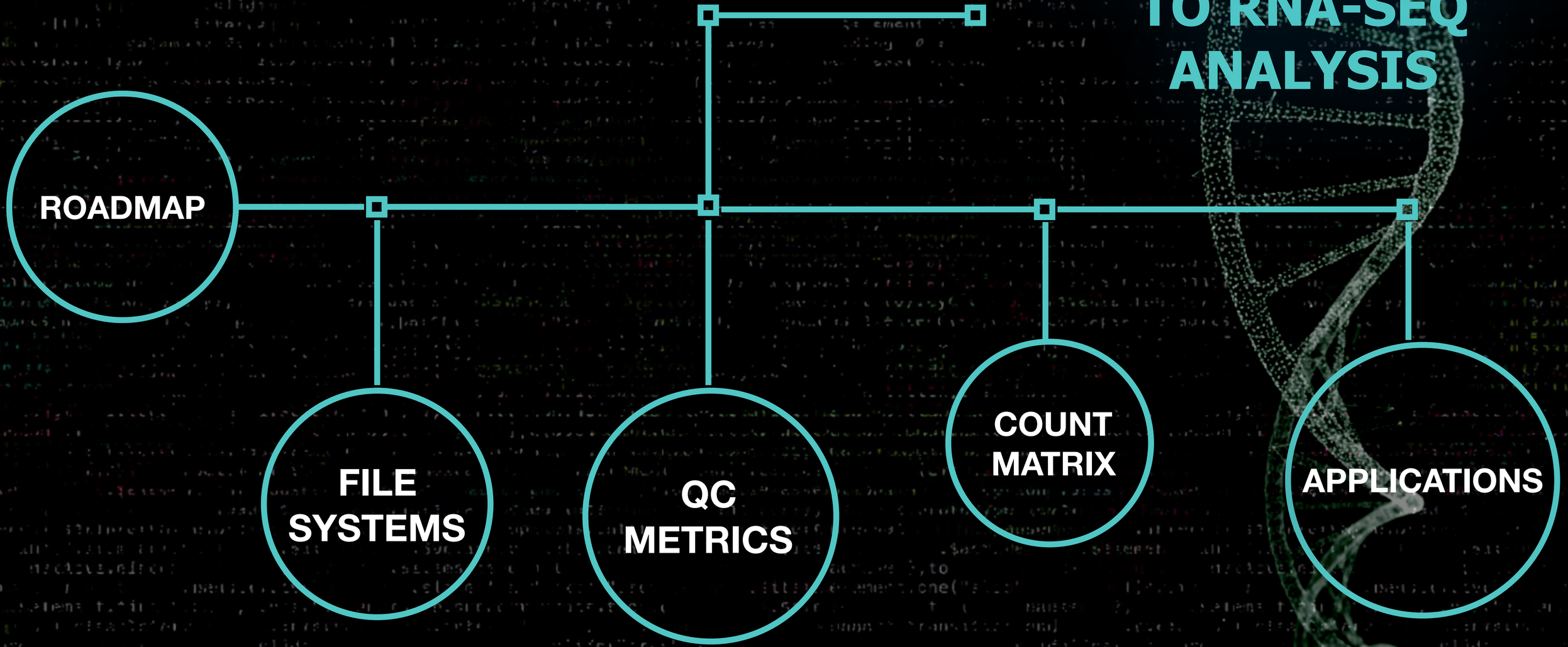
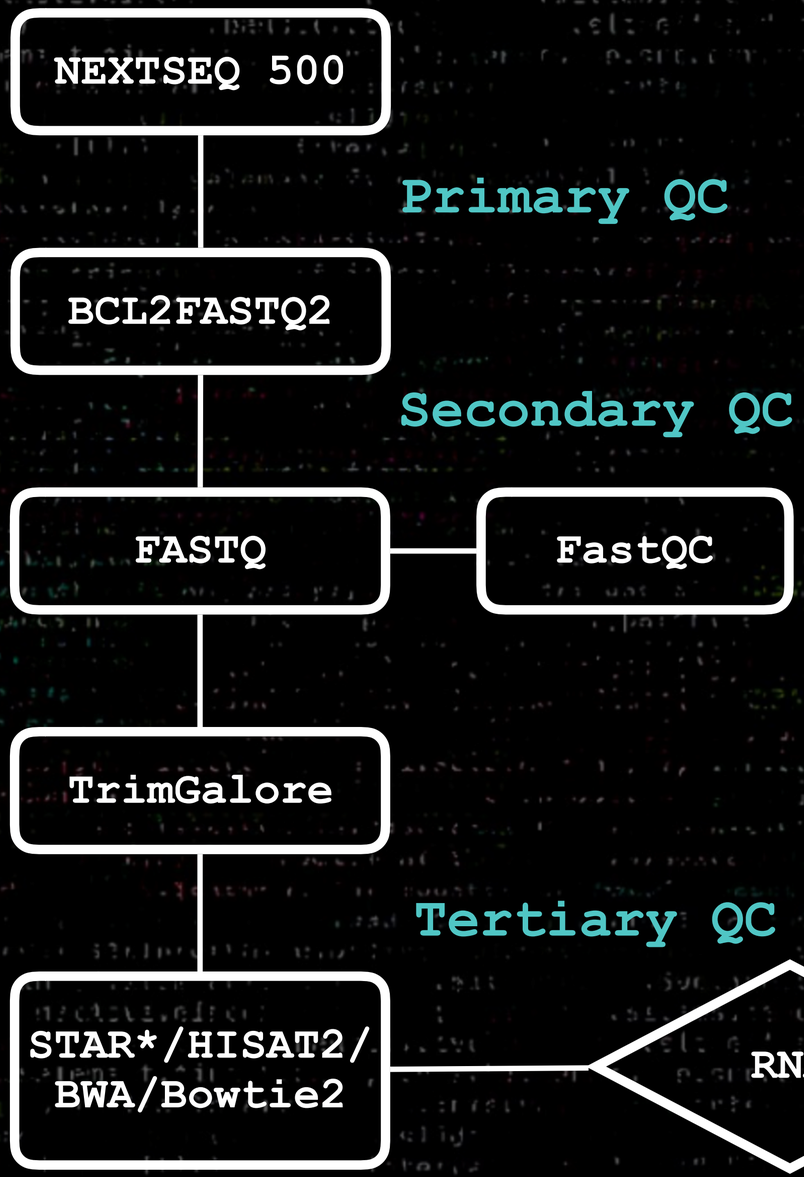
AN INTRODUCTION TO RNA-SEQ ANALYSIS



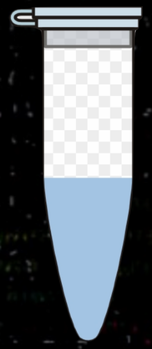
TREX - Faraz Ahmed



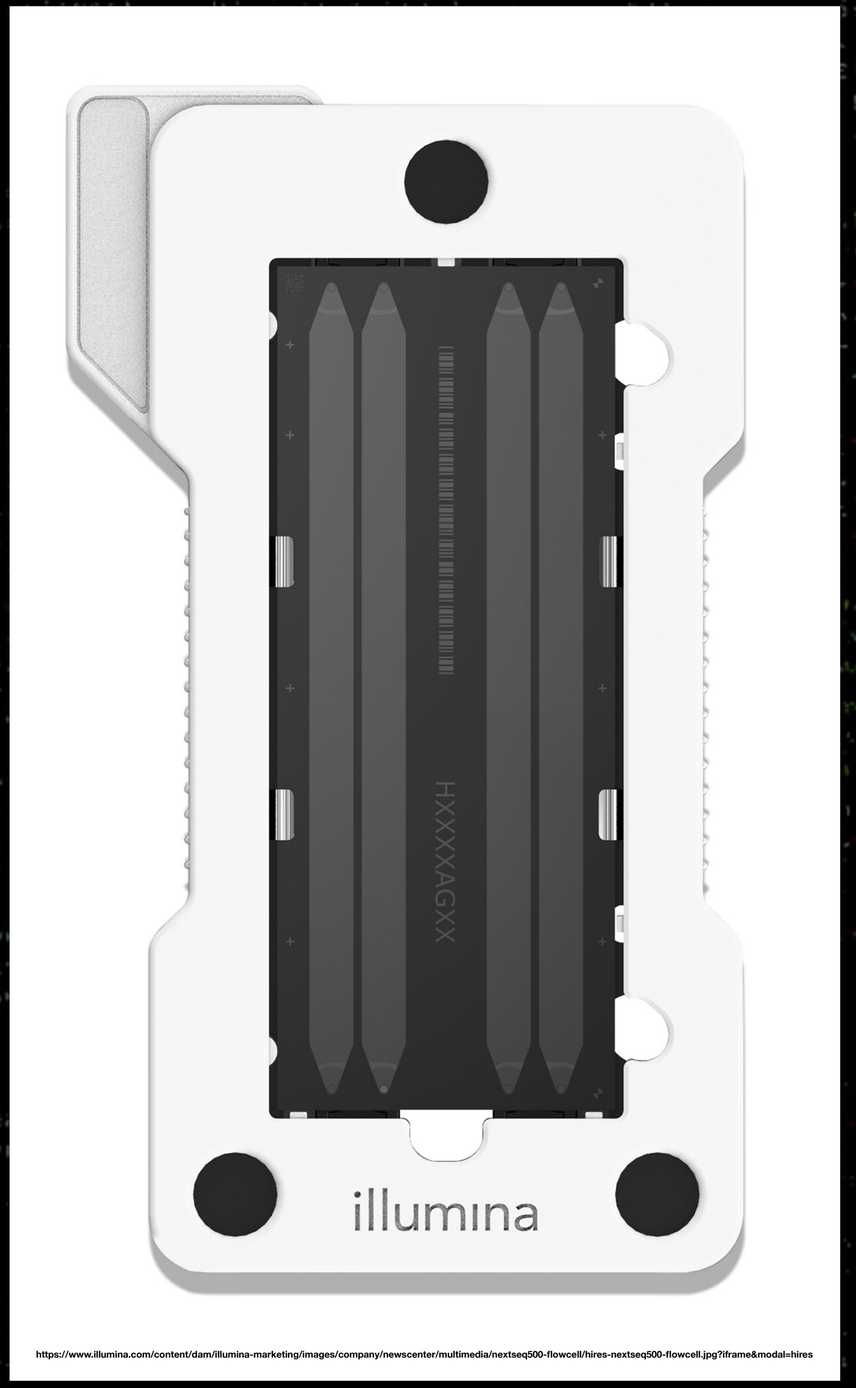
	wt1	wt2	wt3	wt4	wt5	ko1	ko2	ko3	ko4	ko5
gene1	135	148	146	121	140	269	268	227	263	259
gene2	803	797	841	800	874	412	408	388	393	398
gene3	40	25	38	41	35	413	393	417	374	415
gene4	381	383	415	374	354	809	840	859	856	845
gene5	775	766	773	749	784	302	310	324	342	314
gene6	305	313	256	313	315	831	817	832	859	869
gene7	816	819	800	793	790	485	481	429	461	508
gene8	40	22	40	37	32	421	476	479	528	483
gene9	963	935	938	953	948	43	26	41	28	39
gene10	697	749	715	724	715	233	259	284	277	269
gene11	36	50	40	35	44	168	178	168	170	187
gene12	60	66	54	61	71	288	289	293	289	330
gene13	537	517	523	512	515	142	134	145	145	145
gene14	655	615	610	664	606	842	889	827	885	838
gene15	426	439	436	420	432	131	155	159	139	151
gene16	952	976	974	987	947	789	828	825	850	796
gene17	379	446	410	423	394	963	1012	913	968	984
gene18	17	17	14	20	22	131	113	135	127	112
gene19	985	874	896	982	992	848	890	899	896	873
gene20	197	191	202	180	172	765	754	784	791	799
gene21	399	477	414	466	440	686	668	741	754	718

*--quantMode FOR RNA SEQ READS

BCL2FASTQ2

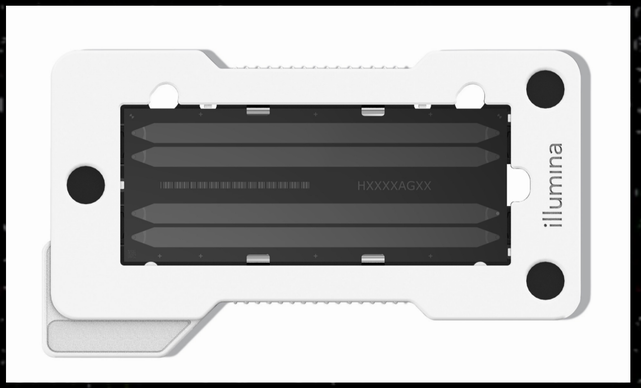


Multiplex Run = Sequencing of Multiple Samples in one RUN



https://www.illumina.com/content/dam/illumina-marketing/images/company/newscenter/multimedia/nextseq500-flowcell/hires-nextseq500-flowcell.jpg?iframe&modal=hires

