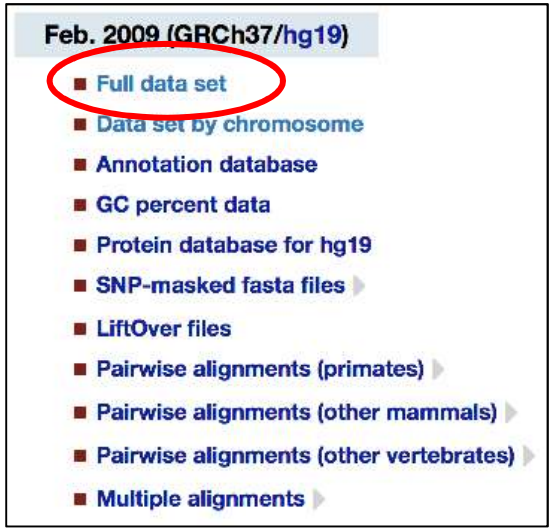
- NCBI. https://www.ncbi.nlm.nih.gov/grc
- UCSC http://hgdownload.soe.ucsc.edu/downloads.html
- Ensemble http://asia.ensembl.org/index.html?redirect=no

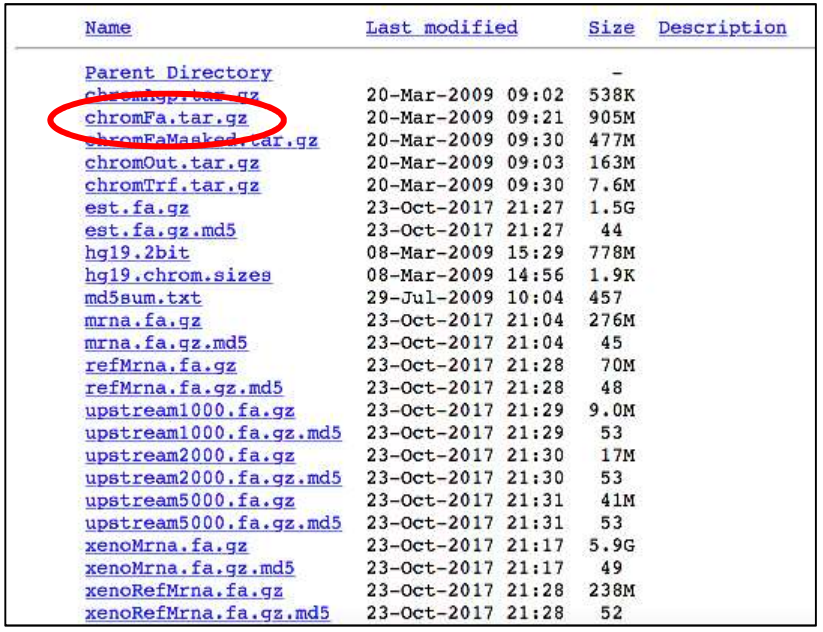
NCBI	UCSC	Ensemble
GRCh36	hg18	ENSEMBL release_52
GRCh37	hg19	ENSEMBL release_59/61/64/68/69/75
GRCh38	hg38	ENSEMBL release_76/77/78/80/81/82

Download GRCh37/hg19 from UCSC









Name	Last modified	Size	Description
Parent Directory		-	
chromFa.tar.gz	20-Mar-2009 09:21	905M	
chromOut.tar.gz	20-Mar-2009 09:03	163M	
chromTrf.tar.gz	20-Mar-2009 09:30	7.6M	
est.fa.gz	23-Oct-2017 21:27	1.5G	
est.fa.gz.md5	23-Oct-2017 21:27	44	
hg19.2bit	08-Mar-2009 15:29	778M	
hg19.chrom.sizes	08-Mar-2009 14:56	1.9K	
md5sum.txt	29-Jul-2009 10:04	457	
mrna.fa.gz	23-Oct-2017 21:04	276M	
mrna.fa.gz.md5	23-Oct-2017 21:04	45	
refMrna.fa.gz	23-Oct-2017 21:28	70M	
refMrna.fa.gz.md5	23-Oct-2017 21:28	48	
upstream1000.fa.gz	23-Oct-2017 21:29	9.0M	
upstream1000.fa.gz.md5	23-Oct-2017 21:29	53	
upstream2000.fa.gz	23-Oct-2017 21:30	17M	
upstream2000.fa.gz.md5	23-Oct-2017 21:30	53	
upstream5000.fa.gz	23-Oct-2017 21:31	41M	
upstream5000.fa.gz.md5	23-Oct-2017 21:31	53	
xenoMrna.fa.gz	23-Oct-2017 21:17	5.9G	
xenoMrna.fa.gz.md5	23-Oct-2017 21:17	49	
xenoRefMrna.fa.gz	23-Oct-2017 21:28	238M	
xenoRefMrna.fa.gz.md5	23-Oct-2017 21:28	52	

Reference genome: FA files

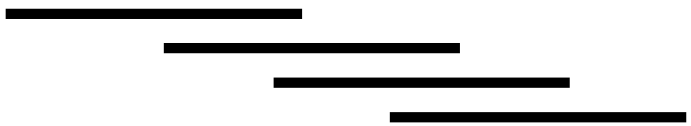
FA

```
>gi|187608668|ref|NM_001043364.2| Bombyx mori moricin (Mor), mRNA
AAACCGCGCAGTTATTTAAAATATGAATATTTTAAAACTTTTCTTTGTTTTTA
TTGTGGCAATGTCTCTGGTGTCATGTAGTACAGCCGCTCCAGCAAAAATACCT
ATCAAGGCCATTAAGACTGTAGGAAAGGCAGTCGGTAAAGGTCTAAGAGCCAT
CAATATCGCCAGTACAGCCAACGATGTTTTCAATTTCTTGAAACCGAAGAAAA
GAAAGCATTAAGAAAAGAAATTGAGTGAATGGTATTAGATATATTACTAAAGG
ATCGATCACAATGATATATAGATAGGTCATAGATGTCAACGTGAATTTATGGA
TTTTTGTTTTCCCCTTTGTAGTACTTACTTATAGTCAGTTCTTAAATTGATTG
CAACGACAACTGTGTACTATTTTTTATATTTGGTTCGAAAAGTTGCATTATTA
ACGATTTTAGAAAATAAAACTACTTTACTTTTACACG
```

Mapping

将reads与参考基因组/转录组的序列进行比较和匹配

Reference genome or transcriptome



Software for mapping:
Bowtie, Bowtie2, tophat2, BWA, HISAT2, STAR...

GTF: Gene transfer format

Structure:
<seqname> <source> <feature> <start> <end> <score> <strand> <frame> [attributes] [comments]

```
381 Twinscan  CDS          380   401   .   +   0   gene_id "001"; transcript_id "001.1";
381 Twinscan  CDS          501   650   .   +   2   gene_id "001"; transcript_id "001.1";
381 Twinscan  CDS          700   707   .   +   2   gene_id "001"; transcript_id "001.1";
381 Twinscan  start_codon  380   382   .   +   0   gene_id "001"; transcript_id "001.1";
381 Twinscan  stop_codon   708   710   .   +   0   gene_id "001"; transcript_id "001.1";
```

ftp://ftp.ensembl.org/pub/release-90/gtf/homo_sapiens/

STAR

```
[zy@rna star]$ STAR --runMode genomeGenerate --runThreadN 10 --genomeDir ./index/ --genomeFa
staFiles /home/genomewide/refgenome/mm10/mm10.fa --sjdbGTFfile /home/genomewide/annotation/m
m10/Mus_musculus.GRCm38.87.chr.gtf --sjdbOverhang 100
Oct 21 22:18:41 ..... started STAR run
Oct 21 22:18:41 ... starting to generate Genome files
```

—runMode：运行程序模式，默认是比对

—runThreadN： 运行的线程数

—genomeDir： 这个参数很重要，是存放你声称index文件路径，需要你事先建立一个有可读写权限的文件夹

—genomeFastaFiles 基因组fasta格式文件

—sjdbGTFfile GTF注释文件

—sjdbOverhang 这个值为你测序read的长度减1，是在注释可变剪切序列的时候使用的最大长度值

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

STAR

```
[zy@rna star]$ STAR --runThreadN 20 --twopassMode Basic --outSAMstrandField intronMotif --outFilterIntronMotifs RemoveNo
ncanonical --genomeDir ./index/ --readFilesIn ./../SRR7469011_1.fastq ./../SRR7469011_2.fastq --outFileNamePrefix ./SRR7
469011 --outSAMtype BAM SortedByCoordinate --quantMode GeneCounts TranscriptomeSAM
Oct 21 23:17:09 ..... started STAR run
Oct 21 23:17:09 ..... loading genome
```

--runThreasN 表示使用20个线程

--twopassMode Basic 进行两次mapping，对新junction更加敏感

--outSAMstrandField intronMotif与--FilterIntronMotif RomoveNoncanonical

对strand进行调整，使得输出的sam或bam文件可以用cufflink进行后续分析

--genomeDir 索引路径

--readFilesIn 需要mapping的fastq文件

--outFileNamePrefix 输出文件前缀名

--outSAMtype BAM SortedByCoordinate 输出文件为bam格式并排序

--quantMode GeneCounts TranscriptomeSAM 输出转录本定量文件

STAR

```
[zy@rna star]$ ls
Log.out                                    SRR7469011Log.out                SRR7469011_STARgenome
SRR7469011Aligned.sortedByCoord.out.bam    SRR7469011Log.progress.out       SRR7469011_STARpass1
SRR7469011Aligned.toTranscriptome.out.bam  SRR7469011ReadsPerGene.out.tab   index
SRR7469011Log.final.out                    SRR7469011SJ.out.tab
```

```
[zy@rna star]$ samtools view SRR7469011Aligned.sortedByCoord.out.bam | head -10
SRR7469011.1251741      163     chr1    3033378 255     18M7S   =       3168538 135195  TTTCAAGTTTCTCTGCATCATTTAN
////66/////////////////6#       NH:i:1  HI:i:1  AS:i:49 nM:i:0
SRR7469011.1328437      419     chr1    3057611 0       25M     =       3057775 214     CTACAAGAGCATGCTAGCTTCCAAT
A/A//6/EEEEEE/E/EAE/EAEAE       NH:i:10 HI:i:4  AS:i:73 nM:i:0
SRR7469011.1328437      339     chr1    3057775 0       50M     =       3057611 -214    CAAGGAGTACCAACCTGGTGCCAACTTAAAAT
AAAAACAACTAAAAAAGC      EEEA<EEEAE/A/AE/AEEAE/AAEEEAEEA6EAAEEEEEAE/AEA/AAA      NH:i:10 HI:i:4  AS:i:73 nM:i:0
SRR7469011.3860969      355     chr1    3057778 0       49M     =       3058037 284     GGAGTACCAACCTGGTGCCAACTTAAAATAAA
AACAACTAAAAAAGCTG       AA/AAEEE/EEEEEEAE6EEEEE/AEAEEEEAEEEEEEEEEE/A/EEEE       NH:i:6  HI:i:4  AS:i:72 nM:i:0
SRR7469011.1865806      355     chr1    3057781 0       50M     =       3058017 258     GTACCAACCTGGTTCCAACTTAAAATAAAAAC
AACTAAAAAAGCGGACTC      AA/AAE6EE/E///E//E6E/EE//EEEEAEEE6EEEA6EE/EE/A/EEE      NH:i:6  HI:i:4  AS:i:66 nM:i:2
SRR7469011.2567446      355     chr1    3057781 0       50M     =       3057944 188     GTACCAACCTGGTGCCAACTTAAAATAAAAAC
AACTAAAAAAGCTGACTC      AAAAAEEEEEEEEEAEEEEEEAEEEAEEEEEEEEEEEEEEEEE/EEEEEE      NH:i:7  HI:i:5  AS:i:73 nM:i:0
SRR7469011.5657774      355     chr1    3057781 0       50M     =       3058017 258     GTACCAACCTGGTGCCAACTTAAAATAAAAAC
AACTAAAAAAGCTGACTC      AAAAA66EE/EAEEEEEEEEEEEEEEE6EAEEEEEEEEEEEEEEEEEEEE      NH:i:6  HI:i:4  AS:i:70 nM:i:0
SRR7469011.2567446      403     chr1    3057944 0       25M     =       3057781 -188    GCTTATTGTTAGATAATTTTAAGAG
EEEEEAAEEEE/EAEEEEEEAAAAA       NH:i:7  HI:i:5  AS:i:73 nM:i:0
SRR7469011.1865806      403     chr1    3058017 0       3S22M   =       3057781 -258    NNNTGTCAATTAACTAATTGTGTAG
###E66E6EA6///E//AEEAAA6A       NH:i:6  HI:i:4  AS:i:66 nM:i:2
SRR7469011.5657774      403     chr1    3058017 0       3S22M   =       3057781 -258    NNNTGTCAATTAACTAATTGTGTAG
###6E/AEA6/AEE6A/66E6AAAA       NH:i:6  HI:i:4  AS:i:70 nM:i:0
```

SAM (BAM)

SAM stands for Sequence Alignment/Map format. It is a TAB-delimited text format consisting of a header section, which is optional, and an alignment section.

Example Header Lines

```
@HD     VN:1.0  SO:coordinate
@SQ     SN:1    LN:249250621    AS:NCBI37       UR:file:/data/local/ref/GATK/human_g1k_v37.fasta        M5:1b22b98cdeb4a9304cb5d48026a85128
@SQ     SN:2    LN:243199373    AS:NCBI37       UR:file:/data/local/ref/GATK/human_g1k_v37.fasta        M5:a0d9851da00400dec1098a9255ac712e
@SQ     SN:3    LN:198022430    AS:NCBI37       UR:file:/data/local/ref/GATK/human_g1k_v37.fasta        M5:fdfd811849cc2fadebc929bb925902e5
@RG     ID:UM0098:1     PL:ILLUMINA     PU:HWUSI-EAS1707-615LHAAXX-L001 LB:80   DT:2010-05-05T20:00:00-0400     SM:SD37743      CN:UMCORE
@RG     ID:UM0098:2     PL:ILLUMINA     PU:HWUSI-EAS1707-615LHAAXX-L002 LB:80   DT:2010-05-05T20:00:00-0400     SM:SD37743      CN:UMCORE
@PG     ID:bwa  VN:0.5.4
@PG     ID:GATK TableRecalibration      VN:1.0.3471     CL:Covariates=[ReadGroupCovariate, QualityScoreCovariate, CycleCovariate, DinucCovariate,
TileCovariate], default_read_group=null, default_platform=null, force_read_group=null, force_platform=null, solid_recal_mode=SET_Q_ZERO, window_size_nqs=5,
homopolymer_nback=7, exception_if_no_tile=false, ignore_nocall_colorspace=false, pQ=5, maxQ=40, smoothing=1
```

In the alignment examples below, you will see that the 2nd alignment maps back to the RG line with ID UM0098.1, and all of the alignments point back to the SQ line with SN:1 because their RNAME is 1.

Example Alignments

This is what the alignment section of a SAM file looks like:

```
1:497:R:-272+13M17D24M  113     1       497     37      37M     15      100338662       0       CGGGTCTGACCTGAGGAGAACTGTGCTCCGCCTTCAG    0;==-
==9;>>>>>=>>>>>>>>>>>=>>>>>>>>>>         XT:A:U  NM:i:0  SM:i:37 AM:i:0  XO:i:1  X1:i:0  XM:i:0  XO:i:0  XG:i:0  MD:Z:37
19:20389:F:275+18M2D19M 99      1       17644   0       37M     =       17919   314     TATGACTGCTAATAATACCTACACATGTTAGAACCAT    >>>>>>>>>>>>>>>>>>>><<>>>
<<>>4:>>:<9     RG:Z:UM0098:1   XT:A:R  NM:i:0  SM:i:0  AM:i:0  X0:i:4  X1:i:0  XM:i:0  XO:i:0  XG:i:0  MD:Z:37
19:20389:F:275+18M2D19M 147     1       17919   0       18M2D19M        =       17644   -314    GTAGTACCAACTGTAAGTCCTTATCTTCATACTTTGT    ;44999;499<8<8<<<8<<>
<<<<><7<;<<<>><<        XT:A:R  NM:i:2  SM:i:0  AM:i:0  X0:i:4  X1:i:0  XM:i:0  XO:i:1  XG:i:2  MD:Z:18^CA19
9:21597+10M2I25M:R:-209 83      1       21678   0       8M2I27M =       21469   -244    CACCACATCACATATACCAAGCCTGGCTGTGTCTTCT    <;9<<5><<<<><<<><><><><><9>>
<>>>9>>><>      XT:A:R  NM:i:2  SM:i:0  AM:i:0  X0:i:5  X1:i:0  XM:i:0  XO:i:1  XG:i:2  MD:Z:35
```

```
HWI-C00138:74:C93K8ANXX:8:2303:18752:95744    147    chr1    13151    1    125M    =    13115    -161    GGAAGGAGAAGGGGATGCACTGTTGGGGAGGCAGCTGTAACTCAAAGCCTTAGCCTCT
```

Col	Field	Type	Regexp/Range	Brief description
1	QNAME	String	`[!-?A-~]{1,254}`	Query template NAME
2	FLAG	Int	$[0,2^{16}-1]$	bitwise FLAG
3	RNAME	String	`\*|[!-()+-<>-~][!-~]*`	Reference sequence NAME
4	POS	Int	$[0,2^{31}-1]$	1-based leftmost mapping POSition
5	MAPQ	Int	$[0,2^{8}-1]$	MAPping Quality
6	CIGAR	String	`\*|([0-9]+[MIDNSHPX=])+`	CIGAR string
7	RNEXT	String	`\*|=|[!-()+-<>-~][!-~]*`	Ref. name of the mate/next read
8	PNEXT	Int	$[0,2^{31}-1]$	Position of the mate/next read
9	TLEN	Int	$[-2^{31}+1,2^{31}-1]$	observed Template LENgth
10	SEQ	String	`\*|[A-Za-z=.]+`	segment SEQuence
11	QUAL	String	`[!-~]+`	ASCII of Phred-scaled base QUALity+33

FLAG: Combination of bitwise FLAGs.[4] Each bit is explained in the following table:

Bit		Description
1	0x1	template having multiple segments in sequencing
2	0x2	each segment properly aligned according to the aligner
4	0x4	segment unmapped
8	0x8	next segment in the template unmapped
16	0x10	SEQ being reverse complemented
32	0x20	SEQ of the next segment in the template being reverse complemented
64	0x40	the first segment in the template
128	0x80	the last segment in the template
256	0x100	secondary alignment
512	0x200	not passing filters, such as platform/vendor quality controls
1024	0x400	PCR or optical duplicate
2048	0x800	supplementary alignment

```
HWI-C00138:74:C93K8ANXX:8:2303:18752:95744      147     chr1    13151   1       125M    =       13115   -161    GGAAGGAGAAGGGGATGCACTGTTGGGGAGGCAGCTGTAACTCAAAGCCTTAGCCTCT
```

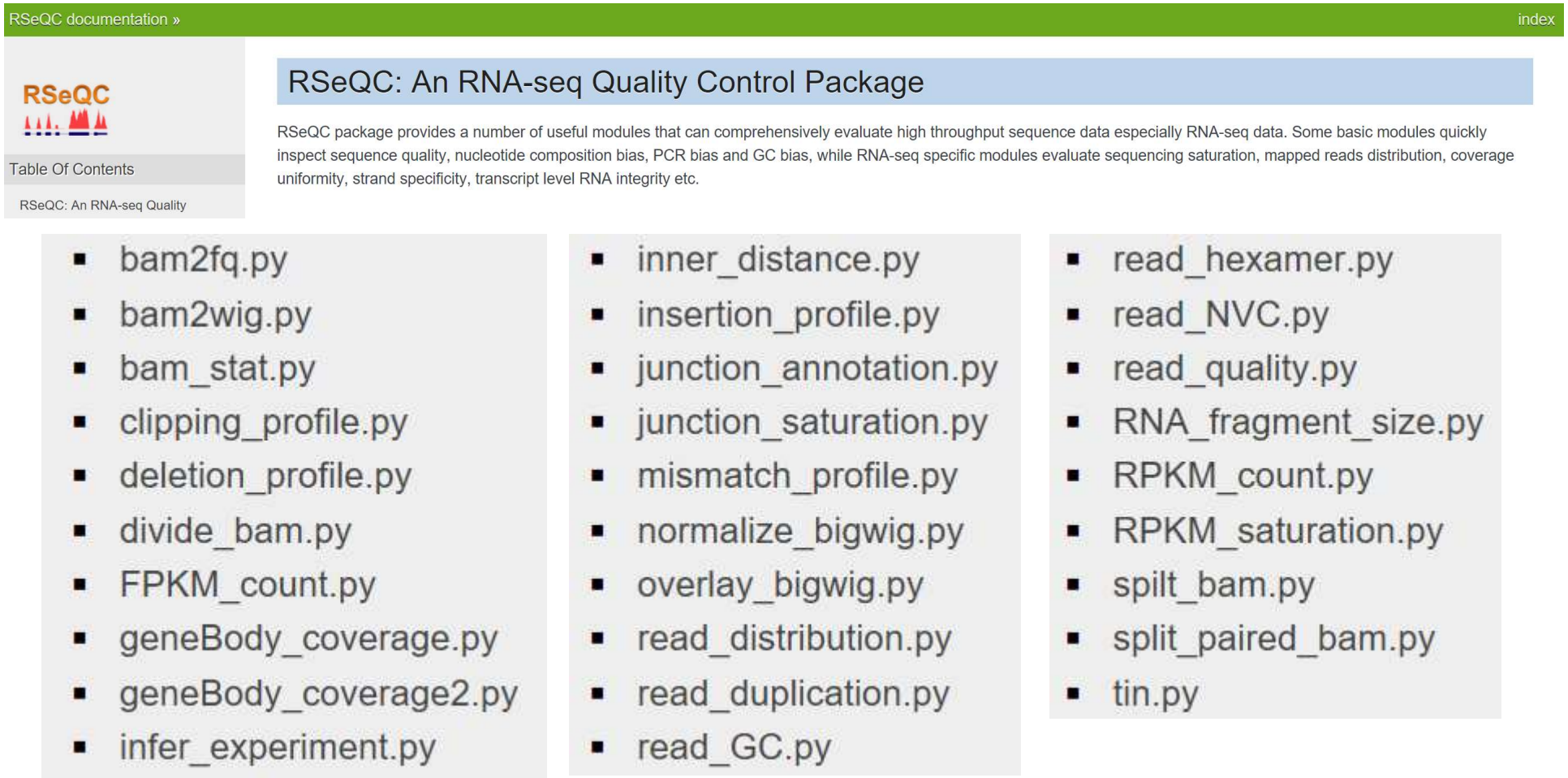
Col	Field	Type	Regexp/Range	Brief description
1	QNAME	String	`[!-?A-~]{1,254}`	Query template NAME
2	FLAG	Int	`[0,2^16-1]`	bitwise FLAG
3	RNAME	String	`\*|[!-()+-<>-~][!-~]*`	Reference sequence NAME
4	POS	Int	`[0,2^31-1]`	1-based leftmost mapping POSition
5	MAPQ	Int	`[0,2^8-1]`	MAPping Quality
6	CIGAR	String	`\*|([0-9]+[MIDNSHPX=])+`	CIGAR string
7	RNEXT	String	`\*|=|[!-()+-<>-~][!-~]*`	Ref. name of the mate/next read
8	PNEXT	Int	`[0,2^31-1]`	Position of the mate/next read
9	TLEN	Int	`[-2^31+1,2^31-1]`	observed Template LENgth
10	SEQ	String	`\*|[A-Za-z=.]+`	segment SEQuence
11	QUAL	String	`[!-~]+`	ASCII of Phred-scaled base QUALity+33

MAPQ : MAPping Quality. It equals $-10 \log_{10} Pr\{\text{mapping position is wrong}\}$, rounded to the nearest integer. A value 255 indicates that the mapping quality is not available.

