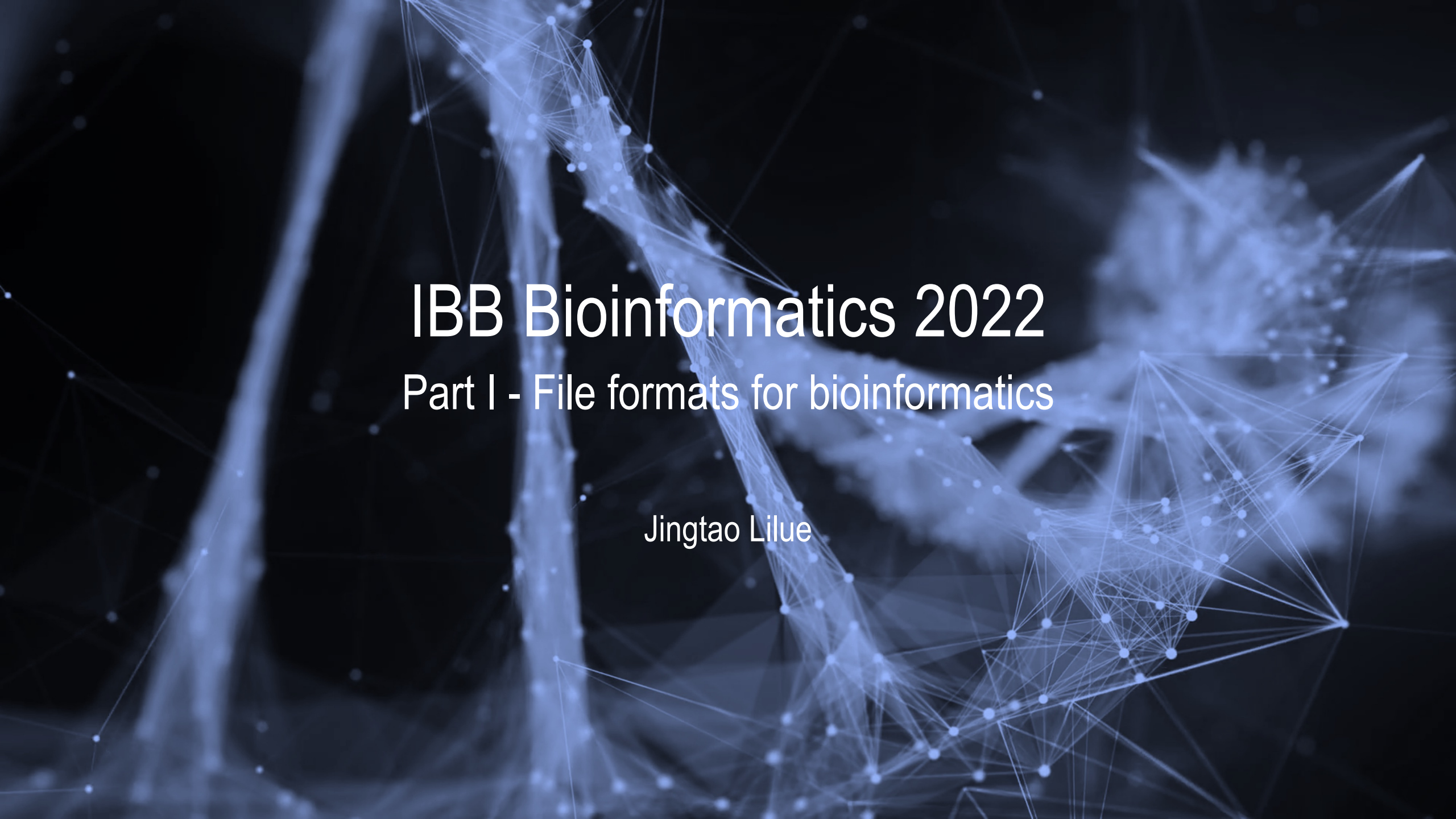


IBB Bioinformatics 2022

Part I - File formats for bioinformatics

Jingtao Lilue

Brief introduction

A brief CV:

2019 – now	Instituto Gulbenkian de Ciência	Head of Bioinformatics facility
2018 – 2019	EMBL-EBI, UK	Research Scientist
2015 – 2018	Wellcome Sanger Institute, UK	Senior Bioinformatician
2013 – 2014	Max-Planck Institute, Germany	Postdoc fellow
2007 – 2012	Universität zu Köln, Germany	Ph.D. Cell biology

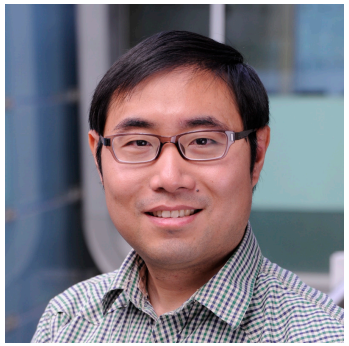
Main interest of research:

Population genetics
Host-pathogen co-evolution
Third generation sequencing and assembly algorithm
Mammalian immunity system

Contact:

jlilue@igc.gulbenkian.pt

Bioinformatics
IGC



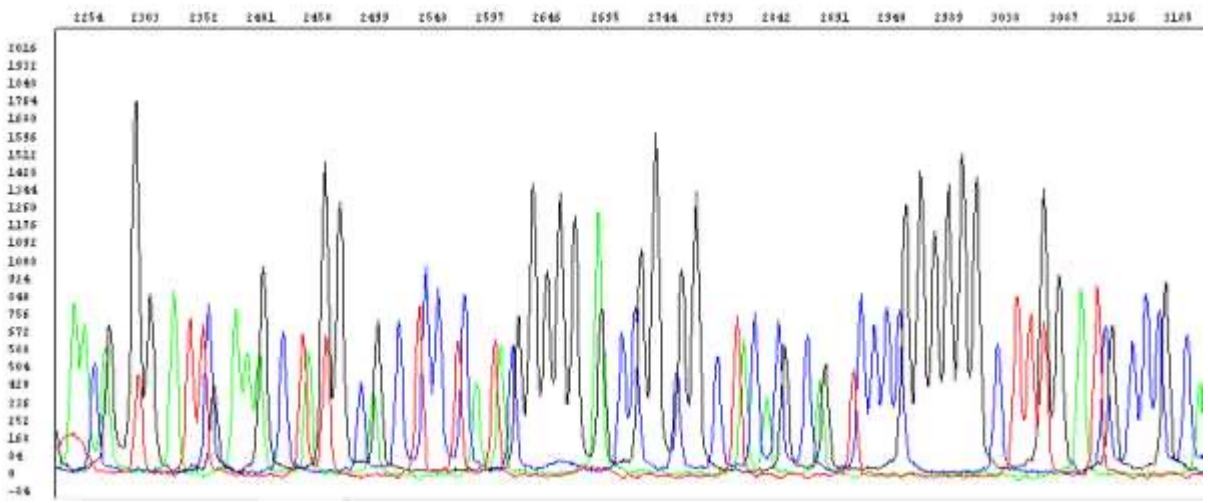
Before you start: what are you working with?


```
rs (reset to no color)
di (directory)
ln (symbolic link)
mh (multi-hardlink)
pi (named pipe, AKA FIFO)
so (socket) do (door)
bd (block device) cd (character device)
or (orphan symlink)

su (set-user-ID)
sg (set-group-ID)
ca (file with capability)
tw (sticky and other-writable directory)
ow (other-writable directory)
st (sticky directory)
ex (executable file)
*.tar *.tgz *.arc *.arj *.taz *.lha *.lz4 *.lzh *.lzma *.tlz *.txz *.tzo *.t7z *
.zip *.z *.Z *.dz *.gz *.lrz *.lz *.lzo *.xz *.bz2 *.bz *.tbz *.tbz2 *.tz *.deb
*.rpm *.jar *.war *.ear *.sar *.rar *.alz *.ace *.zoo *.cpio *.7z *.rz *.cab
*.jpg *.jpeg *.gif *.bmp *.pbm *.pgm *.ppm *.tga *.xbm *.xpm *.tif *.tiff *.png
*.svg *.svgz *.mng *.pcx *.mov *.mpg *.mpeg *.m2v *.mkv *.webm *.ogm *.mp4 *.m4v
 *.mp4v *.vob *.qt *.nuv *.wmv *.asf *.rm *.rmvb *.flc *.avi *.fli *.flv *.gl *.
dl *.xcf *.xwd *.yuv *.cgm *.emf *.axv *.anx *.ogv *.ogx
*.aac *.au *.flac *.mid *.midi *.mka *.mp3 *.mpc *.ogg *.ra *.wav *.axa *.oga *.
spx *.xspf
```

