Gene Quantification

Bulk RNAseq course 2024

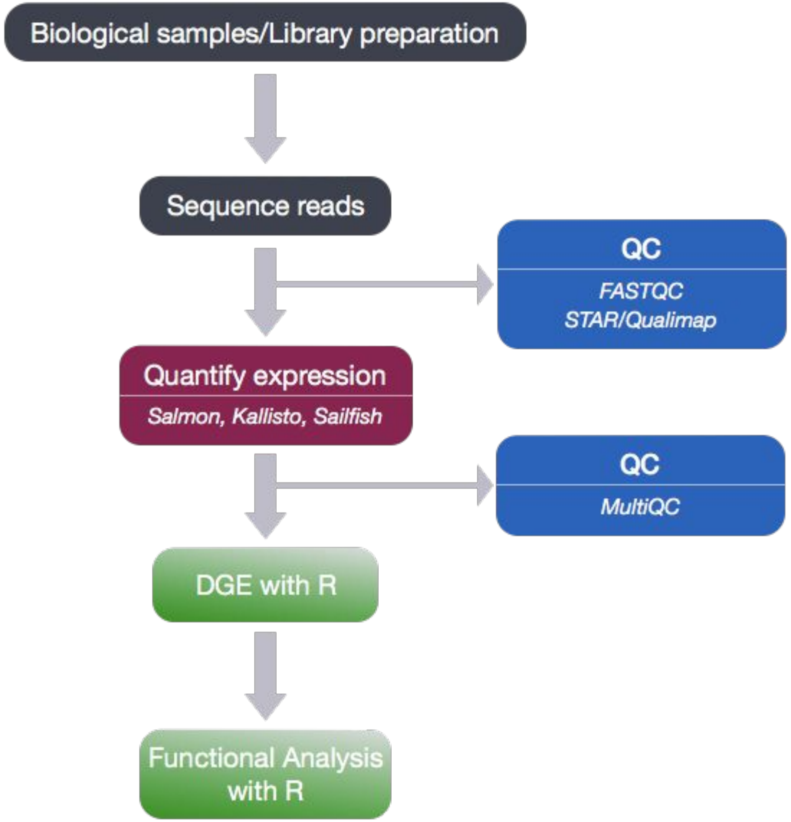
Duy Dao
khuongduying@gmail.com

RNA-SEQ: STEP IN QUANTIFICATION

Definition: Measuring the abundance of transcripts for each gene in a sample

Key Processes:

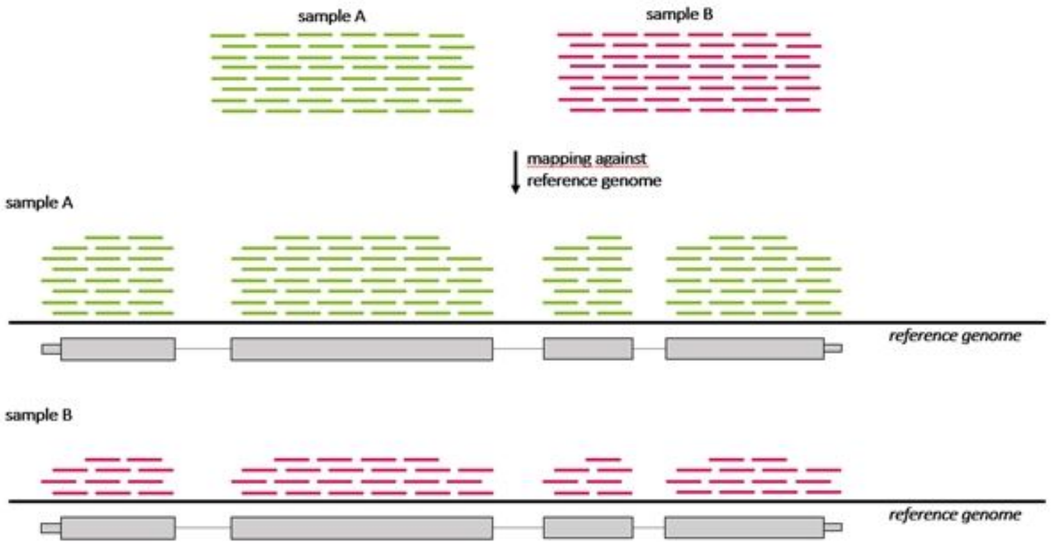
- Read Alignment or Alignment-Free Mapping
- Assigning Reads to Genes
- Counting Reads per Gene


Biological samples/Library preparation
Sequence reads
QC
FASTQC
STAR/Qualimap
Quantify expression
Salmon, Kallisto, Sailfish
QC
MultiQC
DGE with R
Functional Analysis with R

https://hbctraining.github.io/rnaseq-cb321/lessons/analysis_methods.html

RNA-SEQ: QUANTIFICATION

Quantification - Read Count



Count how many reads have mapped to each gene.