Op	BAM	Description
M	0	alignment match (can be a sequence match or mismatch)
I	1	insertion to the reference
D	2	deletion from the reference
N	3	skipped region from the reference
S	4	soft clipping (clipped sequences present in SEQ)
H	5	hard clipping (clipped sequences NOT present in SEQ)
P	6	padding (silent deletion from padded reference)
=	7	sequence match
X	8	sequence mismatch

Compact Idiosyncratic Gapped Alignment Report

RSeQC

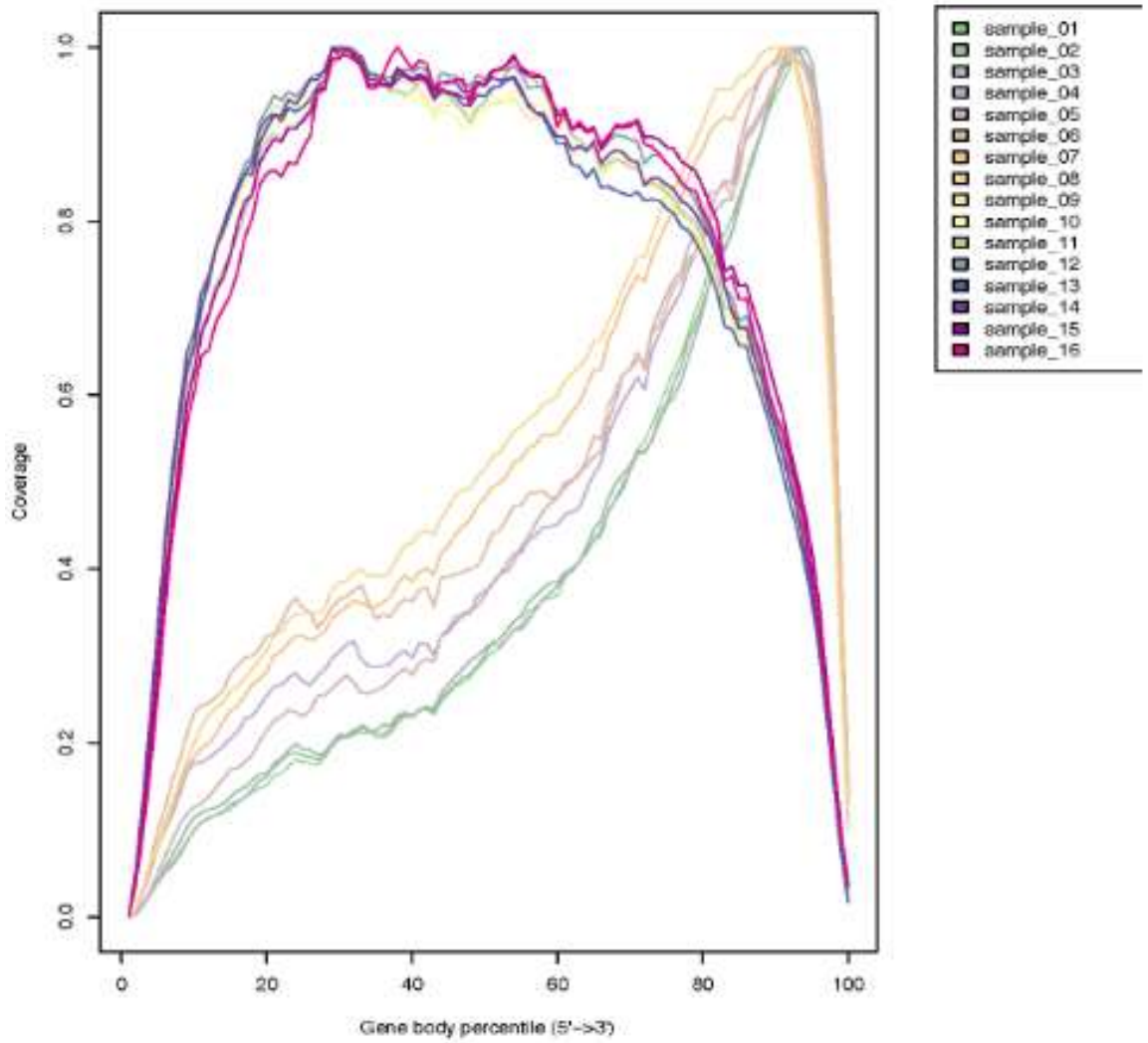


http://rseqc.sourceforge.net/

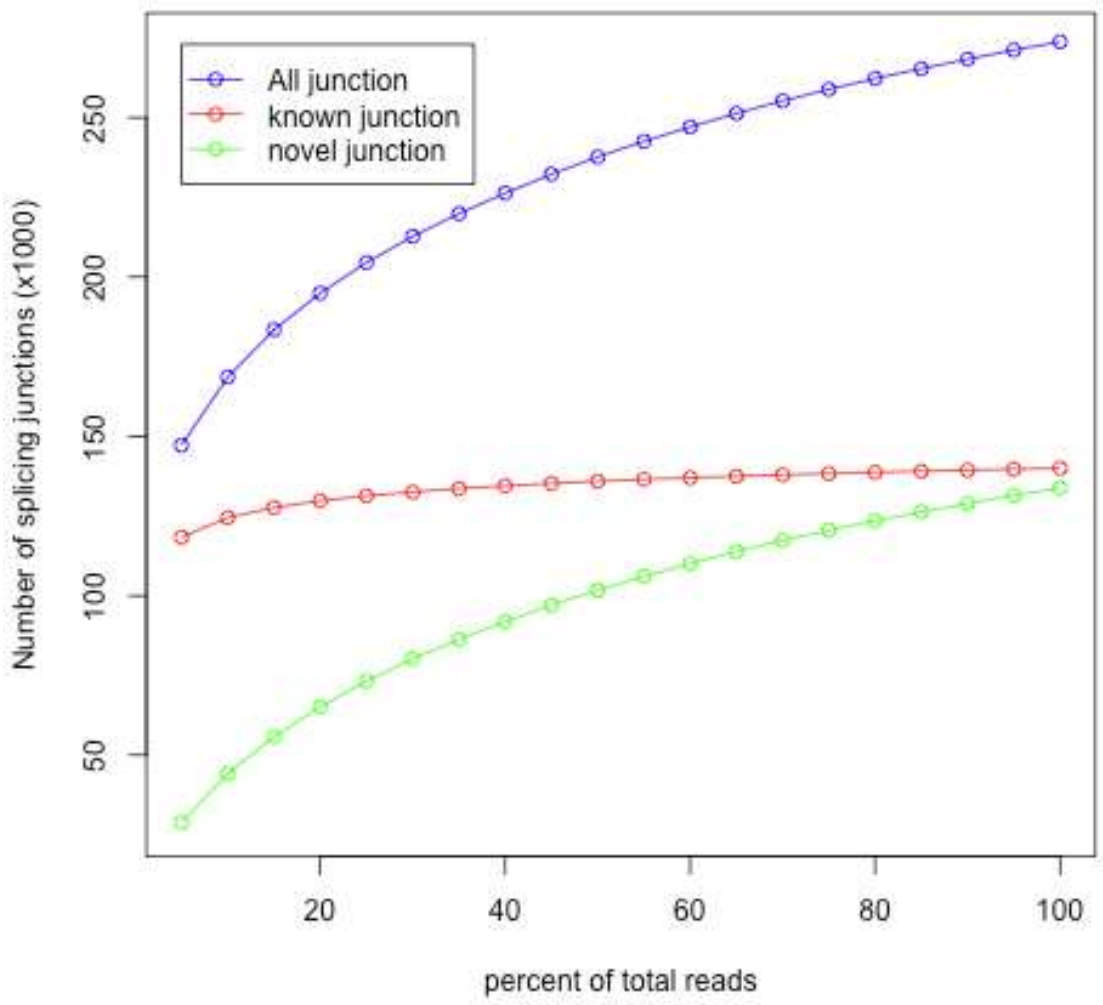
bam_stat.py

```
bam_stat.py  -i Pairend_nonStrandSpecific_36mer_Human_hg19.bam

#Output (all numbers are read count)
#==================================================
Total records:                                 41465027
QC failed:                                     0
Optical/PCR duplicate:                         0
Non Primary Hits                               8720455
Unmapped reads:                                0

mapq < mapq_cut (non-unique):                  3127757
mapq >= mapq_cut (unique):                     29616815
Read-1:                                        14841738
Read-2:                                        14775077
Reads map to '+':                              14805391
Reads map to '-':                              14811424
Non-splice reads:                              25455360
Splice reads:                                  4161455
Reads mapped in proper pairs:                  21856264
Proper-paired reads map to different chrom:    7648
```

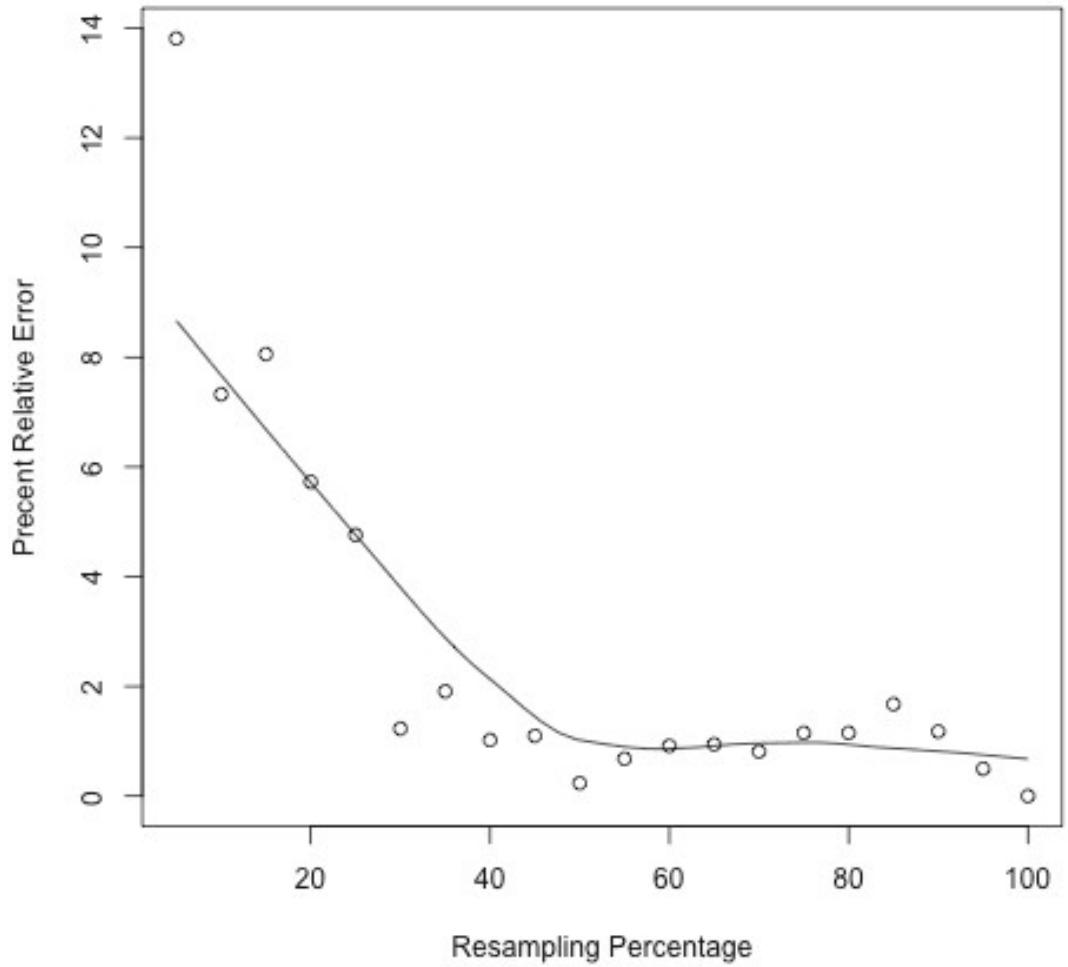
geneBody_coverage.py



junction_saturation.py



RPKM_saturation.py

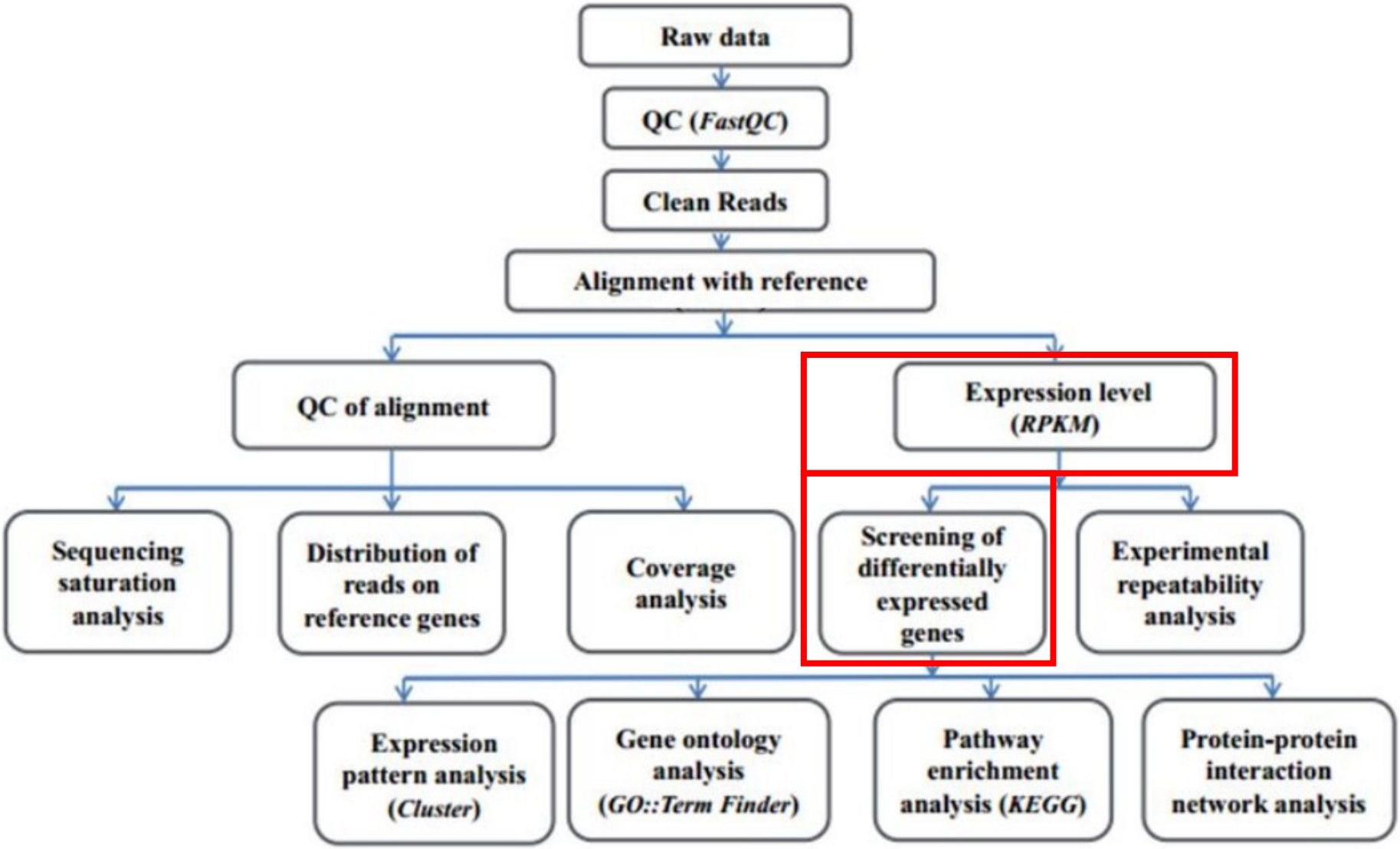


read_distribution.py

Group	Total_bases	Tag_count	Tags/Kb
CDS_Exons	33302033	20002271	600.63
5'UTR_Exons	21717577	4408991	203.01
3'UTR_Exons	15347845	3643326	237.38
Introns	1132597354	6325392	5.58
TSS_up_1kb	17957047	215331	11.99
TSS_up_5kb	81621382	392296	4.81
TSS_up_10kb	149730983	769231	5.14
TES_down_1kb	18298543	266161	14.55
TES_down_5kb	78900674	729997	9.25
TES_down_10kb	140361190	896882	6.39

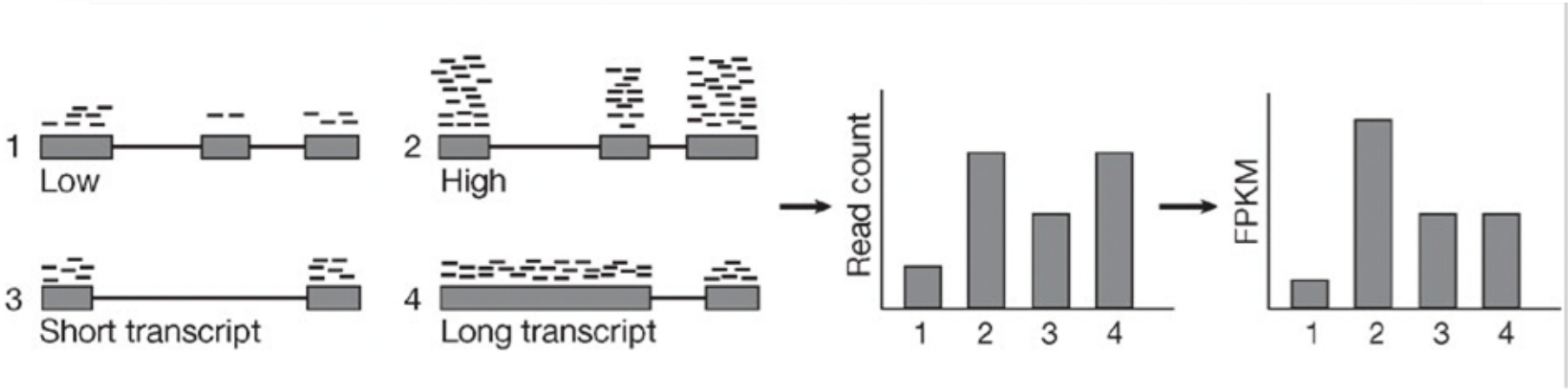
Core-analysis

Quantitative Analysis Pipeline



Normalization

基因名称	样本1	样本2	样本3
A (2kb)	10	12	30
B (4kb)	20	25	60
C (1kb)	5	8	15
D (10kb)	0	0	1
Total reads	35	45	106



Normalization Techniques

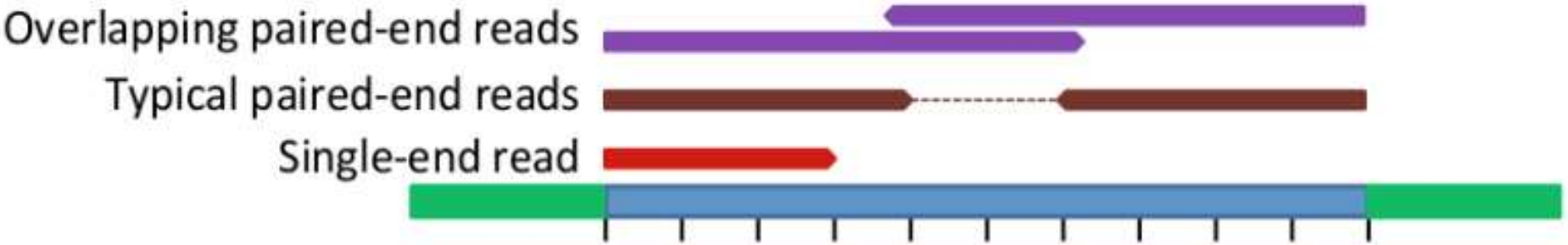
RPKM

Reads Per Kilobase Million reads mapped is a common normalization procedure.

$$\text{RPKM} = \frac{\text{total mapped to gene}}{\text{total mapped to lane (in millions)} \times \text{gene length (in kilobases)}}$$

Difference between RPKM and FPKM

RPKM:
Reads Per Kb of transcript per Million reads
FPKM:
Fragments Per Kb of transcript per Million reads



If we count reads rather than fragments, we might double-count some fragments but not others (second poor quality), leading to a skewed expression value. (Cufflink will use FPKM)

TPM

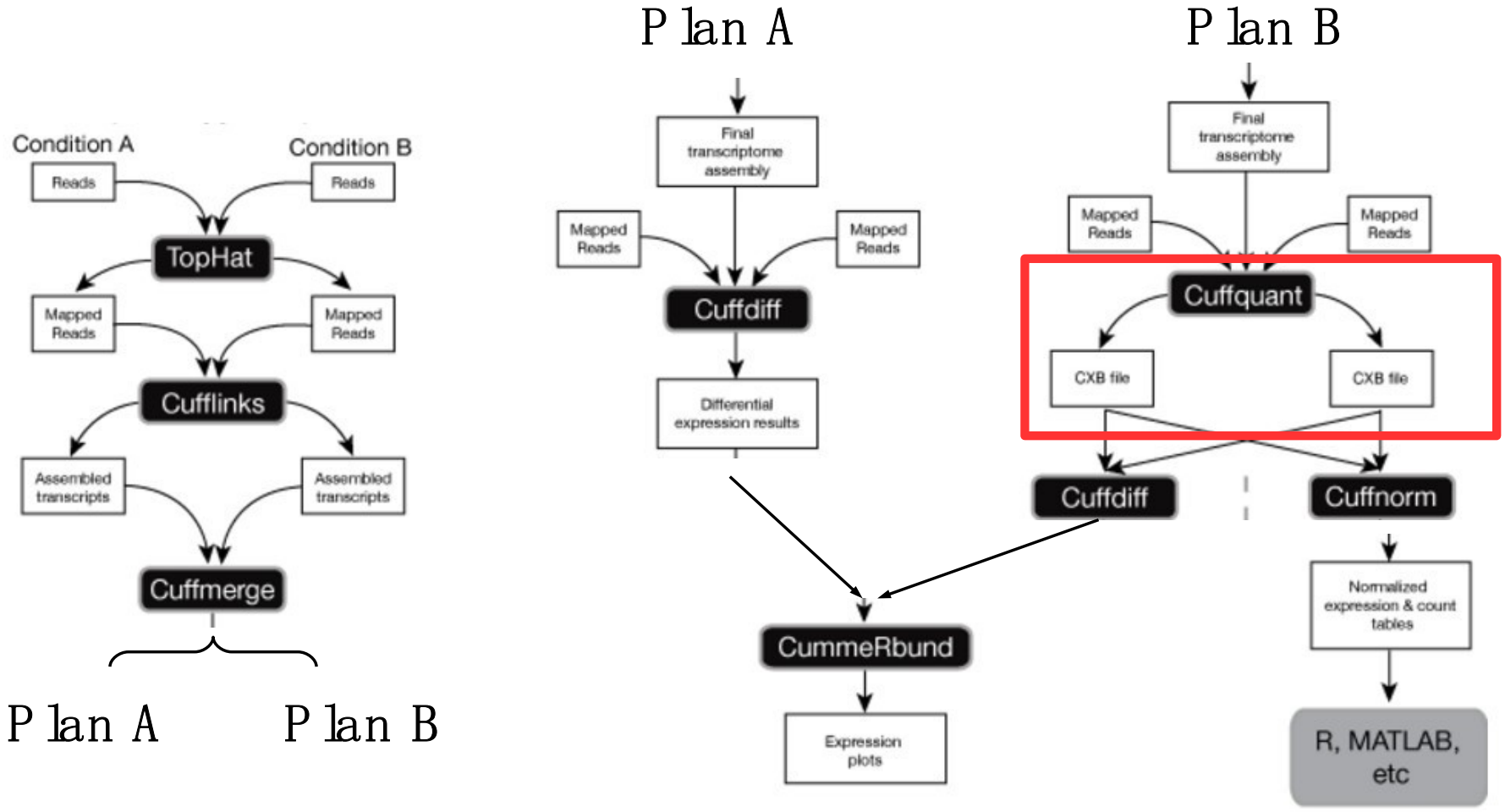
$$TPM = \frac{X_i}{L_i} * \frac{1}{\sum_j \frac{X_i}{L_j}} = \frac{\frac{X_i}{L_i}}{\sum_j \frac{X_i}{L_j}}$$

$$FPKM = \frac{X_i}{L_i} * \frac{1}{\sum_j X_j} = \frac{\frac{X_i}{L_i}}{\sum_j X_j}$$

假如要计算 A 基因的表达量，为了简单，我们这里不考虑可变剪接的情况，X_i 代表比对到 A 基因上的 Fragment 数目,单位为 Millon Fragments；L_i 代表 A 基因外显子的总长度, 单位为 Kilobase。

两个公式的分子时相同的，$\frac{X_i}{L_i}$ 代表外显子长度为 1Kb 的基因 Fragment 的数目,即转录本丰度。

Pipeline (http://cole-trapnell-lab.github.io/cufflinks/manual/)



eg: Plan A (Mouse RNA seq)

a. Build index with bowtie2-build

bowtie2-build /home/genomewide/refgenome/mm10/mm10.fa ./mm10.fa

```
[zy@rna bowtie_idx]$ bowtie2-build /home/genomewide/refgenome/mm10/mm10.fa ./mm10.fa
Settings:
  Output files: "./mm10.fa.*.bt2"
  Line rate: 6
```

```
[zy@rna course_rna_seq]$ bowtie2-build
No input sequence or sequence file specified!
Bowtie 2 version 2.3.4.1 by Ben Langmead (langmea@cs.jhu.edu, www.cs.jhu.edu/~langmea)
Usage: bowtie2-build [options]* <reference_in> <bt2_index_base>
    reference_in            comma-separated list of files with ref sequences
    bt2_index_base          write bt2 data to files with this dir/basename
*** Bowtie 2 indexes work only with v2 (not v1).  Likewise for v1 indexes. ***
Options:
    -f                      reference files are Fasta (default)
    -c                      reference sequences given on cmd line (as
                            <reference_in>)
    --large-index           force generated index to be 'large', even if ref
                            has fewer than 4 billion nucleotides
    -a/--noauto             disable automatic -p/--bmax/--dcv memory-fitting
    -p/--packed             use packed strings internally; slower, less memory
    --bmax <int>            max bucket sz for blockwise suffix-array builder
    --bmaxdivn <int>        max bucket sz as divisor of ref len (default: 4)
    --dcv <int>             diff-cover period for blockwise (default: 1024)
```

```
[zy@rna bowtie_tophat_cuff]$ ls ./bowtie_idx/
logs           mm10.fa.2.bt2  mm10.fa.4.bt2      mm10.fa.rev.2.bt2  mm10_bwt.fa.2.bt2  mm10_bwt.fa.4.bt2
mm10.fa.1.bt2  mm10.fa.3.bt2  mm10.fa.rev.1.bt2  mm10_bwt.fa.1.bt2  mm10_bwt.fa.3.bt2  tmp
```

eg: Plan A (Mouse RNA seq)

b. Mapping with tophat2

$ tophat2 -p 10 -o output_dir ../index/mm9 read1.fq read2.fq

```
[zy@rna bowtie_tophat_cuff]$ cd result_tophat2
[zy@rna result_tophat2]$ ls
SRR7468998  SRR7468999  SRR7469000  SRR7469001  SRR7469002  SRR7469003  logs  prep_reads.info  tmp
[zy@rna result_tophat2]$ cd SRR7468998
[zy@rna SRR7468998]$ ls
accepted_hits.bam  align_summary.txt  deletions.bed  insertions.bed  junctions.bed  logs  prep_reads.info
unmapped.bam
```

