RNA-Seq Workshop
Feb 25, 2021

Anish Bhaswanth Chakka, MS

Bioinformatics Analyst

Cancer Bioinformatics Services (CBS)

Cancer Bioinformatics Services and Genomics Analysis Core

- Uma Chandran, MSIS, PhD - chandran@pitt.edu
- Anish Bhaswanth Chakka, MS
- Rahil Sethi, MS
- Paul Cantalupo, MS
- William Schwarzmann, BS
- Alex Chang, BS
- Vishal Soman, BS

Department of
Biomedical Informatics

Genomics Cores – fee for service

- **Cancer Bionformatics Services (CBS)**
  - http://hillmanresearch.upmc.edu/research/facilities/cancer-bioinformatics/services/
  - For the Hillman Cancer Center
  - Started in 2004
  - Cancer Center Support Grant (CCSG)
  - 3 Master's level analysts
  - Co-authors on 40+ publications
  - Transcriptomics, variant analysis, TCGA, PGRR
- **Genomics Analysis Core (GAC)**
  - https://www.genomicsanalysis.pitt.edu/
  - Two master's level analyst
  - 60 projects to date
  - Work with CRC (Fangping Mu, PhD)


Alex Chang


Anish Chakka


Paul Cantalupo


Rahil Sethi


Vishal Soman


William Schwarzmann

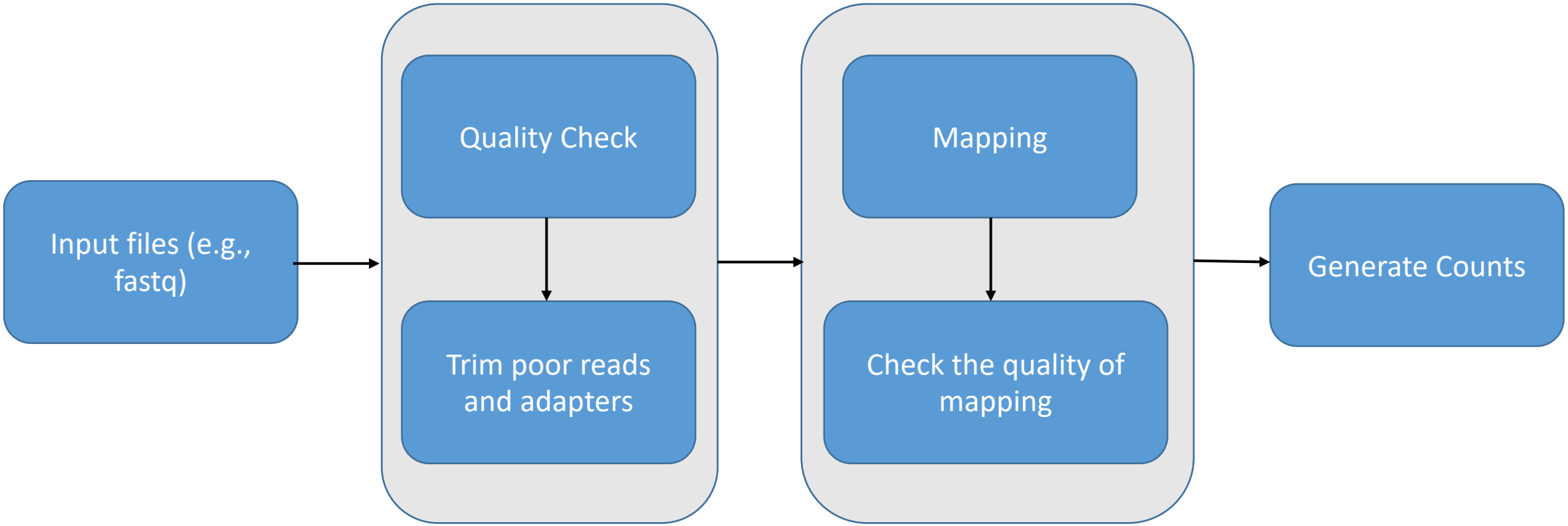
Table of contents

Workshop Handouts

Hackmd document with the workshop material can be found in this link

Table of contents

Pipeline


Input files (e.g., fastq)
Quality Check
Trim poor reads and adapters
Mapping
Check the quality of mapping
Generate Counts

Reality

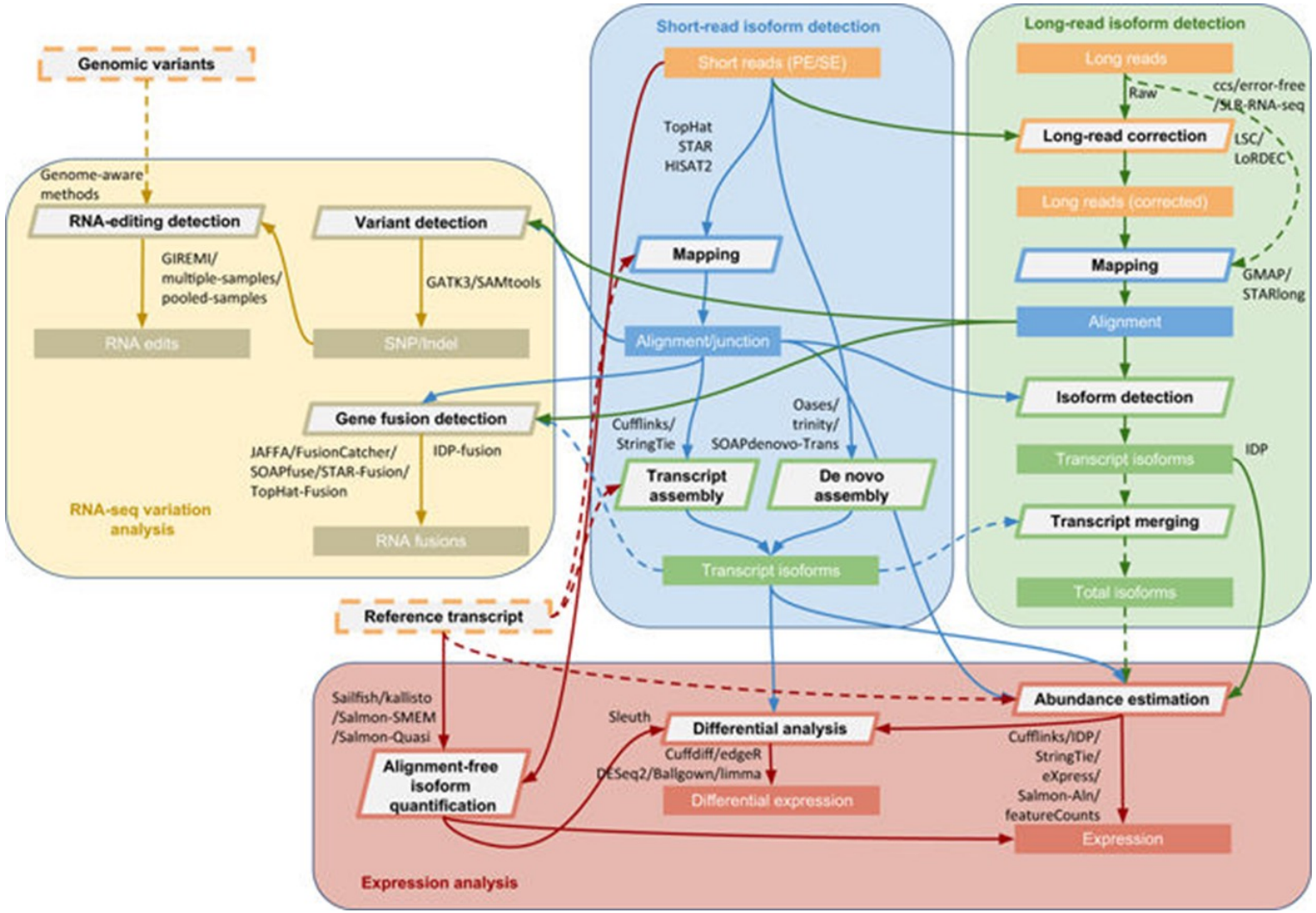
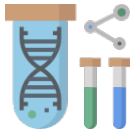

Genomic variants
Genome-aware methods
RNA-editing detection
GIREMI/ multiple-samples/ pooled-samples
RNA edits
Variant detection
GATK3/SAMtools
SNP/Indel
Gene fusion detection
JAFFA/FusionCatcher/ SOAPfuse/STAR-Fusion/ TopHat-Fusion
IDP-fusion
RNA fusions
RNA-seq variation analysis
Short-read isoform detection
Short reads (PE/SE)
TopHat STAR HISAT2
Mapping
Alignment/junction
Cufflinks/ StringTie
Oases/ trinity/ SOAPdenovo-Trans
Transcript assembly
De novo assembly
Transcript isoforms
Long-read isoform detection
Long reads
Raw
ccs/error-free /SLR-RNA-seq
Long-read correction
LSC/ LoRDEC
Long reads (corrected)
Mapping
GMAP/ STARlong
Alignment
Isoform detection
Transcript isoforms
IDP
Transcript merging
Total isoforms
Reference transcript
Sailfish/kallisto /Salmon-SMEM /Salmon-Quasi
Alignment-free isoform quantification
Sleuth
Differential analysis
Cuffdiff/edgeR DESeq2/Ballgown/limma
Differential expression
Abundance estimation
Cufflinks/IDP/ StringTie/ eXpress/ Salmon-Aln/ featureCounts
Expression
Expression analysis

Table of contents

Understanding the data

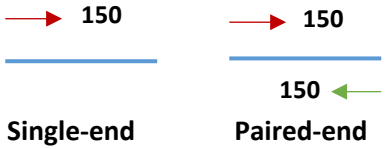
- It’s important to understand the design of the experiment.
- Few things we need to consider before starting the analysis are as below:

What is the objective of the experiment?