From your sequencing provider…

Sanger sequencing: *.fa *.ab1 *.seq

>Sample1
tgcaccaaacatgtctaaagctggaaccaaaattactttctttgaagacaaaaactttca
aggccgccactatgacagcgattgcgactgtgcagatttccacatgtacctgagccgctg
caactccatcagagtggaaggaggcacctgggctgtgtatgaaaggcccaattttgctgg
gtacatgtacatcctaccccggggcgagtatcctgagtaccagcactggatgggcctcaa
cgaccgcctcagctcctgcagggctgttcacctgtctagtggaggccagtataagcttca
gatctttgagaaaggggatttaatggtcagatgcatgagaccacggaagactgcccttc
catcatggagcagttccacatgcgggaggtccactcctgtaaggtgctggagggcgcctg
gatcttctatgagctgcccaactaccgaggcaggcagtacctgctggacaagaaggagta
ccggaagcccgtcgactggggtgcagcttccccagctgtccagtctttccgccgcattgt
ggagtgatgatacagatgcggccaaacgctggctggccttgtcatccaaataagcattat
aaataaaacaattggcatgc



From your sequencing provider…

Illumina: *.fq (*.fastq *.fastq.gz)

Data_S1_L001_R1.fq

```
@IL16_4408:3:5:17860:13258
CTGGCTCACATACAGGCCAGTATAAAGCGTCTCCTTTAAA
+
HHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHE
@IL16_4408:3:14:13276:1210
GTAGAAAACAATATAGAAGCCTTAGGACAGAAAACCCTAA
+
HHHHHHHHHHHHHHHHHDHHHHHHHHBGHHHHHHCDCCF>
@IL16_4408:3:83:5210:21157
AGGCAGCAGGAACTCCAGTTAGGCCATAGTTCATCCTTGC
+
FFFFFFFFFFFFFFFFFFDFFFFFFFFFFFFFFFDCF
```

Data_S1_L001_R2.fq

```
@IL16_4408:3:5:17860:13258
CAGTTTTTTCAGAGCAGTAGCCATTAGGCACAATGTGATT
+
FFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFAFFBFFFF
@IL16_4408:3:14:13276:1210
AGTCAACAGATGTCCTTTGAGCTTAAGAATTCAGCAGAAG
+
FFFFFFFFFFFFFFFFFFFDFFFFFFFFFFFFFFFDCF
@IL16_4408:3:83:5210:21157
GGGCCAGTGTGTCCCCTCCTGCCACTGAAGACCATGCTAT
+
HHHHHHHHHHHHHHHHHDHHHHFFFFFFAFFBFFFFGGGG
```

From your sequencing provider…

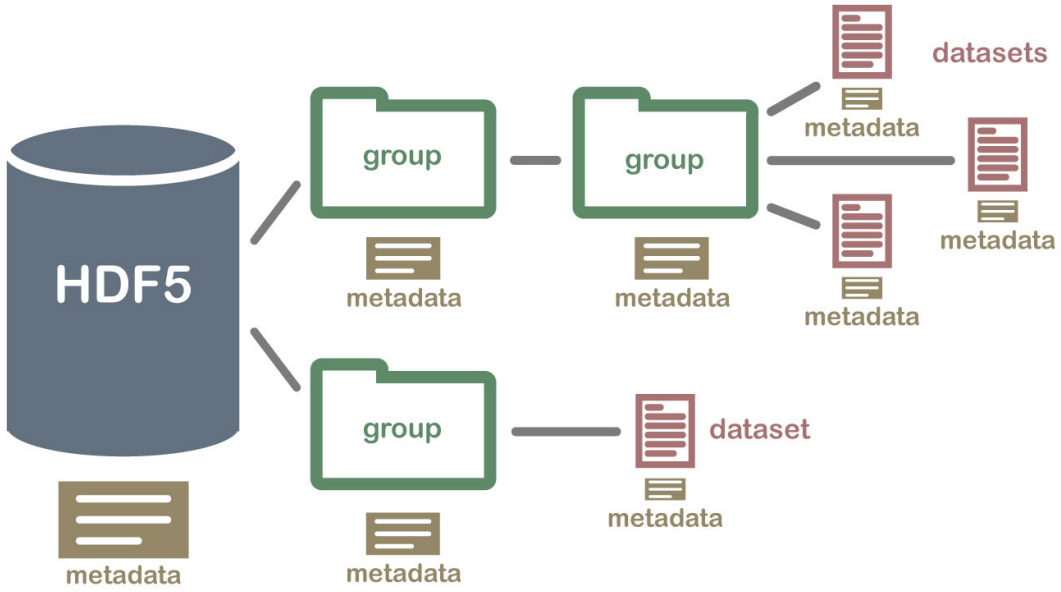
Quality scores of fastq format

```
!"#$%&'()*+´-./0123456789:;<=>?@ABCDEFGHI
0----------------------------------------40
```

Phred Quality Score	Probability of incorrect base call	Base call accuracy
10	1 in 10	90%
20	1 in 100	99%
30	1 in 1000	99.9%
40	1 in 10000	99.99%
50	1 in 100000	99.999%

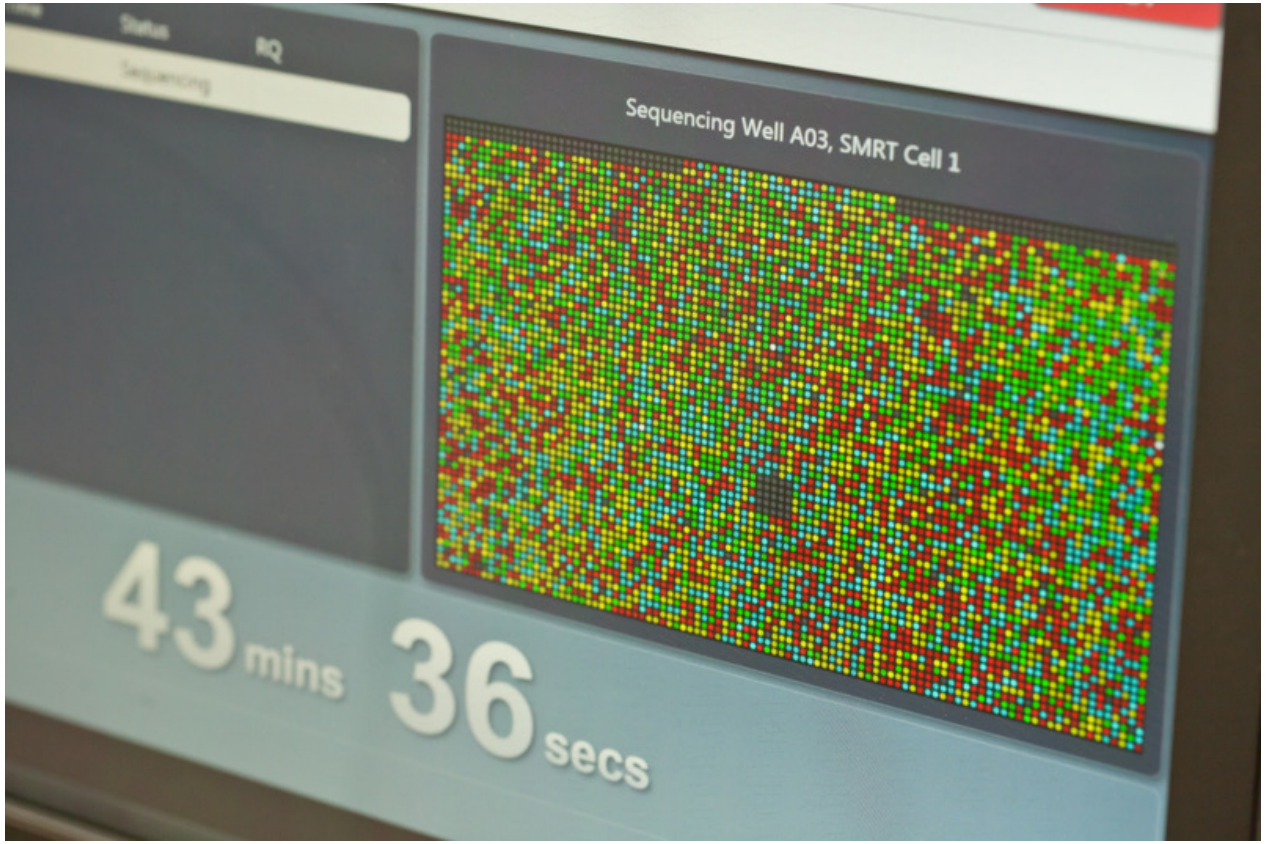
From your sequencing provider…

ONT: *.fq *.hdf5 *.fast5



From your sequencing provider…

PacBio: *.h5 *.fq (*.bam)



From public data base….

*.sra Sequence Read Archive format

↳ *.bam, *.fq, *.ab1 etc.