BCL2FASTQ2



Faster Sequencing and Data Processing Times

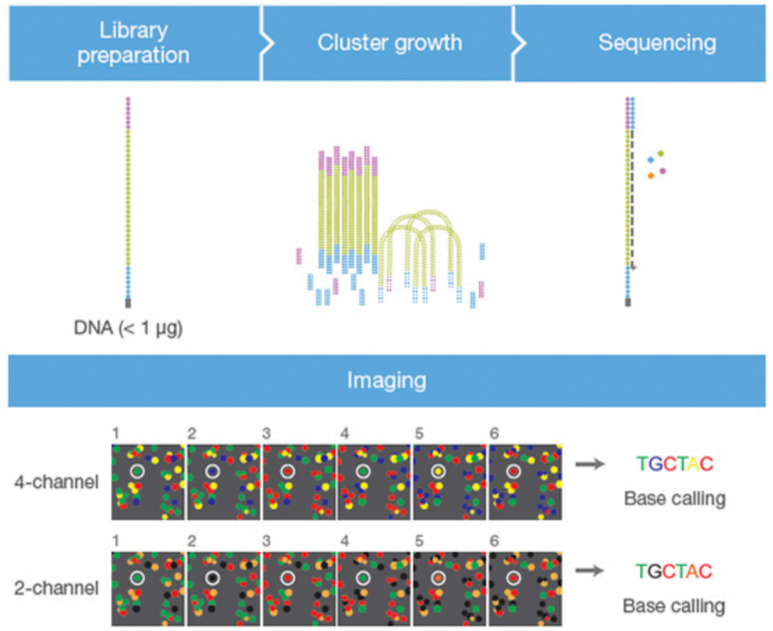


Figure 1. 4-Channel vs. 2-Channel SBS Technology Image Detection and Base Calling.
In 4-channel SBS, 4 images are necessary to capture the unique florescent dyes for each base. In contrast, 2-channel SBS requires only 2 images to determine all 4 base calls.

Faster Sequencing and Data Processing Times

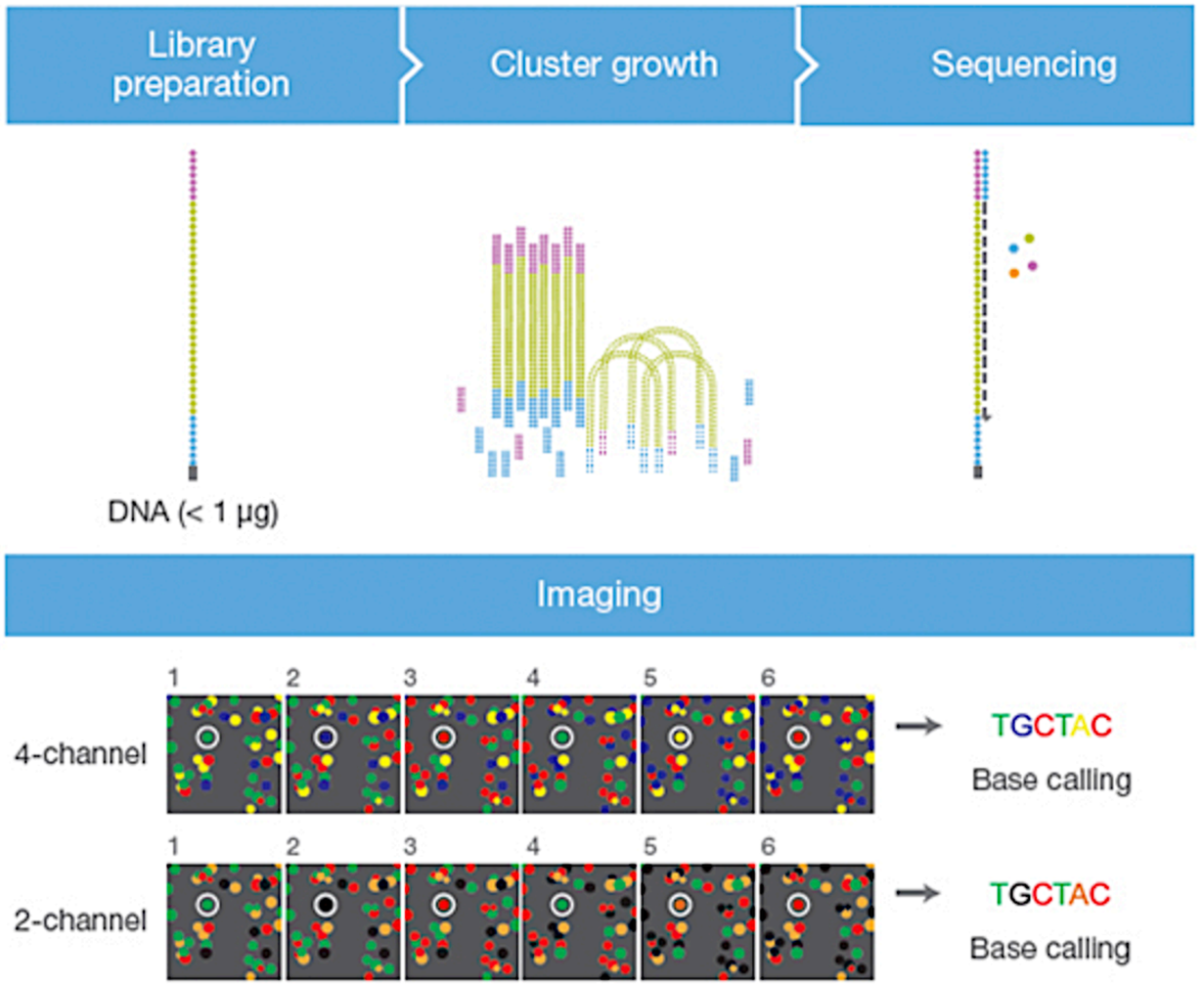


Figure 1. 4-Channel vs. 2-Channel SBS Technology Image Detection and Base Calling.
In 4-channel SBS, 4 images are necessary to capture the unique florescent dyes for each base. In contrast, 2-channel SBS requires only 2 images to determine all 4 base calls.

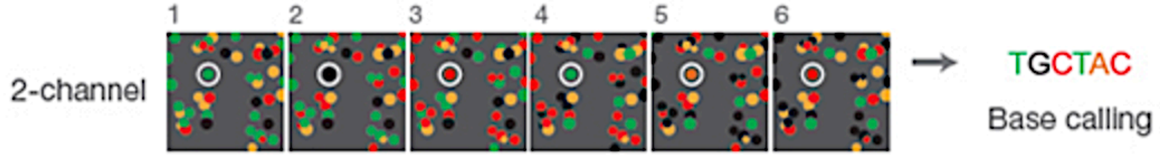
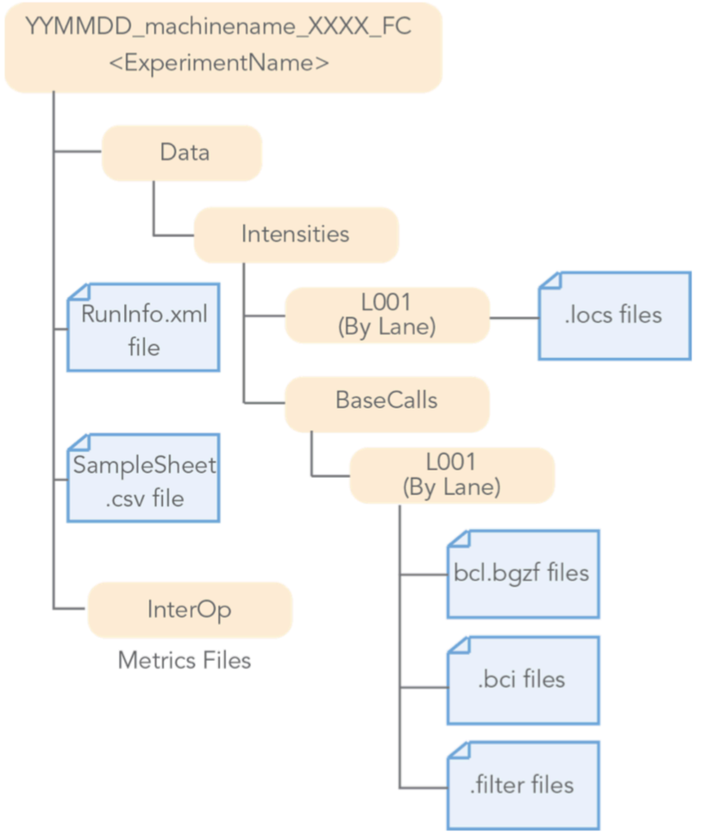


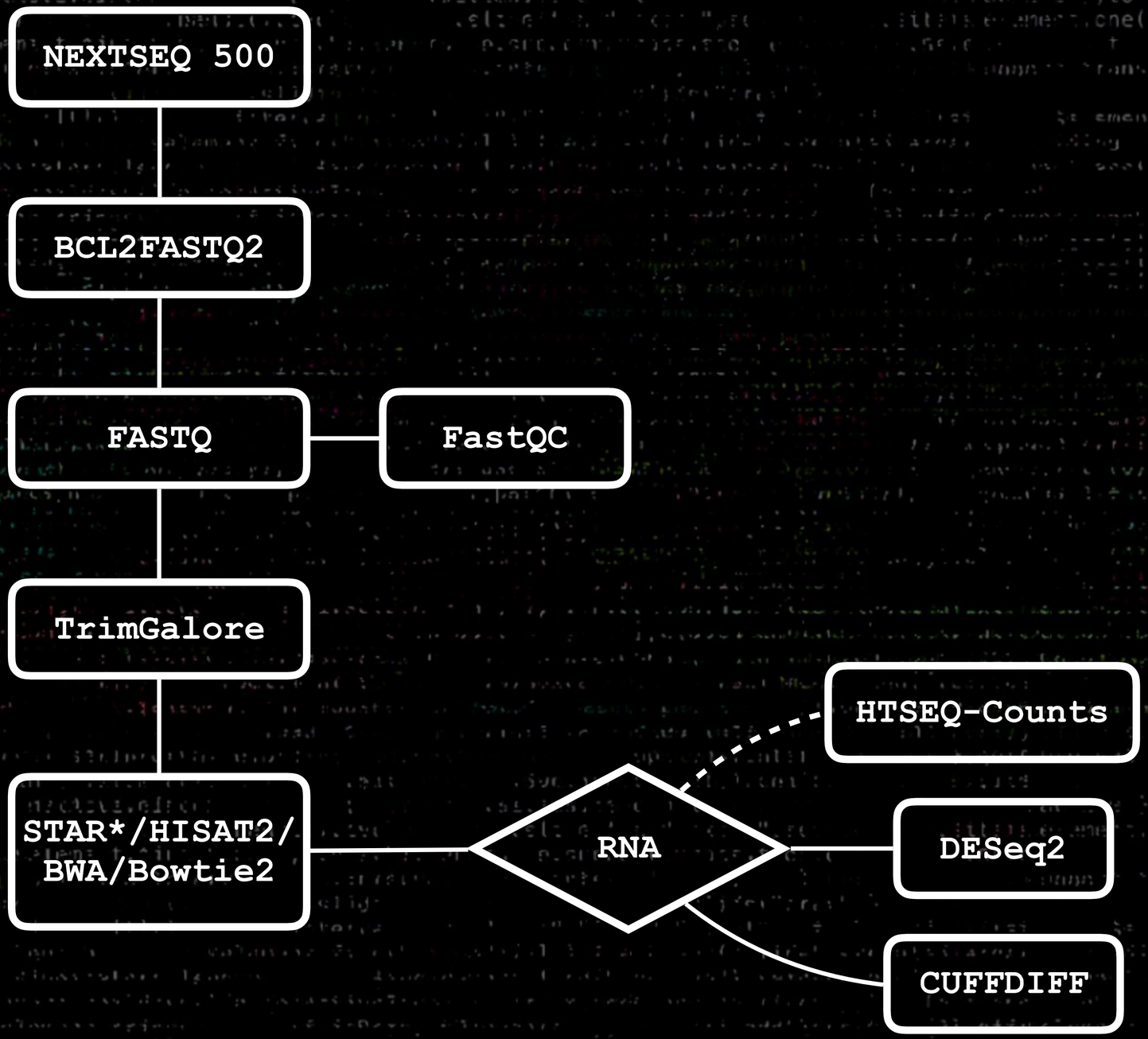
Figure 1. 4-Channel vs. 2-Channel SBS Technology Image Detection and Base Calling.
In 4-channel SBS, 4 images are necessary to capture the unique florescent dyes for each base. In contrast, 2-channel SBS requires only 2 images to determine all 4 base calls.