Which method (total RNA, poly A etc.) was used for sequencing?

Is it single or paired end read?



Strandedness of the data: Forward or reverse strand?

How many replicates were used?

Which organism was used for sequencing?

- We usually send an RNA-Seq questionnaire to the lab before beginning the analysis.

What are reads?

- *Reads* are short sequences produced by the DNA Sequencer.
- DNA sequencers are manufactured and sold by different companies, and they produce reads in different formats.
- The most common file-type to store the reads is *fastq format.*

DNA Sequencer	File format
Illumina	.fastq
SOLiD	.xsq or .csfasta + .qual
Ion-torrent	.fastq

Fastq file and its format

FASTQ format is a text-based format for storing both a biological sequence (usually nucleotide sequence) and its corresponding quality scores.

A **fastq file** typically uses 4 lines per sequence:

Line1: Begins with **'@'** symbol, followed by the sequence identifier and an optional descriptor.

Line2: Includes the **raw sequence**.

Line3: Begins with a **'+'** character.

Line4: Encodes the **quality scores** of the sequence in *Line2*.

Line1	@SIM:1:FCX:1:15:6329:1045 1:N:0:2
Line2	TCGCACTCAACGCCCTGCATATGACAAGACAGAATC
Line3	+
Line4	<>;AA=><9=AAAAAAAAAA9A:<A<;<<<????A=

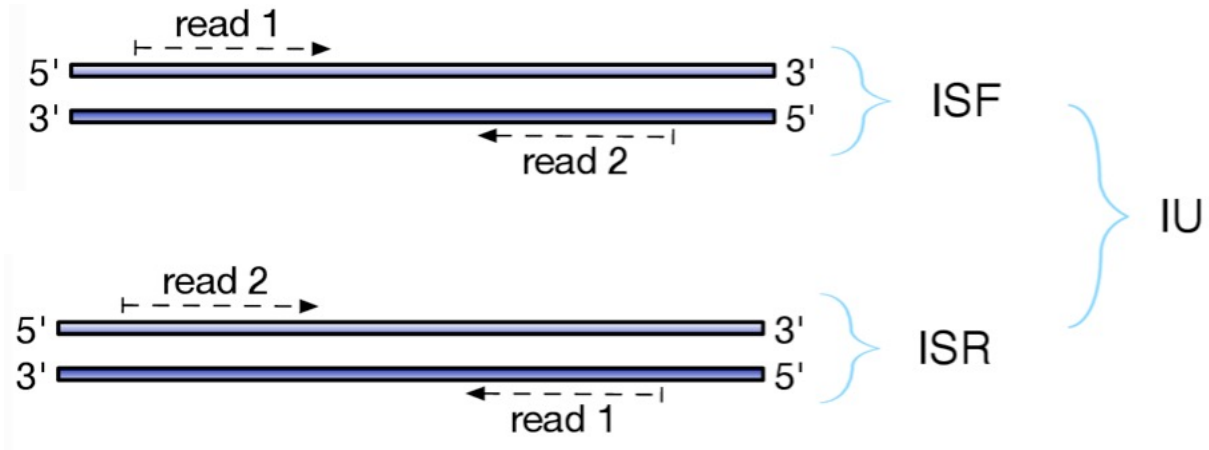
Single and Paired-end reads

- Single-end (SE) reads are sequenced from one end of the DNA fragment, whereas paired-end (PE) reads are sequenced from both ends.
- PE reads improves assembly of repetitive regions and are more effective in identifying InDels.
- However, they are more expensive and time consuming compared to SE reads.

Read Libraries

Tophat	Salmon	
	Paired-end	**Single-end**
`-fr-unstranded`	`-l IU`	`-l U`
`-fr-firststrand`	`-l ISR`	`-l SR`
`-fr-secondstrand`	`-l ISF`	`-l SF`

I = Inward, O = Outward; **S = Stranded, U = Unstranded**; **F = Read 1 from Forward Read, R = Read1 from Reverse Read**



Source: Salmon documentation

Table of contents

Quality scores in a fastq file I

```
Line1 @SIM:1:FCX:1:15:6329:1045 1:N:0:2
Line2 TCGCACTCAACGCCCTGCATATGACAAGACAGAATC
Line3 +
Line4 <>;AA=><9=AAAAAAAAAA9A:<A<;<<<????A=
```

- We do NOT use numbers to represent quality scores, but we use ASCII symbols instead.
- Using one symbol, instead of two digits helps us save space. For e.g., **A** means a quality value of **32**.
- Usually the range is from 0-40, where is 0 means poorest quality and 40 means highest quality

phred+33 score

ASCII Symbol	ASCII Code	Quality Value
<	60	27 **+33 = 60**
>	62	29 **+33 = 62**
;	59	26 **+33 = 59**
A	65	32 **+33 = 65**

ASCII control characters			ASCII printable characters					
00	NULL	(Null character)	32	space	64	@	96	`
01	SOH	(Start of Header)	33	!	65	A	97	a
02	STX	(Start of Text)	34	"	66	B	98	b
03	ETX	(End of Text)	35	#	67	C	99	c
04	EOT	(End of Trans.)	36	$	68	D	100	d
05	ENQ	(Enquiry)	37	%	69	E	101	e
06	ACK	(Acknowledgement)	38	&	70	F	102	f
07	BEL	(Bell)	39	'	71	G	103	g
08	BS	(Backspace)	40	(	72	H	104	h
09	HT	(Horizontal Tab)	41	)	73	I	105	i
10	LF	(Line feed)	42	*	74	J	106	j
11	VT	(Vertical Tab)	43	+	75	K	107	k
12	FF	(Form feed)	44	,	76	L	108	l
13	CR	(Carriage return)	45	-	77	M	109	m
14	SO	(Shift Out)	46	.	78	N	110	n
15	SI	(Shift In)	47	/	79	O	111	o
16	DLE	(Data link escape)	48	0	80	P	112	p
17	DC1	(Device control 1)	49	1	81	Q	113	q
18	DC2	(Device control 2)	50	2	82	R	114	r
19	DC3	(Device control 3)	51	3	83	S	115	s
20	DC4	(Device control 4)	52	4	84	T	116	t
21	NAK	(Negative acknowl.)	53	5	85	U	117	u
22	SYN	(Synchronous idle)	54	6	86	V	118	v
23	ETB	(End of trans. block)	55	7	87	W	119	w
24	CAN	(Cancel)	56	8	88	X	120	x
25	EM	(End of medium)	57	9	89	Y	121	y
26	SUB	(Substitute)	58	:	90	Z	122	z
27	ESC	(Escape)	59	;	91	[	123	{
28	FS	(File separator)	60	<	92	\	124	|
29	GS	(Group separator)	61	=	93	]	125	}
30	RS	(Record separator)	62	>	94	^	126	~
31	US	(Unit separator)	63	?	95	_		
127	DEL	(Delete)						