BED files:

chr1 32 33 G * +

```python
import os
import re
import subprocess


def tophat(path_trim_fq, path_index, path_output, process=2, num_sample=6):
    # read fastq files of paired-end sequencing
    fastq_files = os.listdir(path_trim_fq)
    pair1 = []
    pair2 = []
    for file in fastq_files:
        if file[-5:] == '_1.fq':
            pair1.append(file)
        elif file[-5:] == '_2.fq':
            pair2.append(file)

    # verify matching correctly and join path and filename
    pair1.sort()
    pair2.sort()
    pattern = re.compile(r'SRR(\d+?)_')
    path_pair1 = []
    path_pair2 = []
    for i in range(len(pair2)):
        pair_1 = pattern.search(pair1[i]).group()[3:-1]
        pair_2 = pattern.search(pair2[i]).group()[3:-1]
        assert pair_1 == pair_2
        path_pair1.append(os.path.join(path_trim_fq, pair1[i]))
        path_pair2.append(os.path.join(path_trim_fq, pair2[i]))

    # run tophat2
    subprocesses = []
    for i in range(len(pair1)):
        if i % process == 0:
            for sub in subprocesses:
                sub.wait()
            subprocesses = []
        if i >= num_sample:
            break
        subprocesses.append(subprocess.Popen('tophat2 -p 10 -o ' +
                                             os.path.join(path_output, pattern.search(pair1[i]).group()[:-1])
                                             + ' ' + path_index + ' ' + path_pair1[i]
                                             + ' ' + path_pair2[i], shell=True))
```

eg: Plan A (Mouse RNA seq)

c. Get transcripts and expression level

$ cufflinks -p 5 -u -g mm9.gtf -o output_dir accepted_hits.bam

Output:

```
[zy@rna bowtie_tophat_cuff]$ cd cufflinks_out
[zy@rna cufflinks_out]$ ls
SRR7468998  SRR7468999  SRR7469000  SRR7469001  SRR7469002  SRR7469003
[zy@rna cufflinks_out]$ cd SRR7468998
[zy@rna SRR7468998]$ ls
genes.fpkm_tracking  isoforms.fpkm_tracking  skipped.gtf  transcripts.gtf
```

d. Merge transcripts

$ cuffmerge -g mm10.gtf -s mm10.fa -p 10 -o new_gtf assemblies.txt

Output:

```
[zy@rna cuff_merge]$ ls
genes.fpkm_tracking  isoforms.fpkm_tracking  logs  merged.gtf  skipped.gtf  tmp  transcripts.gtf
```

```python
import os
import subprocess


def cufflinks(bam_path, cufflinks_out, process=3, thread=10):
    # read samples from results of tophat
    samples = os.listdir(bam_path)
    list_srr = []
    for sample in samples:
        if sample[:3] == 'SRR' and os.path.isdir(os.path.join(bam_path, sample)):
            list_srr.append(sample)

    # run cufflinks
    subprocesses = []
    for i in range(len(list_srr)):
        if i % process == 0:
            for sub in subprocesses:
                sub.wait()
            subprocesses = []

        subprocesses.append(subprocess.Popen("cufflinks -p " + str(thread) +
                                             " -o " + os.path.join(cufflinks_out, list_srr[i]) + ' '
                                             + os.path.join(os.path.join(bam_path, list_srr[i]), 'accepted_hits.bam'),
                                             shell=True))

    return
```

```python
def cuffmerge(cufflinks_path, cuffmerge_path, gtf_file, fasta_file, thread=20):
    # build assembiles.txt
    samples = os.listdir(cufflinks_path)
    with open(os.path.join(cufflinks_path, 'assemblies.txt'), 'w') as w_asm:
        for sample in samples:
            if sample[:3] == 'SRR' and os.path.isdir(os.path.join(cufflinks_path, sample)):
                w_asm.write(os.path.isdir(os.path.join(cufflinks_path, sample)) + 'transcripts.gtf' + '\n')

    # run cuffmerge
    os.system("cuffmerge -g " + gtf_file + " -s " + fasta_file + " -o " + cuffmerge_path + " -p " + str(thread) + " " +
              os.path.join(cufflinks_path, 'assemblies.txt'))

    return
```

eg: Plan A (Mouse RNA seq)

e. Differential expression

$ cuffdiff -o output_dir -p 10 -L case,control -u merged.gtf
case_rep1/accepted_hits.bam,case_rep2/accepted_hits.bam
control_rep1/accepted_hits.bam,control_rep2/accepted_hits.bam

Output:

```
[zy@rna cuff_diff]$ ls
bias_params.info          gene_exp.diff             isoforms.fpkm_tracking        tss_group_exp.diff
cds.count_tracking        genes.count_tracking      isoforms.read_group_tracking  tss_groups.count_tracking
cds.diff                  genes.fpkm_tracking       promoters.diff                tss_groups.fpkm_tracking
cds.fpkm_tracking         genes.read_group_tracking read_groups.info              tss_groups.read_group_tracking
cds.read_group_tracking   isoform_exp.diff          run.info                      var_model.info
cds_exp.diff              isoforms.count_tracking   splicing.diff
```

gene_exp.diff

```python
def cuffdiff(bam_path, cuffmerge_path, cuffdiff_path, fasta_file, group_name, num_group1, thread=20):
    # divide into groups
    samples = os.listdir(bam_path)
    all_sample = []
    for sample in samples:
        if sample[:3] == 'SRR' and os.path.isdir(os.path.join(bam_path, sample)):
            all_sample.append(os.path.join(os.path.join(bam_path, sample), 'accepted_hits.bam'))

    group1 = all_sample.sort()[:num_group1]
    group2 = all_sample.sort()[num_group1:]

    # run cuffdiff
    os.system("cuffdiff -o " + cuffdiff_path + " -b " + fasta_file + " -p " + str(thread) +
              " -L " + group_name + " -u " + os.path.join(cuffmerge_path, 'merge.gtf') + ' ' +
              ','.join(group1) + ' ' + ','.join(group2))

    return
```

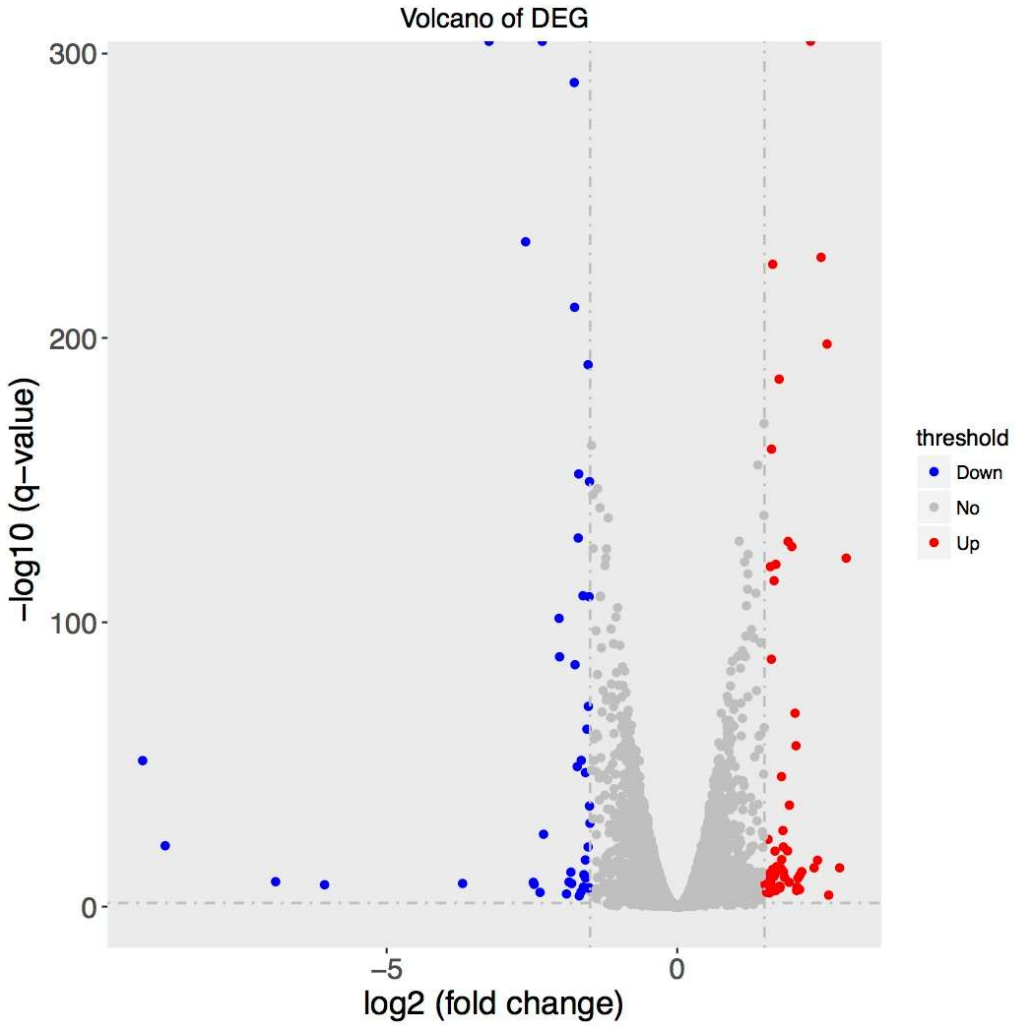
test_id	gene_id	gene	locus	sample_1	sample_2	status	value_1	value_2	log2(fold_change)	test_stat	p_value	q_value	significant
XLOC_000001	XLOC_000001	Lypla1	chr1:4807892-4846735	Control	Brg1_iko	OK	10.7641	10.776	0.00160117	0.00459412	0.99265	0.998533	
XLOC_000002	XLOC_000002	Tcea1	chr1:4857693-4897909	Control	Brg1_iko	OK	5.86869	6.82636	0.218079	0.487167	0.4844	0.984994	
XLOC_000003	XLOC_000003	Atp6v1h	chr1:5083172-5162549	Control	Brg1_iko	OK	74.616	73.2649	-0.0263614	-0.094337	0.89475	0.997713	
XLOC_000004	XLOC_000004	Oprk1	chr1:5588492-5606133	Control	Brg1_iko	NOTEST	0	0	0	0	1	1	no
XLOC_000005	XLOC_000005	Rb1cc1	chr1:6214661-6276104	Control	Brg1_iko	OK	6.38139	5.16199	-0.305942	-0.903573	0.19915	0.895006	
XLOC_000006	XLOC_000006	Fam150a	chr1:6359330-6394731	Control	Brg1_iko	NOTEST	0	0	0	1	1	no	
XLOC_000007	XLOC_000007	St18	chr1:6487230-6860940	Control	Brg1_iko	NOTEST	0.0444639	0.117281	1.39927	0	1	1	no
XLOC_000008	XLOC_000008	Pcmtd1	chr1:7088919-7173628	Control	Brg1_iko	OK	7.87522	6.24575	-0.334445	-0.991377	0.15855	0.873933	
XLOC_000009	XLOC_000009	Rrs1	chr1:9545407-9547455	Control	Brg1_iko	OK	68.8605	65.8784	-0.0638713	-0.213487	0.769	0.997713	
XLOC_000010	XLOC_000010	Adhfe1	chr1:9548045-9631092	Control	Brg1_iko	OK	5.45365	5.48125	0.00728087	0.00439239	0.9929	0.998533	
XLOC_000011	XLOC_000011	3110035E14Rik	chr1:9548045-9631092	Control	Brg1_iko	OK	226.99	197.899	-0.197863	-0.704205	0.3085	0.95476	
XLOC_000012	XLOC_000012	1700034P13Rik	chr1:9718621-9771256	Control	Brg1_iko	OK	11.7949	5.7796	-1.02912	-1.02542	0.1621	0.87501	
XLOC_000013	XLOC_000013	Sgk3	chr1:9798129-9902568	Control	Brg1_iko	OK	5.95722	6.46734	0.118534	0.342095	0.6299	0.995528	
XLOC_000014	XLOC_000014	Mcmdc2	chr1:9908637-9940954	Control	Brg1_iko	NOTEST	0.304122	0.404906	0.412939	0	1	1	
XLOC_000015	XLOC_000015	Cspp1	chr1:10038217-10136768	Control	Brg1_iko	OK	3.15456	3.23183	0.0349142	0.0714437	0.921	0.997713	
XLOC_000016	XLOC_000016	Prex2	chr1:10993464-11303682	Control	Brg1_iko	NOTEST	0.0072481	0.0591681	3.02914	0	1	1	no
XLOC_000017	XLOC_000017	A830018L16Rik	chr1:11414104-11975902	Control	Brg1_iko	NOTEST	0.157796	0.125573	-0.329536	0	1		
XLOC_000018	XLOC_000018	Mir6341	chr1:12425985-12426106	Control	Brg1_iko	NOTEST	0	0	0	0	1	1	no
XLOC_000019	XLOC_000019	Gm17644	chr1:12667562-12673090	Control	Brg1_iko	NOTEST	0	0	0	0	1	1	no
XLOC_000020	XLOC_000020	Sulf1	chr1:12692429-12860372	Control	Brg1_iko	OK	1.92755	1.86713	-0.0459456	-0.0803772	0.9096	0.997713	
XLOC_000021	XLOC_000021	Xkr9	chr1:13668770-13701723	Control	Brg1_iko	NOTEST	0	0	0	0	1	1	no
XLOC_000022	XLOC_000022	Kcnb2	chr1:15312451-15714214	Control	Brg1_iko	OK	1.59325	1.06372	-0.582852	-0.893264	0.205	0.901867	
XLOC_000023	XLOC_000023	Terf1	chr1:15805645-15844052	Control	Brg1_iko	OK	13.8128	12.6411	-0.127889	-0.33132	0.6411	0.995528	
XLOC_000024	XLOC_000024	Rdh10	chr1:16105881-16132550	Control	Brg1_iko	OK	9.06926	18.5476	1.03218	2.53085	0.0014	0.137319	no
XLOC_000025	XLOC_000025	D030040B21Rik	chr1:16657551-16662278	Control	Brg1_iko	NOTEST	0	0	0	0	1	1	no
XLOC_000026	XLOC_000026	Tmem70	chr1:16665190-16678275	Control	Brg1_iko	OK	38.497	50.5297	0.392387	1.27725	0.0682	0.746483	no
XLOC_000027	XLOC_000027	Ly96	chr1:16688455-16709605	Control	Brg1_iko	OK	1.59227	3.98062	1.32191	0.985801	0.1604	0.874785	no
XLOC_000028	XLOC_000028	Gdap1	chr1:17145372-17164270	Control	Brg1_iko	OK	10.8386	11.4984	0.0852504	0.265927	0.7073	0.995528	
XLOC_000029	XLOC_000029	Pi15	chr1:17601900-17630939	Control	Brg1_iko	NOTEST	0	0	0	0	1	1	no
XLOC_000030	XLOC_000030	Crispld1	chr1:17676026-17766207	Control	Brg1_iko	NOTEST	0.432546	0.782959	0.856084	0	1		
XLOC_000031	XLOC_000031	Tfap2d	chr1:19103021-19166332	Control	Brg1_iko	NOTEST	0	0	0	1	1	no	
XLOC_000032	XLOC_000032	Tfap2b	chr1:19208913-19238845	Control	Brg1_iko	NOTEST	0.271816	0.366539	0.431334	0	1	1	
XLOC_000033	XLOC_000033	4930486I03Rik	chr1:20057778-20618057	Control	Brg1_iko	NOTEST	0	0	0	0	1	1	no
XLOC_000034	XLOC_000034	Linc-md1,Mir133b	chr1:20669881-20682958	Control	Brg1_iko	NOTEST	0	0	0	0	1	1	no
XLOC_000035	XLOC_000035	Mir206	chr1:20669881-20682958	Control	Brg1_iko	NOTEST	0	0	0	0	1	1	no
XLOC_000036	XLOC_000036	Il17a	chr1:20730904-20734496	Control	Brg1_iko	NOTEST	0	0	0	0	1	1	no
XLOC_000037	XLOC_000037	Paqr8	chr1:20890621-20938756	Control	Brg1_iko	OK	3.06714	2.98445	-0.0394315	-0.0795528	0.9104	0.997713	
XLOC_000038	XLOC_000038	Efhc1	chr1:20951625-20990841	Control	Brg1_iko	NOTEST	0.411267	0.621886	0.596573	0	1	1	
XLOC_000039	XLOC_000039	Tmem14a	chr1:21218574-21230167	Control	Brg1_iko	OK	16.7748	23.1066	0.462006	1.04956	0.13785	0.872802	no
XLOC_000040	XLOC_000040	Gsta3	chr1:21240584-21265575	Control	Brg1_iko	NOTEST	0.757937	0.445547	-0.766499	0	1	1	
XLOC_000041	XLOC_000041	Khdc1a	chr1:21349676-21352199	Control	Brg1_iko	NOTEST	0	0	0	0	1	1	no
XLOC_000042	XLOC_000042	Khdc1c	chr1:21368330-21369743	Control	Brg1_iko	NOTEST	0	0	0	0	1	1	no
XLOC_000043	XLOC_000043	Khdc1b	chr1:21383555-21386384	Control	Brg1_iko	NOTEST	0	0	0	0	1	1	no
XLOC_000044	XLOC_000044	Mir30a	chr1:23272268-23272339	Control	Brg1_iko	NOTEST	0	0	0	0	1	1	no
XLOC_000045	XLOC_000045	Mir30c-2	chr1:23291700-23291784	Control	Brg1_iko	NOTEST	0	0	0	0	1	1	no
XLOC_000046	XLOC_000046	B3gat2	chr1:23761925-23922381	Control	Brg1_iko	OK	10.2832	9.71283	-0.0823281	-0.156836	0.8211	0.997713	
XLOC_000047	XLOC_000047	Col9a1	chr1:24177609-24252738	Control	Brg1_iko	NOTEST	0.814154	0.775459	-0.0702509	0	1	1	
XLOC_000048	XLOC_000048	Lmbrd1	chr1:24678543-24766301	Control	Brg1_iko	OK	19.8276	24.2612	0.291147	1.02647	0.14355	0.872802	no
XLOC_000049	XLOC_000049	Mir6342	chr1:29421487-29421612	Control	Brg1_iko	NOTEST	0	0	0	0	1	1	no

gene_exp.diff

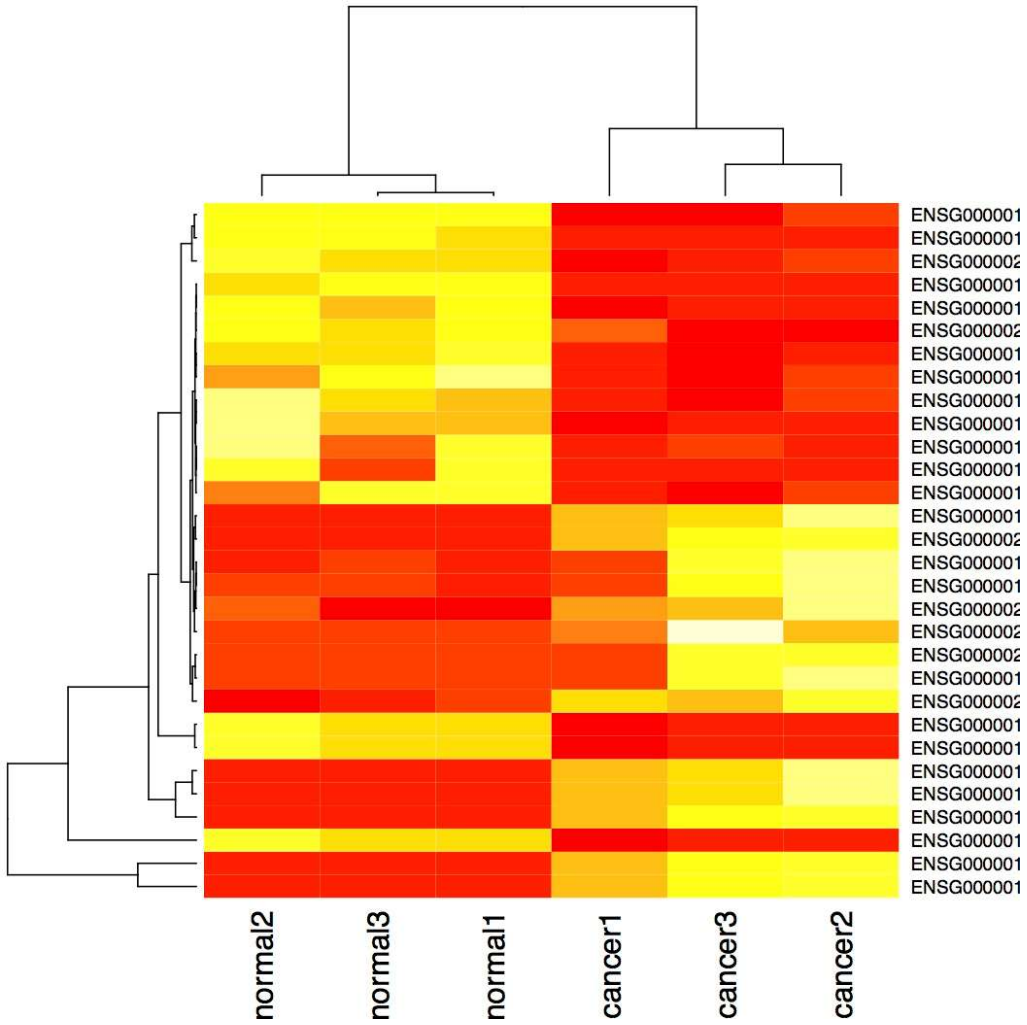
abs(fold_change) >= 1.5, FDR < 0.05

p.adjust(p_value, method="fdr")