Demultiplexing;

Typically done by the sequencing facility

Figure 2 BCL Conversion Input Files from the MiniSeq or NextSeq System





*--quantMode FOR RNA SEQ READS

FASTQ

Light intensities translated to TEXT;

Derived from FASTA;

Very large file, contains every single sequenced read for a given sample;

Number of FASTQ files = Number of Samples (for PE x2)

FASTQ

@<machine_id>:<run number>:<flowcell ID>:<lane>:<tile>:<x-pos>:<y-pos>
<read>:<is filtered>:<control number>:<index sequence>

```
@NB500995:63:HC5GHBGX5:1:11101:22946:1063 1:N:0:CAACTAAT+AGATCTCG
GCAACNTTGATCAGTTCTGACACAGTGTTTTGAACCATATCAGGATCCCTCACATCACACTGAATTGCATGAACCT
+
AAAAA#/EEEEAEEEEEEEEEEEEEEEEEEEEAEEEEEEEEEE/EEEEEEAAEEEE/EEEEEAEEEEEE/E/EEEE
```

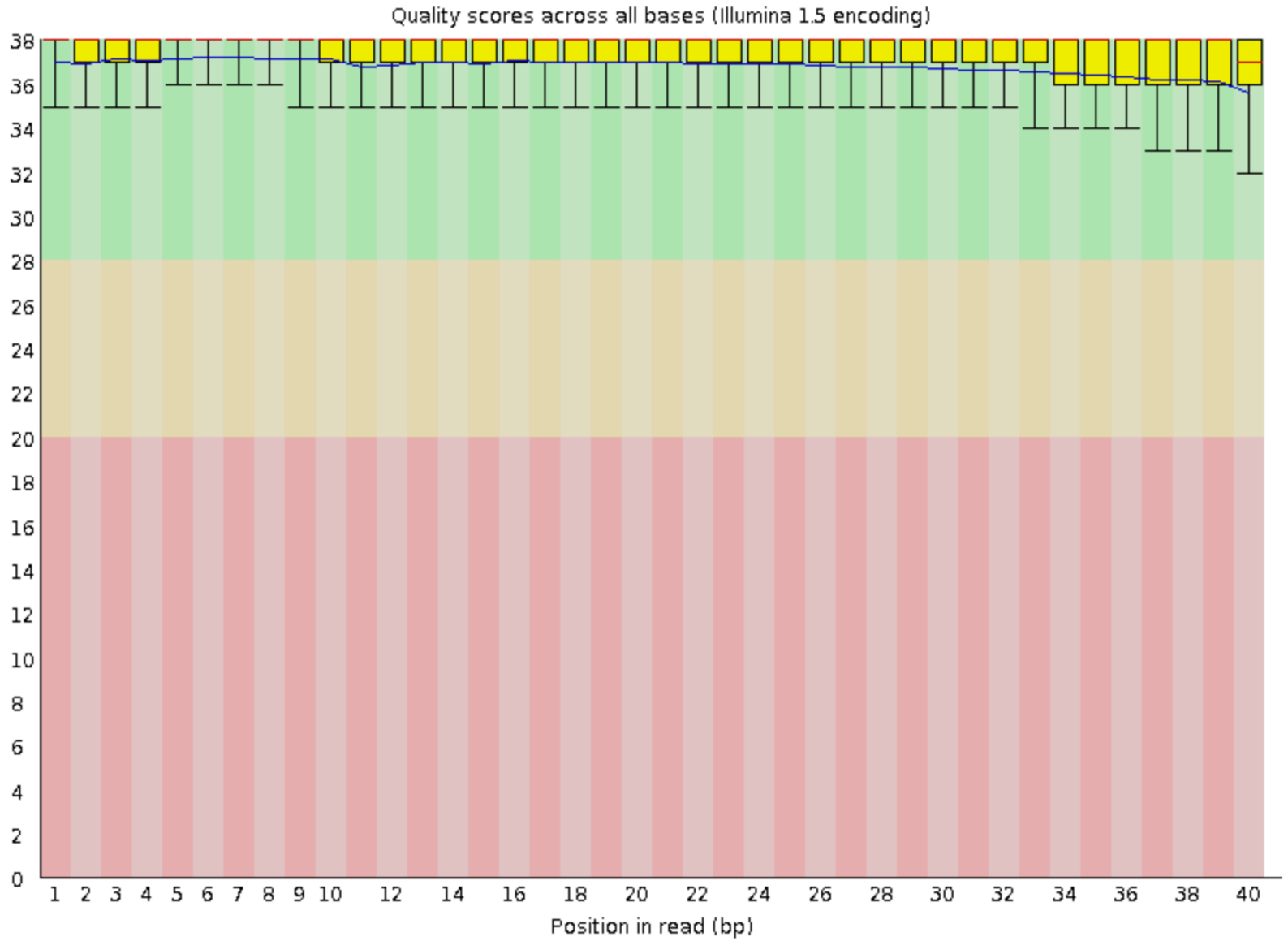
FASTQ

@NB500995:6
GCAACNTTGAT
+
AAAAA#/EEEE

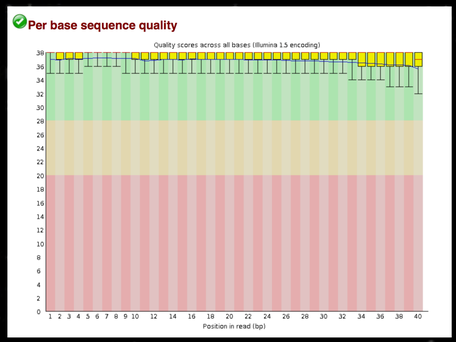
GCATGAACCT

EEE/E/EEEE

Per base sequence quality

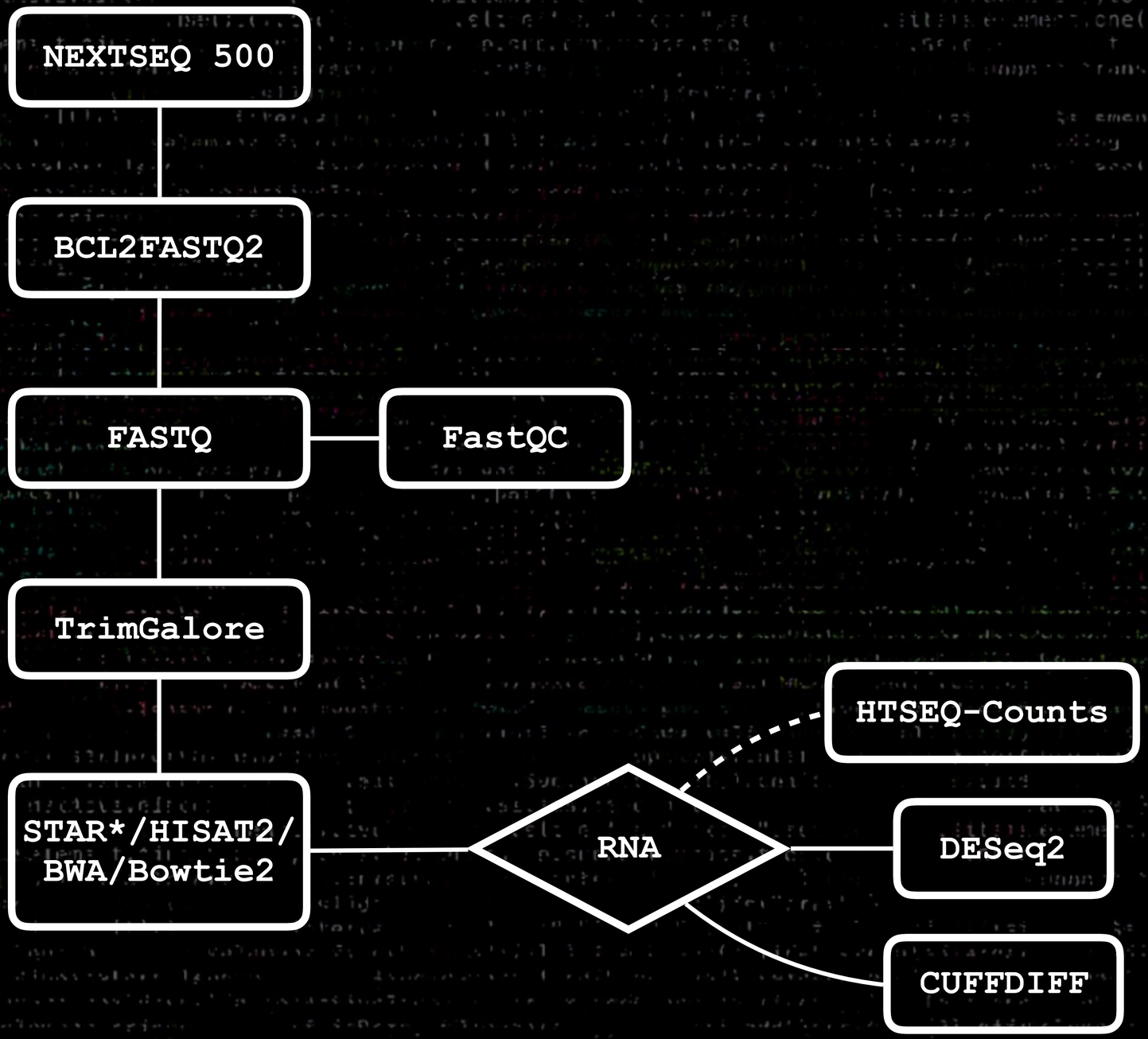


FASTQ → FastQC



A poor RNA-seq run is characterized by:

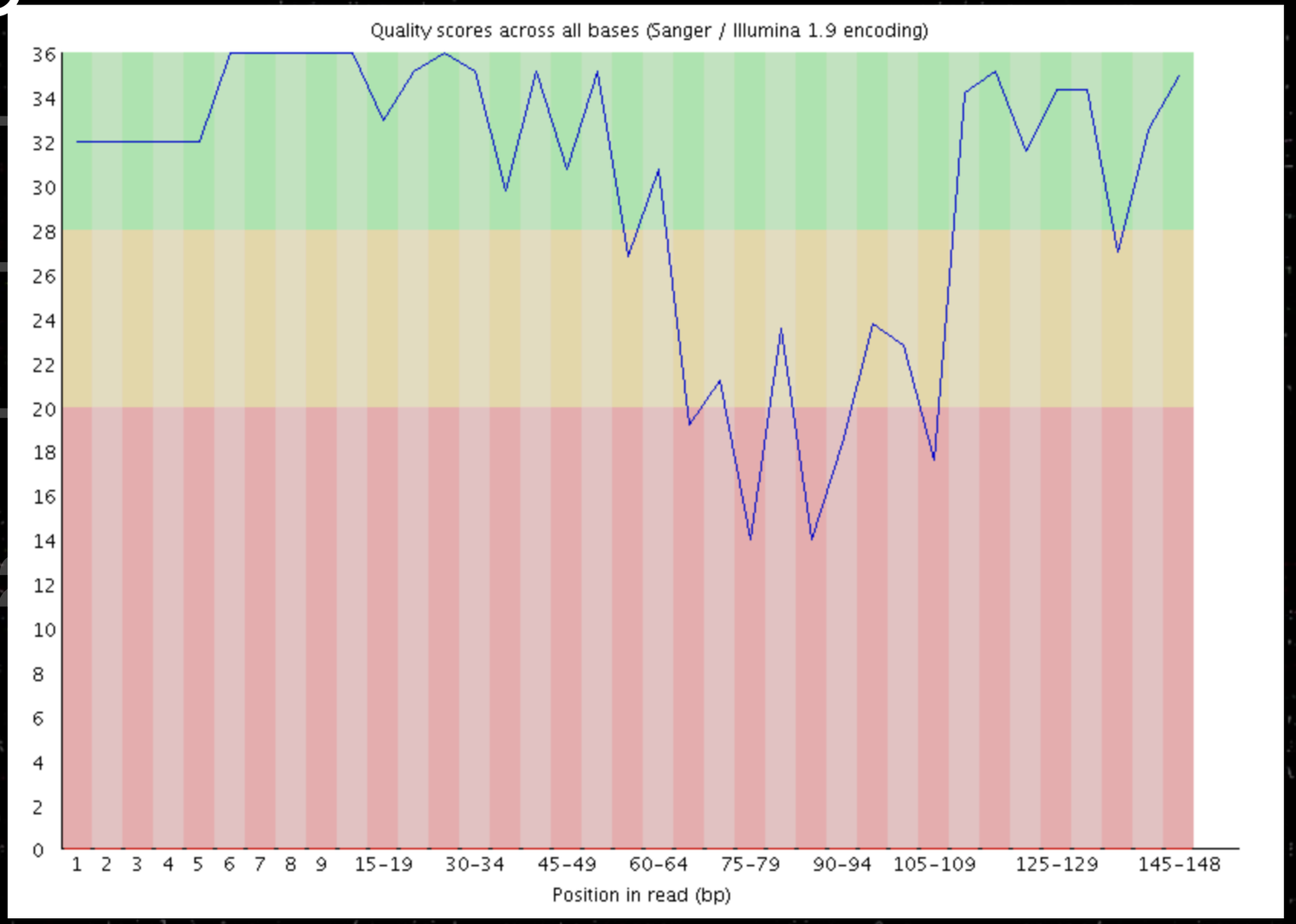
- PCR duplicates
- Adapter contamination
- rRNA and tRNA reads
- Unmappable reads