Quality scores in a fastq file II

- It measures the quality of the identification of the nucleobases generated by automated DNA sequencing.
- It represents an error probability, which can be calculated as:

$$P = 10^{-Q/10} \quad \text{or} \quad Q = -10 \log_{10} P$$

Quality (Q)	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 10,000	99.99%

Quality scores in a fastq file III

```
 SSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSS.....................................
 .........................XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX.....................
 ............................IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII.....................
 ............................JJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJ.....................
 LLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLL.....................................
 !"#$%&'()*+,-./0123456789:;<=>?@ABCDEFGHIJKLMNOPQRSTUVWXYZ[\]^_`abcdefghijklmnopqrstuvwxyz{|}~
 |                         |    |        |                              |                     |
33                        59   64       73                            104                   126
 0........................26...31.......40
                          -5....0........9.............................40
                               0........9.............................40
                                  3.....9.............................41
 0.2......................26...31........41

S - Sanger        Phred+33,  raw reads typically (0, 40)
X - Solexa        Solexa+64, raw reads typically (-5, 40)
I - Illumina 1.3+ Phred+64,  raw reads typically (0, 40)
J - Illumina 1.5+ Phred+64,  raw reads typically (3, 41)
    with 0=unused, 1=unused, 2=Read Segment Quality Control Indicator (bold)
    (Note: See discussion above).
L - Illumina 1.8+ Phred+33,  raw reads typically (0, 41)
```

Source: https://en.wikipedia.org/wiki/FASTQ_format

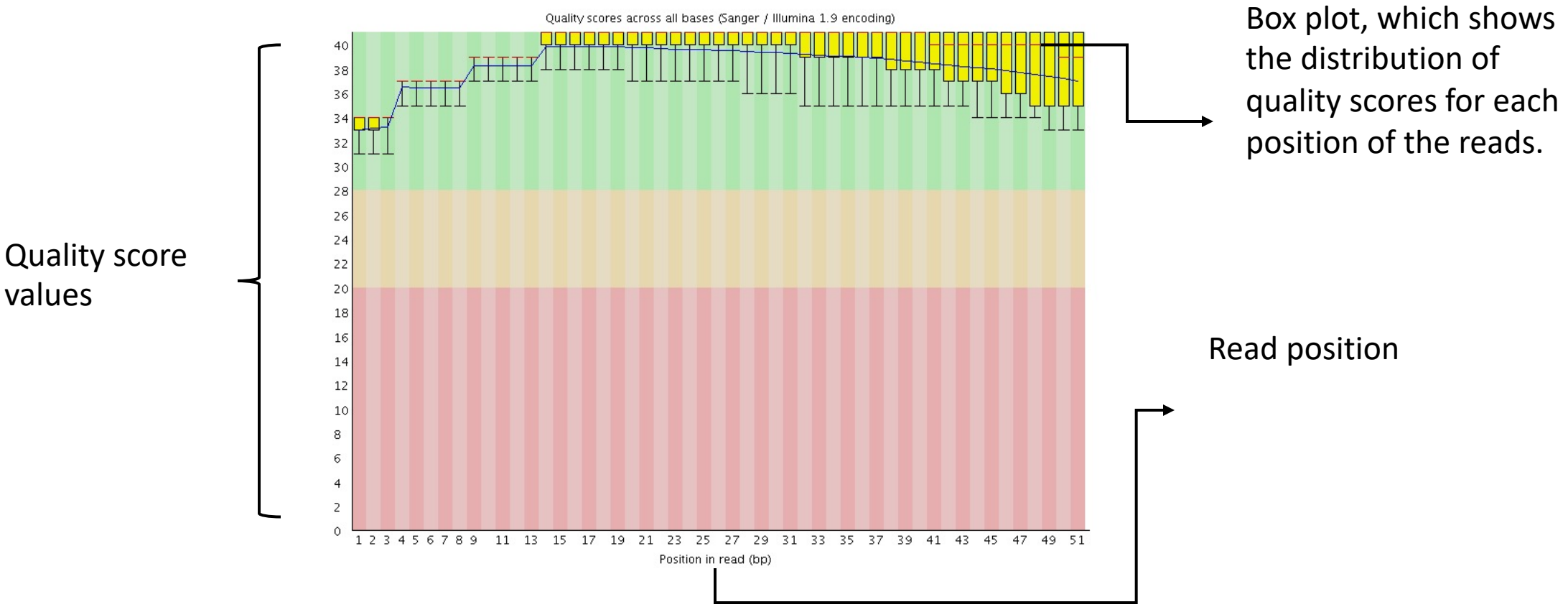
FastQC I

- Each fastq file typically has millions of reads and it's almost impossible to check for quality scores in all these reads manually.
- FastQC is a tool which takes in a fastq file as input and provides summary graphs and tables in html format.
- The tool summarizes the quality scores of all reads in each position and summarizes them in the form of a boxplot.
- It also tells us if there are any adapter sequences in the read.

```
Line1 @SIM:1:FCX:1:15:6329:1045 1:N:0:2
Line2 TCGCACTCAACGCCCTGCATATGACAAGACAGAATC
Line3 +
Line4 <>;AA=><9=AAAAAAAAA9A:<A<;<<<????A=
```

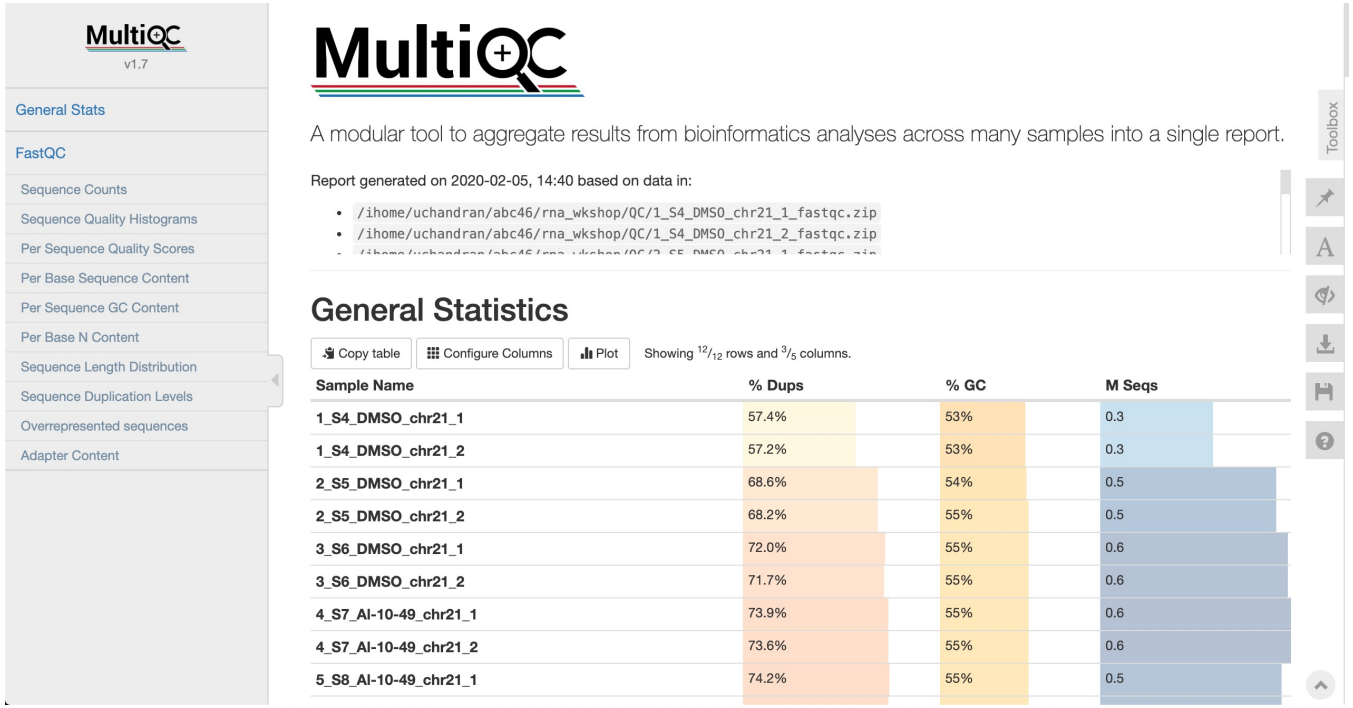
FastQC II

- In this example, each read is 51 bp long and the quality scores of millions of reads are summarized in the form of the Box plots (yellow).