→Using the **featureCounts** tool to get the gene counts

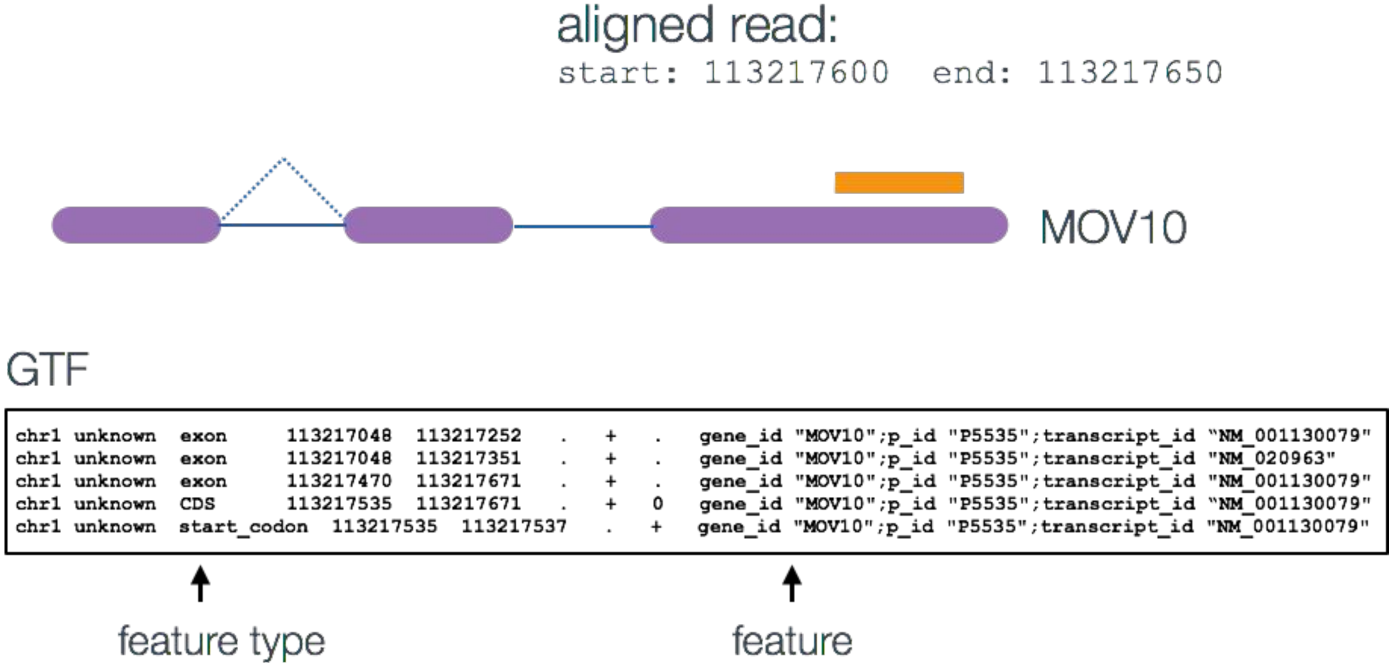
Input: BAM + GTF

Output: Number of reads (counts) associated with each feature of interest (genes, exons, transcript, etc.).