Visualization – Volcano Plot



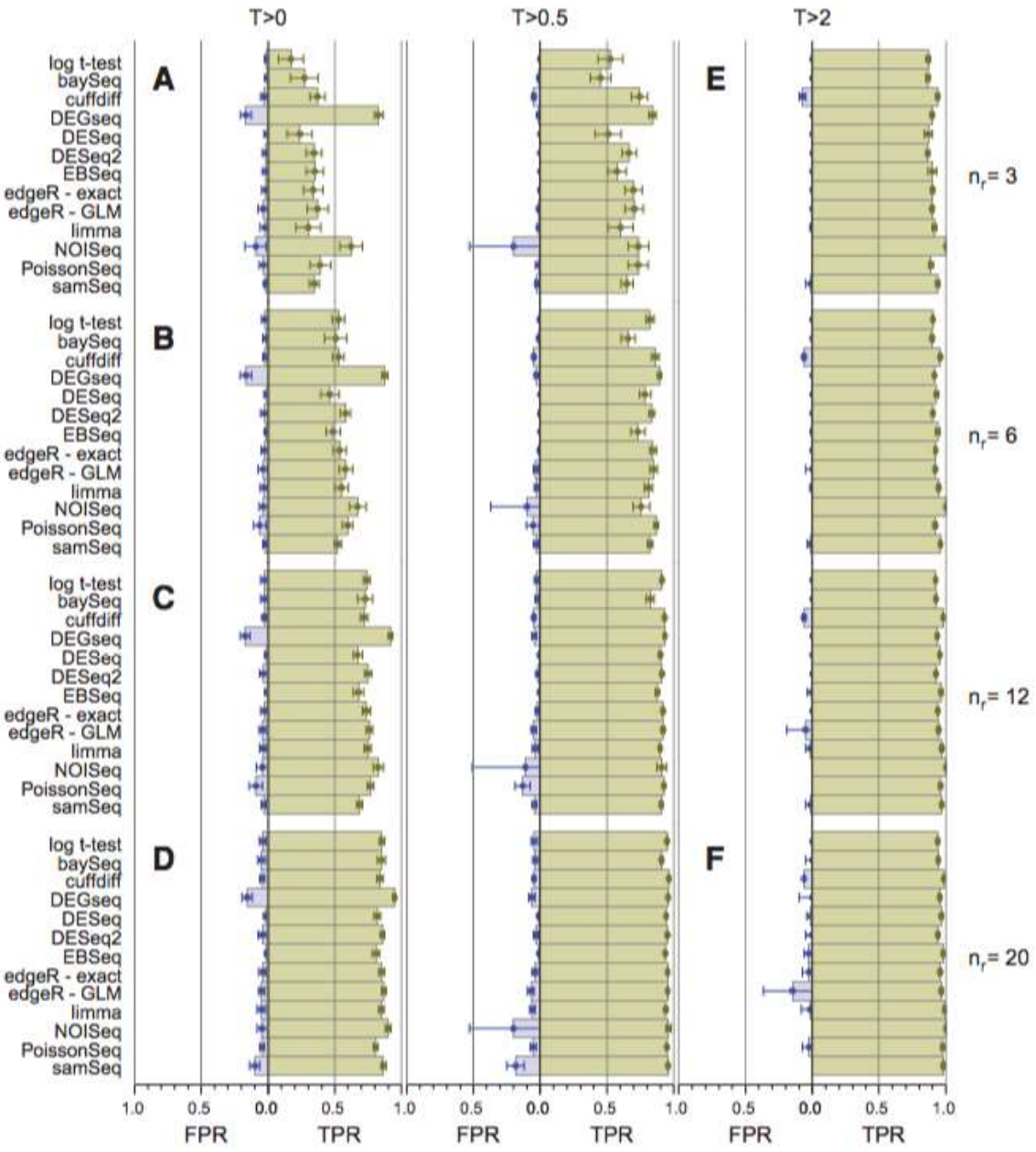
Visualization - heatmap



comparison

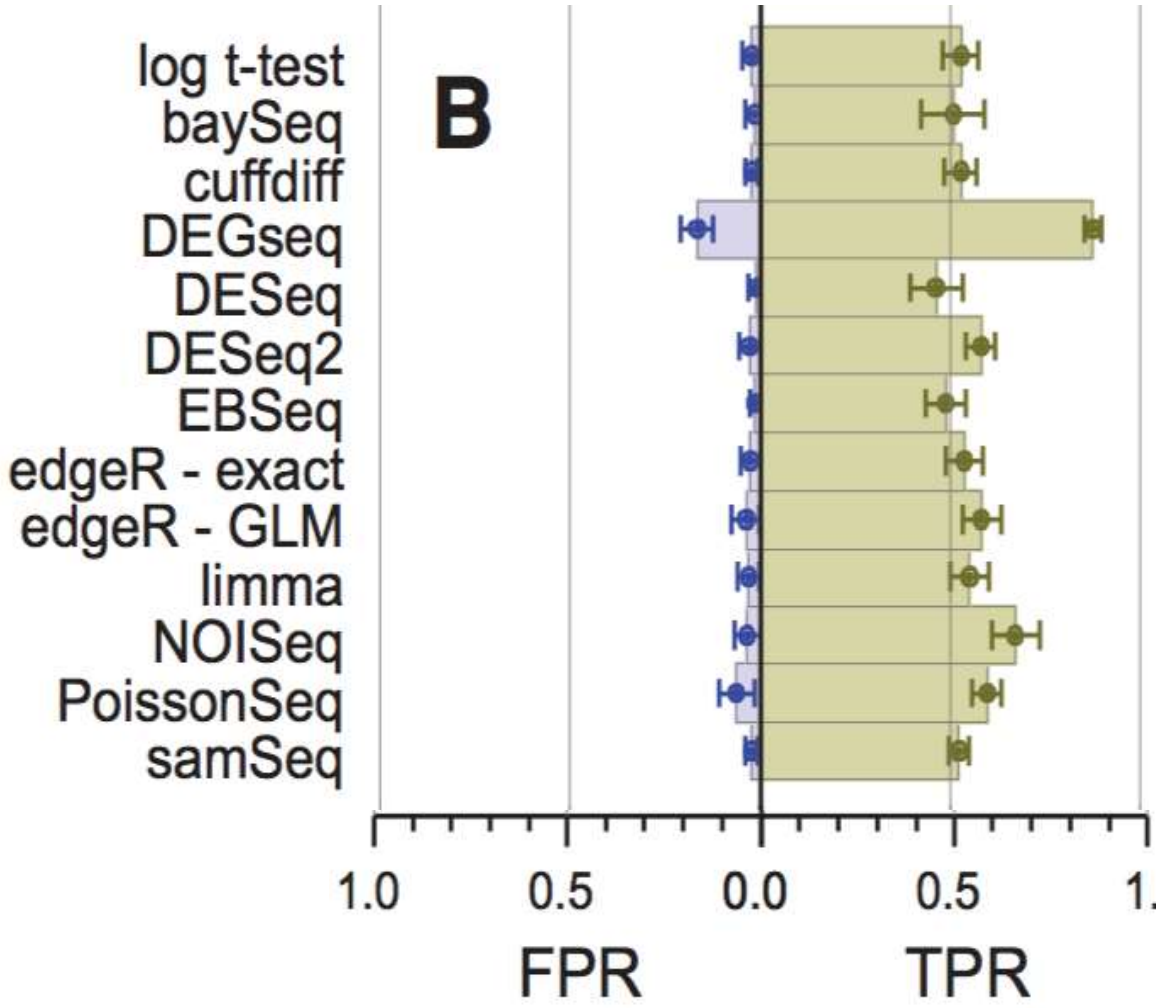
More samples

More confidence

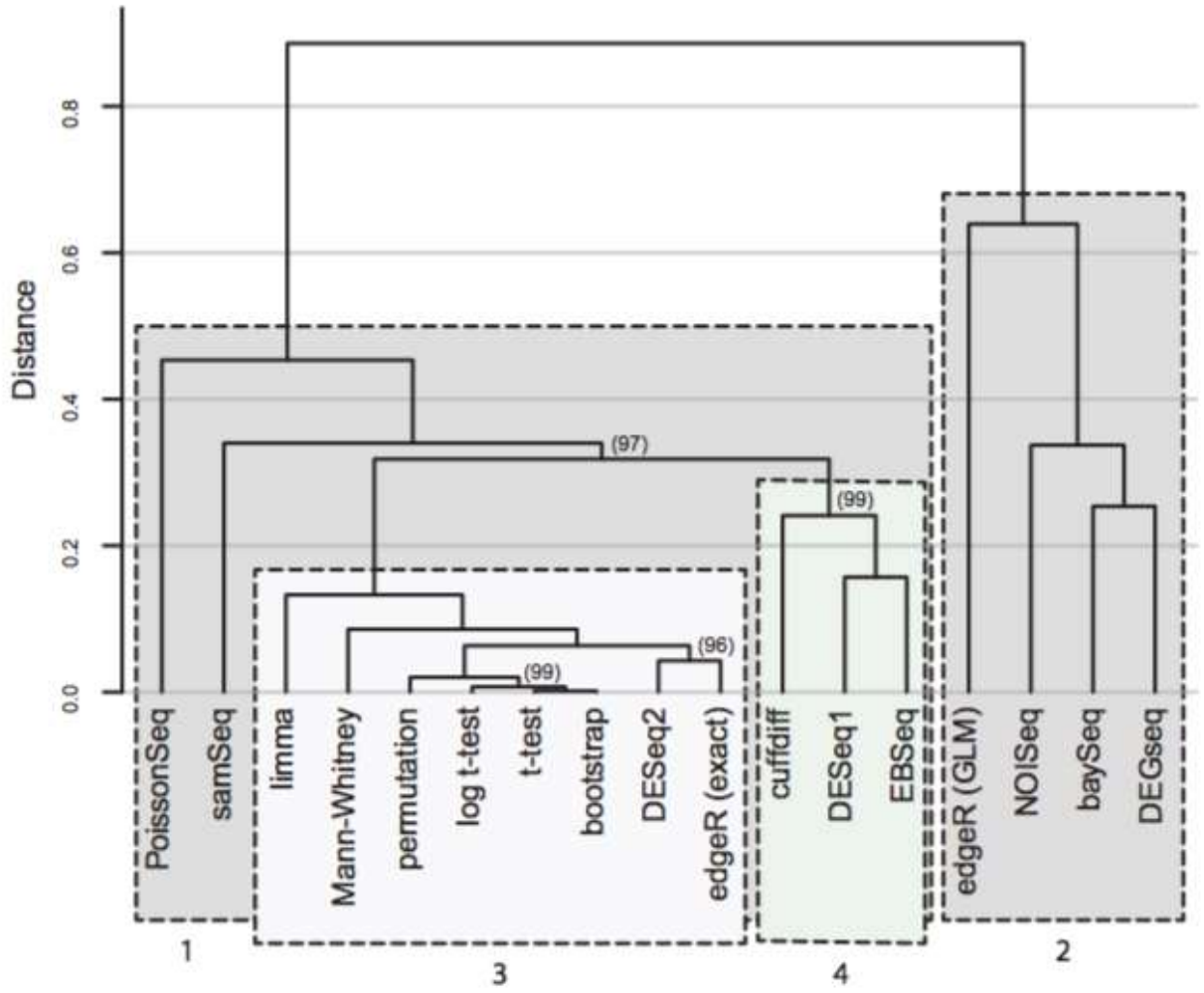


comparison

Little discrepancy among different methods



comparison


Distance
0.0
0.2
0.4
0.6
0.8
(97)
(99)
(96)
(99)
PoissonSeq
samSeq
limma
Mann-Whitney
permutation
log t-test
t-test
bootstrap
DESeq2
edgeR (exact)
cuffdiff
DESeq1
EBSeq
edgeR (GLM)
NOISeq
baySeq
DEGseq
1
3
4
2

comparison

- FPKM / TPM
  - t.test (log(FPKM+1))
  - Cuffdiff
- CPM (Reads counts)
  - DEseq2
  - edgeR

Differential gene expression analysis

- Array-based:

  t-test

  wilcox test

  permutation

- Sequencing-based

  Fisher’s exact test

  Poisson, Negative Binomial

  Likelihood ratio test

Fisher's exact test

- The Fisher's exact test can be used for RNA-seq data without replicates, proceeding on a gene-by-gene basis and organizing the data in a 2*2 contingency table.

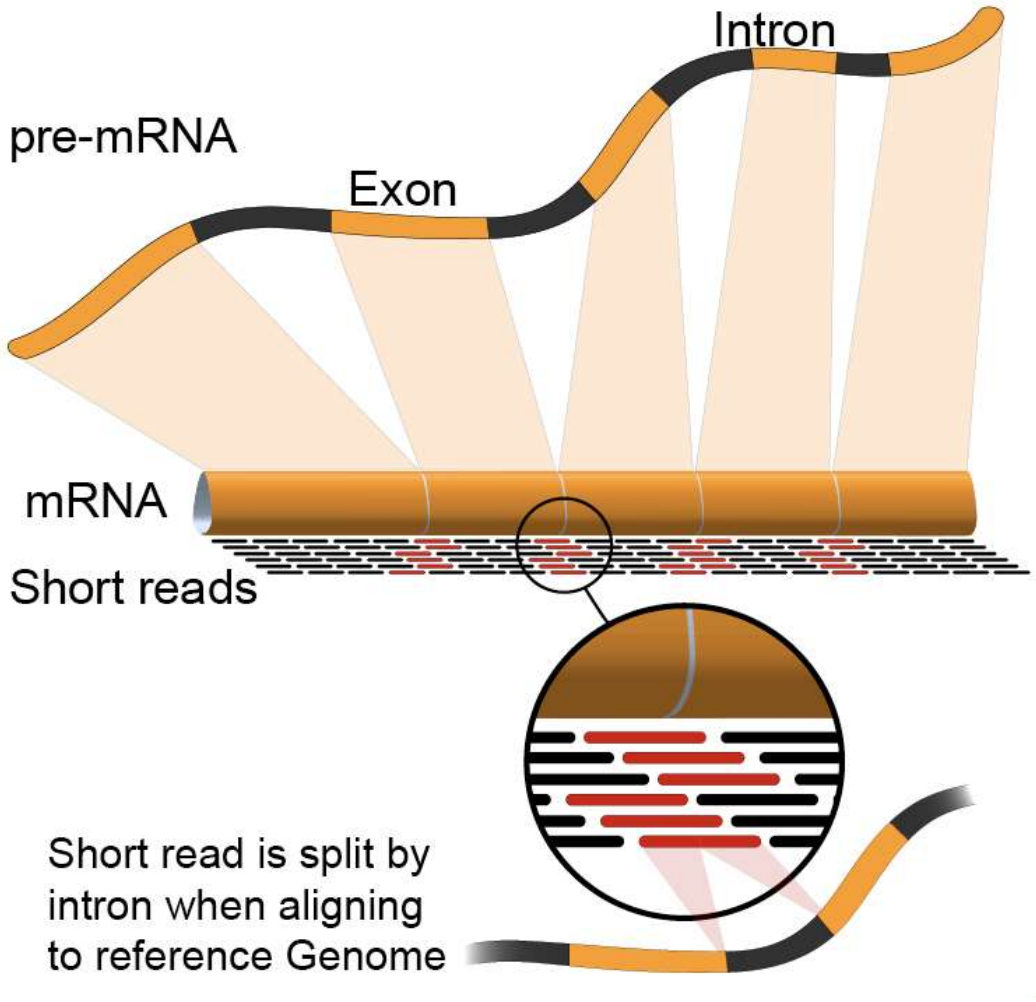
	condition 1	condition 2	Total
Gene x	n_{11}	n_{12}	$n_{11} + n_{12}$
Remaining genes	n_{21}	n_{22}	$n_{21} + n_{22}$
Total	$n_{11} + n_{21}$	$n_{12} + n_{22}$	N

Fisher's exact test

To make a hypothesis test out of this, we need to calculate the probability of observing some number k or more reads n_{11} in order to have a statistical test. In this case, the sum over the tail of the hypergeometric distribution is known as the *Exact Fisher Test*:

$$p(\text{read count} \geq n_{11}) = \sum_{k=n_{11}}^{n_{11}+n_{12}} \frac{\binom{k+n_{12}}{k}\binom{n_{21}+n_{22}}{n_{21}}}{\binom{n}{k+n_{21}}}$$

Poisson Model



Poisson Model

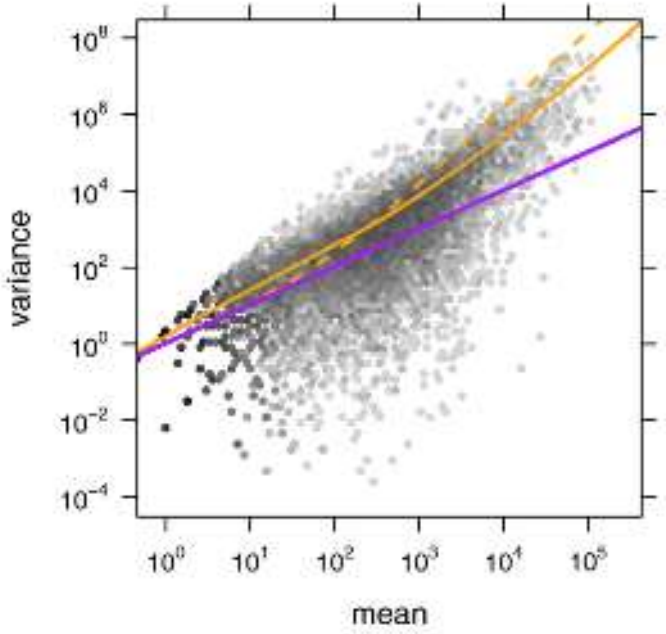
- Imagine we have count data for some list of genes *g1*, *g2*, ... with technical and biological replicates corresponding to two conditions we want to compare.
- We will let $X \sim$ Poisson (λ) be a random variable representing the number of reads falling in *g*.

Problems with Poisson

- Many studies have shown that the variance grows faster than the mean in RNAseq data. This is known as overdispersion.

Orange line: the fitted observed curve.

Purple: the variance implied by the Poisson distribution.



Negative Binomial

- The negative binomial distribution can be used as an alternative to the Poisson distribution. It is especially useful for discrete data over an unbounded positive range whose sample variance exceeds the sample mean.
- The negative binomial has two parameters, the mean $p \in [0,1]$ and $r \in Z$, where p is the probability of a single success and is the total number of successes (here: read counts).

$$NB(K = k) = \binom{k + r - 1}{r - 1} p^r (1 - p)^k$$

And **r=nφ**， φ is divergence

Likelihood ratio test

- The likelihood ratio test is a statistical test that is used by many RNAseq algorithms to assess differential expression. It compares the likelihood of the data assuming no differential expression (null model) against the likelihood of the data assuming differential expression (alternative model).
- It can be shown that D follows a χ^2 distribution, and this can be used to calculate a *p value*.

$$D = -2 \log \frac{\text{likelihood of null model}}{\text{likelihood of alternative model}}$$

Probability of making at least one error

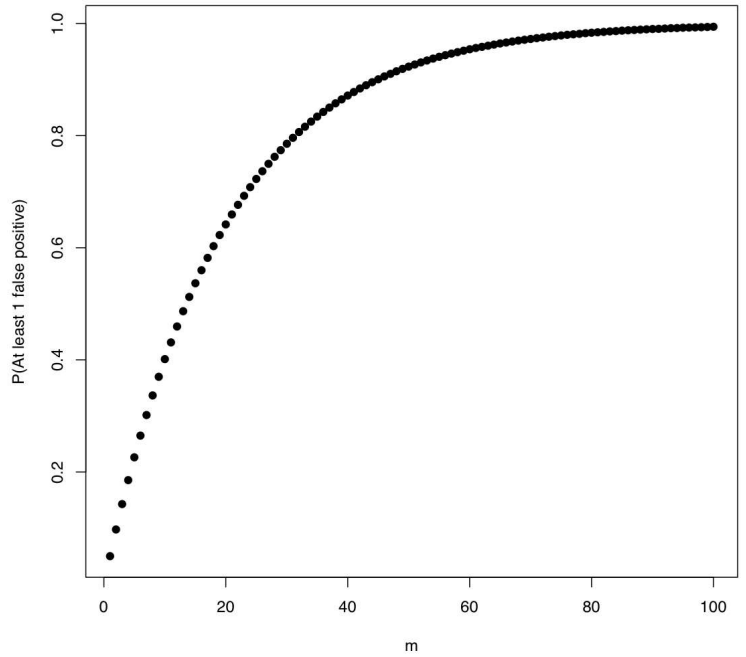
In general, if we perform m hypothesis tests, what is the probability of at least 1 false positive?

P(Making an error) = α

P(Not making an error) = 1 - α

P(Not making an error in m tests) = $(1 - \alpha)^m$

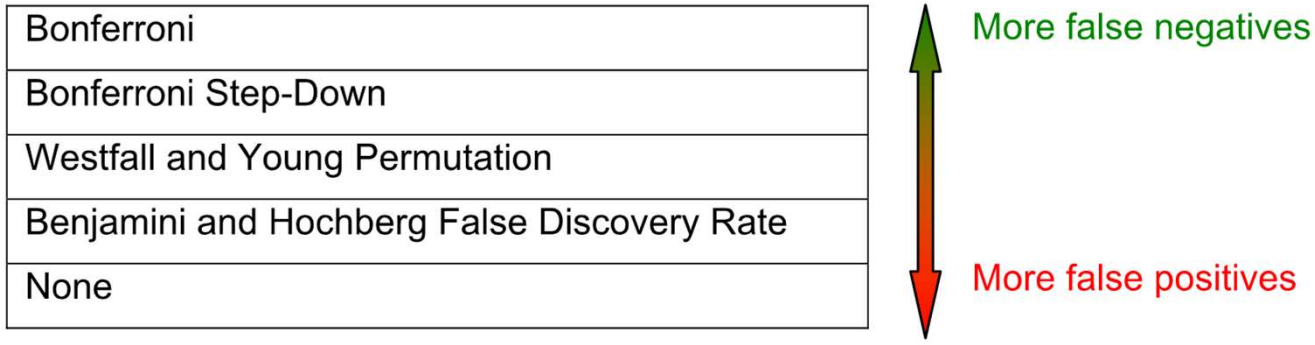
P(Making at least 1 error in m tests) = $1 - (1 - \alpha)^m$



Number of genes tested (N)	False positives incidence	Probability of calling 1 or more false positives by chance ($100(1-0.95^N)$)
1	1/20	5%
2	1/10	10%
20	1	64%
100	5	99.4%

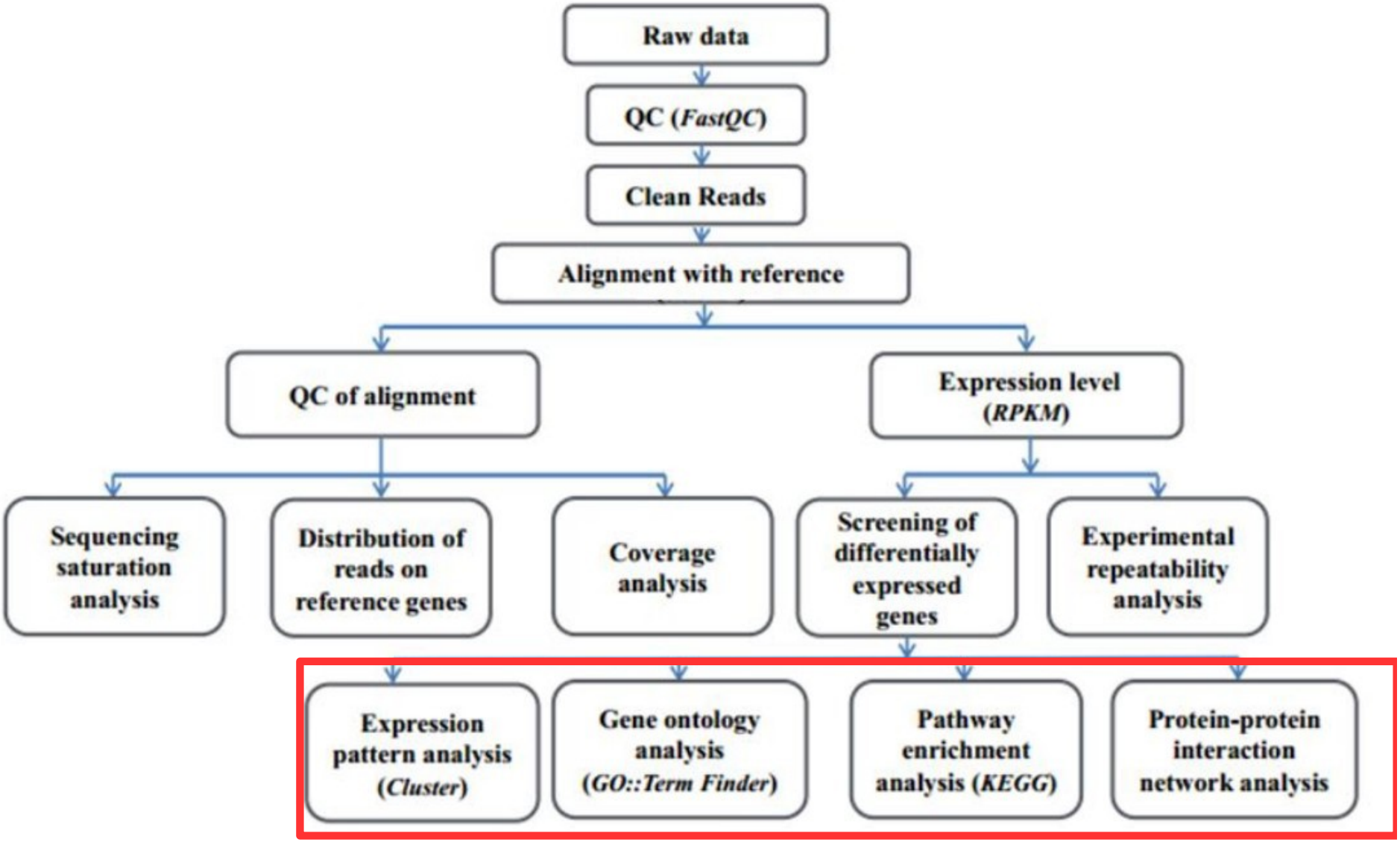
Multiple testing correction

- Bonferroni
- Bonferroni (Holm)
- Westfall and Young Permutation
- Benjamini and Hochberg False Discovery Rate



Advanced-analysis

Quantitative Analysis Pipeline



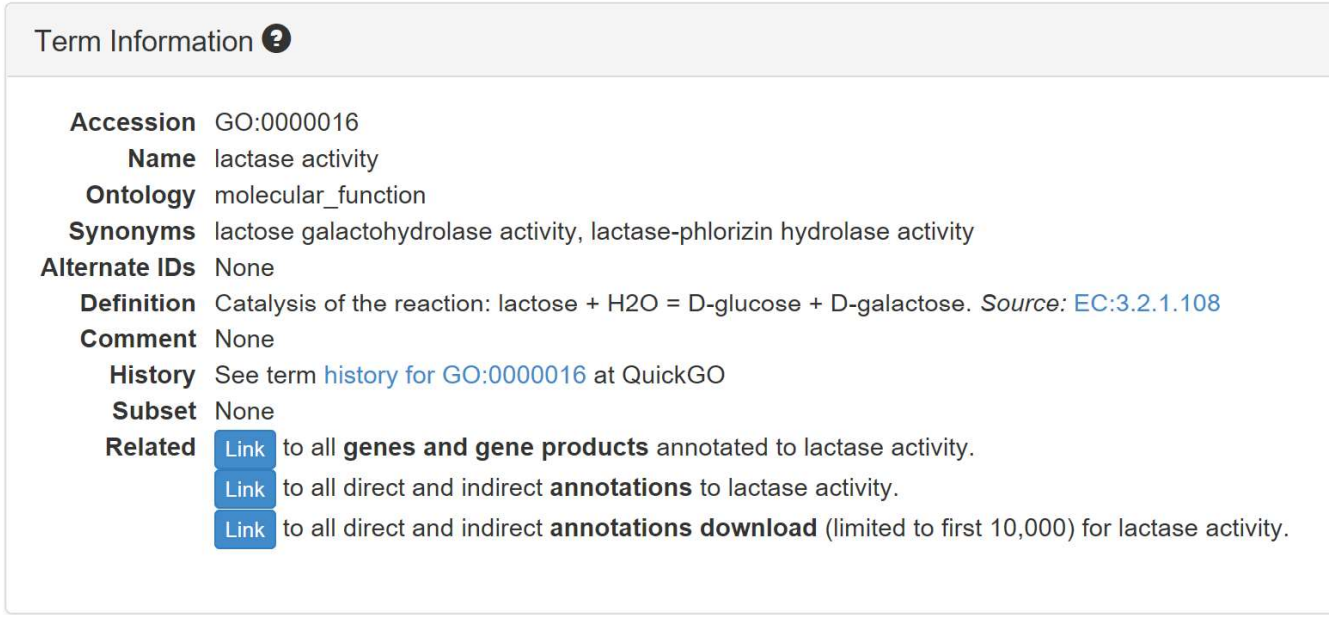
Enrichment analysis

- Why?
- By what? (Databases):
  - Function: Gene ontology
  - Pathway: KEGG & PANTHER
  - Protein domain: InterPro, Pfam
  - Protein interaction: DIP, MINT
- How? :
  - ORA: over-representation analysis
  - FCS: functional class scoring
  - PT: pathway topology
  - NT: network topology

Gene ontology

Gene ontology (GO) is a major bioinformatics initiative to unify the representation of gene and gene product attributes across all species.