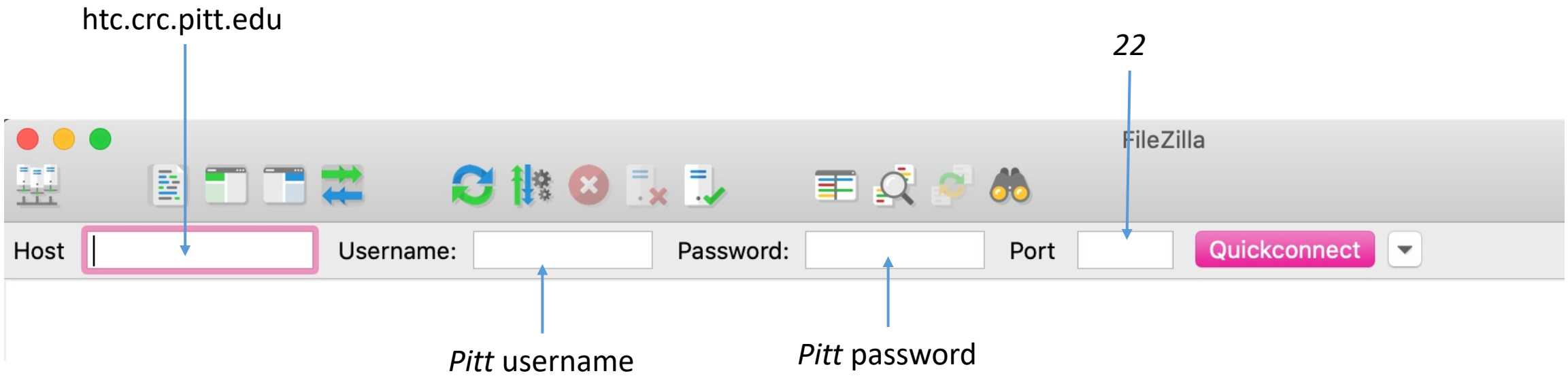
MultiQC

- MultiQC aggregates results from bioinformatics analyses across many samples into a single report
- We get a nice html report which summarizes the FastQC results.
- Then, we will transfer the report to our laptop/desktop using an FTP client.



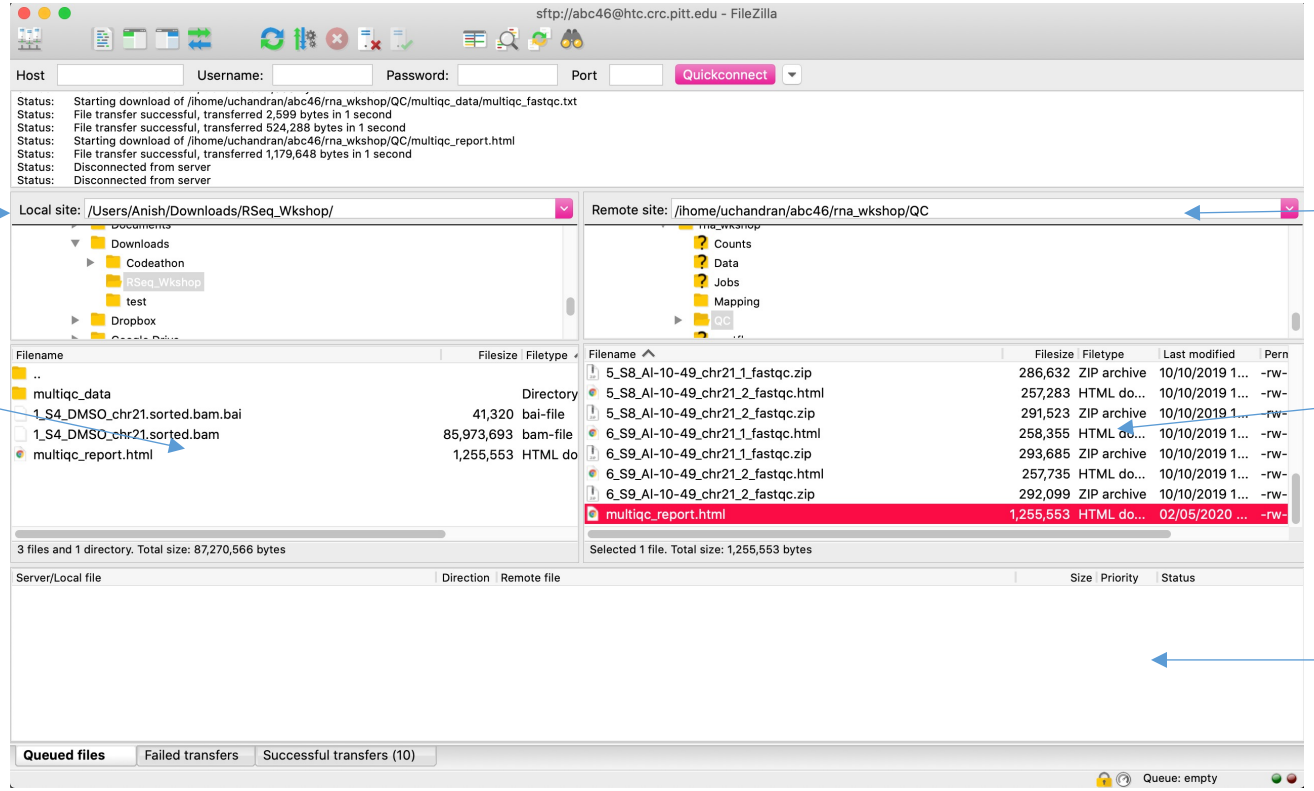
FileZilla

- You can transfer files to your laptop/desktop using free FTP tools like FileZilla.
- It's available for both windows and mac.
- Here is the link: https://filezilla-project.org/



FileZilla

- Install FileZilla: https://filezilla-project.org/
- Login using your Pitt credentials (ensure **sremote** is active).
- Transfer your QC results to your Mac/PC.



Breakout Session-1

Table of contents

Cutadapt

- Bad quality basepairs and adapters should be trimmed.
- **Cutadapt** is a tool which helps in trimming reads based on poor quality score and adapter sequences you provide.
- After trimming if the read is too short, it’s completely removed.
- Paired-end reads should be trimmed simultaneously.
- It’s important to make sure after trimming, the number of reads in both pairs are identical.

Table of contents

What is a reference genome?

- A **reference genome** is a set of nucleic acid sequences assembled as a representative example of a species' genetic material.
- These different reference genomes are constantly updated by Genome Reference Consortium (GRC) and they release newer versions (also called as **builds**).

Challenges choosing a Reference Genome I

- There are subtle differences between human genome versions released by databases like UCSC, NCBI and Ensembl:
  - UCSC uses ‘chr1’ nomenclature, where as Ensembl uses ‘1’ nomenclature. NCBI uses RefSeq accession numbers.
  - The mitochondrion for hg19 in UCSC differs from the one in GRCh37.

Human genome build

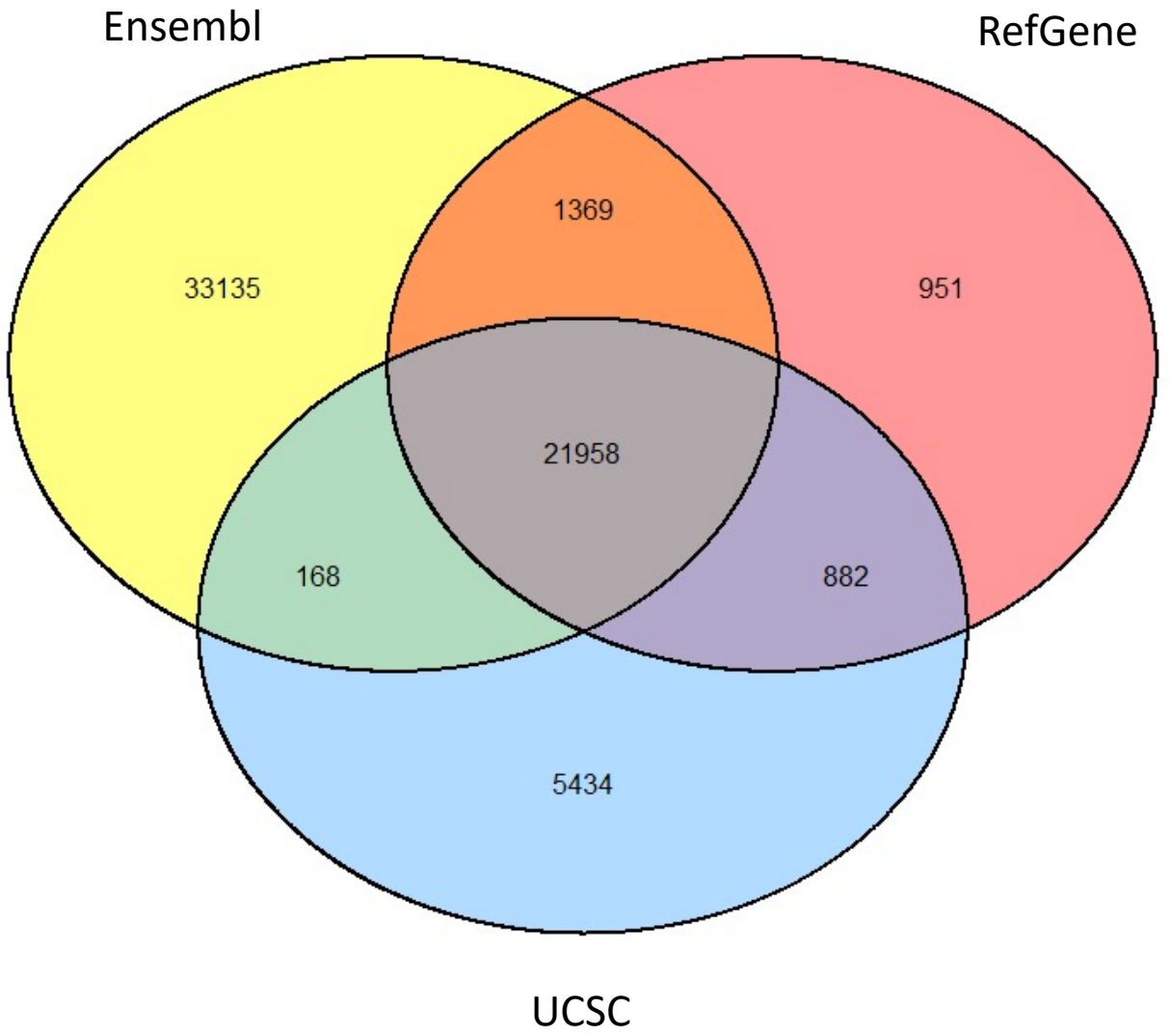
Release name	Date of release	Equivalent UCSC version
GRCh38	Dec 2013	hg38
GRCh37	Feb 2009	hg19
NCBI Build 36.1	Mar 2006	hg18