RNA-SEQ: QUANTIFICATION

Counting reads with featureCounts

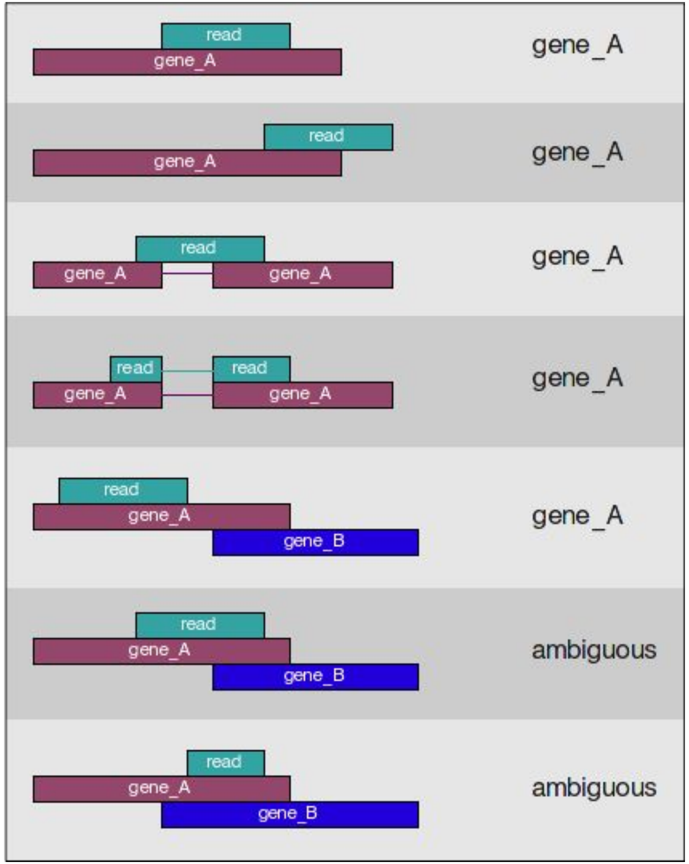
- Accurate, fast and is relatively easy to use
- Counts reads that map to a single location (uniquely mapping) and follows the scheme in the figure below for assigning reads to a gene/exon.



```
chr1 unknown  exon       113217048  113217252   .    +    .   gene_id "MOV10";p_id "P5535";transcript_id "NM_001130079"
chr1 unknown  exon       113217048  113217351   .    +    .   gene_id "MOV10";p_id "P5535";transcript_id "NM_020963"
chr1 unknown  exon       113217470  113217671   .    +    .   gene_id "MOV10";p_id "P5535";transcript_id "NM_001130079"
chr1 unknown  CDS        113217535  113217671   .    +    0   gene_id "MOV10";p_id "P5535";transcript_id "NM_001130079"
chr1 unknown  start_codon  113217535  113217537   .    +   gene_id "MOV10";p_id "P5535";transcript_id "NM_001130079"
```

RNA-SEQ: QUANTIFICATION

Counting reads using featureCounts



- A read is said to overlap a feature if at least one read base is found to overlap the feature.

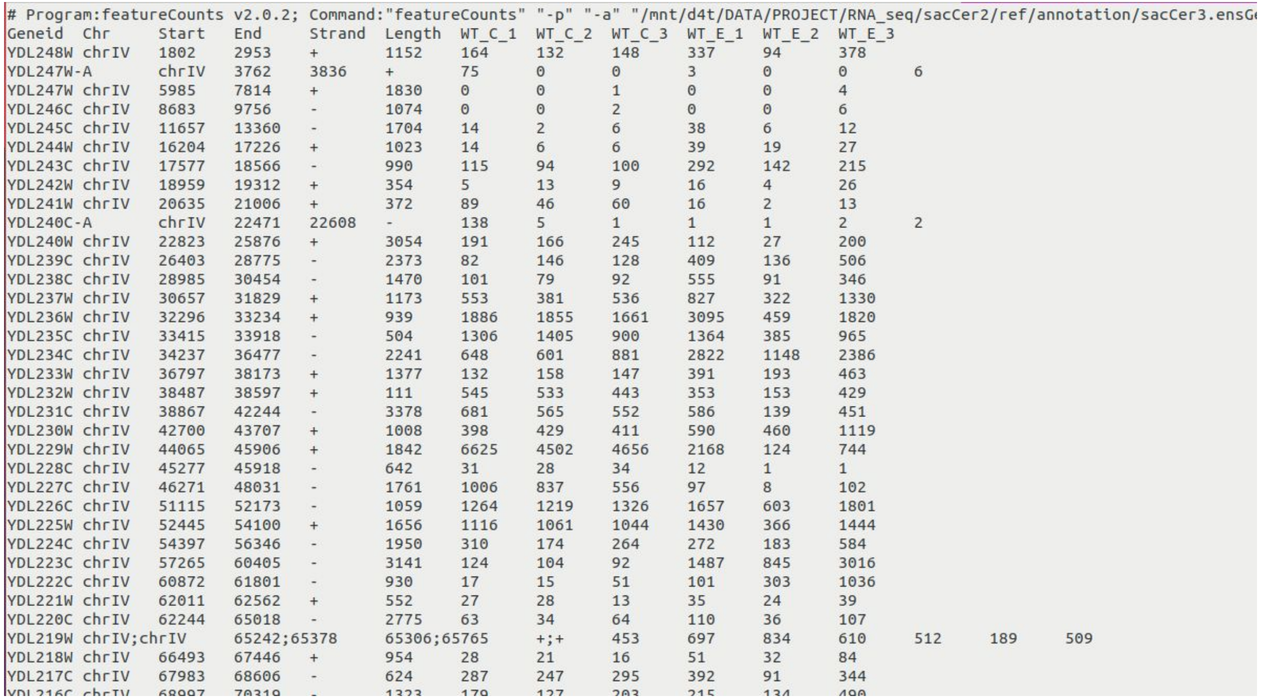
- For paired-end data, a fragment (or template) is said to overlap a feature if any of the two reads from that fragment is found to overlap the feature.