The ontology covers three domains: cellular component, molecular function and biological process,



*www.**geneontology**.org/*

Example GO term

```
id:         GO:0000016
name:       lactase activity
namespace:  molecular_function
def:        "Catalysis of the reaction: lactose + H2O = D-glucose + D-galactose." [EC:3.2.1.108]
synonym:    "lactase-phlorizin hydrolase activity" BROAD [EC:3.2.1.108]
synonym:    "lactose galactohydrolase activity" EXACT [EC:3.2.1.108]
xref:       EC:3.2.1.108
xref:       MetaCyc:LACTASE-RXN
xref:       Reactome:20536
is_a:       GO:0004553 ! hydrolase activity, hydrolyzing O-glycosyl compounds
```

Example annotation

```
Gene product:     Actin, alpha cardiac muscle 1, UniProtKB:P68032
GO term:          heart contraction ; GO:0060047 (biological process)
Evidence code:    Inferred from Mutant Phenotype (IMP)
Reference:        PMID 17611253
Assigned by:      UniProtKB, June 6, 2008
```

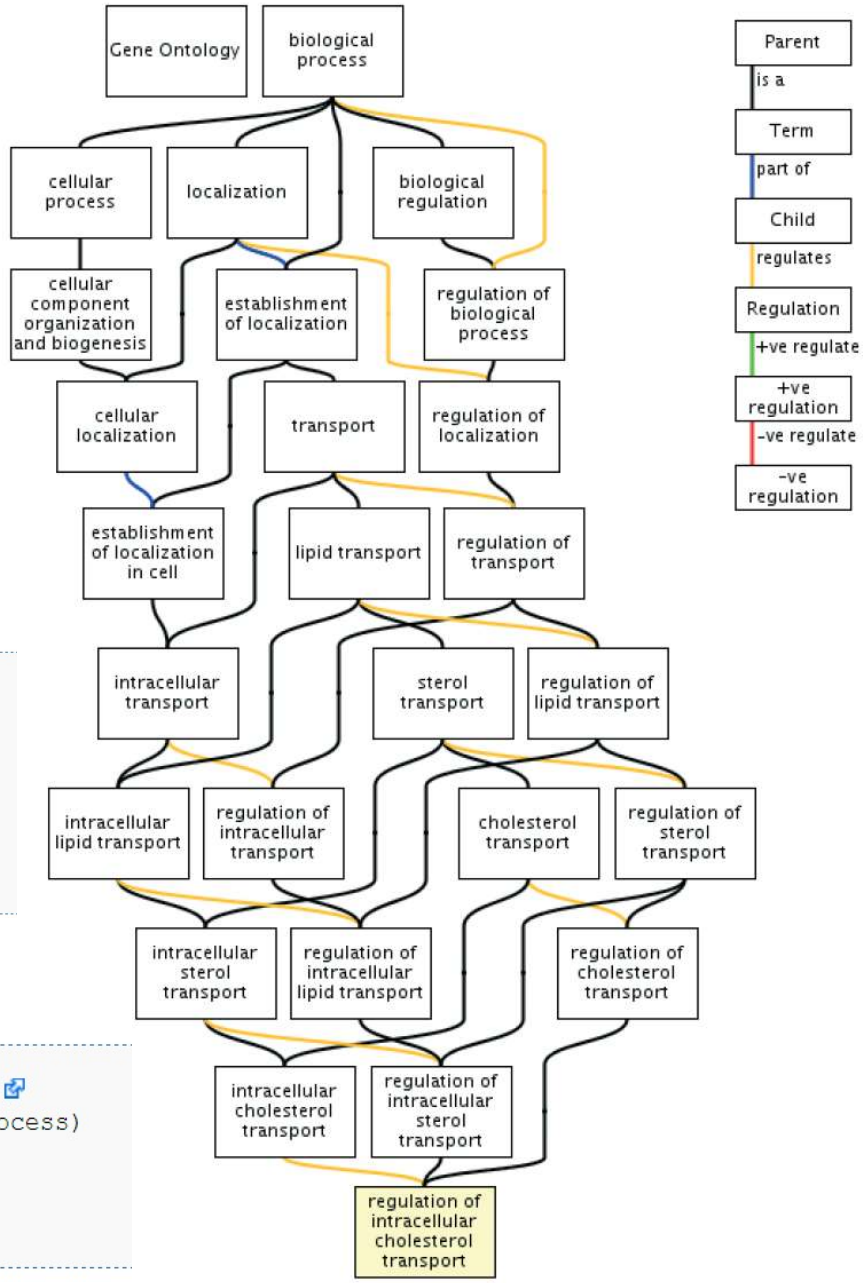
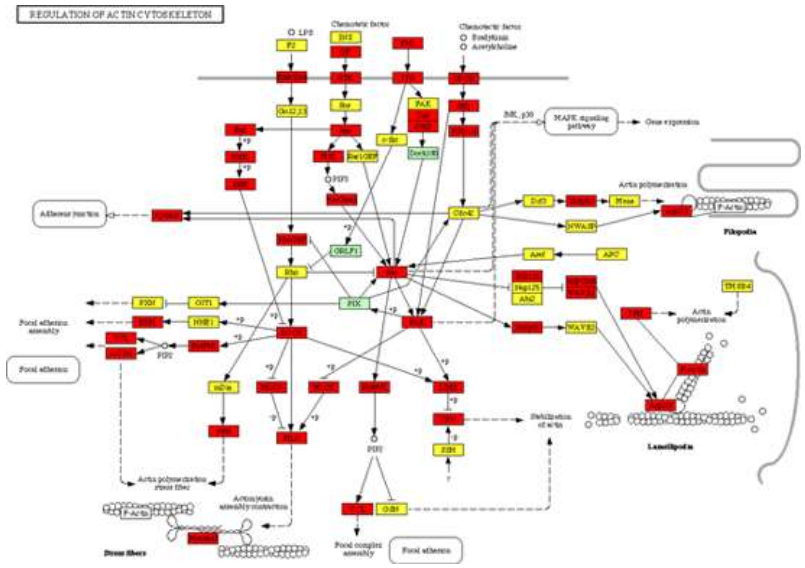


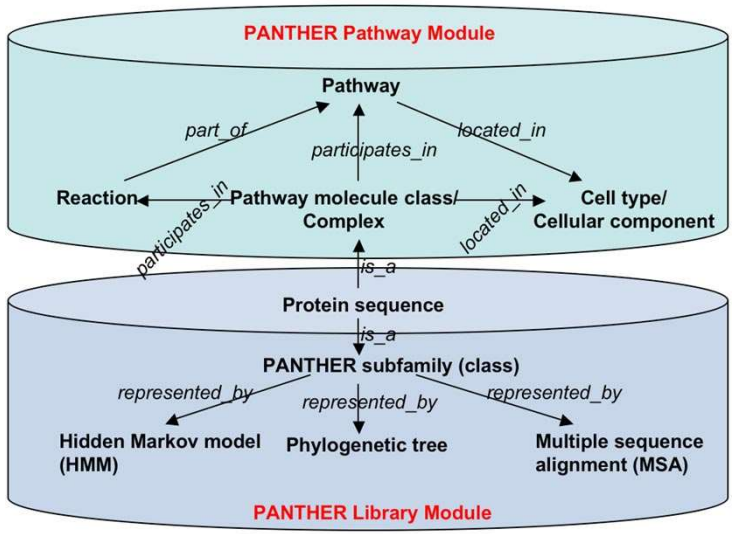
Figure 1. Section of the Gene Ontology for the Biological Process GO term 'regulation of intracellular cholesterol transport' (GO:0032383), showing ancestor terms and the different interconnecting relationships.

KEGG & PANTHER



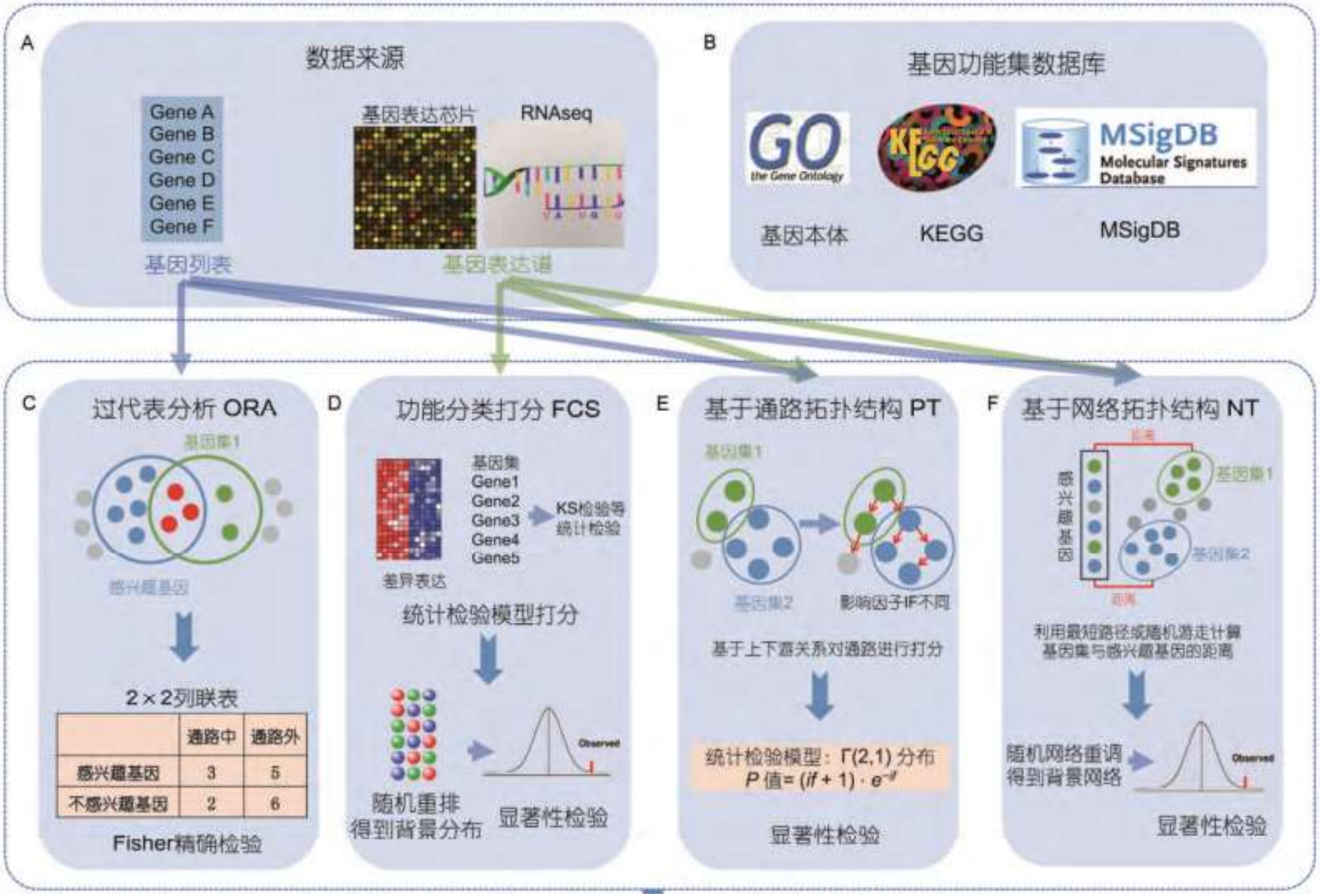
Kyoto Encyclopedia of Genes and Genomes

*www.genome.jp/**kegg**/*



Protein Analysis Through Evolutionary Relationships

*www.**pantherdb**.org/*


A
数据来源
Gene A
Gene B
Gene C
Gene D
Gene E
Gene F
基因列表
基因表达芯片
RNAseq
基因表达谱
B
基因功能集数据库
GO
the Gene Ontology
KEGG
MSigDB
Molecular Signatures Database
基因本体
KEGG
MSigDB
C
过代表分析 ORA
基因集1
感兴趣基因
2×2列联表
通路中
通路外
感兴趣基因
3
5
不感兴趣基因
2
6
Fisher精确检验
D
功能分类打分 FCS
基因集
Gene1
Gene2
Gene3
Gene4
Gene5
KS检验等统计检验
差异表达
统计检验模型打分
Observed
随机重排得到背景分布
显著性检验
E
基于通路拓扑结构 PT
基因集1
基因集2
影响因子IF不同
基于上下游关系对通路进行打分
统计检验模型：Γ(2,1) 分布
P 值= (if + 1) · e^-if
显著性检验
F
基于网络拓扑结构 NT
感兴趣基因
基因集1
基因集2
利用最短路径或随机游走计算基因集与感兴趣基因的距离
随机网络重调得到背景网络
Observed
显著性检验

DAVID Bioinformatics Resources 6.8

National Institute of Allergy and Infectious Diseases (NIAID), NIH

*** Welcome to DAVID 6.8 with updated Knowledgebase (more info). ***
*** If you are looking for DAVID 6.7, please visit our development site. ***

Shortcut to DAVID Tools

Functional Annotation

Gene-annotation enrichment analysis, functional annotation clustering , BioCarta & KEGG pathway mapping, gene-disease association, homologue match, ID translation, literature match and more

Gene Functional Classification

Provide a rapid means to reduce large lists of genes into functionally related groups of genes to help unravel the biological content captured by high throughput technologies. More

Gene ID Conversion

Convert list of gene ID/accessions to others of your choice with the most comprehensive gene ID mapping repository. The ambiguous accessions in the list can also be determined semi-automatically. More

Gene Name Batch Viewer

Display gene names for a given gene list; Search functionally related genes within your list or not in your list; Deep links to enriched detailed information. More

Recommending: A paper published in ***Nature Protocols*** describes step-by-step procedure to use DAVID!

Welcome to DAVID 6.8

2003 - 2017

The **D**atabase for **A**nnotation, **V**isualization and **I**ntegrated **D**iscovery (**DAVID**) v6.8 comprises a full Knowledgebase update to the sixth version of our original web-accessible programs. DAVID now provides a comprehensive set of functional annotation tools for investigators to understand biological meaning behind large list of genes. For any given gene list, DAVID tools are able to:

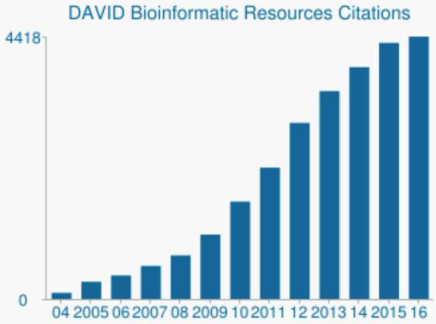
- Identify enriched biological themes, particularly GO terms
- Discover enriched functional-related gene groups
- Cluster redundant annotation terms
- Visualize genes on BioCarta & KEGG pathway maps
- Display related many-genes-to-many-terms on 2-D view.
- Search for other functionally related genes 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

Search

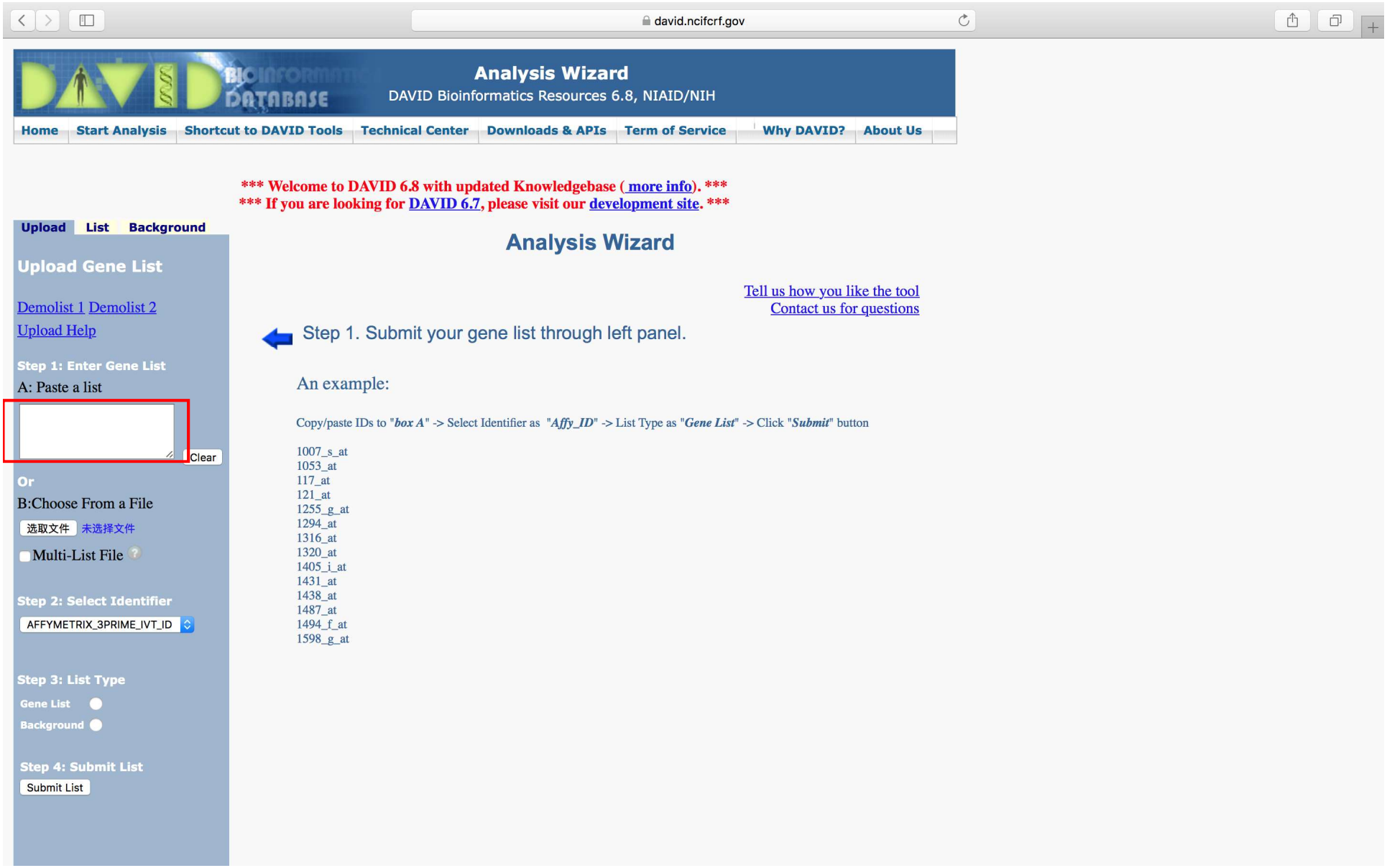
What's Important in DAVID?

- Cite DAVID
- IDs of Affy Exon and Gene arrays supported
- Novel Classification Algorithms
- Pre-built Affymetrix and Illumina backgrounds
- User's customized gene background
- Enhanced calculating speed

Statistics of DAVID



- > 26,000 Citations
- Average Daily Usage: ~2,600 gene lists/sublists from ~800 unique researchers.
- Average Annual Usage: ~1,000,000 gene lists/sublists from >5,000 research institutes


david.ncifcrf.gov
Analysis Wizard
DAVID Bioinformatics Resources 6.8, NIAID/NIH
Home
Start Analysis
Shortcut to DAVID Tools
Technical Center
Downloads & APIs
Term of Service
Why DAVID?
About Us
*** Welcome to DAVID 6.8 with updated Knowledgebase (more info). ***
*** If you are looking for DAVID 6.7, please visit our development site. ***
Upload
List
Background
Upload Gene List
Demolist 1 Demolist 2
Upload Help
Step 1: Enter Gene List
A: Paste a list
Clear
Or
B:Choose From a File
选取文件
未选择文件
Multi-List File
Step 2: Select Identifier
AFFYMETRIX_3PRIME_IVT_ID
Step 3: List Type
Gene List
Background
Step 4: Submit List
Submit List
Analysis Wizard
Tell us how you like the tool
Contact us for questions
Step 1. Submit your gene list through left panel.
An example:
Copy/paste IDs to "box A" -> Select Identifier as "Affy_ID" -> List Type as "Gene List" -> Click "Submit" button
1007_s_at
1053_at
117_at
121_at
1255_g_at
1294_at
1316_at
1320_at
1405_i_at
1431_at
1438_at
1487_at
1494_f_at
1598_g_at

GSVA

```
library(GSVA)
tmp1<-gsva(expr, gset.idx.list, method="gsva",
rnaseq=F)$es.obs
```

表达值count矩阵

基因集（gene，GO）

rnaseq=F 用原始count数

http://www.bioconductor.org/packages/release/bioc/html/GSVA.html

类型	方法	可用性	使用或下载网址
ORA	DAVID[10]	在线工具	https://david.ncifcrf.gov
	GOstat[11]	在线工具	http://gostat.wehi.edu.au
	GenMAPP[12]	在线工具	http://www.genmapp.org
	GoMiner[13]	在线工具	http://discover.nci.nih.gov/gominer
	Onto-Express[14]	在线工具	http://vortex.cs.wayne.edu
FCS	GSEA[8,15]	Java 软件, R 语言包	http://software.broadinstitute.org/gsea
	GSA[16]	R 语言包	https://cran.r-project.org/web/packages/GSA/index.html
	PADOG[17]	R 语言包	www.bioconductor.org/packages/release/bioc/html/PADOG.html
	SAFE[18]	R 语言包	http://www.bios.unc.edu/~fwright/SAFE
	Globaltest[19,20]	R 语言包	http://www.bioconductor.org/packages/2.0/bioc/html/globaltest.html
	Sigpathway[21]	R 语言包	http://bioconductor.org/packages/release/bioc/html/sigPathway.html
	GAGE[22]	R 语言包	www.bioconductor.org/packages/release/bioc/html/gage.html
	GSVA[23]	R 语言包	www.bioconductor.org/packages/release/bioc/html/GSVA.html
	PLAGE[24]	R 语言包	http://dulci.biostat.duke.edu/pathways/misc.html
	ZSCORE[25]	R 语言包	www.bioconductor.org/packages/release/bioc/html/limma.html
	SSGSEA[26]	R 语言包	http://www.broadinstitute.org/cancer/software/genepattern/modules/docs/ssGSEAProjection/
	MRGSE[27]	R 语言包	www.bioconductor.org/packages/release/bioc/html/limma.html
	ANCOVA[28]	R 语言包	https://cran.r-project.org/web/packages/fANCOVA/index.html
	CAMERA[29]	R 语言包	www.bioconductor.org/packages/release/bioc/html/limma.html
PT	MetaCore	商业软件	http://www.genego.com/metacore.php
	Pathway-Express[30]	在线工具, R 语言包	vortex.cs.wayne.edu/projects.htm
	SPIA[31]	R 语言包	www.bioconductor.org/packages/release/bioc/html/SPIA.html
	TopoGSA[32]	在线工具	www.topogsa.org
	CePa[33]	R 语言包	https://cran.rstudio.com/web/packages/CePa/index.html
	ToPASeq[34]	R 语言包	www.bioconductor.org/packages/release/bioc/html/ToPASeq.html
	NetGSA[35]	R 语言包	https://cran.r-project.org/web/packages/netgsa/index.html
	DEGraph[36]	R 语言包	www.bioconductor.org/packages/release/bioc/html/DEGraph.html
	BPA[37]	软件包	http://bumil.boun.edu.tr/bpa
	ACST[38]	R 语言包	http://omictools.com/analysis-of-consistent-signal-transduction-tool
NT	NEA[39]	R 语言包	https://r-forge.r-project.org/projects/nea2
	EnrichNet[40]	在线工具, R 语言包	www.enrichnet.org
	GANPA[41]	R 语言包	https://cran.r-project.org/web/packages/GANPA/index.html
	LEGO[42]	在线工具, R 语言包	lego.tianlab.cn
	NOA[43]	在线工具	app.aporc.org/NOA
	GOGANPA[44]	R 语言包	https://cran.r-project.org/web/packages/GOGANPA/index.html

Thank you!

scRNA-seq

Bulk RNA-seq 优缺点

- Useful for comparative transcriptomics

(物种间比较)

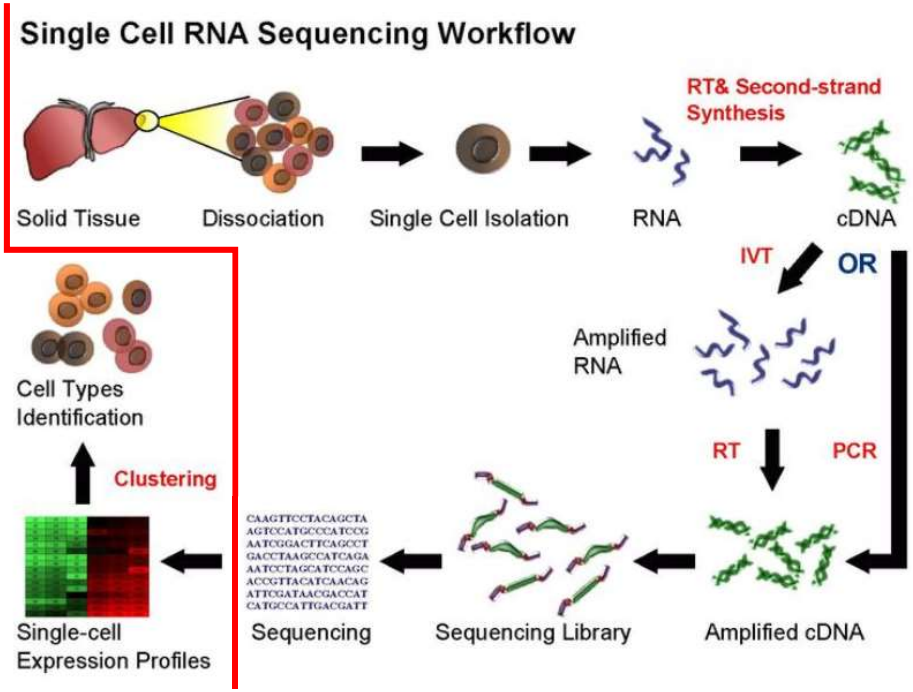
- Useful for quantifying expression signatures from ensembles(疾病研究)
- **Insufficient** for heterogeneous systems
  - early development studies
  - complex tissues
- **Insufficient** for stochastic nature of gene expression

What is scRNA-seq

- A new technology, first publication by (Tang et al. 2009)
- gain widespread popularity ~2014
- Measures the **distribution of expression levels** (Bulk RNA-seq?)
- **cell-specific changes studies**
  - cell type identification(细胞分类)
  - heterogeneity of cell responses(细胞反应的异质性)
  - stochasticity of gene(基因随机性)

How to do scRNA-seq

Experimental design



Bioinformatics analyze

Workflow (Wikipedia)