SAM/BAM and CRAM files - head

File format	File size (GB)
SAM	7.4
BAM	1.9
CRAM lossless	1.4
CRAM 8 bins	0.8
CRAM no quality	0.26

1000 Genomes BAM file head

```
@HD     VN:1.0  GO:none SO:coordinate
@SQ     SN:1    LN:249250621    AS:NCBI37       UR:file:/lustre/scratch102/projects/g1k/ref/main_project/human_g1k_v37.fasta    M5:1b22b98cdeb4a9304cb5d48026a85128
@SQ     SN:2    LN:243199373    AS:NCBI37       UR:file:/lustre/scratch102/projects/g1k/ref/main_project/human_g1k_v37.fasta    M5:a0d9851da00400dec1098a9255ac712e
@SQ     SN:3    LN:198022430    AS:NCBI37       UR:file:/lustre/scratch102/projects/g1k/ref/main_project/human_g1k_v37.fasta    M5:fdfd811849cc2fadebc929bb925902e5
@SQ     SN:4    LN:191154276    AS:NCBI37       UR:file:/lustre/scratch102/projects/g1k/ref/main_project/human_g1k_v37.fasta    M5:23dccd106897542ad87d2765d28a19a1
@SQ     SN:5    LN:180915260    AS:NCBI37       UR:file:/lustre/scratch102/projects/g1k/ref/main_project/human_g1k_v37.fasta    M5:0740173db9ffd264d728f32784845cd7
@SQ     SN:6    LN:171115067    AS:NCBI37       UR:file:/lustre/scratch102/projects/g1k/ref/main_project/human_g1k_v37.fasta    M5:1d3a93a248d92a729ee764823acbbc6b
@SQ     SN:7    LN:159138663    AS:NCBI37       UR:file:/lustre/scratch102/projects/g1k/ref/main_project/human_g1k_v37.fasta    M5:618366e953d6aaad97dbe4777c29375e
@SQ     SN:8    LN:146364022    AS:NCBI37       UR:file:/lustre/scratch102/projects/g1k/ref/main_project/human_g1k_v37.fasta    M5:96f514a9929e410c6651697bded59aec
@SQ     SN:9    LN:141213431    AS:NCBI37       UR:file:/lustre/scratch102/projects/g1k/ref/main_project/human_g1k_v37.fasta    M5:3e273117f15e0a400f01055d9f393768
@SQ     SN:10   LN:135534747    AS:NCBI37       UR:file:/lustre/scratch102/projects/g1k/ref/main_project/human_g1k_v37.fasta    M5:988c28e000e84c26d552359af1ea2e1d
@SQ     SN:11   LN:135006516    AS:NCBI37       UR:file:/lustre/scratch102/projects/g1k/ref/main_project/human_g1k_v37.fasta    M5:98c59049a2df285c76ffb1c6db8f8b96
@SQ     SN:12   LN:133851895    AS:NCBI37       UR:file:/lustre/scratch102/projects/g1k/ref/main_project/human_g1k_v37.fasta    M5:51851ac0e1a115847ad36449b0015864
@SQ     SN:13   LN:115169878    AS:NCBI37       UR:file:/lustre/scratch102/projects/g1k/ref/main_project/human_g1k_v37.fasta    M5:283f8d7892baa81b510a015719ca7b0b
@SQ     SN:14   LN:107349540    AS:NCBI37       UR:file:/lustre/scratch102/projects/g1k/ref/main_project/human_g1k_v37.fasta    M5:98f3cae32b2a2e9524bc19813927542e
@SQ     SN:15   LN:102531392    AS:NCBI37       UR:file:/lustre/scratch102/projects/g1k/ref/main_project/human_g1k_v37.fasta    M5:e5645a794a8238215b2cd77acb95a078
@SQ     SN:16   LN:90354753     AS:NCBI37       UR:file:/lustre/scratch102/projects/g1k/ref/main_project/human_g1k_v37.fasta    M5:fc9b1a7b42b97a864f56b348b06095e6
@SQ     SN:17   LN:81195210     AS:NCBI37       UR:file:/lustre/scratch102/projects/g1k/ref/main_project/human_g1k_v37.fasta    M5:351f64d4f4f9ddd45b35336ad97aa6de
@SQ     SN:18   LN:78077248     AS:NCBI37       UR:file:/lustre/scratch102/projects/g1k/ref/main_project/human_g1k_v37.fasta    M5:b15d4b2d29dde9d3e4f93d1d0f2cbc9c
@SQ     SN:19   LN:59128983     AS:NCBI37       UR:file:/lustre/scratch102/projects/g1k/ref/main_project/human_g1k_v37.fasta    M5:1aacd71f30db8e561810913e0b72636d
@SQ     SN:20   LN:63025520     AS:NCBI37       UR:file:/lustre/scratch102/projects/g1k/ref/main_project/human_g1k_v37.fasta    M5:0dec9660ec1efaaf33281c0d5ea2560f
@SQ     SN:21   LN:48129895     AS:NCBI37       UR:file:/lustre/scratch102/projects/g1k/ref/main_project/human_g1k_v37.fasta    M5:2979a6085bfe28e3ad6f552f361ed74d
@SQ     SN:22   LN:51304566     AS:NCBI37       UR:file:/lustre/scratch102/projects/g1k/ref/main_project/human_g1k_v37.fasta    M5:a718acaa6135fdca8357d5bfe94211dd
@SQ     SN:X    LN:155270560    AS:NCBI37       UR:file:/lustre/scratch102/projects/g1k/ref/main_project/human_g1k_v37.fasta    M5:7e0e2e580297b7764e31dbc80c2540dd
@SQ     SN:Y    LN:59373566     AS:NCBI37       UR:file:/lustre/scratch102/projects/g1k/ref/main_project/human_g1k_v37.fasta    M5:1fa3474750af0948bdf97d5a0ee52e51
@SQ     SN:MT   LN:16569        AS:NCBI37       UR:file:/lustre/scratch102/projects/g1k/ref/main_project/human_g1k_v37.fasta    M5:c68f52674c9fb33aef52dcf399755519
@SQ     SN:GL000207.1   LN:4262 AS:NCBI37       UR:file:/lustre/scratch102/projects/g1k/ref/main_project/human_g1k_v37.fasta    M5:f3814841f1939d3ca19072d9e89f3fd7
@RG     ID:ERR001268    PL:ILLUMINA     LB:NA12878.1    PI:200  DS:SRP000032    SM:NA12878      CN:MPIMG
@RG     ID:ERR001269    PL:ILLUMINA     LB:NA12878.1    PI:200  DS:SRP000032    SM:NA12878      CN:MPIMG
@RG     ID:ERR001698    PL:ILLUMINA     LB:g1k-sc-NA12878-CEU-1 PI:200  DS:SRP000032    SM:NA12878      CN:SC
@RG     ID:SRR001114    PL:ILLUMINA     LB:Solexa-3620  PI:0    DS:SRP000032    SM:NA12878      CN:BI
@RG     ID:SRR001115    PL:ILLUMINA     LB:Solexa-3623  PI:0    DS:SRP000032    SM:NA12878      CN:BI
@PG     ID:GATK TableRecalibration.4    VN:v2.2.16      CL:Covariates=[ReadGroupCovariate, QualityScoreCovariate, DinucCovariate, CycleCovariate], use_original_quals=true,
default_read_group=DefaultReadGroup, default_platform=ILLUMINA, 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, pQ=5, maxQ=40, smoothing=1
@PG     ID:bwa  VN:0.5.5
```

SAM/BAM and CRAM files – data lines

readID43G:1:1202:19894/1 256 contig43 613960 1 65M * 0 0 CCAGCGCGAACGAAATCCGCATGCGTCTGGTCGTTGCACGGAACGGCGGCGGTGTGATGCACGGC
EDDEEDEE=EE?DE??DDDBADEBEFFFDBEFFEBCBC=?BEEEE@=:?::?7?:8-6?7?@??# AS:i:0 XS:i:0 XN:i:0 XM:i:0 XO:i:0 XG:i:0 NM:i:0 MD:Z:65 YT:Z:UU

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	`\*|[:rname:^*=][:rname:]*`	Reference sequence NAME[10]
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	`\*|=|[:rname:^*=][:rname:]*`	Reference 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