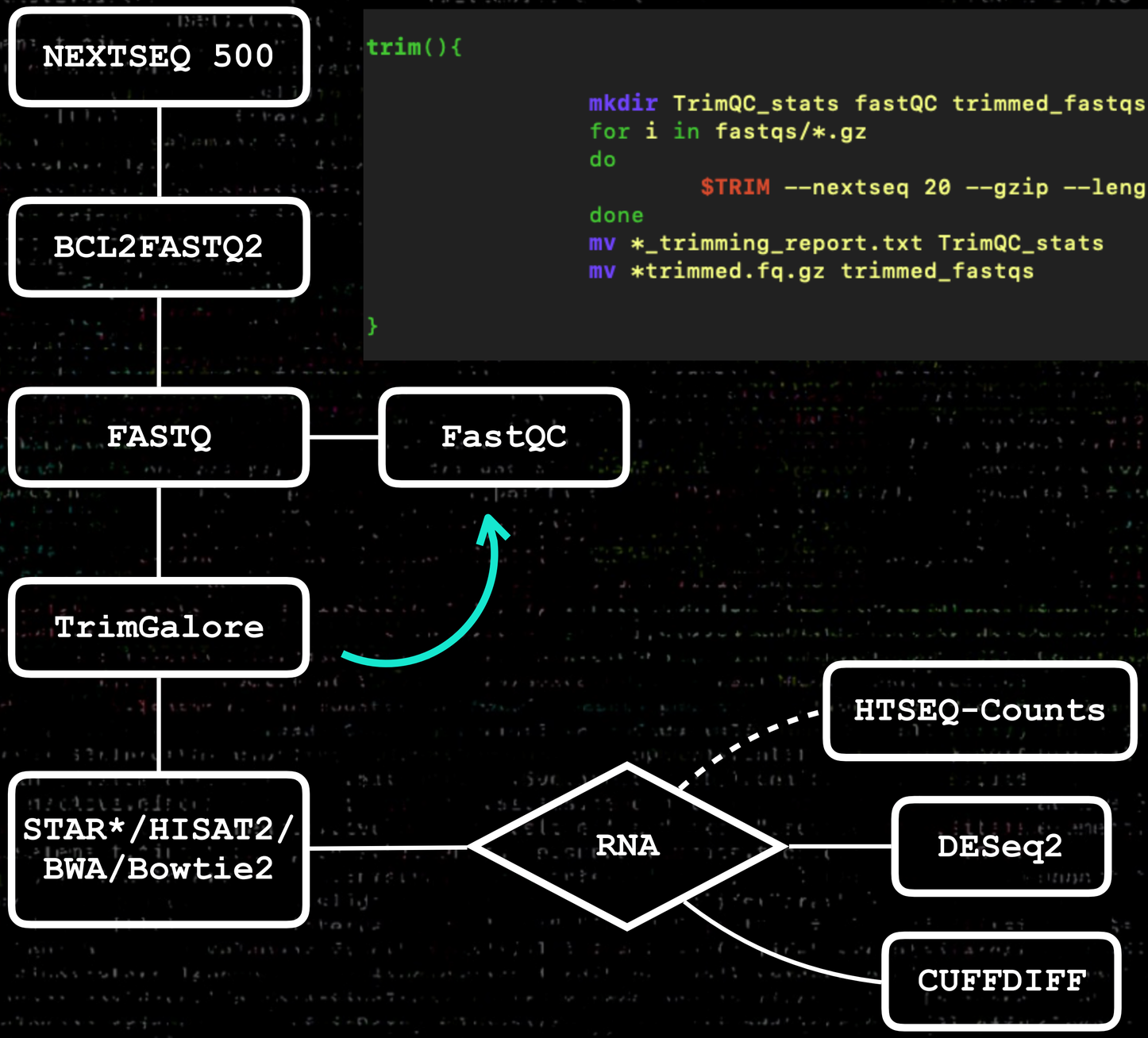

NEXTSEQ 500
BCL2FASTQ2
FASTQ
FastQC
TrimGalore
STAR*/HISAT2/
BWA/Bowtie2
RNA
HTSEQ-Counts
DESeq2
CUFFDIFF

TrimGalore

- Trim for low quality reads;
- Trim for adapter sequences;
- Trim for noisy short fragments;
- 2 color chemistry bias;

TrimGalore


Quality scores across all bases (Sanger / Illumina 1.9 encoding)
36
34
32
30
28
26
24
22
20
18
16
14
12
10
8
6
4
2
0
1 2 3 4 5 6 7 8 9 15-19 30-34 45-49 60-64 75-79 90-94 105-109 125-129 145-148
Position in read (bp)

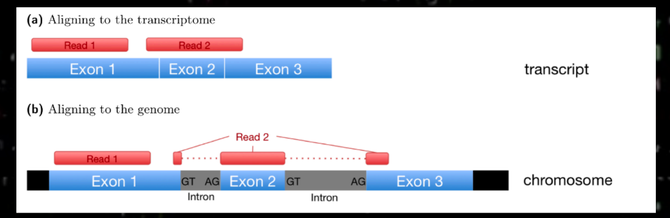



```
trim(){

        mkdir TrimQC_stats fastQC trimmed_fastqs
        for i in fastqs/*.gz
        do
                $TRIM --nextseq 20 --gzip --length 50  --fastqc --fastqc_args "-t 4 --outdir ./fastQC" $i
        done
        mv *_trimming_report.txt TrimQC_stats
        mv *trimmed.fq.gz trimmed_fastqs

}
```

STAR*/HISAT2/
BWA/Bowtie2



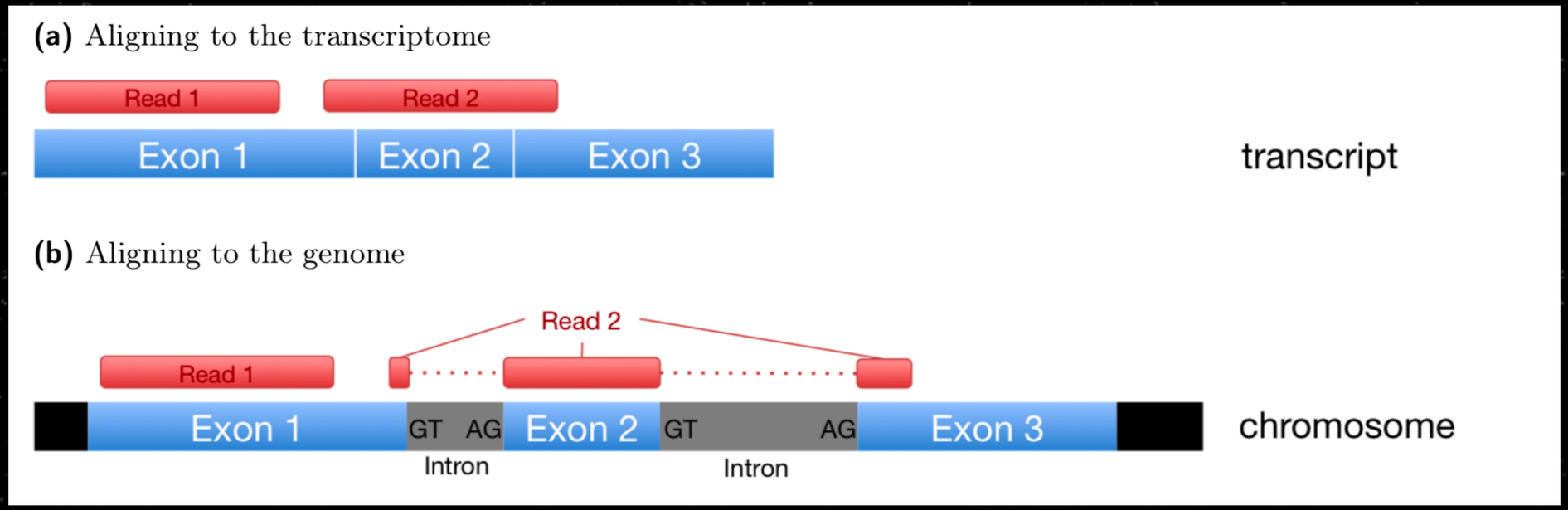
MAPPING/ALIGNMENT :

Assignment of FASTQ reads to most likely locus of origin in the REFERENCE GENOME

This step has to be done for all FASTQ files

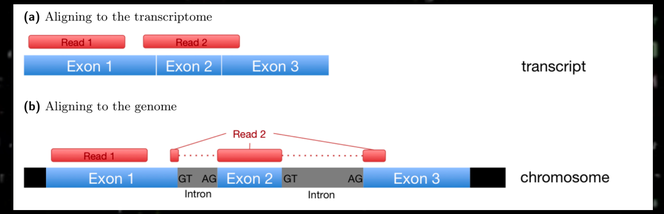
*--quantMode FOR RNA SEQ READS

STAR*/HISAT2/ BWA/Bowtie2



*--quantMode FOR RNA SEQ READS

STAR*/HISAT2/
BWA/Bowtie2



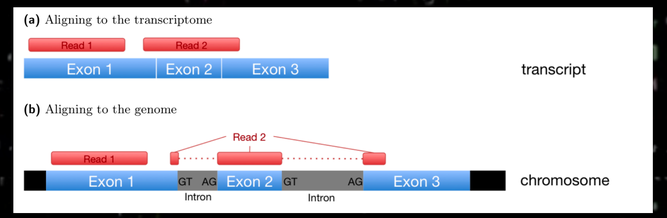
Preliminary Step for all Aligners:

Generate genome index (per genome type)

Allows for computationally efficient mapping

*--quantMode FOR RNA SEQ READS

STAR*/HISAT2/ BWA/Bowtie2

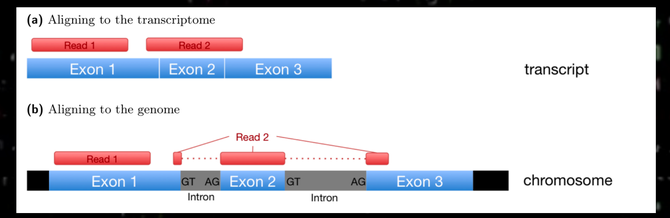



```
STAR \
    --runThreadN 12 \
    --runMode genomeGenerate \
    --genomeDir /path/to/genomeIndex/ \
    --genomeFastaFiles referenceGenome.fasta \
    --sjdbGTFfile referenceAnnotation.gtf \
    --sjdbOverhang 100 \
    --limitGenomeGenerateRAM 152003700778
```

```
STAR \
        --runThreadN 12 \
        --genomeDir /path/to/genomeIndex \
        --readFilesIn sample.fastq.gz \
        --readFilesCommand gunzip -c \
        --outSAMstrandField intronMotif \
        --outFilterIntronMotifs RemoveNoncanonical \
        --outSAMtype BAM SortedByCoordinate \
        --outFileNamePrefix prefix. \
        --limitBAMsortRAM 61675612266 \
        --quantMode GeneCounts
```