Methods and Protocols in scRNA-seq

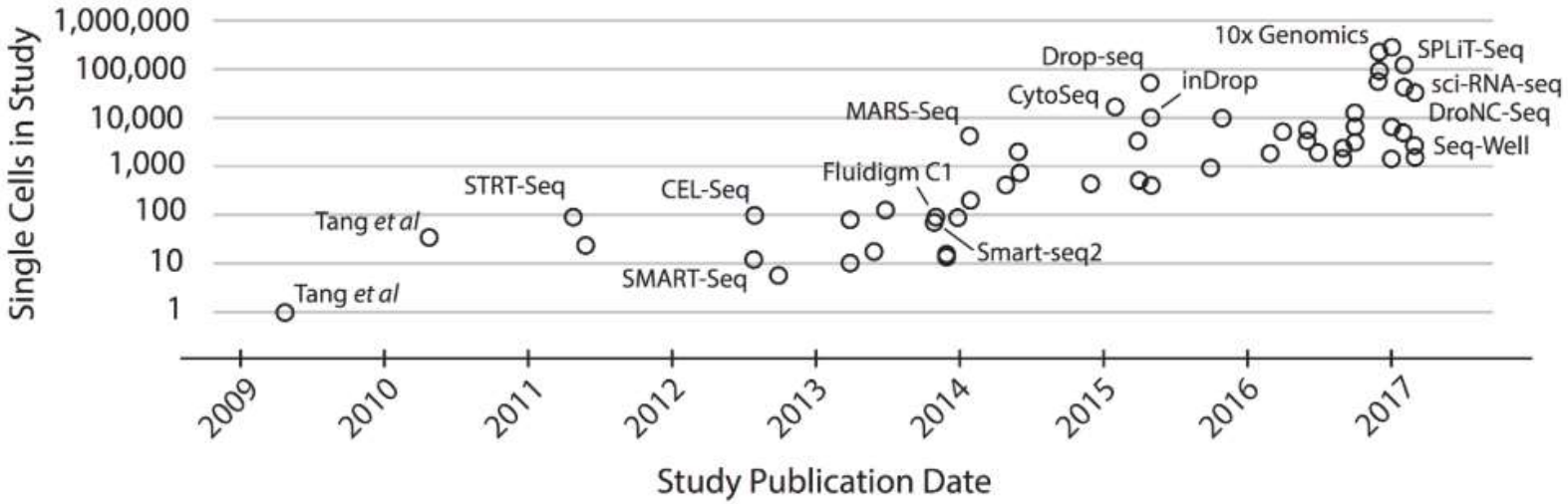


image taken from [Svensson et al]

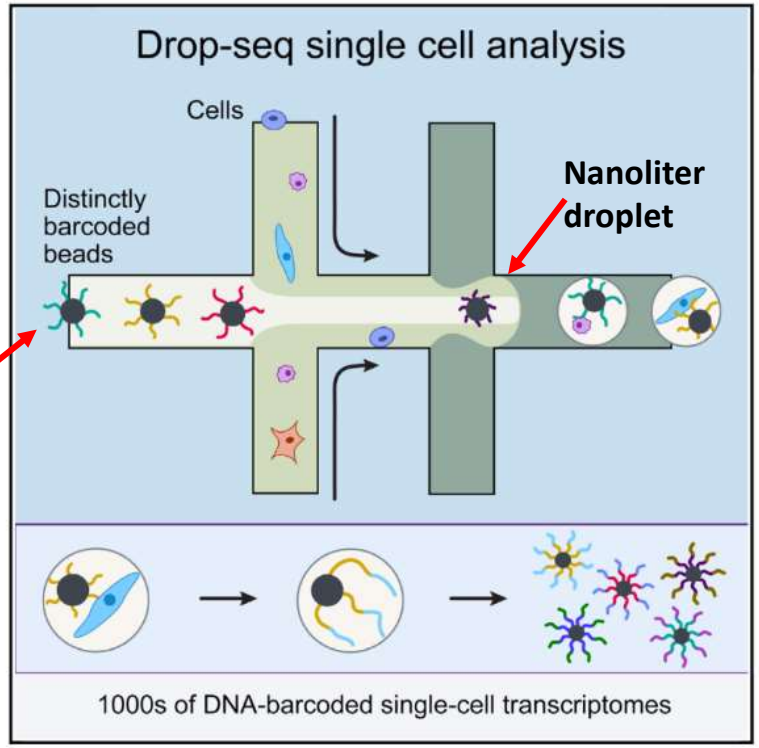
scRNA-seq example

	Smart-seq	CEL-seq	SCRB-seq	DROP-seq
Unique Molecular identifier(UMI)	No	Yes	Yes	Yes
Transcript quantification	Full-length	Tag-based(3')	Tag-based(3')	Tag-based(3')
Sensitivity(Recall)	Max			
Efficiency			Max	Max

Choice depend on biological question!

Full-length: studying different isoforms
Tag-based: combine with UMI to facilitate gene-level quantification

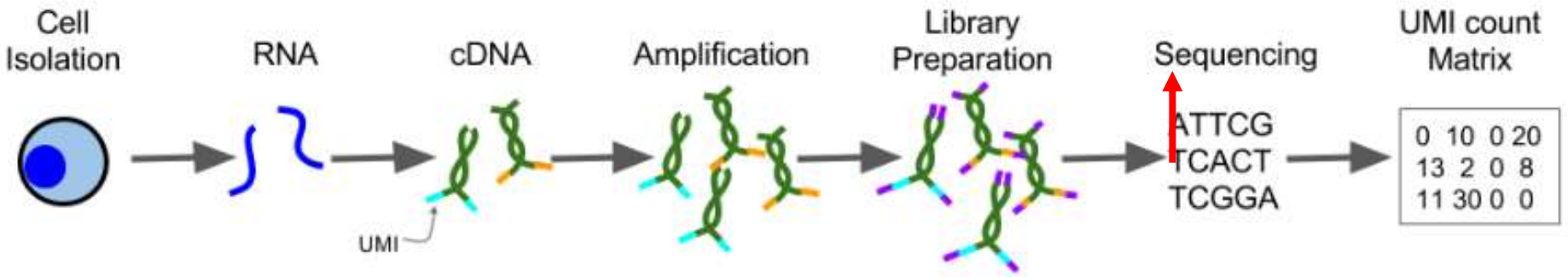
DROP-seq



- droplet based methods
- sequenced reads can be assigned
- highest throughput
- the coverage is low

(only a few thousand different transcripts)

From Sequencing to Expression Matrix



FastQC

- check the quality of the reads you have sequenced
- can be used for both bulk and single-cell RNA-seq data
- https://www.bioinformatics.babraham.ac.uk/projects/fastqc/

- $ fastqc –o results 1.fastq 2.fastq #paired-end
- And check the result html file

Trimming Reads

- trim sequencing adapters
- Trim low quality reads from the ends of reads.

- http://www.bioinformatics.babraham.ac.uk/projects/trim_galore/
- FastQC report→”Adapter Content”→**Trim Galore!** →check again

Demultiplexing_1

- identifying and removing the cell-barcode sequence from one or both reads.
- compare each cell-barcode with expected barcodes→remove
- For UMI containing data: we should keep the UMI code info

Demultiplexing_2

Identifying cell-containing droplets/microwells

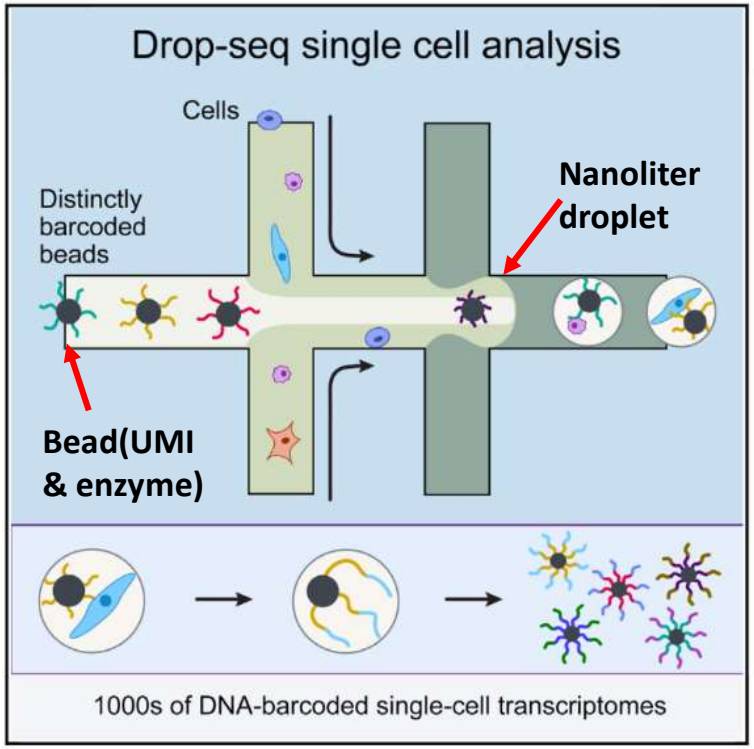
∵ Some RNA will leak out of dead/damaged cells

∴ 1.Some droplets include intact cell

2.Other capture a small amount of the ambient RNA

(will contaminate the library and final output)

How to distinguish this situation?

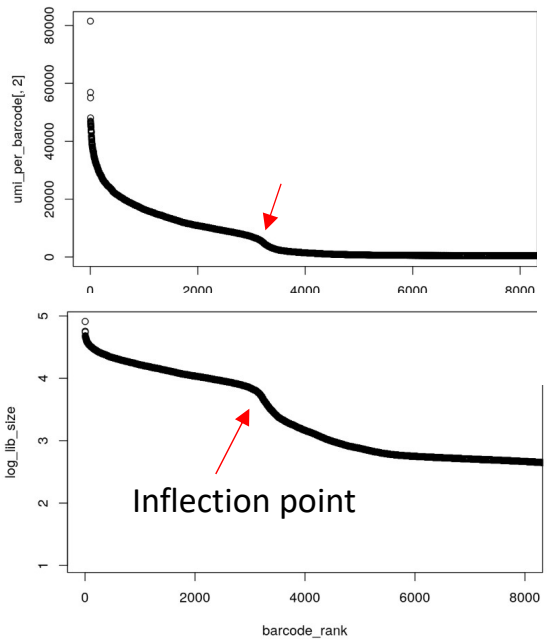


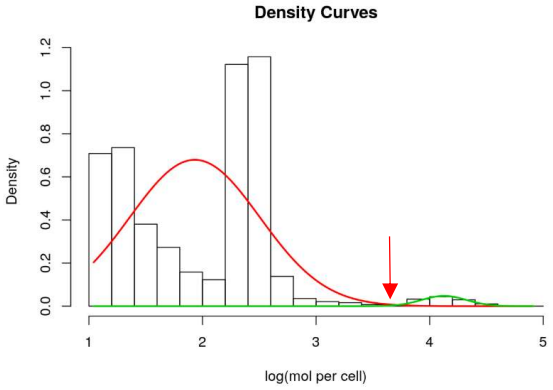
Demultiplexing_2

Identifying cell-containing droplets/microwells

- We use **total molecules per barcode** to measure

  - Inflection point
  - Mix model
  - Expected num of cells

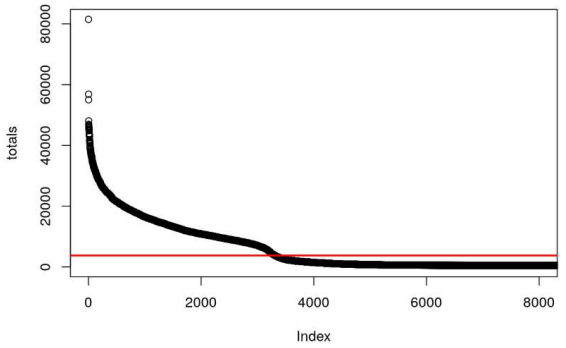




GMM(混合高斯模型)

$$p(x) = \sum_{i=1}^{K} \phi_i \frac{1}{\sqrt{2\sigma_i^2 \pi}} e^{-\frac{(x-\mu_i)^2}{2\sigma_i^2}}$$

EM Algorithm




```
thresh = totals[round(0.01*n_cells)]/10
```

Align Reads

- **STAR**

2 steps:

(i) create a genome index

$ STAR --runThreadN 4 --runMode genomeGenerate --genomeDir indices/STAR --genomeFastaFiles **reference.transcripts.fa(save resource)**

(ii) Align

$ STAR --runThreadN 4 --genomeDir indices/STAR --readFilesIn 1.fastq 2.fastq --outFileNamePrefix results/STAR/

Align Reads

- **Kallisto**

pseudo-aligner(For large full-transcript datasets from well annotated organisms)

1. K-mer strategy
   - pseudo-alignment much faster than traditional aligners
   - cope better with sequencing errors than traditional aligners.
2. Specially designed mode for scRNA-seq
   - psuedo-aligns to a reference transcriptome
   - deal with splice isoforms
   - aligns to equivalence classes

$ kallisto index -i indices/Kallisto/transcripts.idx reference.transcripts.fa

$ kallisto pseudo -i indices/Kallisto/transcripts.idx -o results/Kallisto -b batch.txt

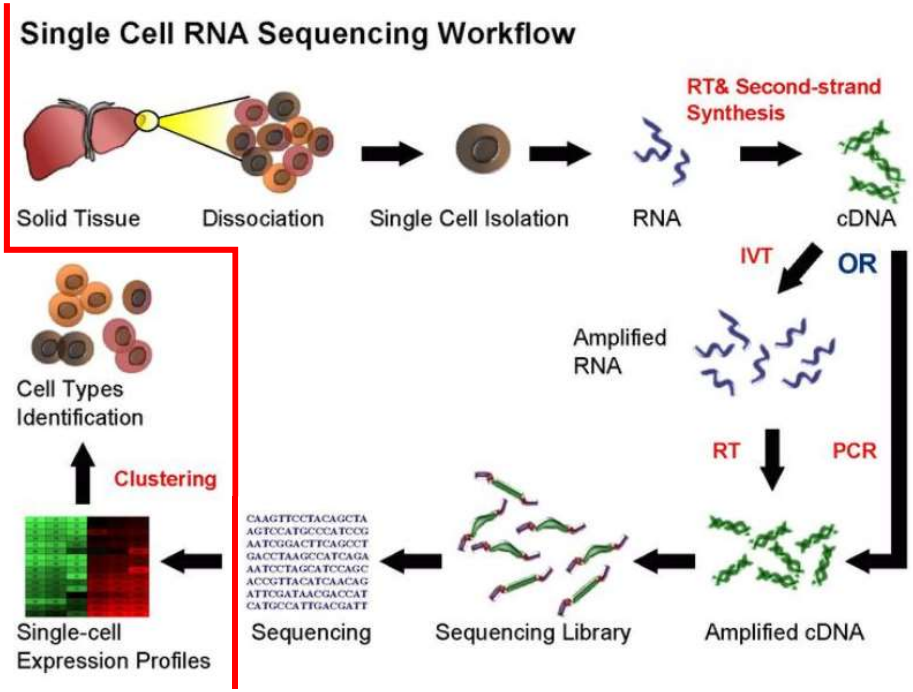
Construction of Expression Matrix

- UMI count→Nums of transcript for a gene in a cell
- Get MATRIX:

	Cell1	Cell2	Cell3	Cell4	…
Gene1	UMI count				
Gene2					
Gene3					
Gene4					
…					

How to do scRNA-seq

Bioinformatics analyze



Experimental design

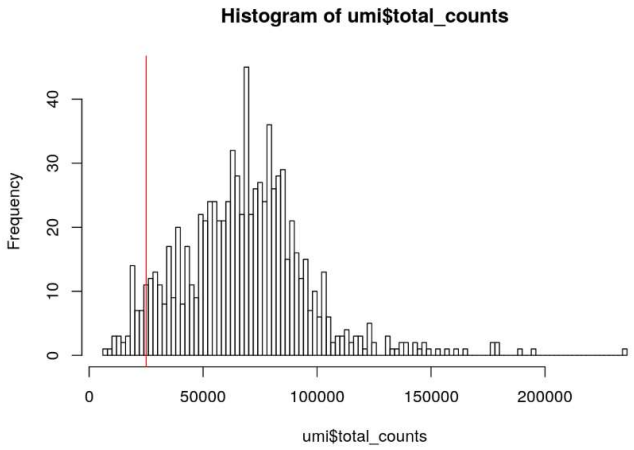
Workflow (Wikipedia)

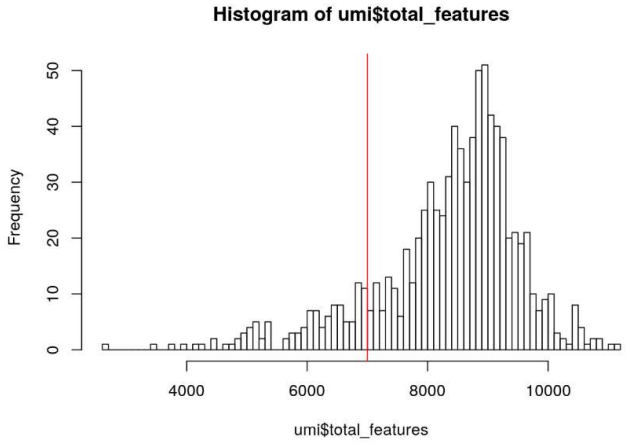
Bioinfomatics analyze

- Expression QC
- Normalization
- Choose variable genes
- Dimensional reduction
- Clustering
- tSNE
- Find differentially expressed genes

Expression QC

- Remove genes that are not expressed in any cell(UMI count > 0)
- Distribution of the total number of RNA molecules detected per sample
- Distribution of the total number of unique genes detected in each sample





Normalization

(Some quantification method incorporates this info automatically)

- Due to different library size

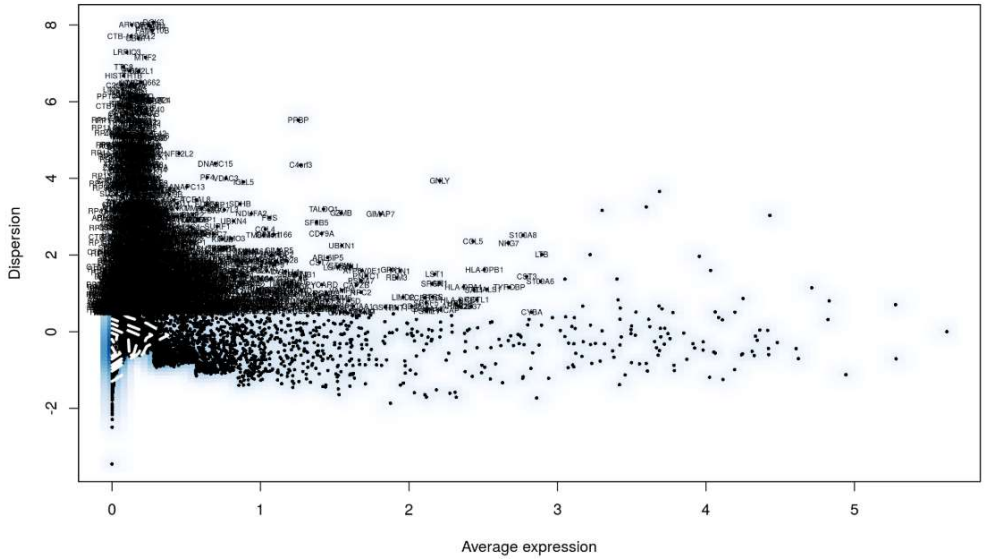
- “LogNormalize” method
- (i) normalizes by the total expression
- (ii) multiply a scale factor
- (iii) log-transform

Choose variable genes

Average expression

Variance

	Cell1	Cell2	Cell3	Cell4	…
Gene1	UMI count				
Gene2					
Gene3					
Gene4					
…					



Dimensional reduction

- linear dimensional reduction→PCA

- Reason:
  - Determine statistically significant principal components
  - Helpful for clustering

- What and How to do PCA ?

Dimensional reduction(PCA)

- Principal Component Analysis(主成分分析)
- 高维空间→低维空间
- Optimization objective:
  - Max→ $Var(a) = \frac{1}{m}\sum_{i=1}^{m} a_i^2$
  - Min → $Cov(a,b) = \frac{1}{m}\sum_{i=1}^{m} a_i b_i$

即使降维后主成分的方差最大而成分之间的协方差最小
→ 协方差矩阵

Dimensional reduction(PCA)

1. 将原始数据整合成(m,n)矩阵，m代表样本数,n代表维数
2. 将每一行作零均值化
3. 求协方差矩阵 $C=\frac{1}{m}XX^{\mathsf{T}}$
4. 求该矩阵的特征值和特征向量
5. 根据特征值大小排序，取对应的k个特征向量组成矩阵P
6. Y=PX即为降维后的数据

$$X=\begin{pmatrix} a_1 & a_2 & \cdots & a_m \\ b_1 & b_2 & \cdots & b_m \end{pmatrix}$$

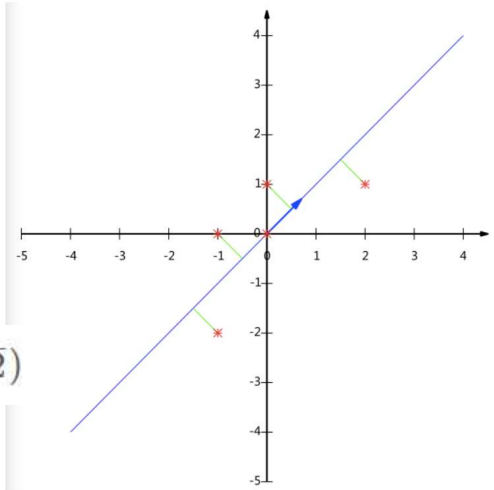
$$\frac{1}{m}XX^{\mathsf{T}}=\begin{pmatrix} \frac{1}{m}\sum_{i=1}^{m}a_i^2 & \frac{1}{m}\sum_{i=1}^{m}a_ib_i \\ \frac{1}{m}\sum_{i=1}^{m}a_ib_i & \frac{1}{m}\sum_{i=1}^{m}b_i^2 \end{pmatrix}$$

设Y=PX

$$\begin{aligned} D &= \frac{1}{m}YY^{\mathsf{T}} \\ &= \frac{1}{m}(PX)(PX)^{\mathsf{T}} \\ &= \frac{1}{m}PXX^{\mathsf{T}}P^{\mathsf{T}} \\ &= P(\frac{1}{m}XX^{\mathsf{T}})P^{\mathsf{T}} \\ &= PCP^{\mathsf{T}} \end{aligned}$$

$$\begin{pmatrix} -1 & -1 & 0 & 2 & 0 \\ -2 & 0 & 0 & 1 & 1 \end{pmatrix}$$

$$\begin{pmatrix} -3/\sqrt{2} & -1/\sqrt{2} & 0 & 3/\sqrt{2} & -1/\sqrt{2} \end{pmatrix}$$



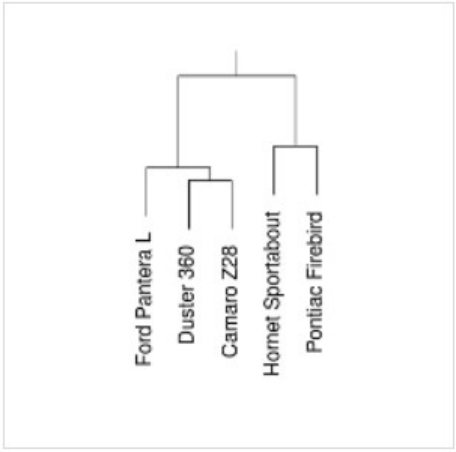
Clustering

- **Hierarchical clustering**
- **k-means**
- **Graph-based methods**

- Find a consistent pattern to cluster cells

Hierarchical clustering

```
# find distance matrix
> d <- dist(as.matrix(mtcars))
# apply hirarchical clustering
> hc <-
 hclust(d)
# plot the dendrogram
> plot(hc)
```



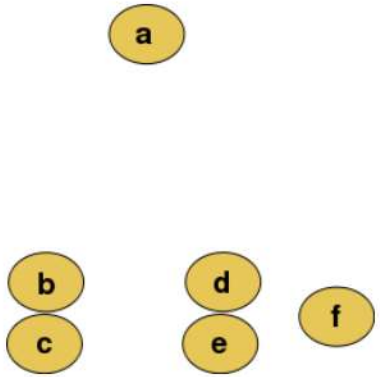


Figure 8.1: Raw data

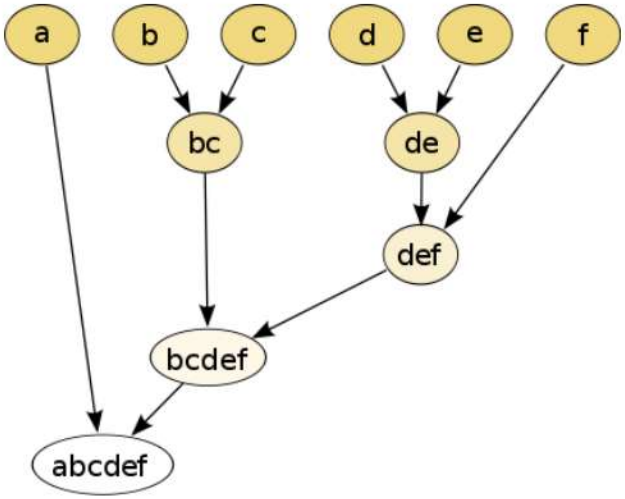


Figure 8.2: The hierarchical clustering dendrogram

k-means

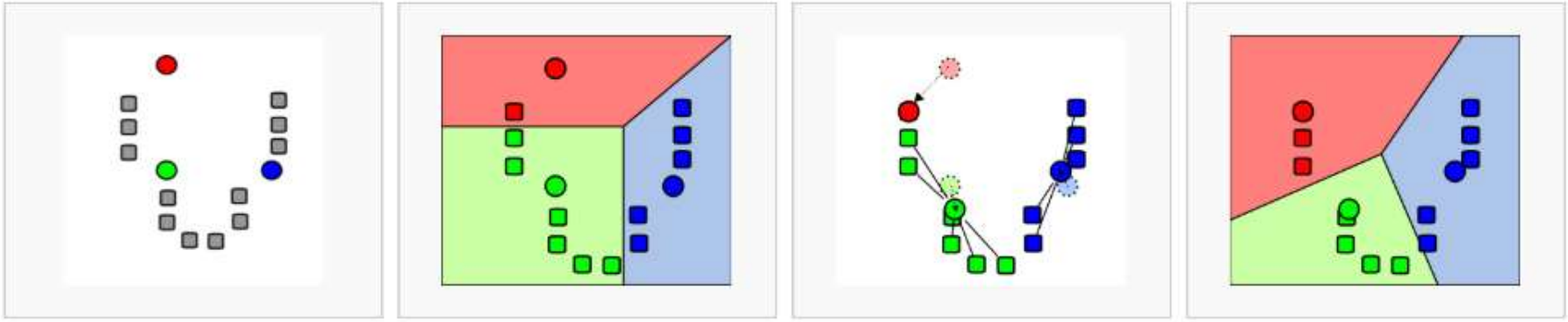


Figure 8.3: Schematic representation of the k-means clustering

kmeans(data,center)