SAM/BAM and CRAM files – flag

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

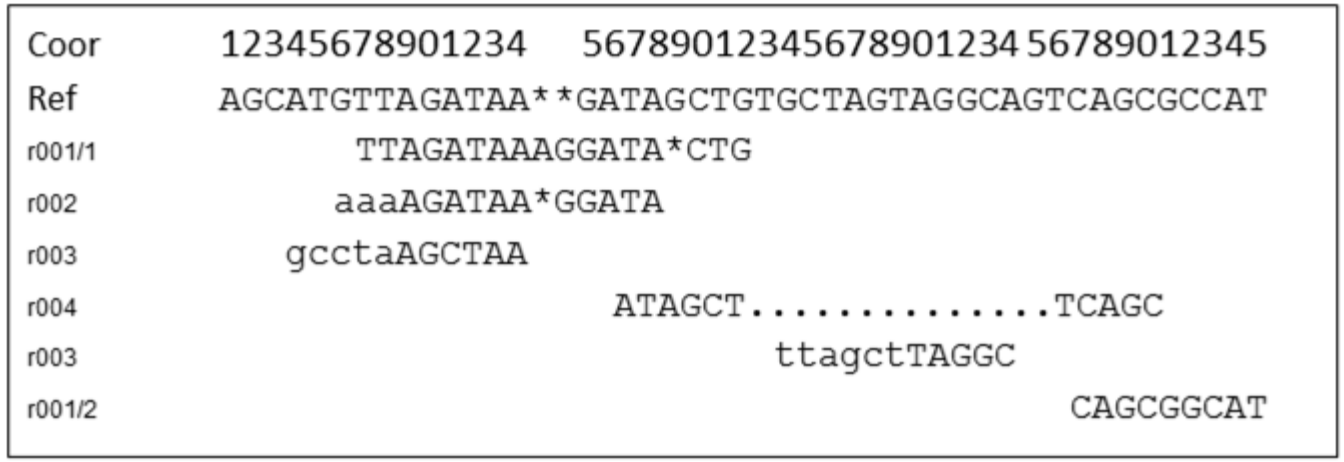
https://www.samformat.info/sam-format-flag

73 ⇄ **Paired: 133** Reset

#	Decimal	Description of first read
1	1	Read paired
2	2	Read mapped in proper pair
3	4	Read unmapped
4	8	Mate unmapped
5	16	Read reverse strand
6	32	Mate reverse strand
7	64	First in pair
8	128	Second in pair
9	256	Not primary alignment
10	512	Read fails platform/vendor quality checks
11	1024	Read is PCR or optical duplicate
12	2048	Supplementary alignment
Sum	73	

Decimal	Description of second read
1	Read paired
2	Read mapped in proper pair
4	Read unmapped
8	Mate unmapped
16	Read reverse strand
32	Mate reverse strand
64	First in pair
128	Second in pair
256	Not primary alignment
512	Read fails platform/vendor quality checks
1024	Read is PCR or optical duplicate
2048	Supplementary alignment
133	

SAM/BAM and CRAM files – cigar

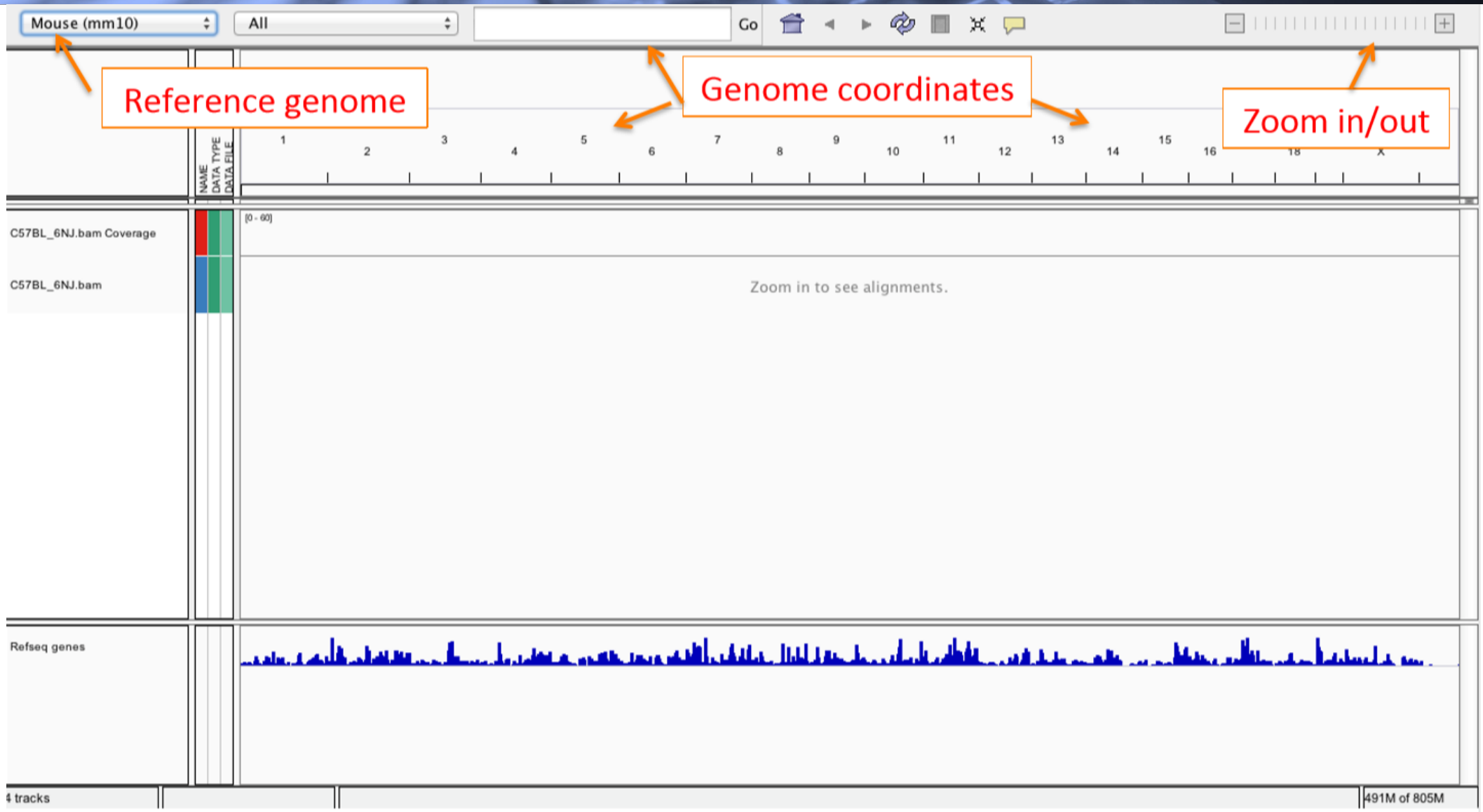


```
Coor      12345678901234   5678901234567890123456789012345
Ref       AGCATGTTAGATAA**GATAGCTGTGCTAGTAGGCAGTCAGCGCCAT
r001/1          TTAGATAAAGGATA*CTG
r002          aaaAGATAA*GGATA
r003        gcctaAGCTAA
r004                        ATAGCT..............TCAGC
r003                              ttagctTAGGC
r001/2                                          CAGCGGCAT
```