*--quantMode FOR RNA SEQ READS

STAR*/HISAT2/ BWA/Bowtie2




```
STAR \
    --runThreadN 12 \
    --runMode genomeGenerate \
    --genomeDir /path/to/genomeIndex/ \
    --genomeFastaFiles referenceGenome.fasta \
    --sjdbGTFfile referenceAnnotation.gtf \
    --sjdbOverhang 100 \
    --limitGenomeGenerateRAM 152003700778
```

```
STAR \
        --runThreadN 12 \
        --genomeDir /path/to/genomeIndex \
        --readFilesIn sample.fastq.gz \
        --readFilesCommand gunzip -c \
        --outSAMstrandField intronMotif \
        --outFilterIntronMotifs RemoveNoncanonical \
        --outSAMtype BAM SortedByCoordinate \
        --outFileNamePrefix prefix. \
        --limitBAMsortRAM 61675612266 \
        --quantMode GeneCounts
```

*--quantMode FOR RNA SEQ READS

SEQUENCE ALIGNMENT MAP: SAM/BAM

Consensus format to store alignment records;

All aligners will generate results in the SAM format;

There are 2 sections in this file:

- Header Section
- Alignment Section

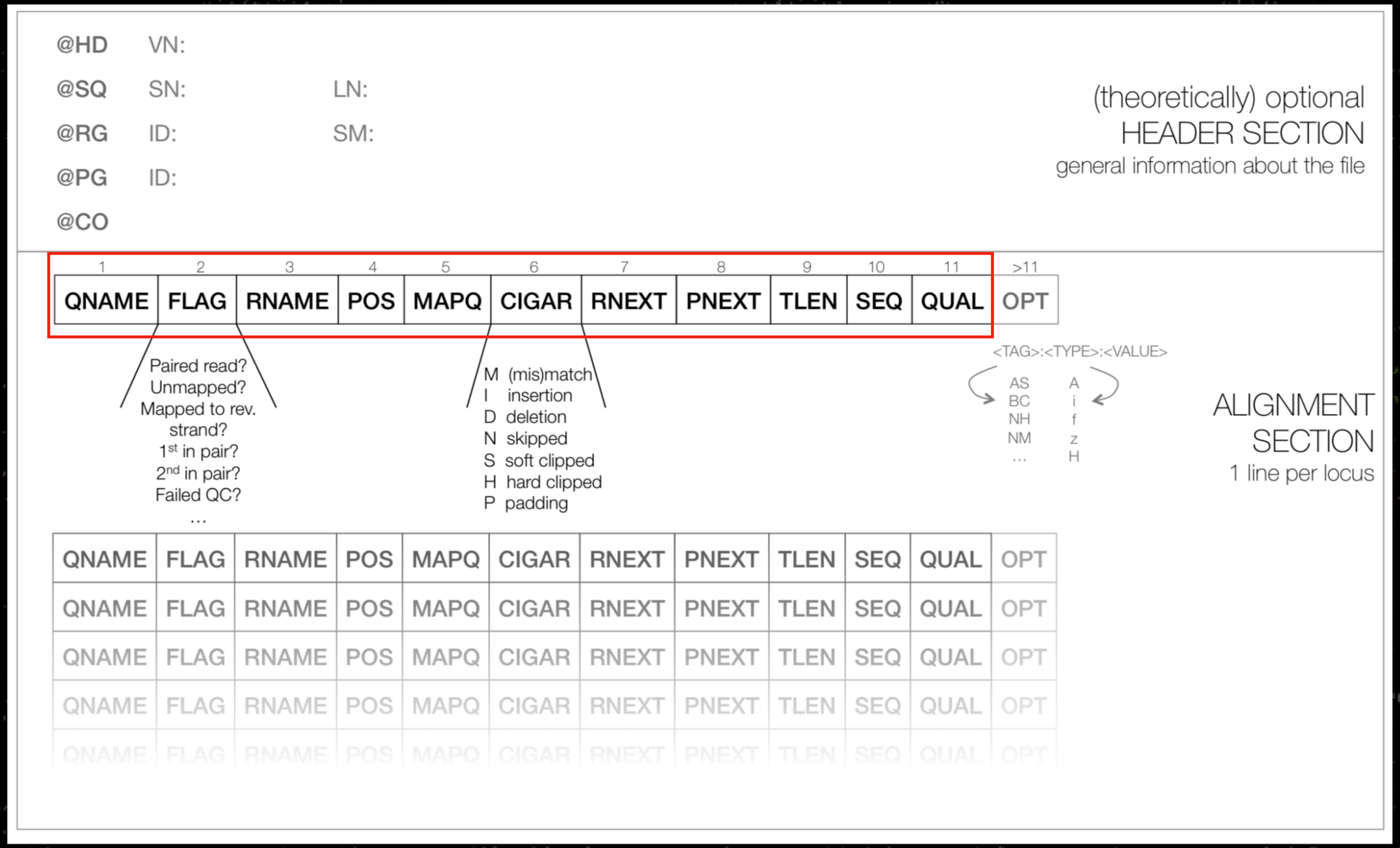
SEQUENCE ALIGNMENT MAP: SAM/BAM

Consensus format to store alignment records;

All aligners will generate results in the SAM format;

There are 2 sections in this file:

- Header Section
- Alignment Section



<QNAME> <FLAG> <RNAME> <POS> <MAPQ> <CIGAR> <MRNM> <MPOS> <ISIZE> <SEQ> <QUAL>

Pos.	Field	Example entry	Description	NA value
1	QNAME	Read1	Query template (= read) name (PE: read pair name)	required
2	FLAG	83 	Information about the read's mapping properties encoded as bit-wise flags (see next section and Table 5).	required
3	RNAME	chrI	Reference sequence name. This should match a `@SQ` line in the header.	*
4	POS	15364	1-based leftmost mapping position of the first matching base. Set as 0 for an unmapped read without coordinates.	0
5	MAPQ	30	Mapping quality of the alignment. Should be a Phred-scaled posterior probability that the position of the read is incorrect, but the value is completely dependent on the alignment program. Some tools set this to 0 if multiple alignments are found for one read.	0
6	CIGAR	51M	Detailed information about the alignment (see below).	*
7	RNEXT	=	PE reads: reference sequence name of the next read. Set to "=" if both mates are mapped to the same chromosome.	*
8	PNEXT	15535	PE reads: leftmost mapping position of the next read.	0
9	TLEN	232	PE reads: inferred template length (fragment size).	0
10	SEQ	CCA...GGC	The sequence of the aligned read on the forward strand (not including indels).	*
11	QUAL	BBH...1+B	Base quality (same as the quality string in the `FASTQ` format, but always in Sanger format [ASCII+33]).	*
12ff	OPT	NM:i:0	Optional fields (format: `<TAG>:<TYPE>:<VALUE>`; see below).	

<QNAME> <FLAG> <RNAME> <POS> <MAPQ> <CIGAR> <MRNM> <MPOS> <ISIZE> <SEQ> <QUAL>

Pos.	Field	Example entry	Description	NA value
1	QNAME	Read1	Query template (= read) name (PE: read pair name)	required
2	FLAG	83		
3	RNAME	chrI		
4	POS	1536		
5	MAPQ	30		
6	CIGAR	51M		
7	RNEXT	=		
8	PNEXT	1553		
9	TLEN	232		
10	SEQ	CCA		
11	QUAL	BBF		
12ff	OPT	NM:		

SAM Flag: 4 Explain

Switch to mate Toggle first in pair / second in pair

Find SAM flag by property:

To find out what the SAM flag value would be for a given combination of properties, tick the boxes for those that you'd like to include. The flag value will be shown in the SAM Flag field above.

- [ ] read paired
- [ ] read mapped in proper pair
- [x] read unmapped
- [ ] mate unmapped
- [ ] read reverse strand
- [ ] mate reverse strand
- [ ] first in pair
- [ ] second in pair
- [ ] not primary alignment
- [ ] read fails platform/vendor quality checks
- [ ] read is PCR or optical duplicate
- [ ] supplementary alignment

Summary:

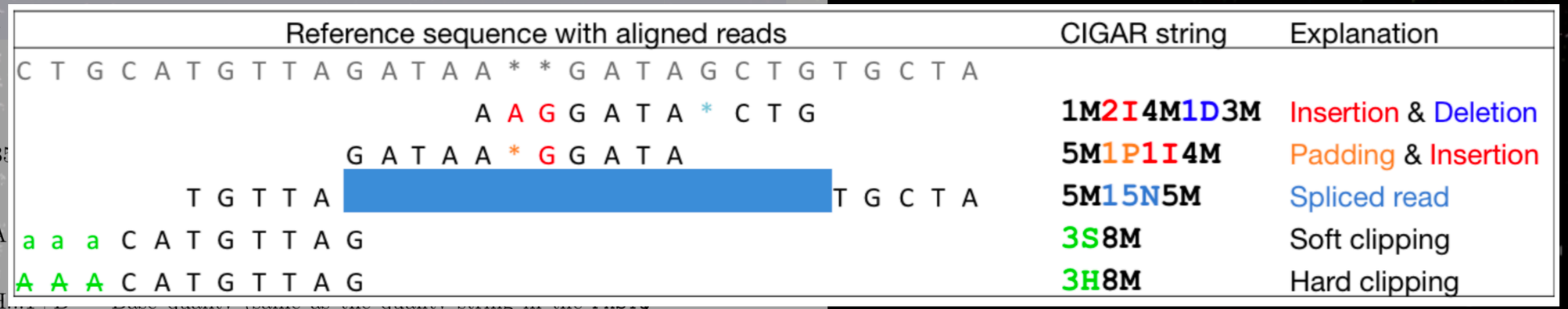
read unmapped (0x4)

<QNAME> <FLAG> <RNAME> <POS> <MAPQ> <CIGAR> <MRNM> <MPOS> <ISIZE> <SEQ> <QUAL>

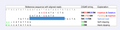
Pos.	Field	Example entry	Description	NA value
1	QNAME	Read1	Query template (= read) name (PE: read pair name)	required
2	FLAG	83 	Information about the read's mapping properties encoded as bit-wise flags (see next section and Table 5).	required
3	RNAME	chrI	Reference sequence name. This should match a `@SQ` line in the header.	*
4	POS	15364	1-based leftmost mapping position of the first matching base. Set as 0 for an unmapped read without coordinates.	0
5	MAPQ	30	Mapping quality of the alignment. Should be a Phred-scaled posterior probability that the position of the read is incorrect, but the value is completely dependent on the alignment program. Some tools set this to 0 if multiple alignments are found for one read.	0
6	CIGAR	51M 	Detailed information about the alignment (see below).	*
7	RNEXT	=	PE reads: reference sequence name of the next read. Set to "=" if both mates are mapped to the same chromosome.	*
8	PNEXT	15535	PE reads: leftmost mapping position of the next read.	0
9	TLEN	232	PE reads: inferred template length (fragment size).	0
10	SEQ	CCA...GGC	The sequence of the aligned read on the forward strand (not including indels).	*
11	QUAL	BBH...1+B	Base quality (same as the quality string in the `FASTQ` format, but always in Sanger format [ASCII+33]).	*
12ff	OPT	NM:i:0	Optional fields (format: `<TAG>:<TYPE>:<VALUE>`; see below).	

<QNAME> <FLAG> <RNAME> <POS> <MAPQ> <CIGAR> <MRNM> <MPOS> <ISIZE> <SEQ> <QUAL>

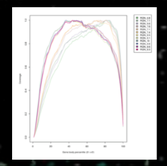
Pos.	Field	Example entry	Description	NA value
1	QNAME	Read1	Query template (= read) name (PE: read pair name)	required
2	FLAG	83	Information about the read's mapping properties encoded as bit-wise flags (see next section and Table 5).	required
3	RNAME	chrI	Reference sequence name. This should match a `@SQ` line in the header.	*
4	POS	15364	1-based leftmost mapping position of the first matching base. Set as 0 for an unmapped read without coordinates.	0
5	MAPQ	30	Mapping quality of the alignment. Should be a Phred-scaled posterior probability that the position of the read is incorrect, but the value is completely dependent on the alignment program. Some tools set this to 0 if multiple alignments are found for one read.	0
6	CIGAR	51M		
7	RNEXT	=		
8	PNEXT	1553[illegible]		
9	TLEN	232		
10	SEQ	CCA		
11	QUAL	BBH[illegible]	Base quality (same as the quality string in the FASTQ format, but always in Sanger format [ASCII+33]).	
12ff	OPT	NM:i:0	Optional fields (format: `<TAG>:<TYPE>:<VALUE>`; see below).	