Graph-based methods

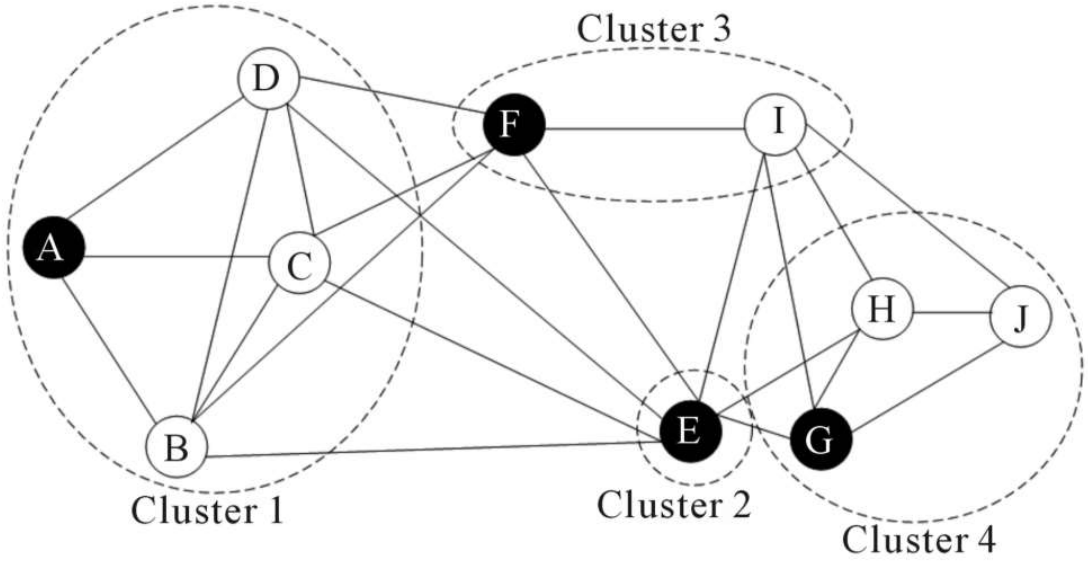


Figure 8.4: Schematic representation of the graph network

Modularity:

$$M = \frac{1}{2L}\sum_{i,j=1}^{N}(A_{ij} - \frac{k_i k_j}{2L})\delta(c_i, c_j)$$ max

L:num of edges
ki:degree of node I
Aij:(i,j) value in Adjacency matrix
δ for Kronecker-delta function

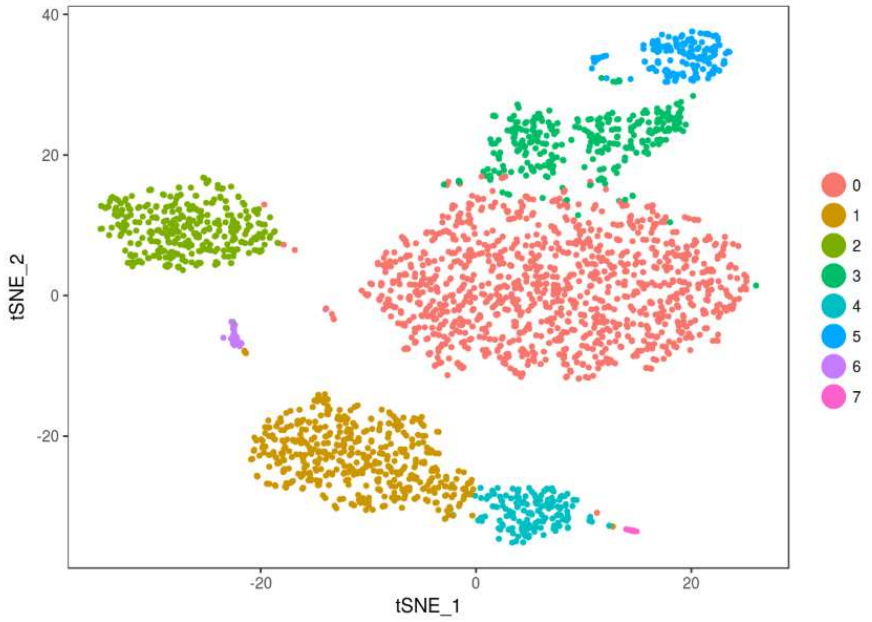
Fast-Greedy Modularity-Maximization

tSNE for visualization

- t-distributed stochastic neighbor embedding
  - t分布随机邻节点嵌入
- (i) machine learning to classify
- (ii) Euclidean distance→Prob
- (ii) tSNE

(Non-linear dimensional reduction)

多维→二维以可视化

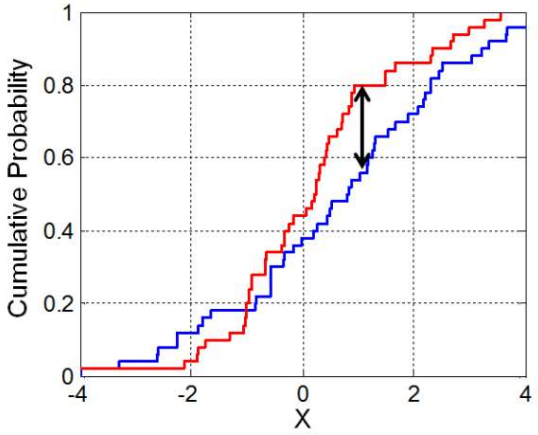


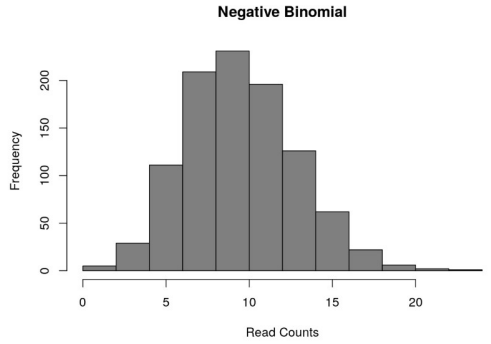
Find differentially expressed genes

Due to a large number of samples (i.e. cells)

Bulk RNA-seq : comparing estimates of mean-expression
scRNA-seq : identify differences between distribution of groups

(i) From a existing statistical model(NB model)
(ii)Non-parametric test (KS-test)

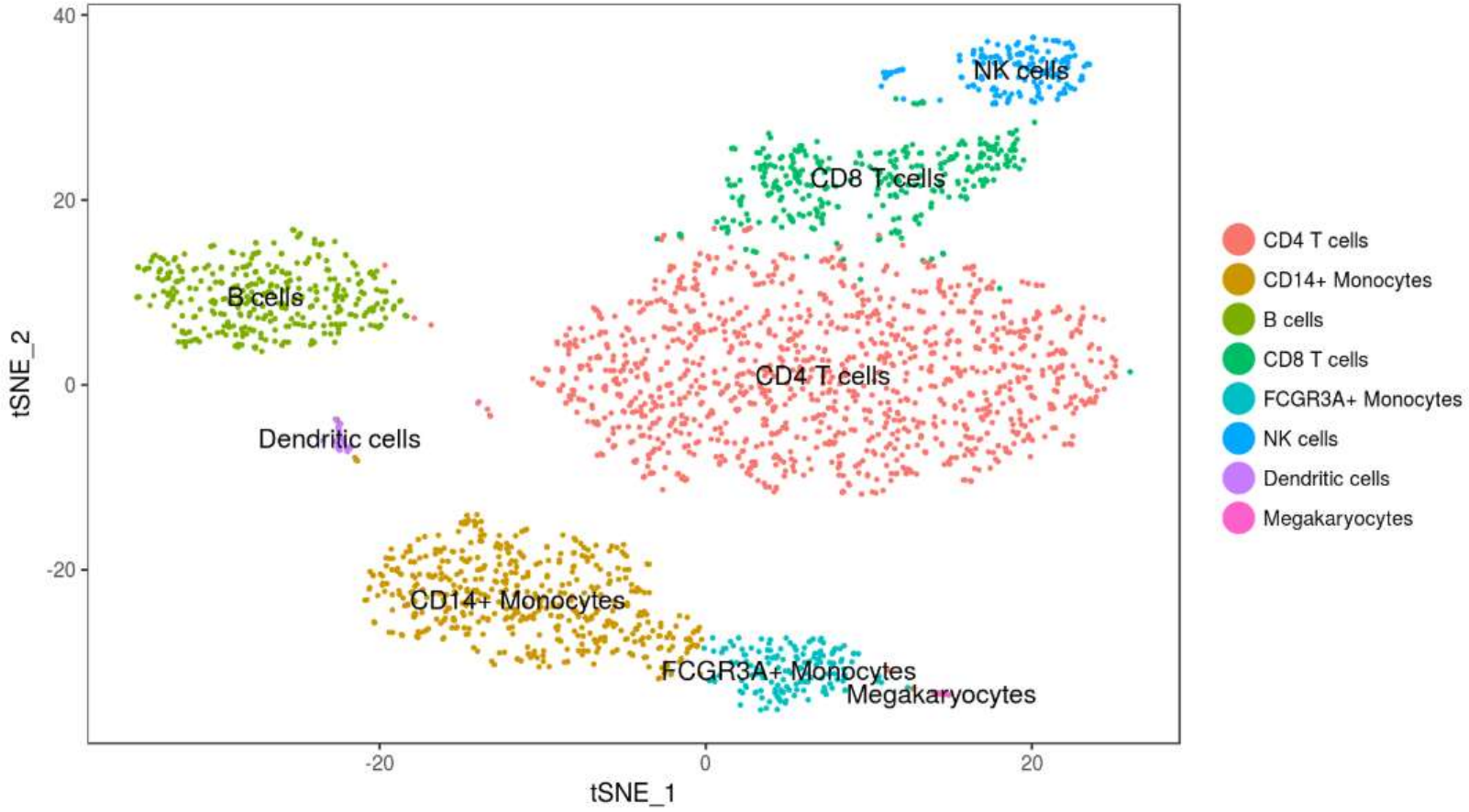




Find differentially expressed genes

- Some methods or tools in R for DE(differential expression)
- $ ks.test
  - (popular non-parametric, cant deal with zero-problem)
- $ wilcox.test
  - Wilcox-rank-sum test(another non-parametric based on difference in median)
- edgeR
  - based on a negative binomial model
- Monocle
  - use several different models for DE based on type of expression data

Assigning cell type identity to clusters

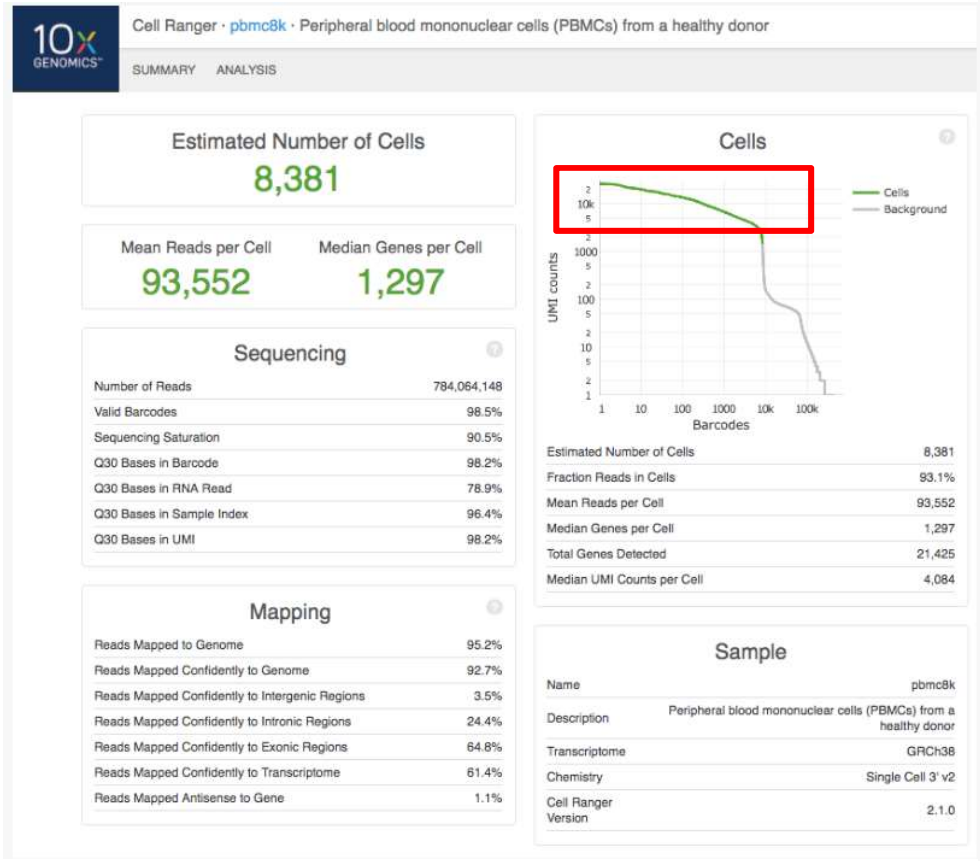


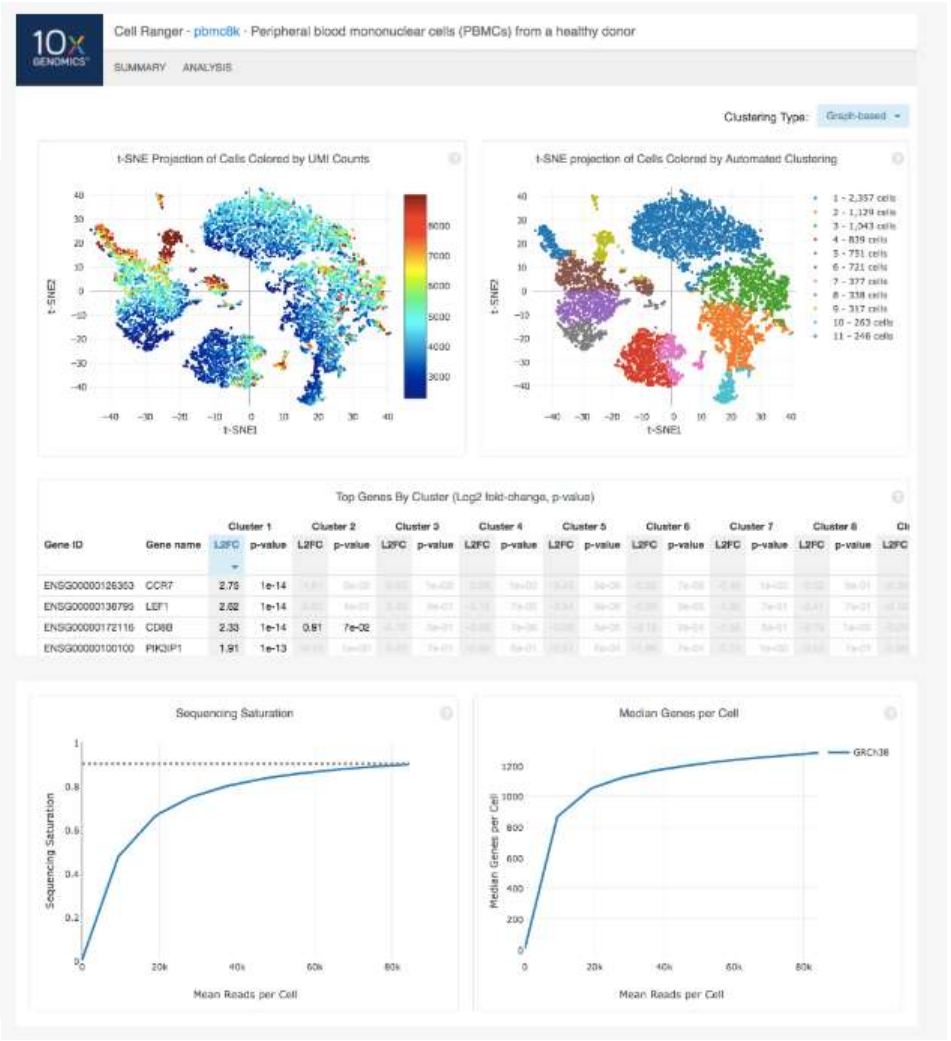
Cellranger

- An integration tool for scRNA-seq analysis

```
$ tar -xzvf cellranger-2.2.0.tar.gz
$ tar -xzvf refdata-cellranger-GRCh38-1.2.0.tar.gz
$ cellranger testrun --id=tiny
$ cellranger count --id=sample345 \
                   --transcriptome=/opt/refdata-cellranger-GRCh38-1.2.0 \
                   --fastqs=/home/jdoe/runs/HAWT7ADXX/outs/fastq_path \
                   --sample=mysample \
                   --expect-cells=1000
```

summary HTML file





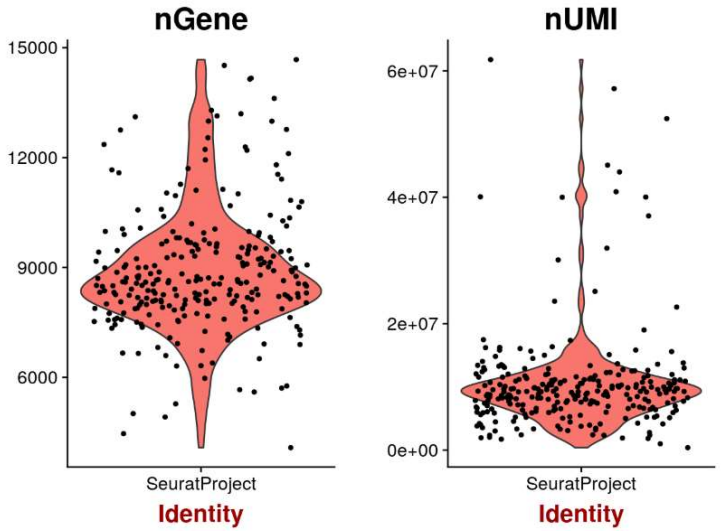
Seurat SEURAT

Seurat is a popular **R** package that can perform QC, analysis, and exploration of scRNA-seq data(**Deng** dataset)

```
#library import
>library(SingleCellExperiment)
>library(Seurat)
>library(mclust)
>library(dplyr)
#read data
>seuset <- CreateSeuratObject(
   raw.data = counts(deng),
   min.cells = 3,
   min.genes = 200
)
```

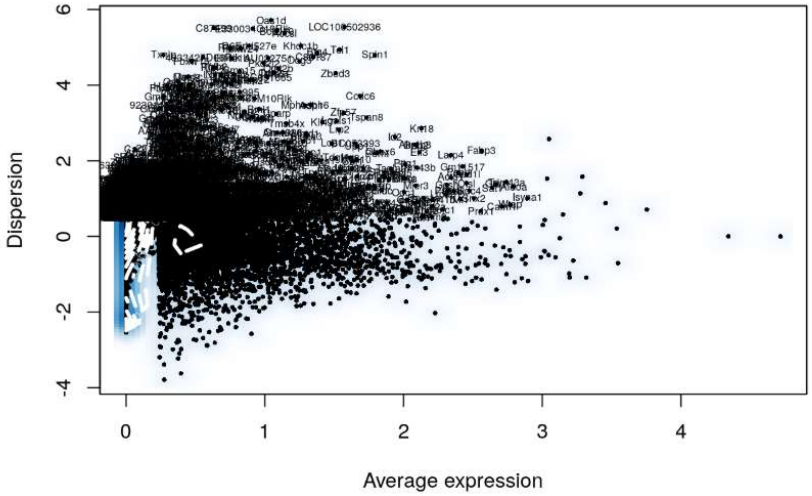
SEURAT

```
#explore QC metrics
>VlnPlot(
    object = seuset,
    features.plot = c("nGene", "nI
    nCol = 2
)
#filter cells
>seuset <- FilterCells(
    object = seuset,
    subset.names = c("nUMI"),
    high.thresholds = c(2e7)
)
```



SEURAT

```
#Normalize the data
>seuset <- NormalizeData(
    object = seuset,
    normalization.method = "LogN
    scale.factor = 10000
)
#Find highly variable genes
>seuset <- FindVariableGenes(
    object = seuset,
    mean.function = ExpMean,
    dispersion.function = LogVMR,
    x.low.cutoff = 0.0125,
    x.high.cutoff = 3,
    y.cutoff = 0.5
)
```



SEURAT

```
#PCA
>seuset <- RunPCA(
   object = seuset,
   pc.genes = seuset@var.genes,
   do.print = TRUE,
   pcs.print = 1:5,
   genes.print = 5
)
#Visualize PCA
>PrintPCA(object = seuset, pcs.print = 1:5, genes.print = 5,
use.full = FALSE)
>PCAPlot(object = seuset, dim.1 = 1, dim.2 = 2)
...
```

SEURAT

```
#JackStraw find significant PCs
>seuset <- JackStraw(
   object = seuset,
   num.replicate = 100,
   do.print = FALSE
)
#Visualize
>JackStrawPlot(object = seuset, PCs = 1:9)
>PCElbowPlot(object = seuset)
```

SEURAT

```
#clustering the cell
>seuset <- FindClusters(
   object = seuset,
   reduction.type = "pca",
   dims.use = 1:8,
   resolution = 1.0,
   print.output = 0,
   save.SNN = TRUE
)
#visualize the cluster result
>seuset <- RunTSNE(
   object = seuset,
   dims.use = 1:8,
   do.fast = TRUE
)
>TSNEPlot(object = seuset)
```

SEURAT

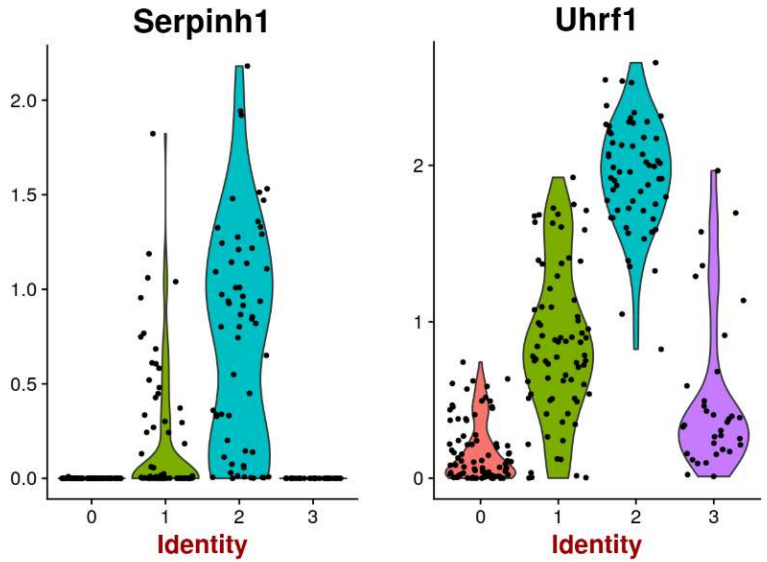
#Find marker gene for cluster 2

>markers2 <- FindMarkers(seuset, 2)

#then visualize

>VlnPlot(object = seuset, features.

rownames(markers2)[1:2])



#Find markers for all clusters

>markers <- FindAllMarkers(

object = seuset,

only.pos = TRUE,

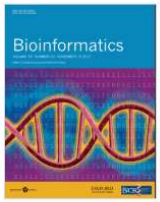
min.pct = 0.25,

thresh.use = 0.25

)

Introduction——SPRINT

https://academic.oup.com/bioinformatics/article/33/22/3538/4004872



Volume 33, Issue 22
15 November 2017

Article Contents

< Previous Next >

SPRINT: an SNP-free toolkit for identifying RNA editing sites

Feng Zhang, Yulan Lu, Sijia Yan, Qinghe Xing, Weidong Tian

Bioinformatics, Volume 33, Issue 22, 15 November 2017, Pages 3538–3548,
https://doi.org/10.1093/bioinformatics/btx473
Published: 24 July 2017 **Article history**

Abstract

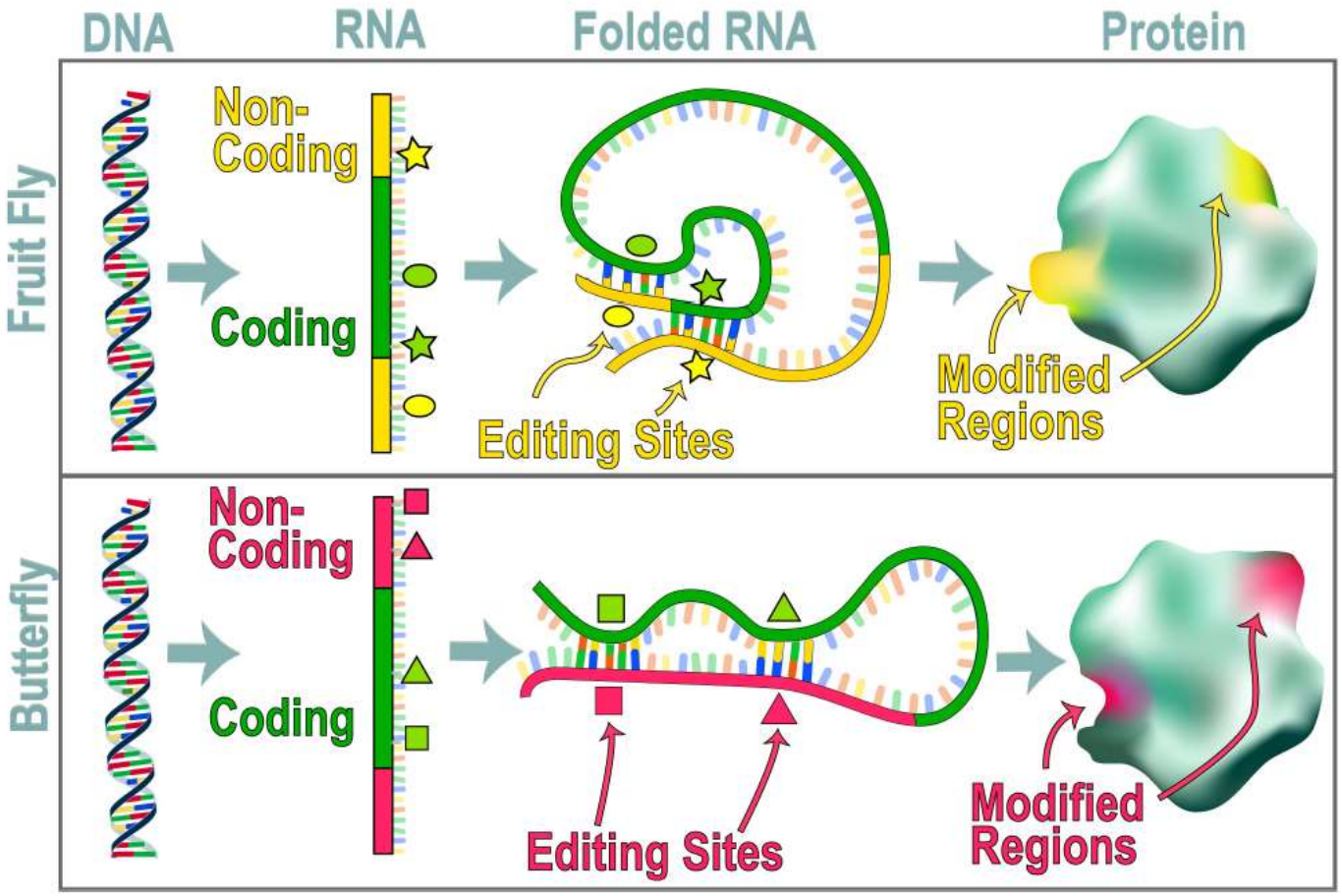
Motivation

RNA editing generates post-transcriptional sequence alterations. Detection of RNA editing sites (RESs) typically requires the filtering of SNVs called from RNA-seq data using an SNP database, an obstacle that is difficult to overcome for most organisms.