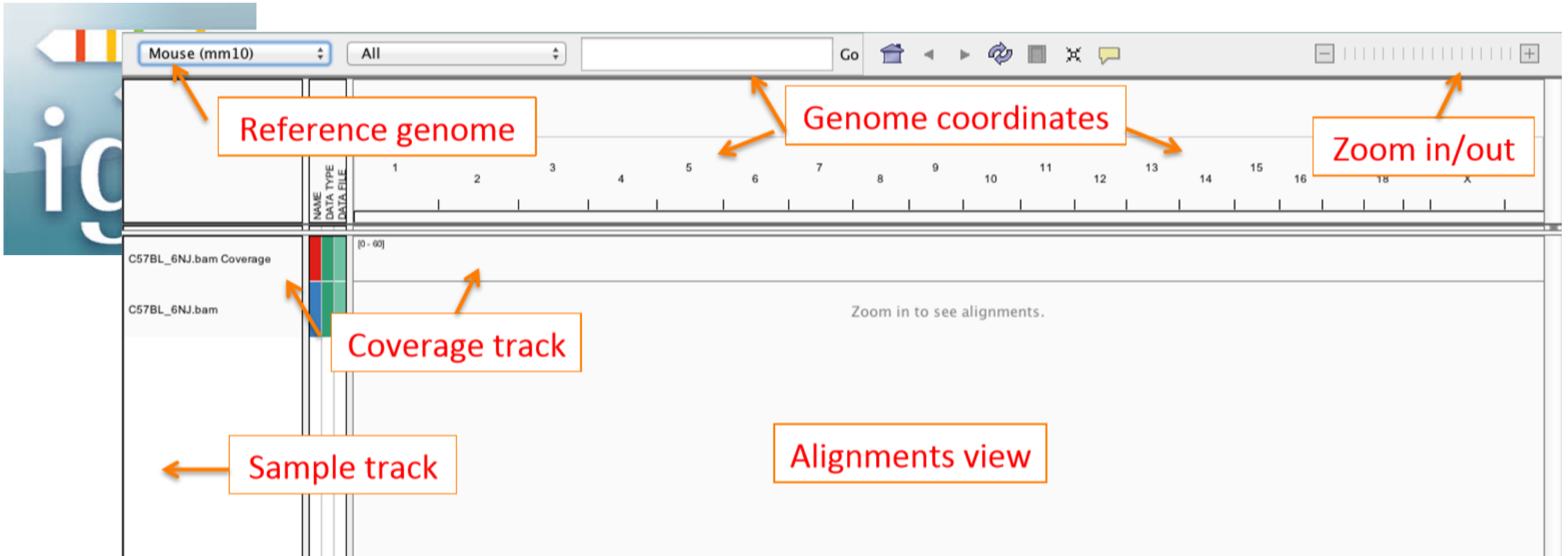
```
@SQ SN:ref LN:45
r001  99    ref  7   30  8M2I4M1D3M  =  37  39   TTAGATAAAGGATACTG  *
r002  0     ref  9   30  3S6M1P1I4M  *  0   0    AAAAGATAAGGATA     *
r003  0     ref  9   30  5S6M        *  0   0    GCCTAAGCTAA        *
r004  0     ref  16  30  6M14N5M     *  0   0    ATAGCTTCAGC        *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

D	Deletion
I	Insertion
H	Hard clipping
S	Soft clipping
M	Match
N	Skipped region
P	Padding, padded area in the read

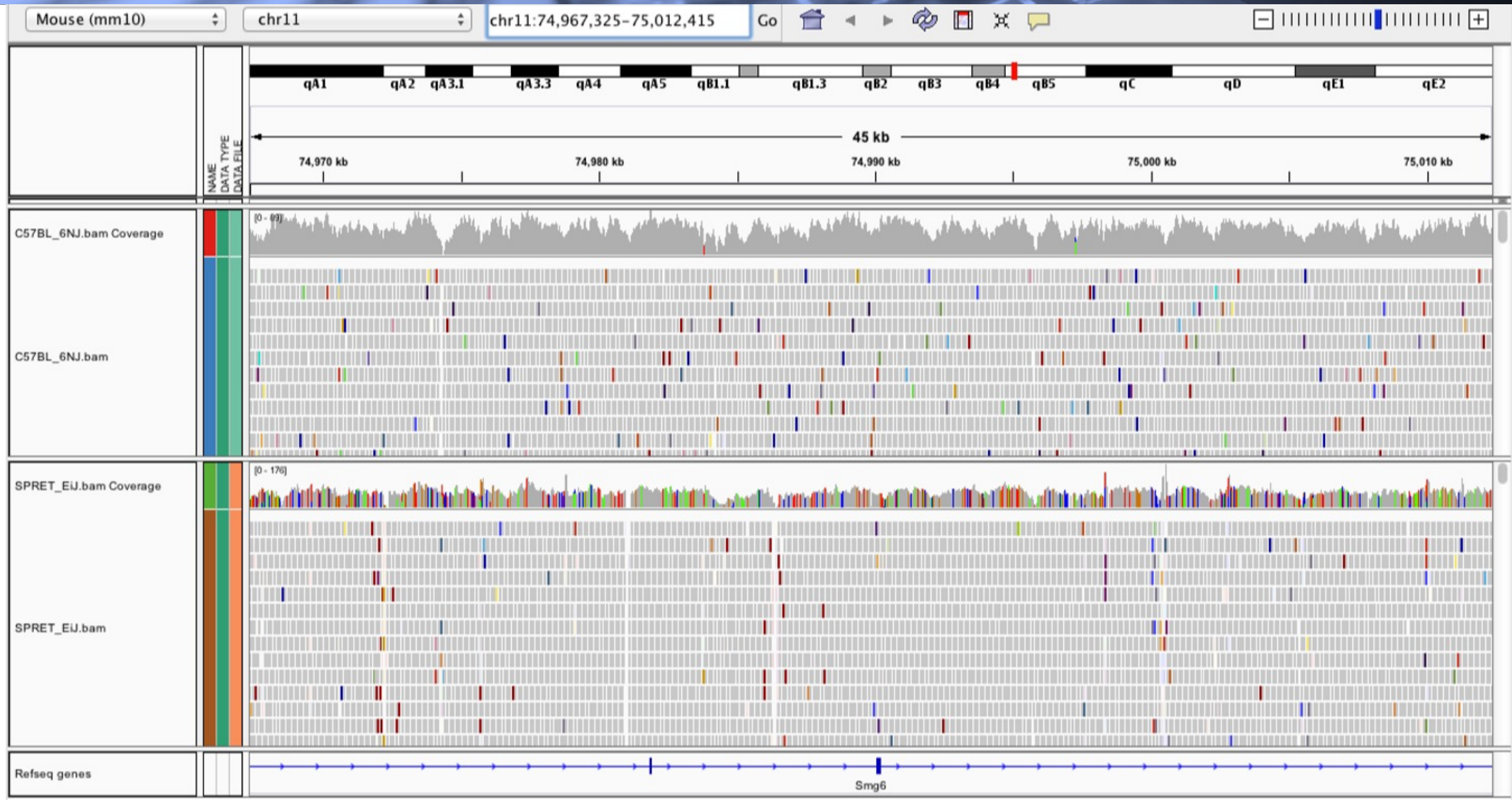
SAM/BAM viewer - IGV



SAM/BAM viewer - IGV



SAM/BAM viewer - IGV

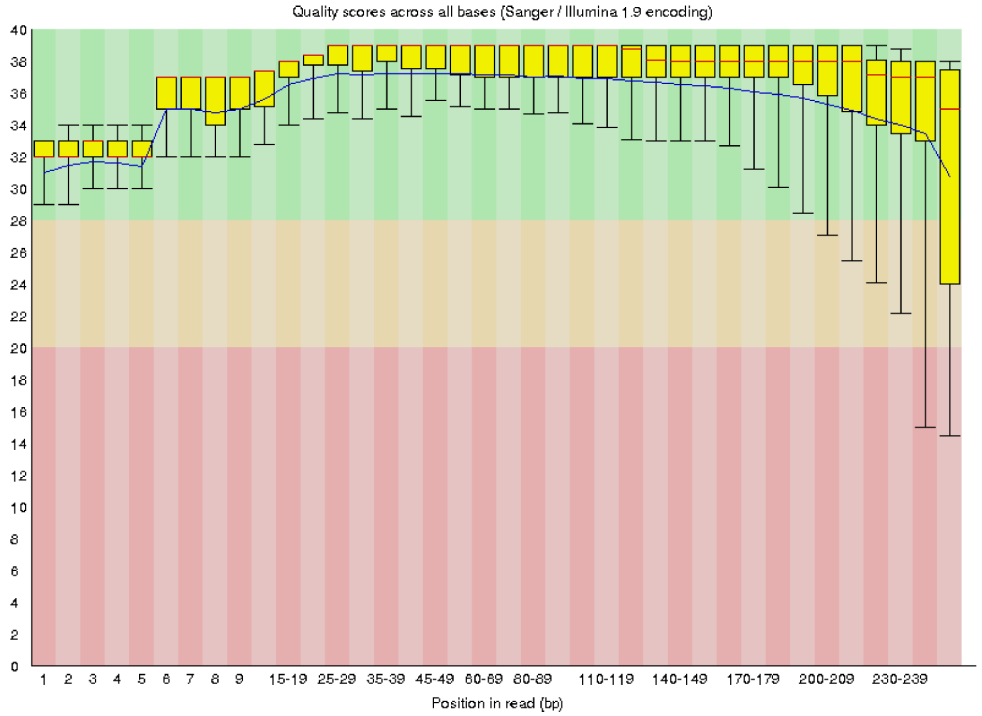


SAM/BAM viewer - IGV

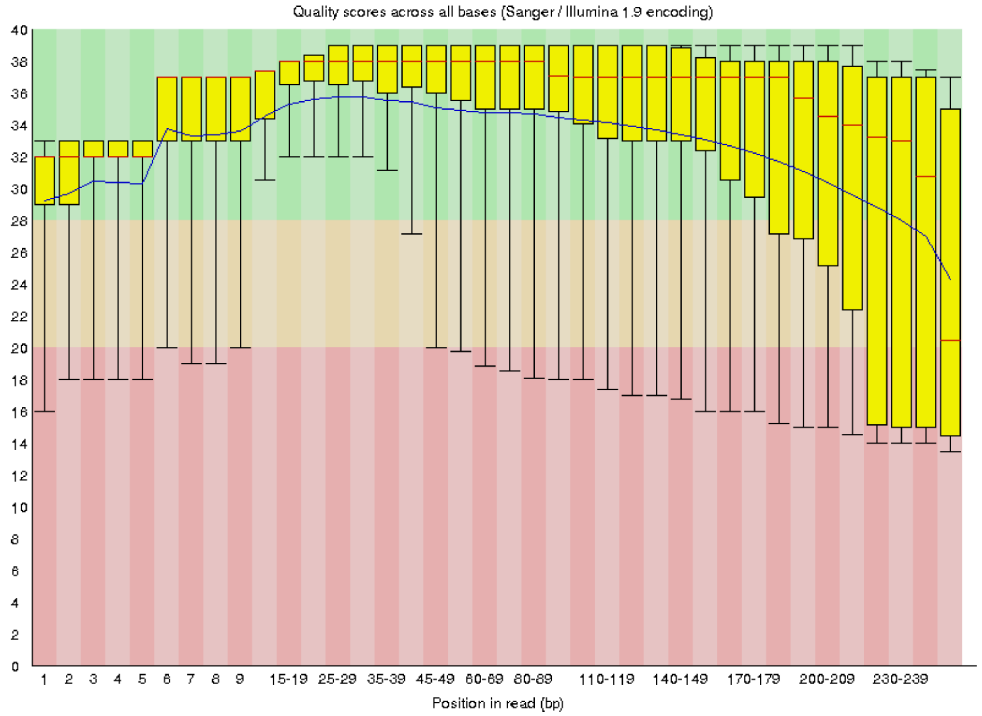
SAM files utility – SAMtools