<QNAME> <FLAG> <RNAME> <POS> <MAPQ> <CIGAR> <MRNM> <MPOS> <ISIZE> <SEQ> <QUAL>

Pos.	Field	Example entry	Description	NA value
1	QNAME	Read1	Query template (= read) name (PE: read pair name)	required
2	FLAG	83 	Information about the read's mapping properties encoded as bit-wise flags (see next section and Table 5).	required
3	RNAME	chrI	Reference sequence name. This should match a `@SQ` line in the header.	*
4	POS	15364	1-based leftmost mapping position of the first matching base. Set as 0 for an unmapped read without coordinates.	0
5	MAPQ	30	Mapping quality of the alignment. Should be a Phred-scaled posterior probability that the position of the read is incorrect, but the value is completely dependent on the alignment program. Some tools set this to 0 if multiple alignments are found for one read.	0
6	CIGAR	51M 	Detailed information about the alignment (see below).	*
7	RNEXT	=	PE reads: reference sequence name of the next read. Set to "=" if both mates are mapped to the same chromosome.	*
8	PNEXT	15535	PE reads: leftmost mapping position of the next read.	0
9	TLEN	232	PE reads: inferred template length (fragment size).	0
10	SEQ	CCA...GGC	The sequence of the aligned read on the forward strand (not including indels).	*
11	QUAL	BBH...1+B	Base quality (same as the quality string in the `FASTQ` format, but always in Sanger format [ASCII+33]).	*
12ff	OPT	NM:i:0	Optional fields (format: `<TAG>:<TYPE>:<VALUE>`; see below).	

STAR*/HISAT2/
BWA/Bowtie2



Could most reads be aligned?

Are there any obvious biases of the read distributions?

Are the replicate samples as similar to each other as expected?

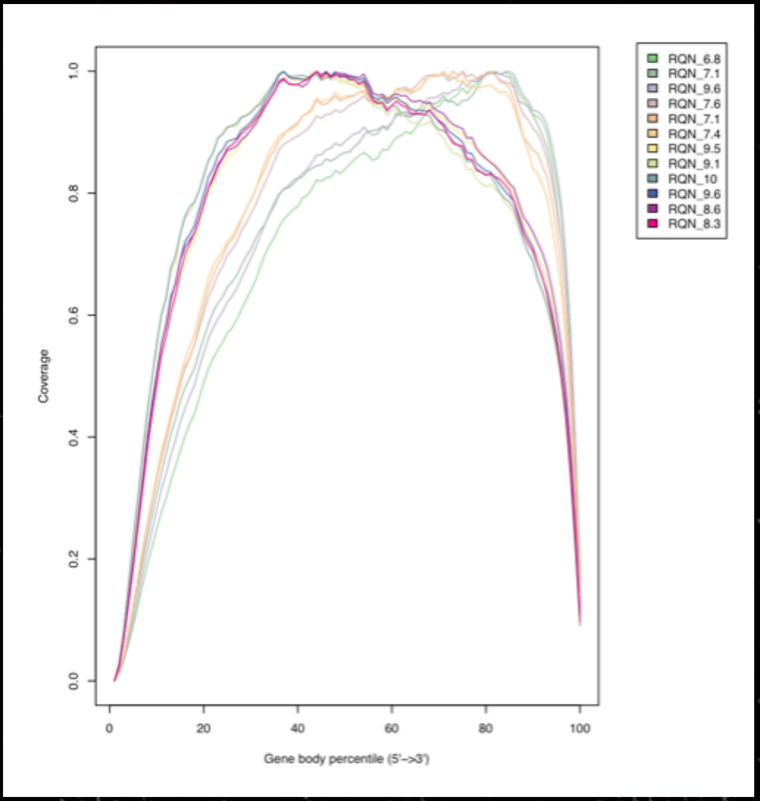
*--quantMode FOR RNA SEQ READS

STAR*/HISAT2/ BWA/Bowtie2

Could most reads be aligned?

Are there any obvious biases of the read distributions?

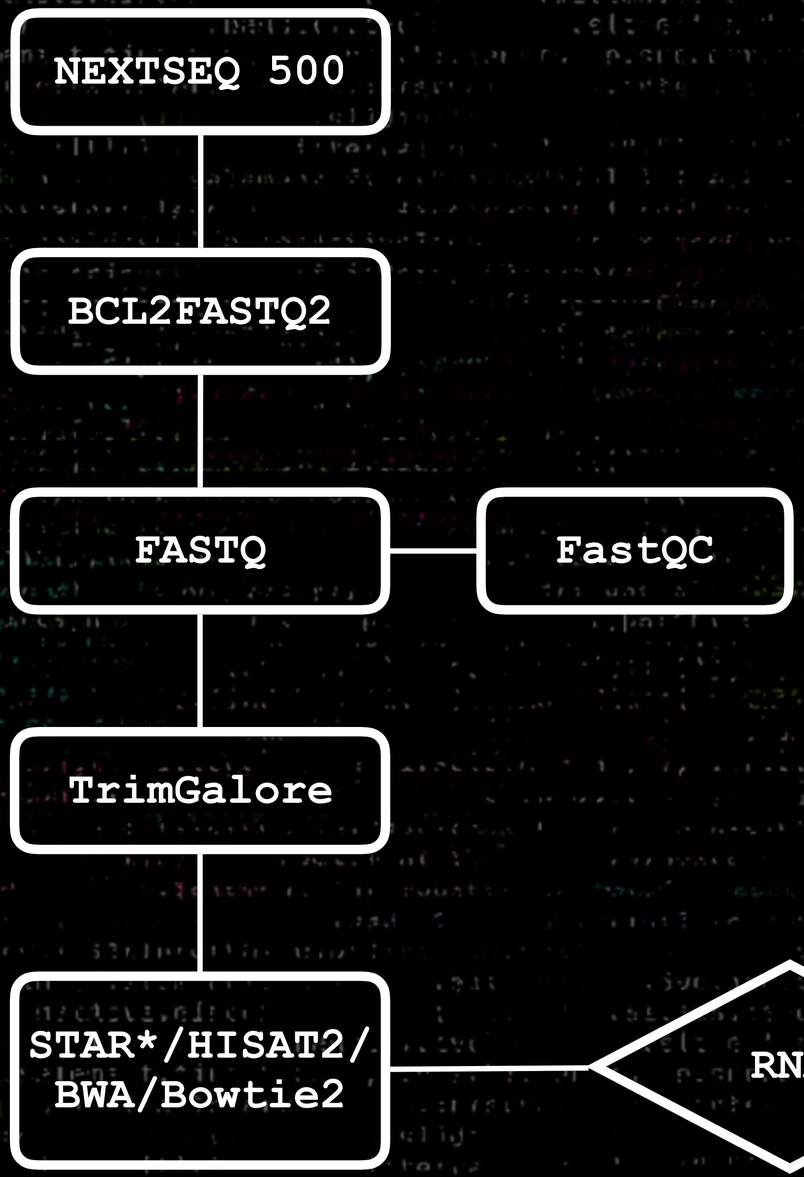
Are the replicate samples as similar to each other as expected?



Copy table Showing 12/12 rows and 2/2 columns.

Sample Name	% Aligned	M Aligned
RQN_10	88.7%	32.9
RQN_9.6_2	88.3%	29.6
RQN_8.3	87.7%	31.6
RQN_7.6	87.4%	32.1
RQN_8.6	87.3%	20.9
RQN_9.1	87.1%	28.8
RQN_9.6_1	85.0%	32.0
RQN_7.4	83.0%	29.9
RQN_9.5	80.2%	27.7
RQN_7.1_1	77.3%	26.6
RQN_6.8	70.7%	24.8
RQN_7.1_2	68.1%	23.9

*--quantMode FOR RNA SEQ READS



	wt1	wt2	wt3	wt4	wt5	ko1	ko2	ko3	ko4	ko5
gene1	135	148	146	121	140	269	268	227	263	259
gene2	803	797	841	800	874	412	408	388	393	398
gene3	40	25	38	41	35	413	393	417	374	415
gene4	381	383	415	374	354	809	840	859	856	845
gene5	775	766	773	749	784	302	310	324	342	314
gene6	305	313	256	313	315	831	817	832	859	869
gene7	816	819	800	793	790	485	481	429	461	508
gene8	40	22	40	37	32	421	476	479	528	483
gene9	963	935	938	953	948	43	26	41	28	39
gene10	697	749	715	724	715	233	259	284	277	269
gene11	36	50	40	35	44	168	178	168	170	187
gene12	60	66	54	61	71	288	289	293	289	330
gene13	537	517	523	512	515	142	134	145	145	145
gene14	655	615	610	664	606	842	889	827	885	838
gene15	426	439	436	420	432	131	155	159	139	151
gene16	952	976	974	987	947	789	828	825	850	796
gene17	379	446	410	423	394	963	1012	913	968	984
gene18	17	17	14	20	22	131	113	135	127	112
gene19	985	874	896	982	992	848	890	899	896	873
gene20	197	191	202	180	172	765	754	784	791	799
gene21	399	477	414	466	440	686	668	741	754	718

*--quantMode FOR RNA SEQ READS

DESeq2

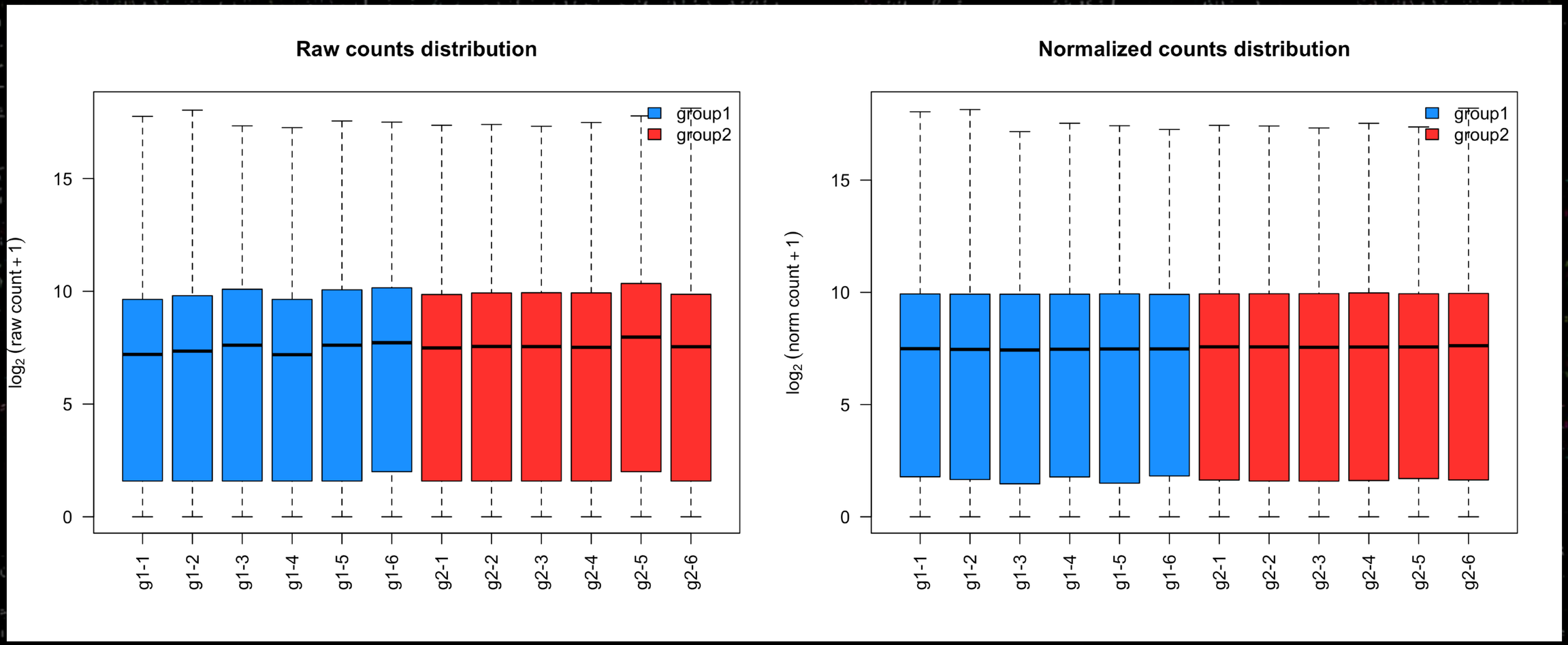
Number of reads mapped to a gene depends on:

- **Its own expression level;**
- **Its length;**
- **The sequencing depth;**
- **Expression of all other genes within the sample;**

Normalization is done to eliminate systematic effects;

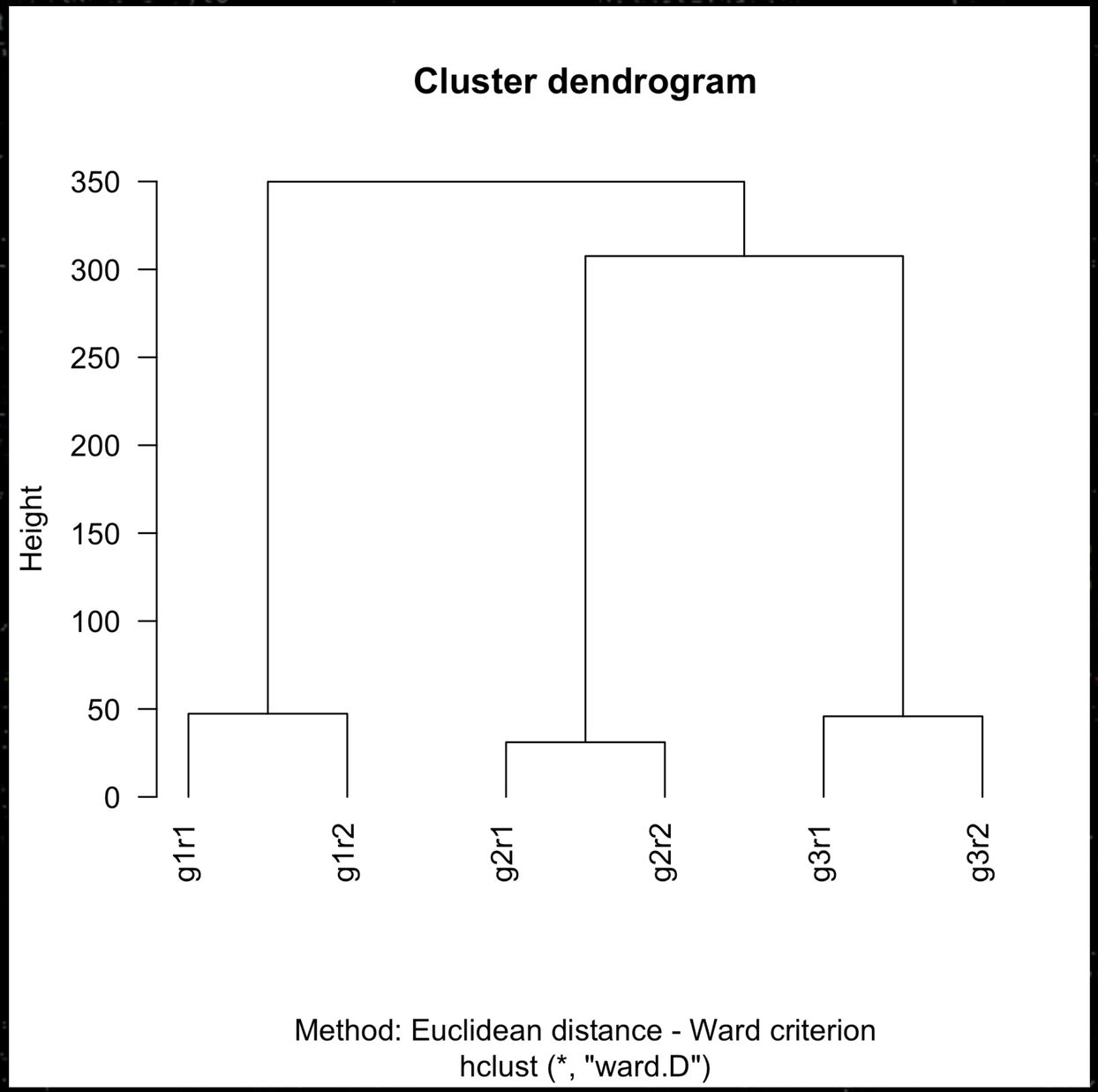
```
DESeq2::estimateSizeFactors( )
```

DESeq2


Raw counts distribution
log2 (raw count + 1)
group1
group2
0
5
10
15
g1-1
g1-2
g1-3
g1-4
g1-5
g1-6
g2-1
g2-2
g2-3
g2-4
g2-5
g2-6
Normalized counts distribution
log2 (norm count + 1)
group1
group2
0
5
10
15
g1-1
g1-2
g1-3
g1-4
g1-5
g1-6
g2-1
g2-2
g2-3
g2-4
g2-5
g2-6

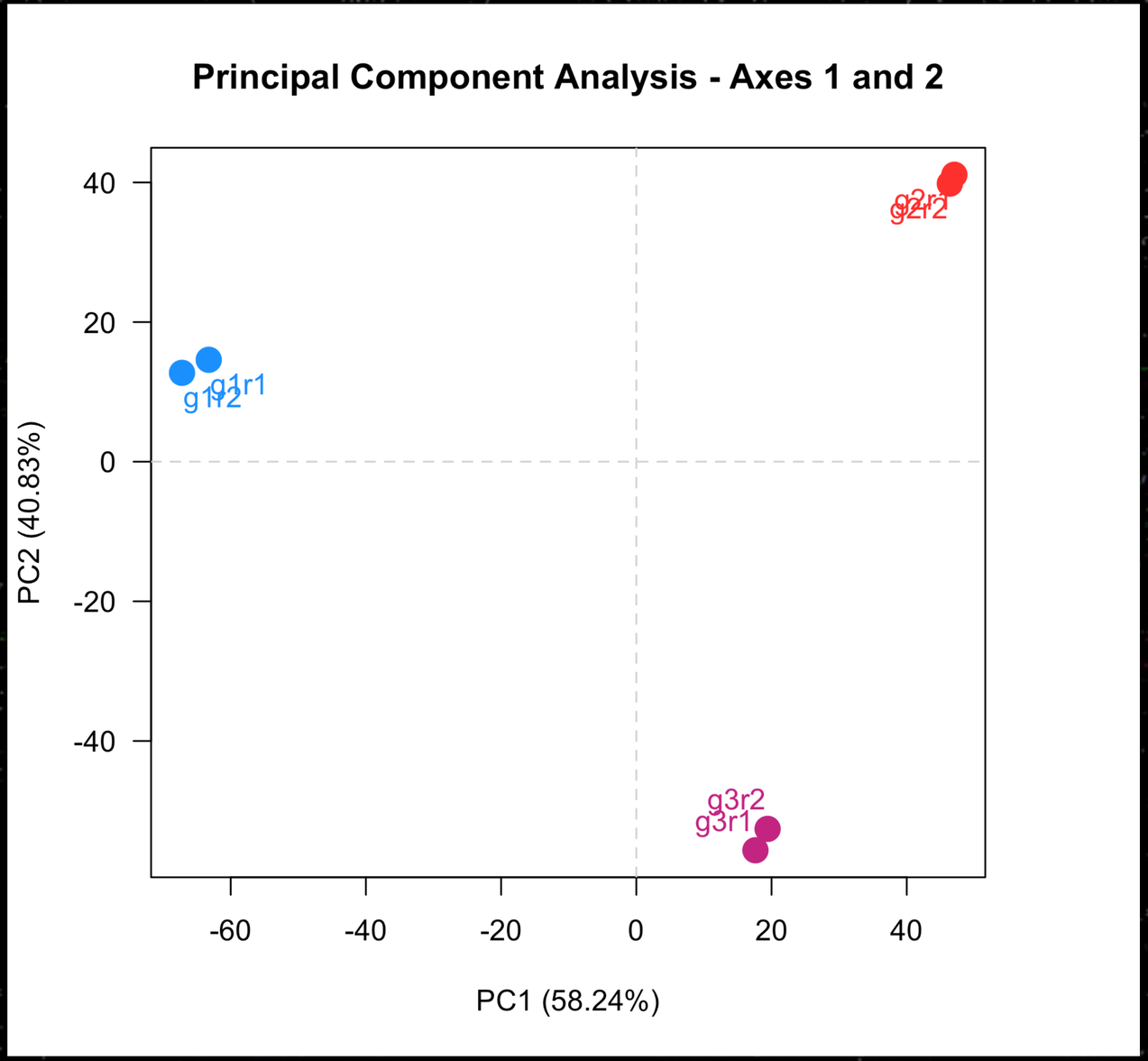
DESeq2

Exploration of Global normalized read count patterns:



DESeq2

Exploration of Global normalized read count patterns:



RNA SEQ DATA SETS

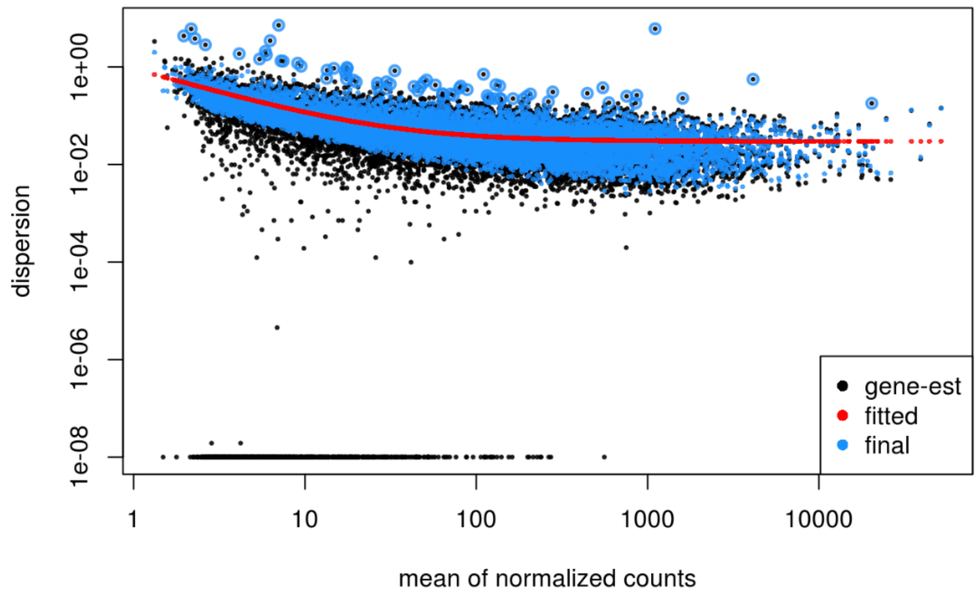
How do we best model this data 🤔 ???

Count based (Discrete) RNA-seq data, suffers from non-uniform mean-variance relationships

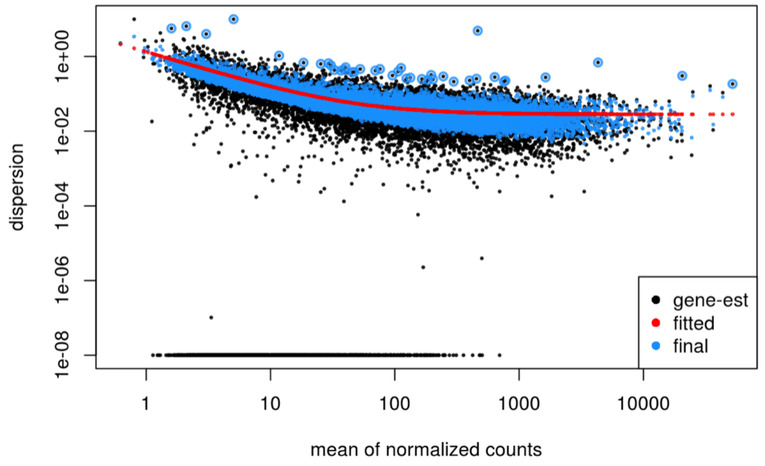
Heteroscedasticity;