Results

Here, we present a novel method named SPRINT that identifies RESs without the need to filter out SNPs. SPRINT also integrates the detection of hyper RESs from remapped reads, and has been fully automated to any RNA-seq data with reference genome sequence available. We have rigorously validated SPRINT's effectiveness in detecting RESs using RNA-seq data of samples in which genes encoding RNA editing enzymes are knock down or over-expressed, and have also demonstrated its superiority over current methods. We have applied SPRINT to investigate RNA editing across tissues and species, and also in the development of mouse embryonic central nervous system. A web resource (http://sprint.tianlab.cn) of RESs identified by SPRINT has been

RNA-Editing

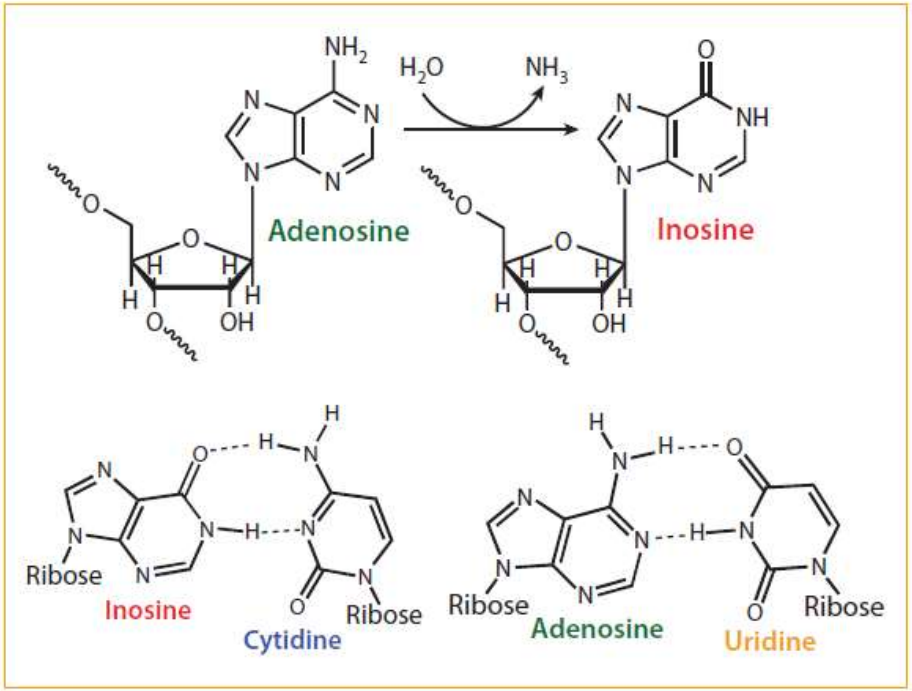


https://www.nsf.gov/news/mmg/media/images/1_rna_editing_h1.jpg

RNA Editing : RNA editing generates post-transcriptional sequence alterations, primarily the modification of RNA nucleotides[1]。

Primary types : adenosine-to-inosine (A-to-I, detected as A-to-G)[2]、 cytosine- to-uracil (C-to-U, detected as C-to-T)[3]

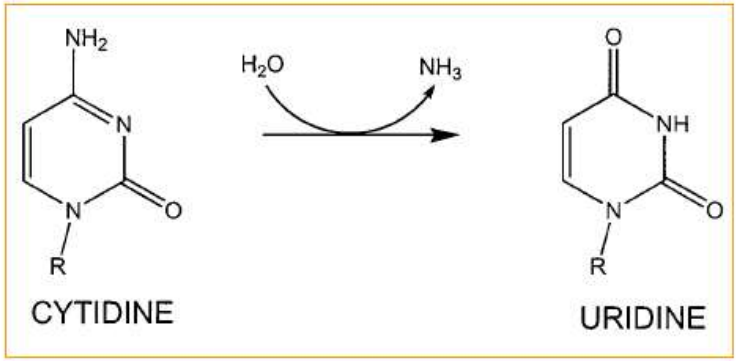
A-to-I editing



ADARs (Adenosine Deaminases Acting on RNA)

Nishikura K, Annu Rev Biochem. 2010;79:321-49.

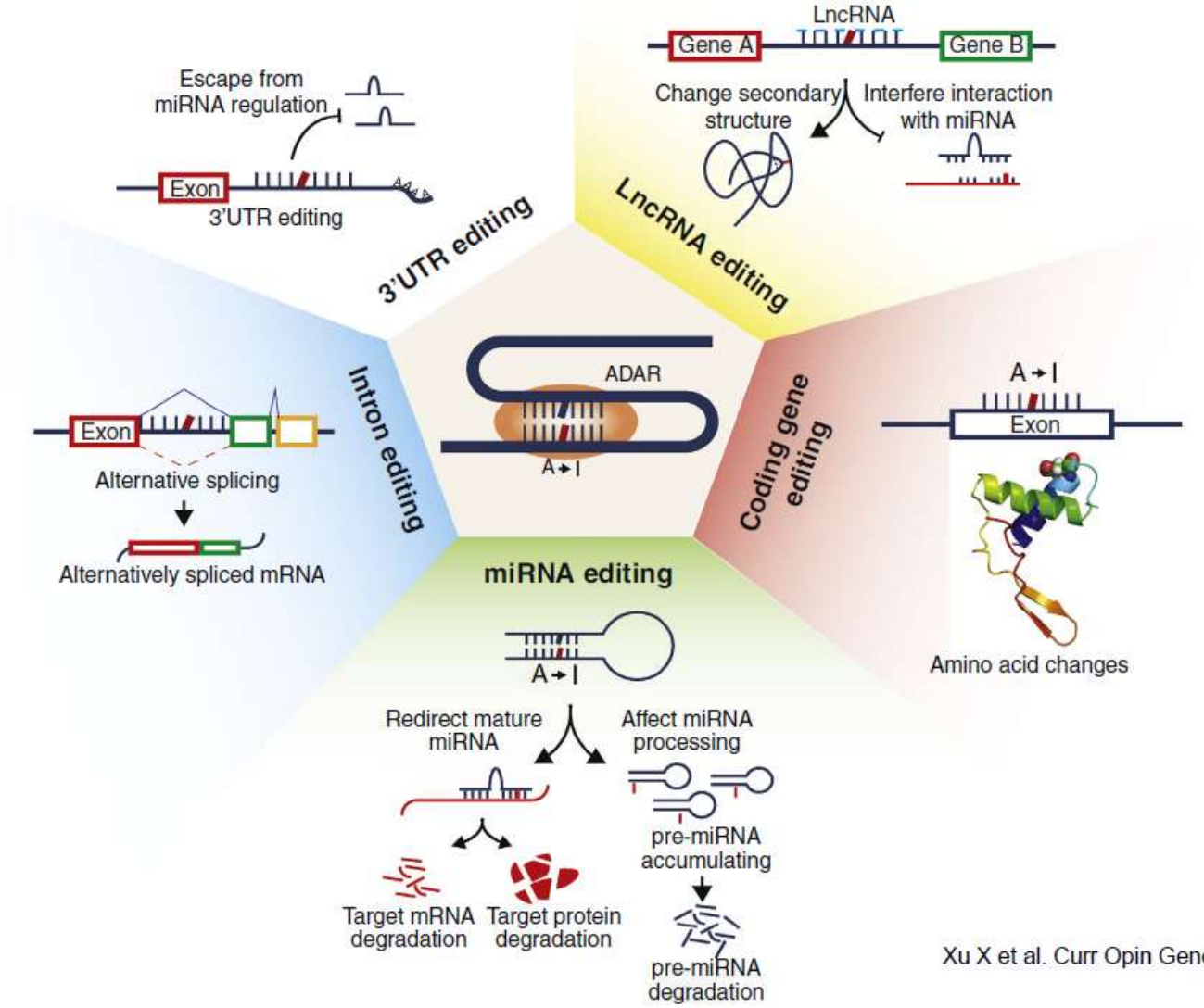
C-to-U editing



APOBEC1 (Apolipoprotein B MRNA Editing Enzyme Catalytic Subunit 1)

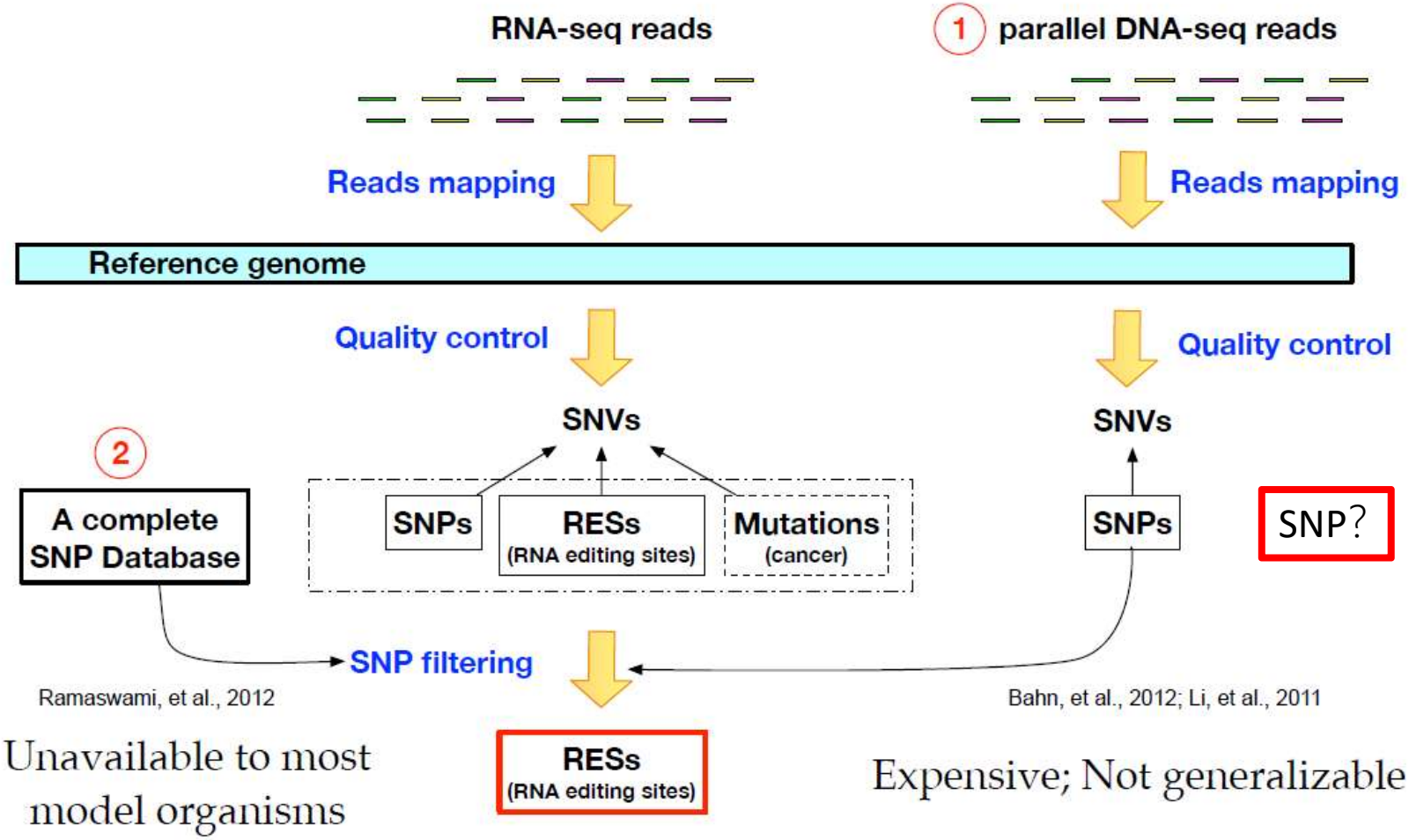
Gott J & Emerson R, Annu Rev Genet. 2000;34:4991-531.

Target of RNA editing

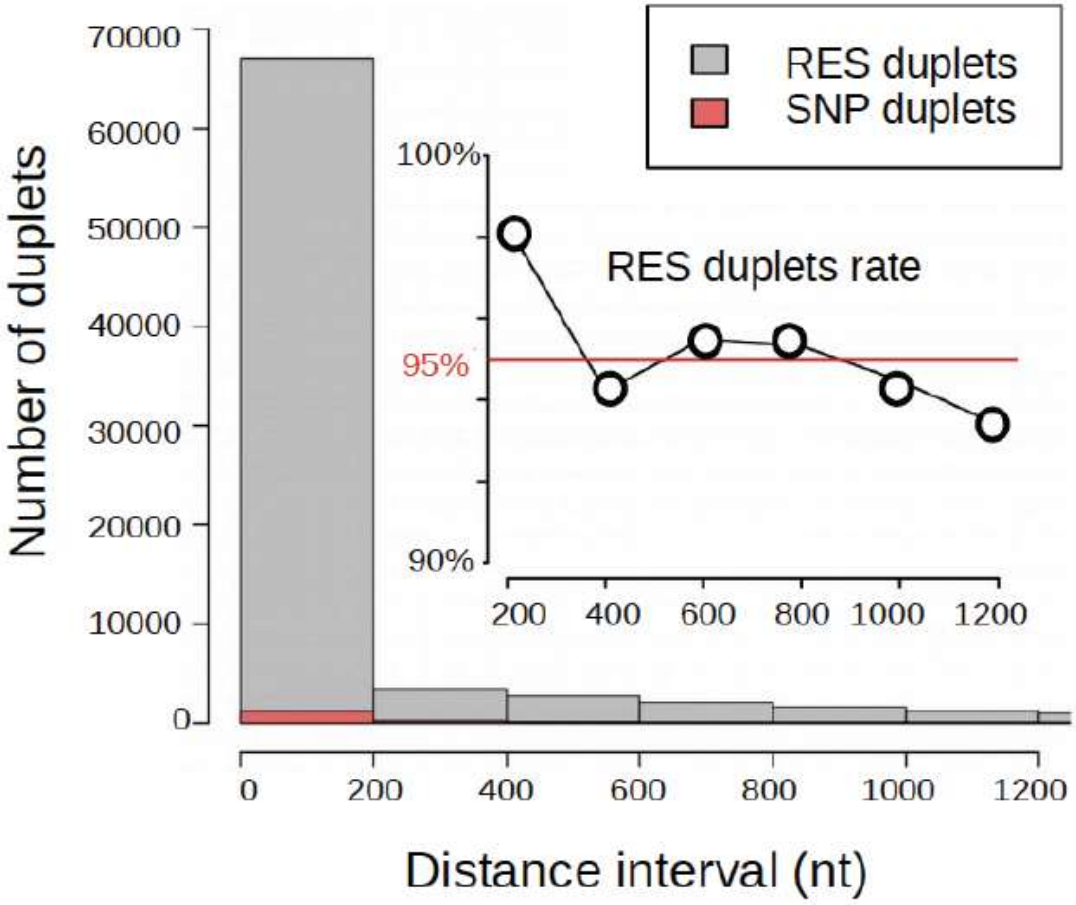


Xu X et al. Curr Opin Genet Dev. (2018)

Typical procedures to detect RNA editing

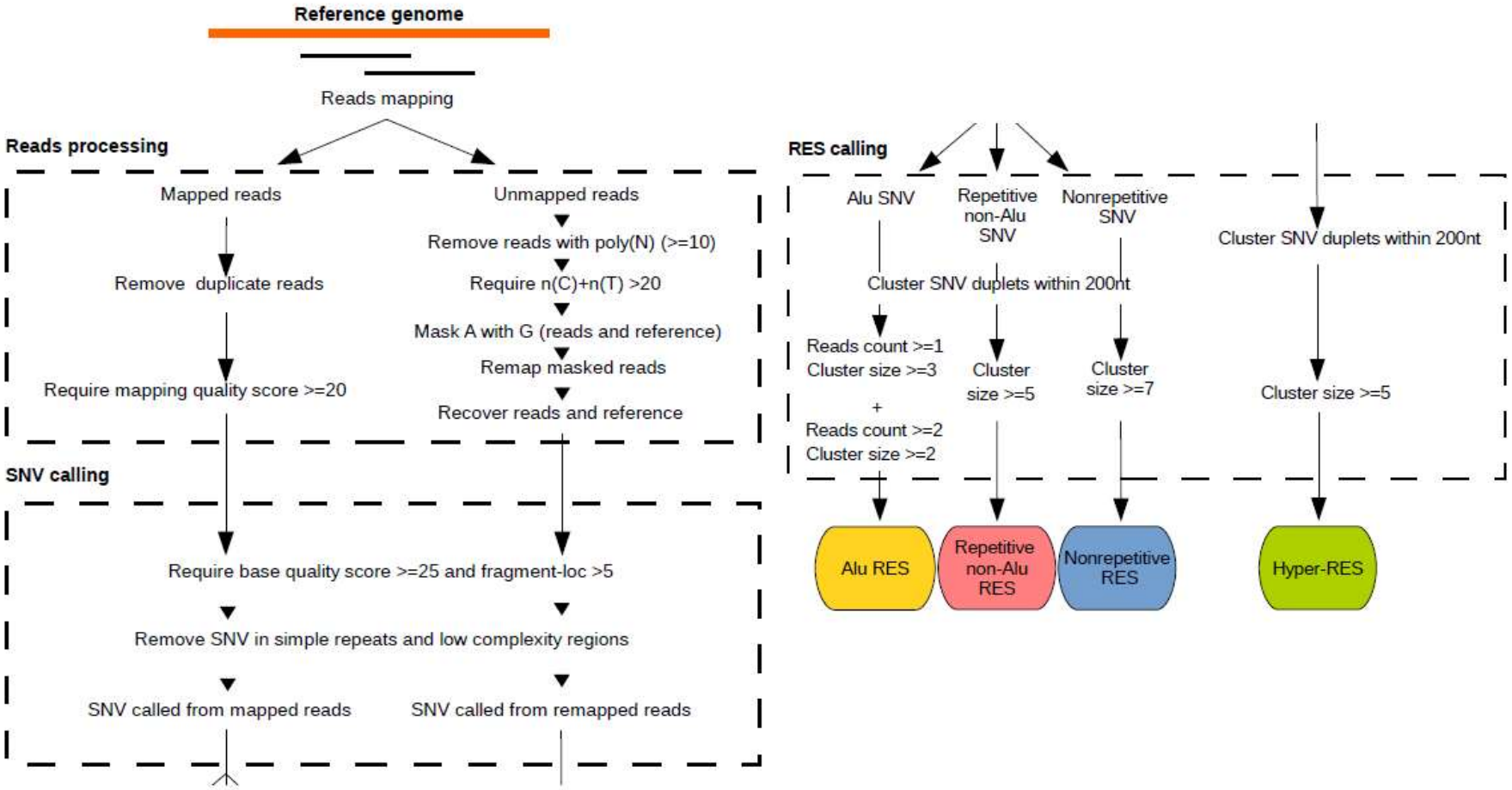


SNV duplet: two adjacent SNVs of same type of variation (A-to-G variation)



Zhang, et al. SPRINT: an SNP-free toolkit for identifying RNA editing sites. Bioinformatics. 2017 Nov 15;33(22):3538-3548.

SPRINT pipeline



Zhang, et al. SPRINT: an SNP-free toolkit for identifying RNA editing sites. Bioinformatics. 2017 Nov 15;33(22):3538-3548.

SPRINT Performance

				Alu sites			Repetitive non-Alu sites				Nonrepetitive sites			
	Cell lines	Tools	Known SNP (%)	Total	A-to-G (%)	Precision (%)	FDR (%)	Total	A-to-G (%)	Precision (%)	FDR (%)	Total	A-to-G (%)	Precision (%)	FDR (%)
Regular-RESs	GM12878, cell	SPRINT	0	336,304	97.7	96.5	-	14,019	87.8	97.2	-	5,407	49.8[a]	96.8	-
		Ramaswami, et al.*	100	147,029	95.8	-	-	2,385	97.4	-	-	1,451	86.6	-	-
	GM12878, cytosolic	SPRINT	0	359,725	98.9	96.9	-	5,469	96	96.8	-	2,081	75.9	95.5	-
		GIREMI*	70	36,131	99	99.4	-	267	83.7	84.3	-	1,193	82.8	73.8	-
		GIREMI*	100	39,757	99.7	-	-	260	88.6	-	-	1,010	73.5	-	-
	U87MG	SPRINT	0	48,085	99.6	96.2	3.2	988	99.5	97.1	4.5	296	87.8	91.2	0
		GIREMI	100	2,152	99.8	-	0.7	114	96.5	-	9	509	88.6	-	53
		RNAEditor	100	62,979	-	-	8.2	6,142	-	-	42.3	155	-	-	55.5
		REDItools	100	313	98	-	2.3	116	38.8	-	100	9,945	40	-	100

			Total	A-to-G (%)
Hyper-RESs	GM12878, cell	SPRINT	328,762	97.9
		Porath et al.*	157,077	96
	U87MG	SPRINT	57,913	96
		Porath et al.*	27,124	94.6

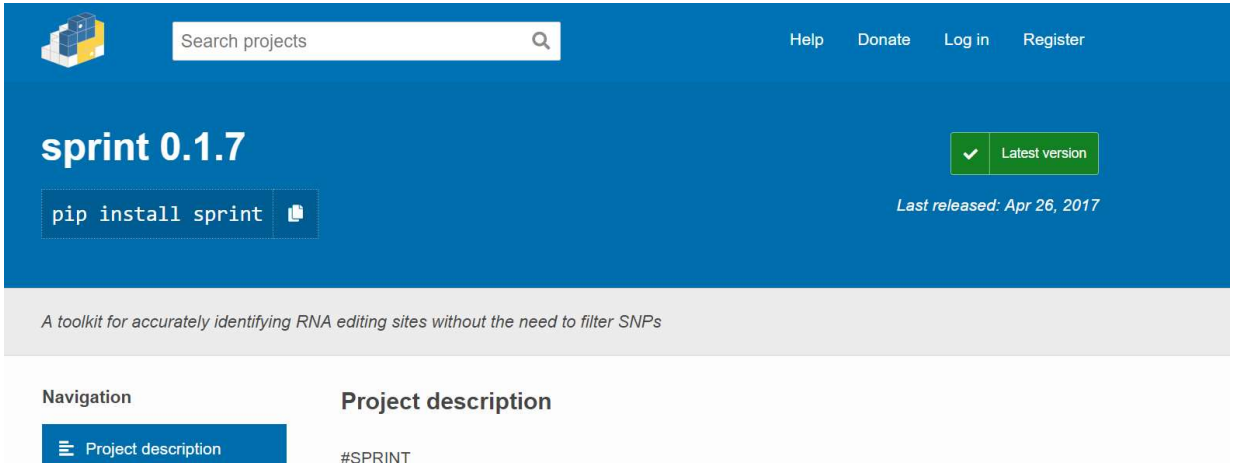
Zhang, et al. SPRINT: an SNP-free toolkit for identifying RNA editing sites. Bioinformatics. 2017 Nov 15;33(22):3538-3548.

Package——SPRINT

Environment: **Python2.7**

Installation: **pip install sprint**

For more info: https://pypi.org/project/sprint/



Thank you!