```
Program: samtools (Tools for alignments in the SAM format)
Version: 1.9 (using htslib 1.9)

Usage:   samtools <command> [options]

Commands:
  -- Indexing
     dict           create a sequence dictionary file
     faidx          index/extract FASTA
     fqidx          index/extract FASTQ
     index          index alignment

  -- Editing
     calmd          recalculate MD/NM tags and '=' bases
     fixmate        fix mate information
     reheader       replace BAM header
     targetcut      cut fosmid regions (for fosmid pool only)
     addreplacerg   adds or replaces RG tags
     markdup        mark duplicates

  -- File operations
     collate        shuffle and group alignments by name
     cat            concatenate BAMs
     merge          merge sorted alignments
     mpileup        multi-way pileup
     sort           sort alignment file
     split          splits a file by read group
     quickcheck     quickly check if SAM/BAM/CRAM file appears intact
     fastq          converts a BAM to a FASTQ
     fasta          converts a BAM to a FASTA

  -- Statistics
     bedcov         read depth per BED region
     depth          compute the depth
     flagstat       simple stats
     idxstats       BAM index stats
```

QC for fastq files with fastqc

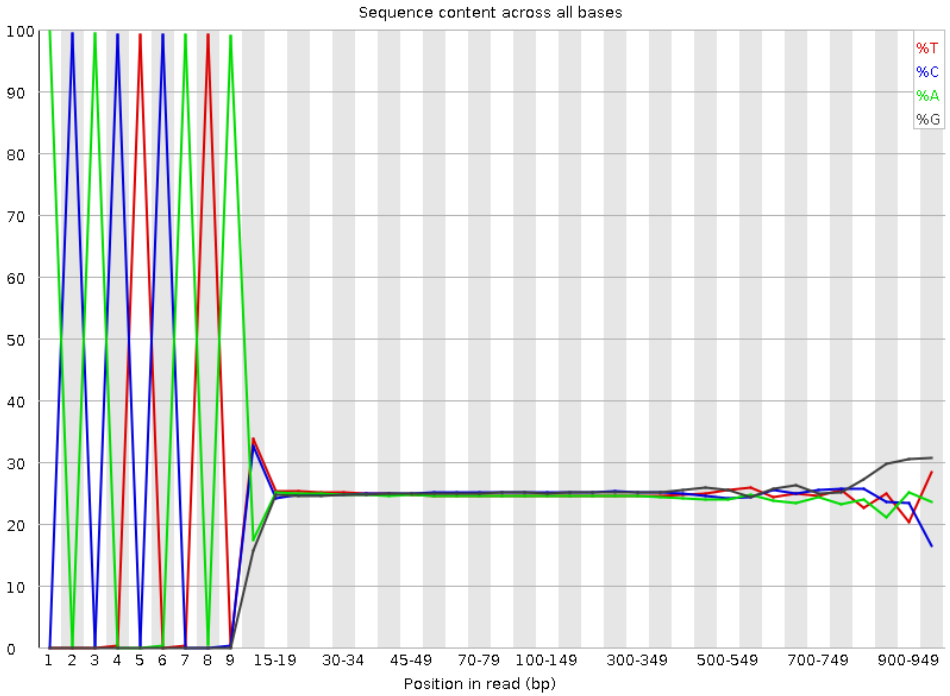


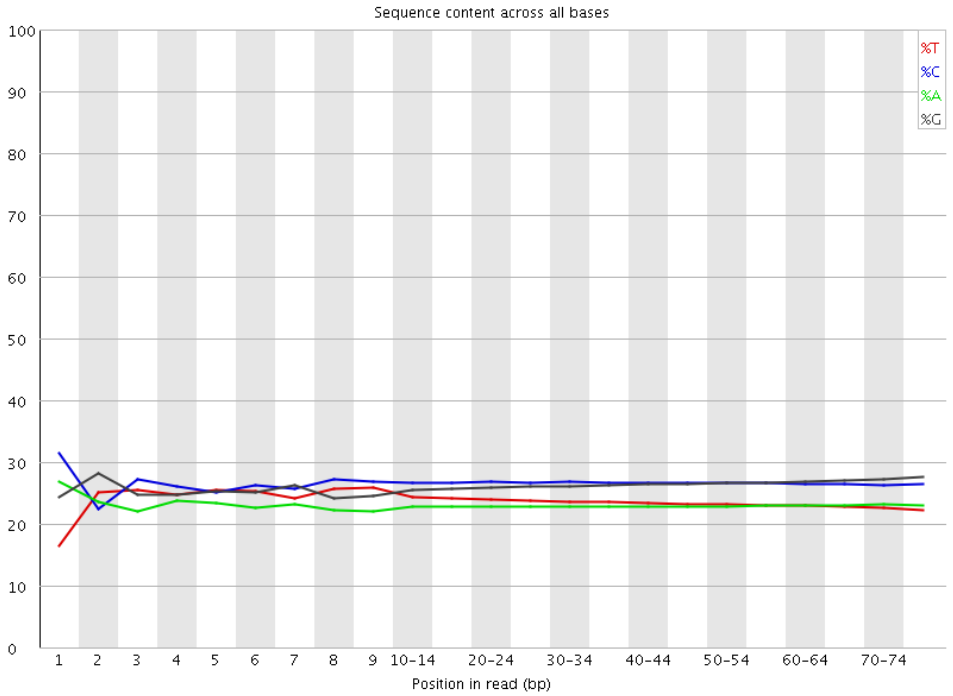
Forward reads



Reverse reads