Mouse genome build

Release name	Date of release	Equivalent UCSC version
GRCm38	Dec 2011	mm10
NCBI Build 37	Jul 2007	mm9
NCBI Build 36	Feb 2006	mm8

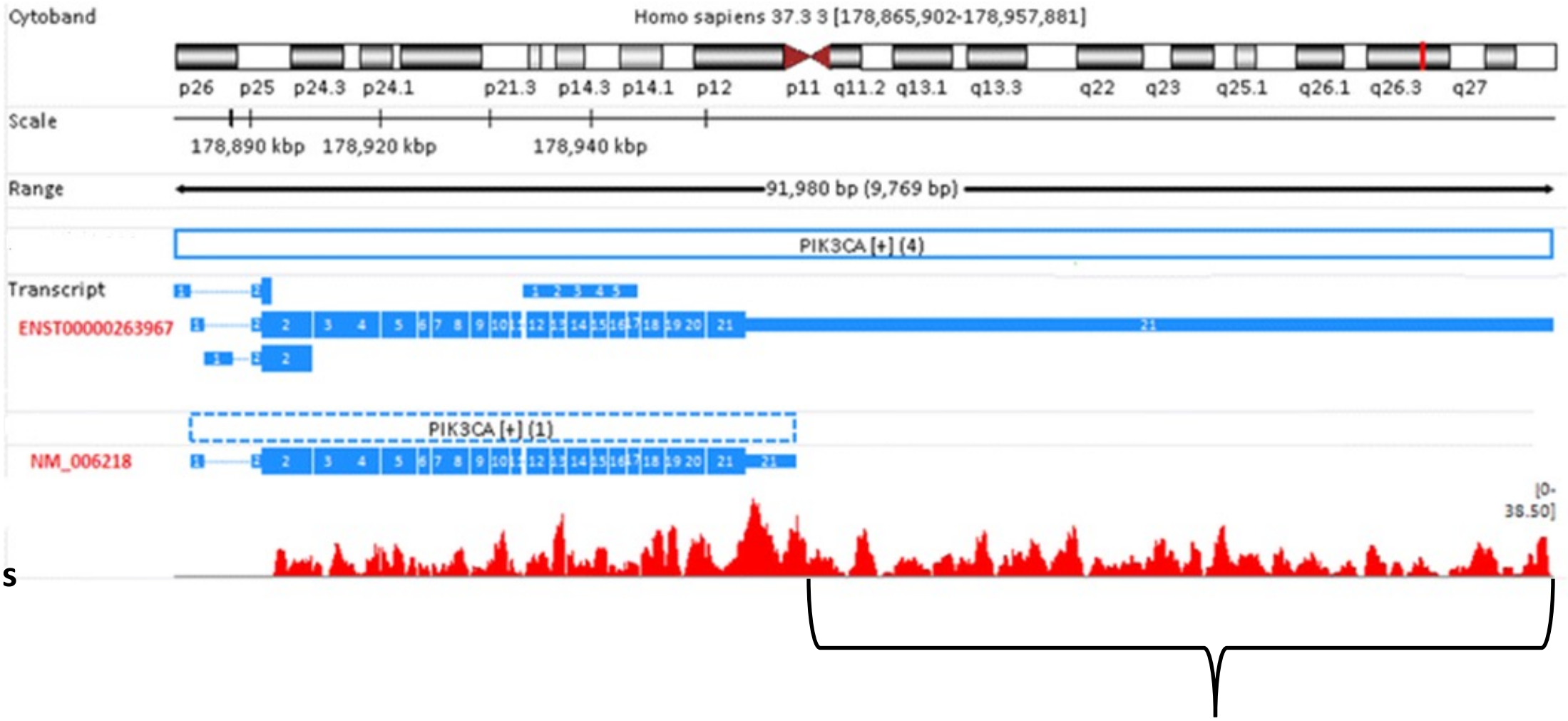
Challenges choosing a Reference Genome II

- Total number of unique genes will differ between reference genomes
- Quantitation and then DE will be affected by reference genome, mapping strategy and quantitation methods
- Output from different pipelines can differ considerably



Zhao and Zhang: BMC Genomics 2015

Challenges choosing a Reference Genome III



Ensembl

RefGene

Liver Reads

Zhao et al., BMC Genomics

Covered by Ensembl but not RefGene

What is mapping?

- Reads do not come with position information and hence, we do not know which part of the genome they belong to.
- We need to use the sequence of the read to find the corresponding region in the reference genome.
- This process of assigning the reads to specific location in the genome is called mapping.
- There are different tools which can do mapping. Few of them are:
  - Tophat2
  - HISAT
  - STAR
  - Salmon
  - Kallisto
- A mapper typically takes a *reference genome* and *reads* as the input.

Tool Ecosystem Reality

- "**TopHat2** correctly aligned more than 98% of the reads, which was **higher than with any of the other methods**, whose accuracy ranged from 88 to 97%"
- "**STAR** **outperforms other aligners by a factor of >50 in mapping speed**, aligning to the human genome 550 million 2 × 76 bp paired-end reads per hour on a modest 12-core server, while at the same time **improving alignment sensitivity and precision**"
- "Tests on real and simulated data sets showed that **HISAT** is the **fastest system currently available**, with **equal or better accuracy than any other method**"
- "We present **Kallisto**, an RNA-seq quantification program that is **two orders of magnitude faster** than previous approaches and **achieves similar accuracy**."
- "We show that **Salmon** typically **outperforms both kallisto and eXpress in terms of accuracy**"

Hisat2

- **HISAT2** is a fast and sensitive alignment program for mapping next-generation sequencing reads.
- The tool outputs alignments in **SAM** format.
- **Steps involved**
  a. Build an index of the genome using *hisat2-build*
  b. Run *hisat2* for alignment, using the index generated above

SAM format

- SAM stands for **Sequence Alignment/Map format**.
- The output of mapping tool is typically in a SAM format, which includes information about the alignment and quality of the mapping.
- SAM file can be compressed into BGZF format called as **BAM (Binary SAM)**.
- It is a TAB-delimited text format consisting of a **header section**, which is optional, and an **alignment section**.

Header

```
@HD     VN:1.0  SO:coordinate
@SQ     SN:21   LN:46709983
@PG     ID:hisat2       PN:hisat2       VN:2.1.0        CL:"/ihome/crc/install/hisat2/hisat2-2.1.0/hisat2-align-s --wrapper basic-0 -x /bgfs/genomics/workshop
s/rnaseq_2019f/Refs/GRCh38_index_chr21/GRCh38_index_chr21 -S /ihome/uchandran/abc46/rna_wkshop/Mapping/1_S4_DMSO_chr21.sam -p 3 --dta -1 /ihome/uchandran/abc4
```

```
SRR5861494.11050870     99      21      5011256 1       90M     =       5011321 155     TCAGCCCCAAAAGGCGAGATATCTTGAAGCTTACAGGTTATAGGTGGGTTTAGGGATTCTTTAGTTGGCA
GTTGGTTGAAAGAGTTGAGC    CCCFFFFFHHHHHJJJIJJJJJIJJJIIIJJIJJIJJDHHIGJJDFHJ8FHHGIJG@AEHHHFFCDFFEEDCCDDDB@?CCDCDADCCCC      AS:i:0  ZS:i:0  XN:i:0  XM:i:0  XO:i:0
XG:i:0  NM:i:0  MD:Z:90 YS:i:0  YT:Z:CP NH:i:2
SRR5861494.11050870     147     21      5011321 1       90M     =       5011256 -155    TGGCAGTTGGTTGAAAGAGTTGAGCTTTGCCTAAAAACTGGGAGTCAGTAGAAAGGAATGCTTGAGTTAA
AATAAGGAGGTCTGCTGTCT    CCEEEDEFFHHFHHHHJIHIJJJICIHIJIJJJJIHDGFBIIIIIIIJJJJJJIJIJJGIJJIGJJJJJJIJJJJJIHHHHGFFFFFCCC      AS:i:0  ZS:i:0  XN:i:0  XM:i:0  XO:i:0
XG:i:0  NM:i:0  MD:Z:90 YS:i:0  YT:Z:CP NH:i:2
```

Alignment

SAM: Header section

- If present, the header must be prior to the alignments. Header lines start with ‘@’, while alignment lines do not.