RNA SEQ DATA SETS

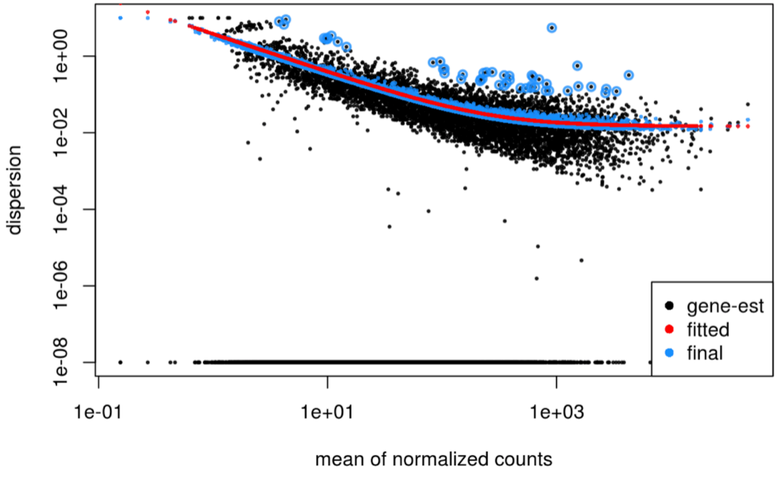
10



5



2



DESeq2

Once, mean-variance relationship is modeled, Wald test is used to report Differentially Expressed Genes!

```
## log2 fold change (MAP): condition treated vs untreated
## Wald test p-value: condition treated vs untreated
## DataFrame with 9921 rows and 5 columns
##                     baseMean       log2FoldChange             lfcSE
##                    <numeric>            <numeric>         <numeric>
## FBgn0000008 95.1442917575889  0.00119919675662286  0.151896997597845
## FBgn0000014 1.05652281859341 -0.00473412281044922  0.205467617376393
## FBgn0000017 4352.55356876647   -0.189899902298335  0.120376617165947
## FBgn0000018  418.61048415965  -0.06995753111158887  0.123900600388886
## FBgn0000024   6.406199980976   0.0175271520689073  0.198632752197541
## ...                      ...                  ...               ...
## FBgn0261570 3208.38861003698    0.241102900991117  0.124446879845224
## FBgn0261572 6.19718814545467  -0.0657617344183244   0.21413513713 68
## FBgn0261573 2240.97951122377   0.0100061908254208 0.0993764053703328
## FBgn0261574 4857.68037348332  0.00843552221427279  0.140826652378679
## FBgn0261575 10.6825203335563  0.00809100502438704  0.201470391594341
##                       pvalue              padj
##                    <numeric>         <numeric>
## FBgn0000008  0.991881656848254  0.99721076667093
## FBgn0000014  0.817298682951798                NA
## FBgn0000017 0.0575591059082212 0.288001711413016
## FBgn0000018  0.480855815353124 0.826833683766374
## FBgn0000024  0.759787936488384 0.943501114514859
## ...                      ...               ...
## FBgn0261570 0.0203070137750051 0.144240002513885
## FBgn0261572  0.216202637789157 0.607847805203262
## FBgn0261573  0.910614550167166 0.982656666760864
## FBgn0261574  0.936290772501261 0.988179230260622
## FBgn0261575   0.86052160317937  0.96792800379094
```

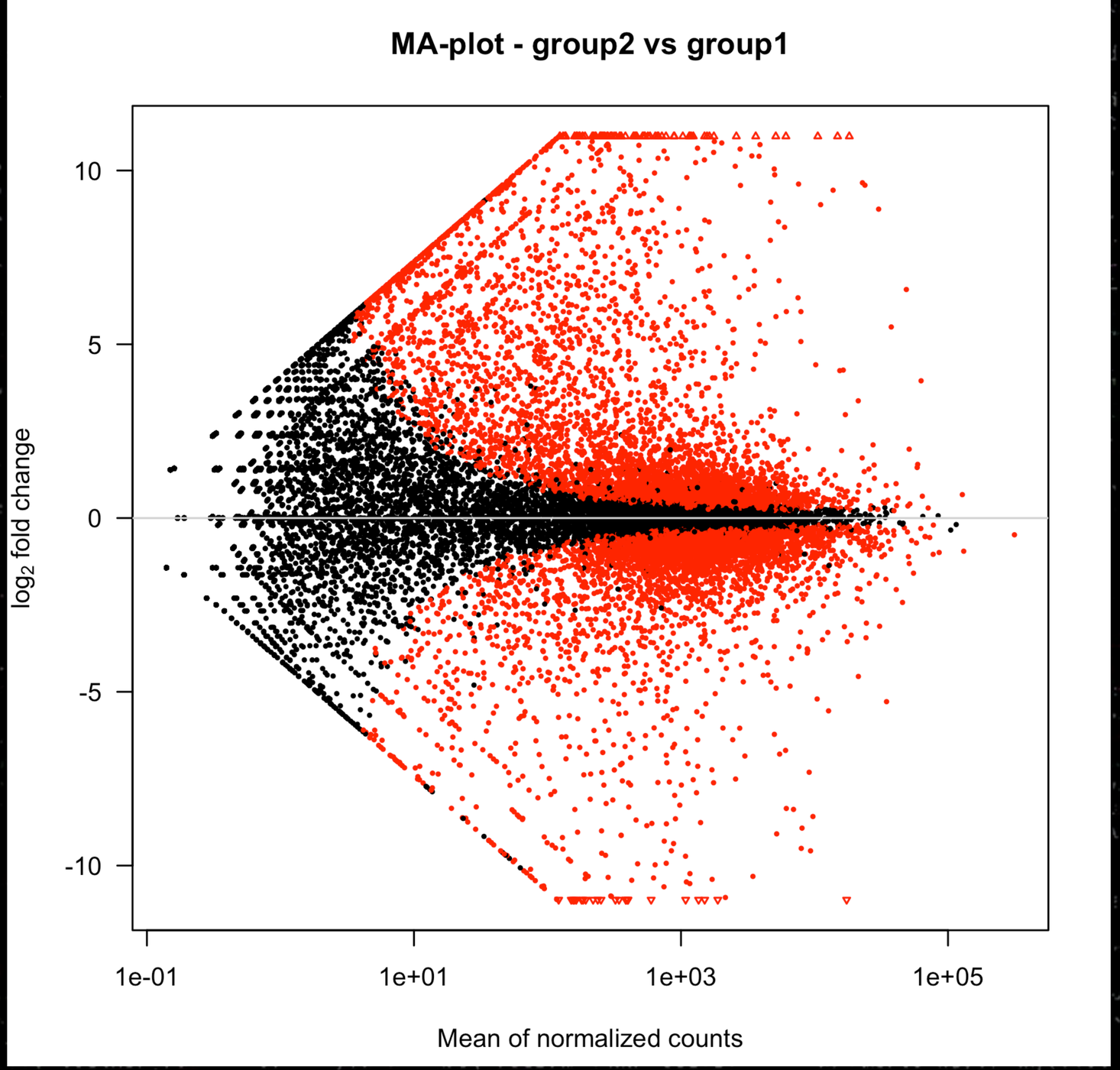
DESeq2

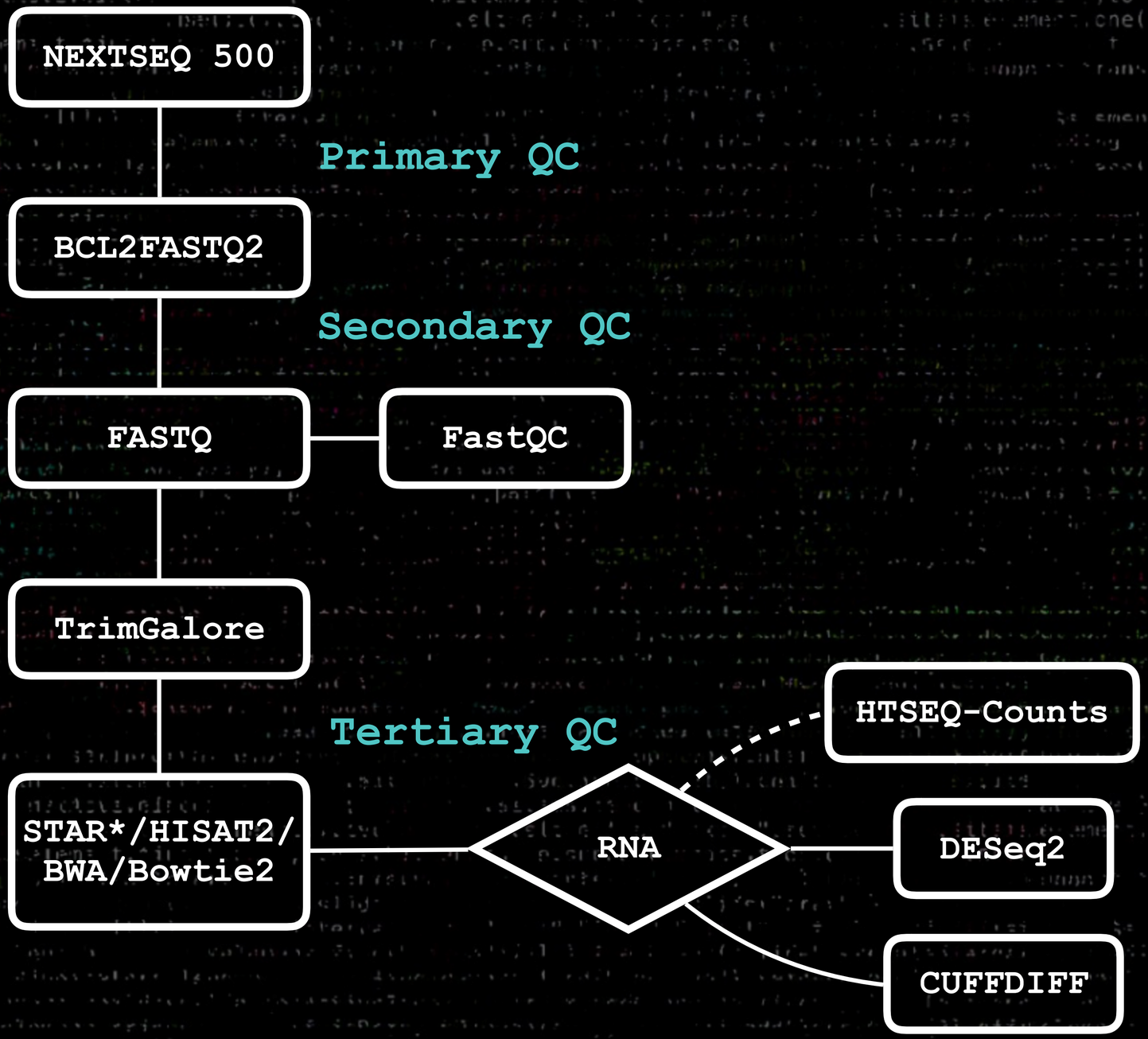
```
## log2 fold change (MAP): condition treated vs untreated
## Wald test p-value: condition treated vs untreated
## DataFrame with 9921 rows and 5 columns
##                    baseMean       log2FoldChange              lfcSE
##                   <numeric>            <numeric>          <numeric>
## FBgn0000008 95.1442917575889  0.00119919675662286  0.151896997597845
## FBgn0000014 1.05652281859341 -0.00473412281044922  0.205467617376393
## FBgn0000017 4352.55356876647   -0.189899902298335  0.120376617165947
## FBgn0000018  418.61048415965  -0.0699575311158887  0.123900600388886
## FBgn0000024   6.406199980976   0.0175271520689073  0.198632752197541
## ...                      ...                  ...                ...
## FBgn0261570 3208.38861003698    0.241102900991117  0.124446879845224
## FBgn0261572 6.19718814545467  -0.0657617344183244    0.2141351371368
## FBgn0261573 2240.97951122377   0.0100061908254208 0.0993764053703328
## FBgn0261574 4857.68037348332  0.00843552221427279  0.140826652378679
## FBgn0261575 10.6825203335563  0.00809100502438704  0.201470391594341
##                      pvalue              padj
##                   <numeric>         <numeric>
## FBgn0000008  0.991881656848254  0.99721076667093
## FBgn0000014  0.817298682951798                NA
## FBgn0000017 0.0575591059082212 0.288001711413016
## FBgn0000018  0.480855815353124 0.826833683766374
## FBgn0000024  0.759787936488384 0.943501114514859
## ...                      ...               ...
## FBgn0261570 0.0203070137750051 0.144240002513885
## FBgn0261572  0.216202637789157 0.607847805203262
## FBgn0261573  0.910614550167166 0.982656666760864
## FBgn0261574  0.936290772501261 0.988179230260622
## FBgn0261575   0.86052160317937  0.96792800379094
```

DESeq2

MA-plot

Differentially expressed features are highlighted in red





*--quantMode FOR RNA SEQ READS

THANK YOU FOR LISTENING!

Contact Info: http://rnaseqcore.vet.cornell.edu/

E-mail List Serve: TREX-GENEREG-L

Jen Grenier: jgrenier@cornell.edu

Faraz Ahmed: fahmed@cornell.edu

Christine Butler: cab18@cornell.edu

Ann Tate: aef93@cornell.edu

Scan me