QC for fastq files with fastqc

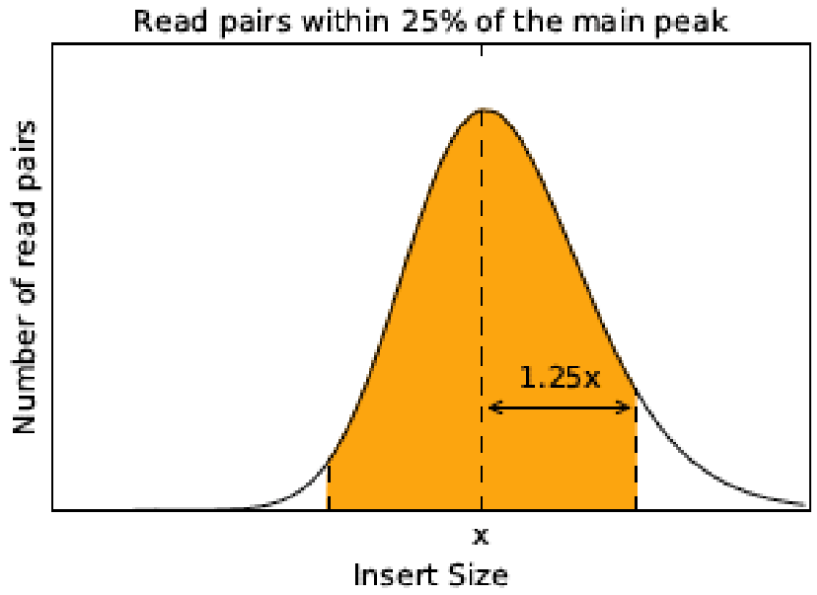




QC for bam files with samtools

A good example for QC data (human)

Minimum number of mapped bases	90%
Maximum error rate	0.02%
Maximum number of duplicate reads	5%
Minimum number of mapped reads which are properly paired	80%
Maximum number of duplicated bases due to overlapping	4%
Minimum number of reads within 25% of the main peak	80%



Test yourself

Get test file from NCBI (wget)

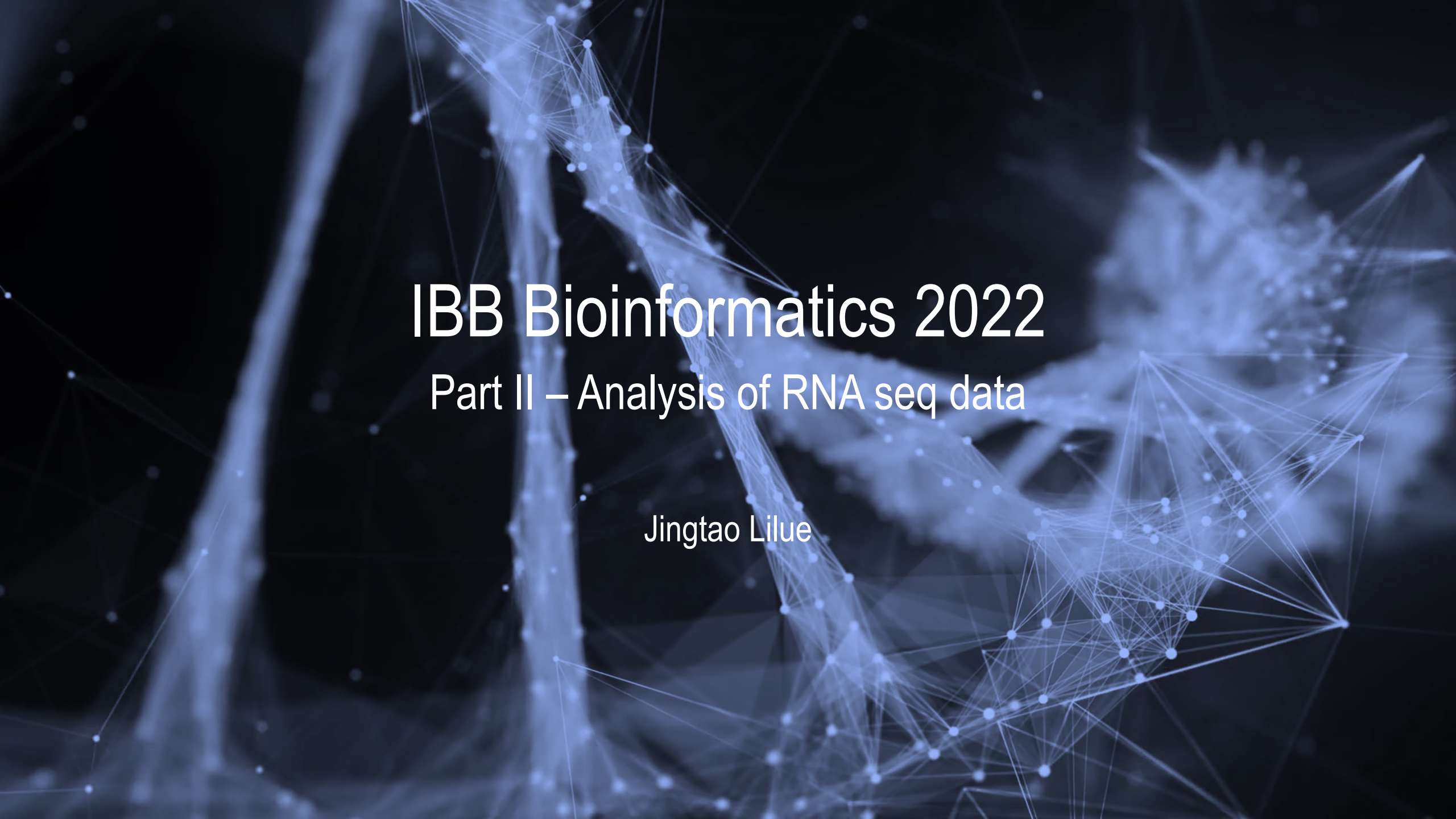
https://ftp.ncbi.nlm.nih.gov/toolbox/gbench/tutorial/Tutorial6/BAM_Test_Files/scenario2_no_index_unsorted_need_id_mapping/

Sort and index the file with samtools (samtools sort, samtools index)

View its head and first lines (samtools view)

See the statistics (samtools stats)

Test fastq files with fastqc

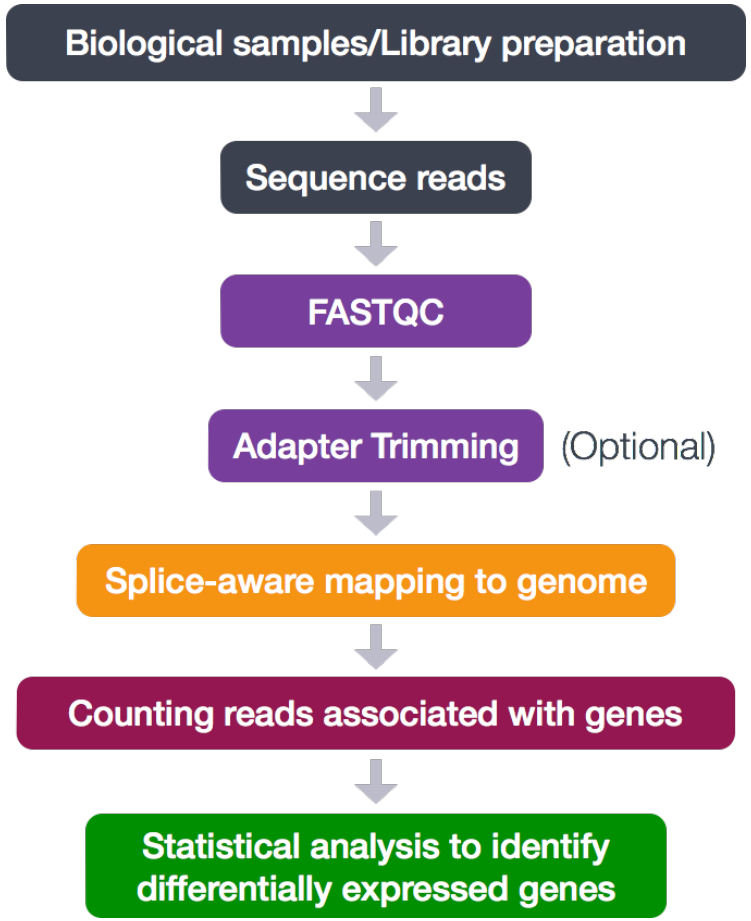


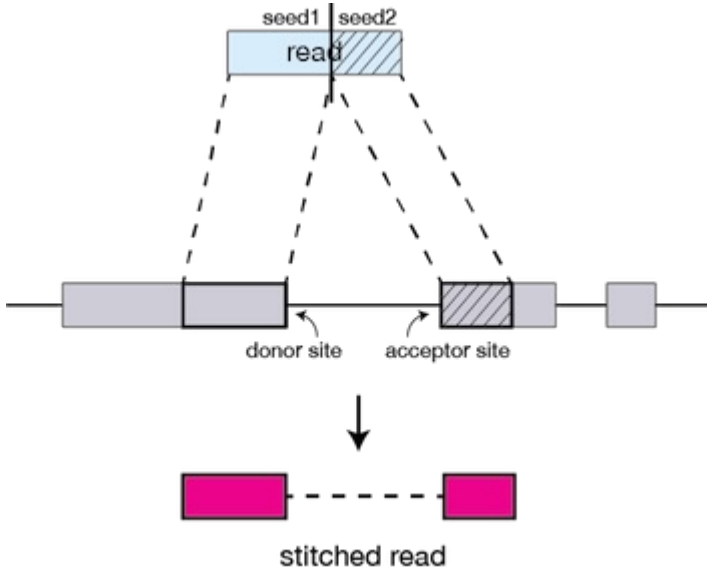
IBB Bioinformatics 2022

Part II – Analysis of RNA seq data

Jingtao Lilue

Analysis of RNA-seq data – overview





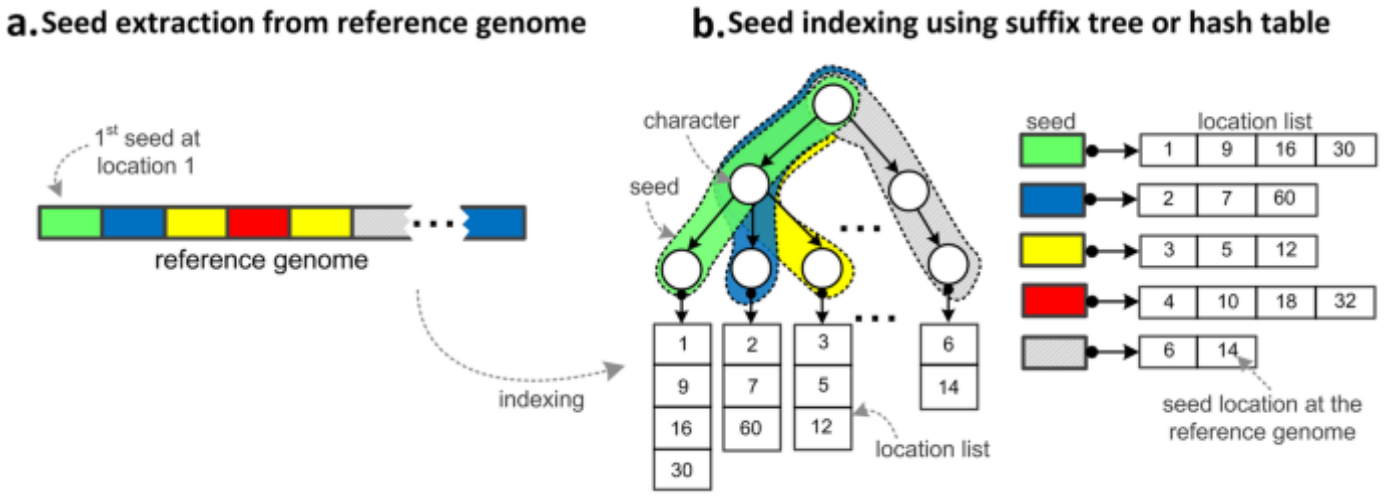
Analysis of RNA-seq data - Aligners

Aligner	Mapping speed: million read pairs/hour		Peak physical RAM, GB	
STAR	309.2	549.9	27.0	28.4
STAR sparse	227.6	423.1	15.6	16.0
TopHat2	8.0	10.1	4.1	11.3
RUM	5.1	7.6	26.9	53.8
MapSplice	3.0	3.1	3.3	3.3
GSNAP	1.8	2.8	25.9	27.0

STAR aligner for RNA seq

- Build index for reference genome