- It gives some important information like:
  - Is the SAM file sorted?
    - If NO => **SO:unsorted**
    - If YES => **SO:coordinate** or **SO:queryname**
  - Sequence information
  - Some information about the tool which was used to generate it

SAM: Alignment section

- Each **alignment line has 11 mandatory fields** for essential alignment information such as mapping position, and variable number of optional fields for flexible or aligner specific information.

Col No.	Field	Brief description
1	QNAME	Query template NAME
2	FLAG	bitwise FLAG
3	RNAME	Reference sequence NAME
4	POS	1-based leftmost mapping POSition
5	MAPQ	MAPping Quality
6	CIGAR	CIGAR string
7	RNEXT	Reference name of the mate/next read
8	PNEXT	Position of the mate/next read
9	TLEN	observed Template LENgth
10	SEQ	segment SEQuence
11	QUAL	ASCII of Phred-scaled base QUALity+33

Salmon

- Salmon is a tool for **wicked-fast** transcript quantification from RNA-seq data.
- What input files do we need to run?
  - A FASTA file containing your **reference transcripts**
  - A (set of) **FASTA/FASTQ** file(s) containing your reads.
  - Optionally, Salmon can make use of pre-computed alignments (in the form of a **SAM/BAM** file) to the transcripts rather than the raw reads.
- **Steps**
  a. Indexing (transcript file)
  b. Quantification (on specific reads)

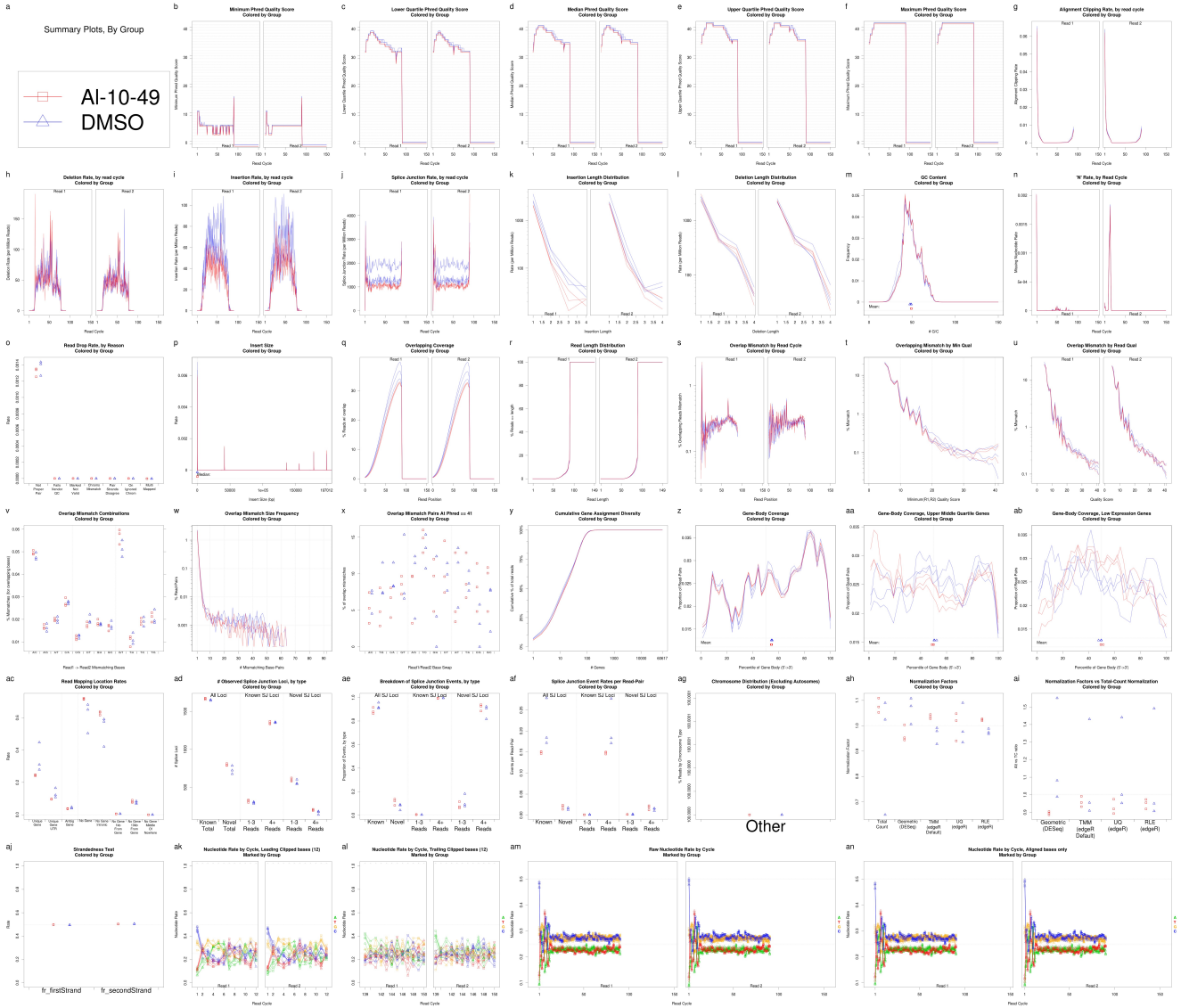
Salmon Output

Name	Length	EffectiveLength	TPM	NumReads
ENST00000390534.1	62	6.012	0	0
ENST00000390535.2	66	8.652	0	0
ENST00000390536.2	62	6.012	0	0
ENST00000361390.2	956	783.447	14.486598	1
ENST00000361453.3	1042	869.422	13.054061	1
ENST00000361624.2	1542	1369.422	33.15117	4
ENST00000361739.1	684	511.537	44.374035	2
ENST00000361851.1	207	61.099	185.757128	1
ENST00000361899.2	681	508.541	0	0
ENST00000362079.2	784	611.476	0	0
ENST00000361227.2	346	178.693	0	0
ENST00000361335.1	297	134.033	0	0
ENST00000361381.2	1378	1205.422	9.415363	1
ENST00000361567.2	1812	1639.422	20.768572	3
ENST00000361681.2	525	352.91	64.319326	2
ENST00000361789.2	1141	968.422	23.439135	2
ENST00000619842.4	2598	2425.422	0	0
ENST00000618826.4	2471	2298.422	0	0
ENST00000620311.1	2392	2219.422	0	0
ENST00000617983.1	2404	2231.422	0	0
ENST00000617299.4	2859	2686.422	0	0
ENST00000620917.4	2795	2622.422	0	0
ENST00000618874.1	1913	1740.422	0	0
ENST00000618686.1	2237	2064.422	5.497657	1

QC on mapped reads I

- After the mapping has been performed, it's important to check the quality of our mapping results.
- We typically check for:
  - Percentage of overall reads mapped
  - Mapping percentage on intron, exons and intergenic regions
  - GC %
  - No. of genes mapped in each sample
- To check the quality of BAM files we use tools like
  - QoRTs
  - Samtools
  - Picard Tools

QC on mapped reads II



Breakout Session-2

Table of contents

Counts

- RNA-Seq data is usually analyzed by creating a count matrix data of genes per sample.
- We can do this by counting how many reads from a given alignment file (BAM) map to a list of genomic features (e.g., gene).
- There are different ways to do this and it also depends on annotations used.
- Two common tools for counting the data are:
  - HT-Seq
  - featureCounts
- In our current analysis, we will use HT-Seq.

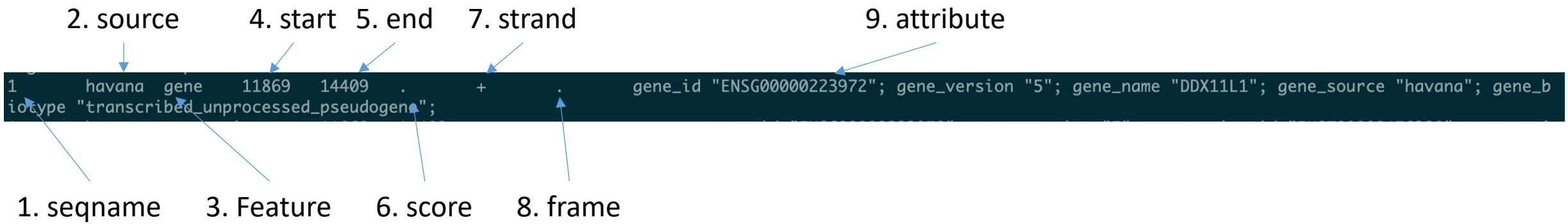
GTF file I

- The **Gene transfer format** (**GTF**) is a tab delimited text file format, which is used to hold information about gene structure.
- Each line consists of 9 columns of data/fields and they are as follows:

Field	Description
seqname	Name of the chromosome or scaffold
source	Data source
feature	Feature type name. E.G., Gene, exon, transcript etc.
start	Start position of the feature
end	End position of the feature
score	A floating point value
strand	+ (Forward) or – (reverse)
frame	One of '0', '1' or '2'. '0' indicates that the first base of the feature is the first base of a codon, '1' that the second base is the first base of a codon, and so on
attribute	A semicolon-separated list of tag-value pairs, providing additional information about the feature

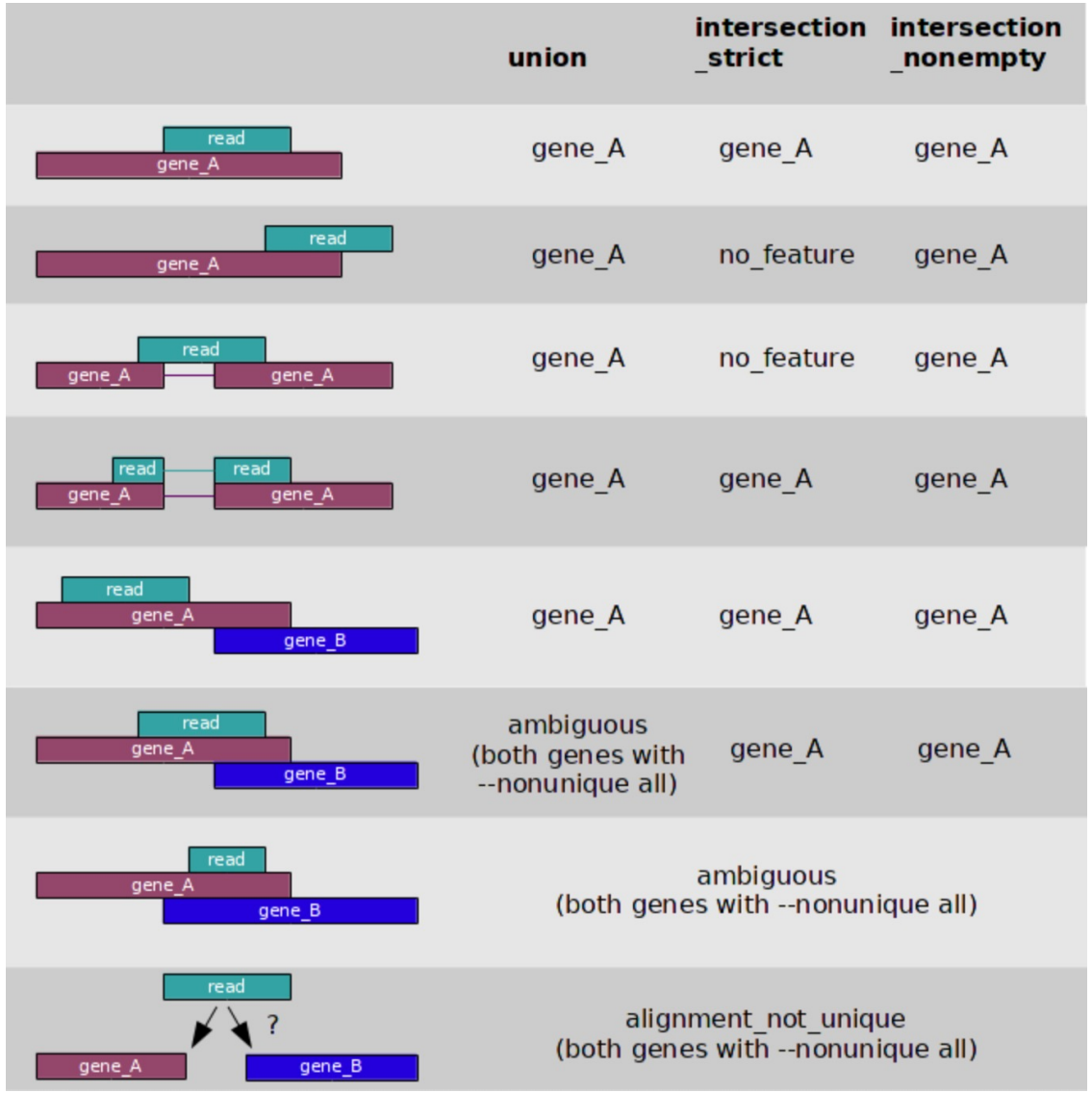
GTF file II

Here is an example GTF file of human downloaded from ensembl database



HT-Seq

- The htseq-count script allows to choose between three modes:
  - The **union** of all the sets *s(i)* for mode union. This mode is recommended for most use cases.
  - The intersection of all the sets *s(i)* for mode **intersection-strict**.
  - The intersection of all non-empty sets *s(i)* for mode **intersection-nonempty**.
- Union mode is the most appropriate and recommended method by the authors.
- The tool takes a BAM file and GTF file as input.



Example HT-Seq Count File

Genes

Counts

```
ENSG00000160201 50
ENSG00000160202 0
ENSG00000160207 12
ENSG00000160208 2518
ENSG00000160209 1090
ENSG00000160211 0
ENSG00000160213 885
ENSG00000160214 1252
ENSG00000160216 1709
ENSG00000160218 881
ENSG00000160219 0
ENSG00000160221 267
ENSG00000160223 229
ENSG00000160224 0
ENSG00000160226 128
ENSG00000160229 0
ENSG00000160233 55
ENSG00000160255 3682
ENSG00000160256 234
ENSG00000160271 0
ENSG00000160282 1
ENSG00000160284 130
```

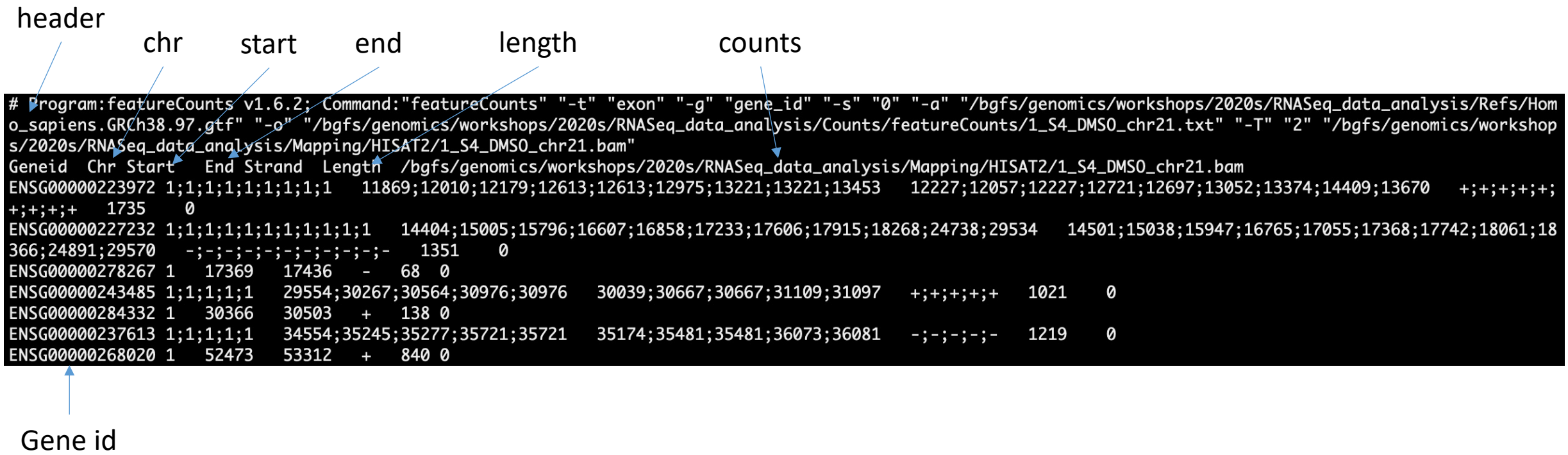
Summary at the bottom of the page

```
ENSG00000288109 0
ENSG00000288110 0
ENSG00000288111 0
__no_feature    16530
__ambiguous     12753
__too_low_aQual 4384
__not_aligned   90
__alignment_not_unique  167805
```

featureCounts

- featureCounts runs much faster compared to HT-Seq counts.
- It takes the following files as input:
  - A BAM/SAM file
  - An annotation file with chromosome coordinates of features (typically a GTF file).
- With default parameters both HT-Seq and featureCounts should give the same output for single-end reads.