```
STAR --runThreadN 2 --runMode genomeGenerate --genomeDir ./ --genomeFastaFiles
./genome.fa
```



STAR aligner for RNA seq

- Index building

```
STAR --runThreadN 2 --runMode genomeGenerate --genomeDir ./ --genomeFastaFiles ./genome.fa
```

- Align reads to the reference

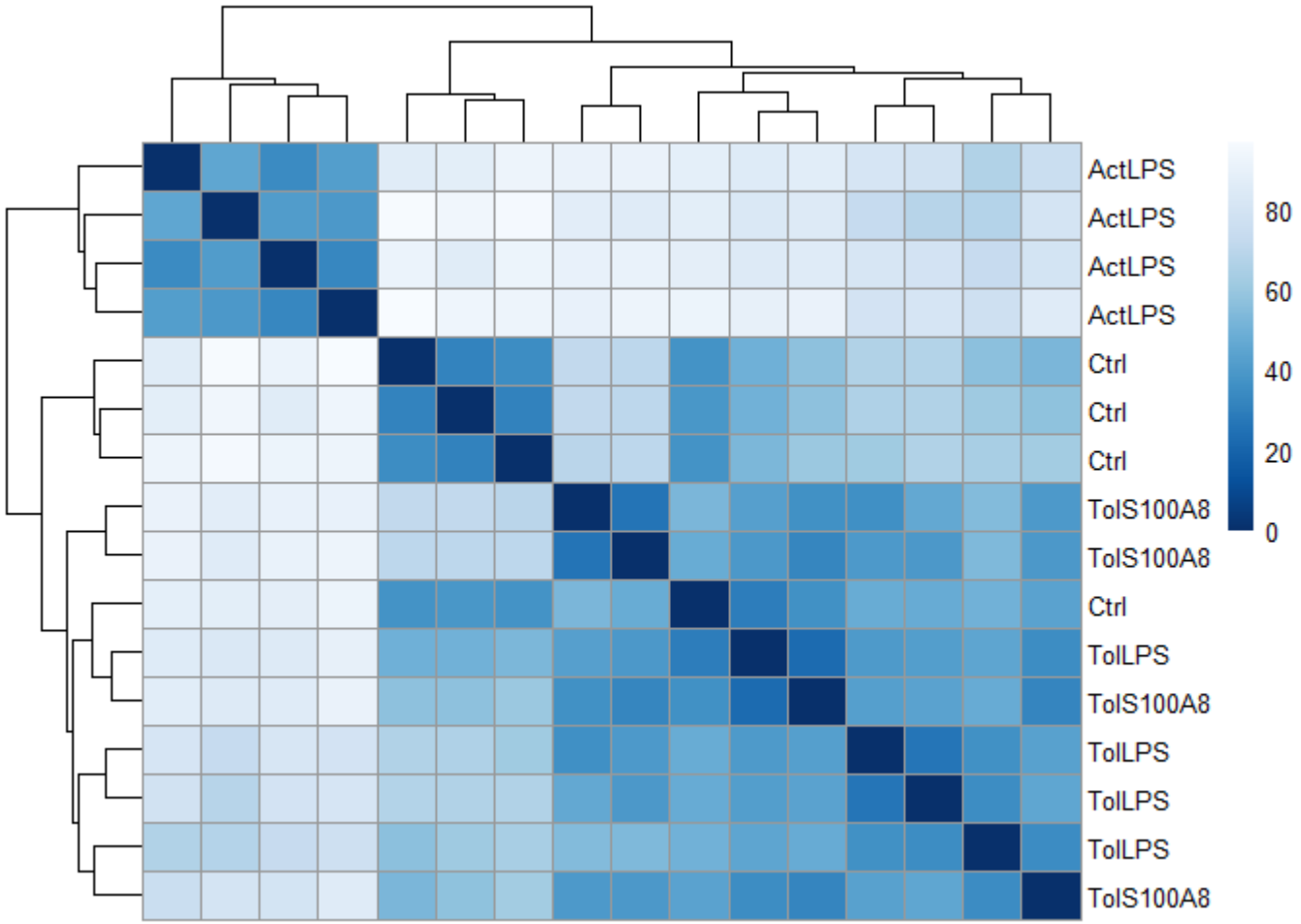
```
STAR --genomeDir ./ --readFilesIn file1.fq --alignIntronMax 500000 --outSAMtype BAM SortedByCoordinate --outFileNamePrefix file1 --sjdbOverhang 99 --sjdbGTFfile ./genes.gtf --runThreadN 2
```

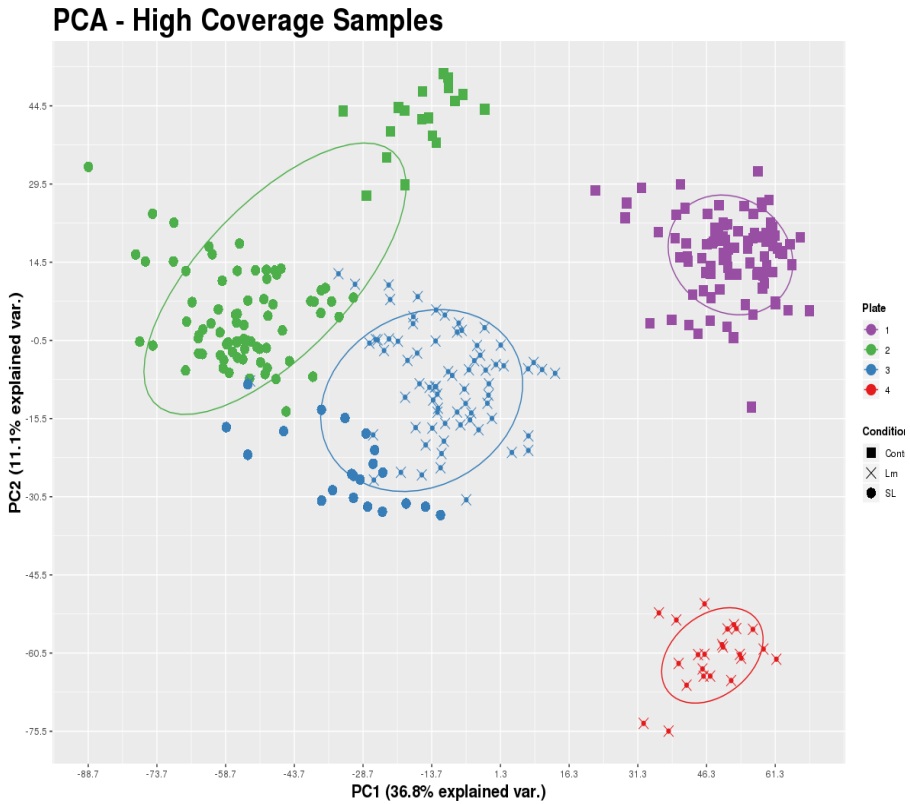
STAR aligner for RNA seq

- Check the Log output
- Build Bam index file
- Check stats of the bam file
- View with IGV, check the splicing, and expression level of genes (Ide)
- Feature count from Subread package

```
featureCounts -a ./genes.gtf -o ~/output.fc input.bam input2.bam -Q20 --primary --ignoreDup -C -t CDS -g 'gene' -T 4
```

Differential gene expression analysis with deSeq2





Differential gene expression analysis with deSeq2