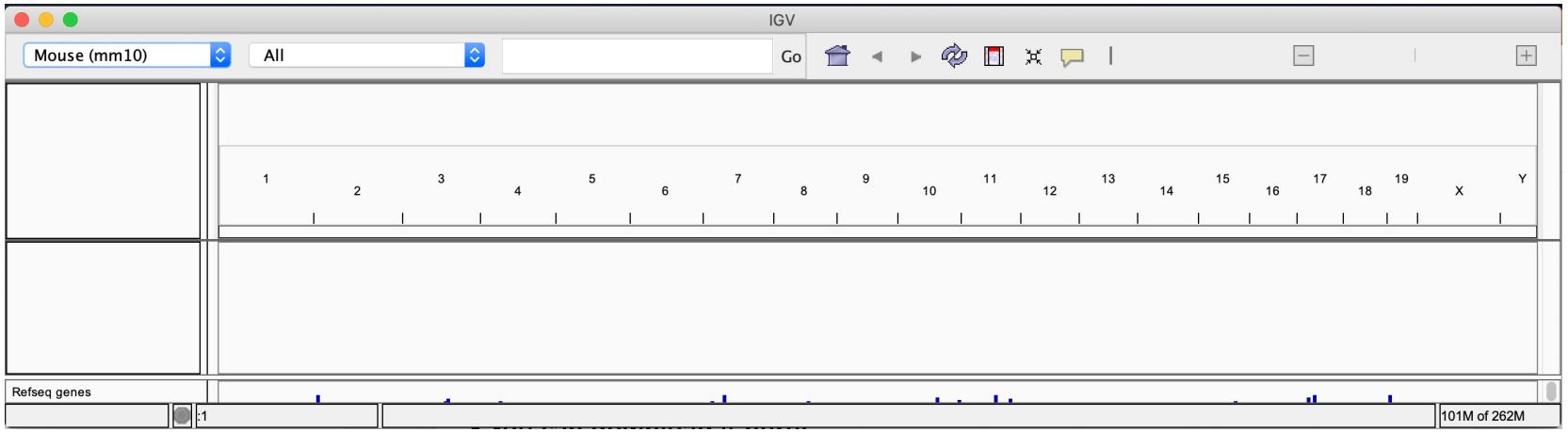
Example featureCount File



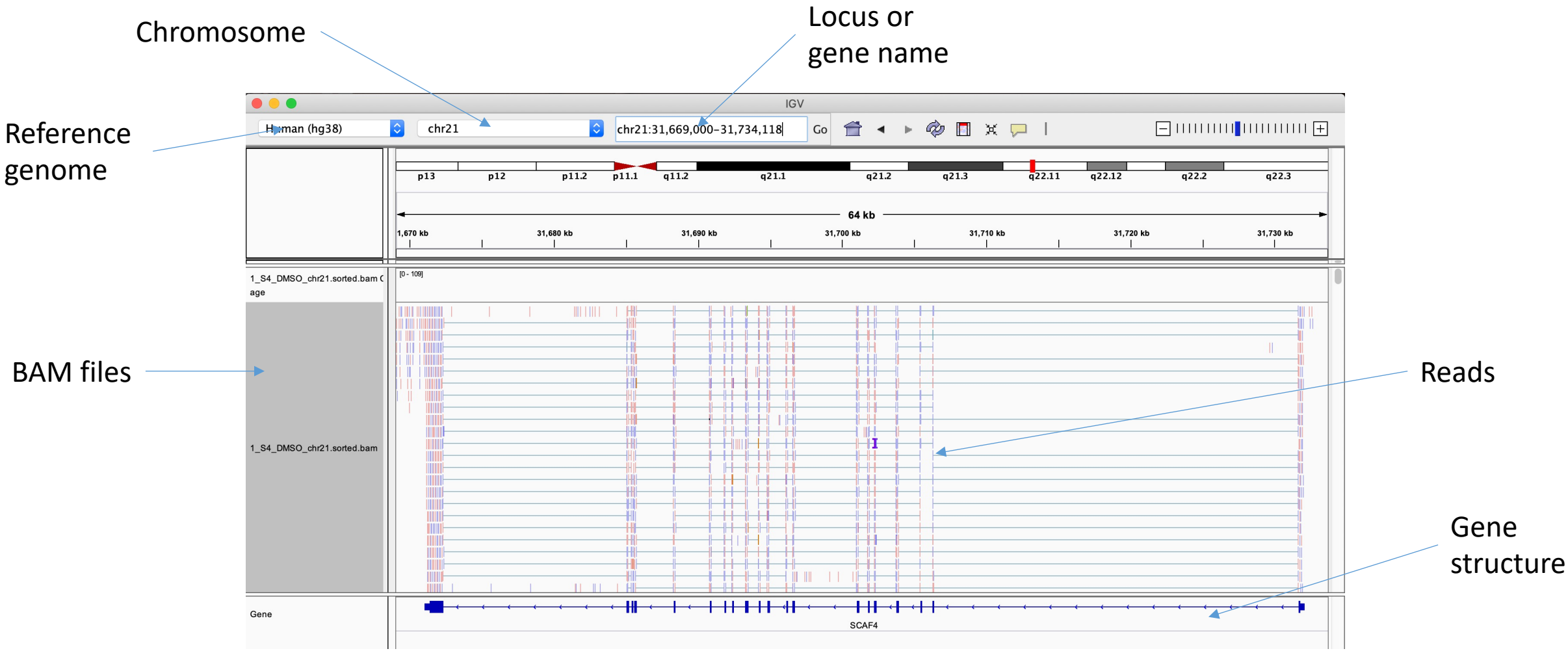
IGV Browser I

- The **Integrative Genomics Viewer (IGV)** is a high-performance visualization tool for interactive exploration of large, integrated genomic datasets.

- We can use this tool visualize our BAM files.



IGV Browser II

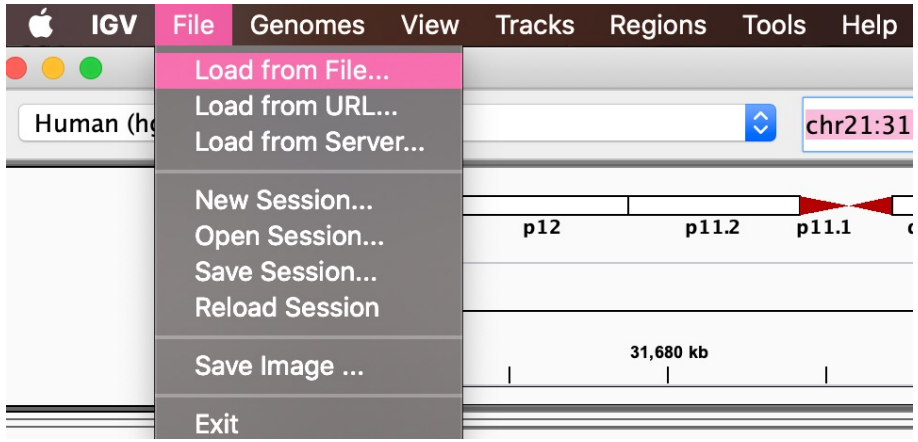


!! BAM files need to be indexed before loading them in IGV browser

Tool Demo: IGV (BAMs)

- You can download IGV browser from this link: https://software.broadinstitute.org/software/igv/download
- Open FileZilla, to transfer your BAM files (.bam and .bai)
- Open IGV browser and load your reference genome.

- Load you BAM file into IGV browser



Breakout Session-3

Table of contents

Concluding Remarks

- It’s important to review output from each step carefully.
- For e.g.,
  - Are the reads of good quality?
  - Did the trimming work?
  - What is the percentage of mapping?
  - What does the HT seq count look like: no features, no aligned
- After we get the counts, we perform differential expression analysis using Bioconductor R packages like edgeR, DESeq etc.

Thank you!