- If strandedness is specified, then in addition to considering the genomic coordinates it will also take the strand into account for counting.

RNA-SEQ: QUANTIFICATION

Counting reads using featureCounts

gene | Location | Strand | Length | Count



```
# Program:featureCounts v2.0.2; Command:"featureCounts" "-p" "-a" "/mnt/d4t/DATA/PROJECT/RNA_seq/sacCer2/ref/annotation/sacCer3.ensG
Geneid  Chr     Start   End     Strand  Length  WT_C_1  WT_C_2  WT_C_3  WT_E_1  WT_E_2  WT_E_3
YDL248W chrIV   1802    2953    +       1152    164     132     148     337     94      378
YDL247W-A       chrIV   3762    3836    +       75      0       0       3       0       0       6
YDL247W chrIV   5985    7814    +       1830    0       0       1       0       0       4
YDL246C chrIV   8683    9756    -       1074    0       0       2       0       0       6
YDL245C chrIV   11657   13360   -       1704    14      2       6       38      6       12
YDL244W chrIV   16204   17226   +       1023    14      6       6       39      19      27
YDL243C chrIV   17577   18566   -       990     115     94      100     292     142     215
YDL242W chrIV   18959   19312   +       354     5       13      9       16      4       26
YDL241W chrIV   20635   21006   +       372     89      46      60      16      2       13
YDL240C-A       chrIV   22471   22608   -       138     5       1       1       1       2       2
YDL240W chrIV   22823   25876   +       3054    191     166     245     112     27      200
YDL239C chrIV   26403   28775   -       2373    82      146     128     409     136     506
YDL238C chrIV   28985   30454   -       1470    101     79      92      555     91      346
YDL237W chrIV   30657   31829   +       1173    553     381     536     827     322     1330
YDL236W chrIV   32296   33234   +       939     1886    1855    1661    3095    459     1820
YDL235C chrIV   33415   33918   -       504     1306    1405    900     1364    385     965
YDL234C chrIV   34237   36477   -       2241    648     601     881     2822    1148    2386
YDL233W chrIV   36797   38173   +       1377    132     158     147     391     193     463
YDL232W chrIV   38487   38597   +       111     545     533     443     353     153     429
YDL231C chrIV   38867   42244   -       3378    681     565     552     586     139     451
YDL230W chrIV   42700   43707   +       1008    398     429     411     590     460     1119
YDL229W chrIV   44065   45906   +       1842    6625    4502    4656    2168    124     744
YDL228C chrIV   45277   45918   -       642     31      28      34      12      1       1
YDL227C chrIV   46271   48031   -       1761    1006    837     556     97      8       102
YDL226C chrIV   51115   52173   -       1059    1264    1219    1326    1657    603     1801
YDL225W chrIV   52445   54100   +       1656    1116    1061    1044    1430    366     1444
YDL224C chrIV   54397   56346   -       1950    310     174     264     272     183     584
YDL223C chrIV   57265   60405   -       3141    124     104     92      1487    845     3016
YDL222C chrIV   60872   61801   -       930     17      15      51      101     303     1036
YDL221W chrIV   62011   62562   +       552     27      28      13      35      24      39
YDL220C chrIV   62244   65018   -       2775    63      34      64      110     36      107
YDL219W chrIV;chrIV     65242;65378     65306;65765     +;+     453     697     834     610     512     189     509
YDL218W chrIV   66493   67446   +       954     28      21      16      51      32      84
YDL217C chrIV   67983   68606   -       624     287     247     295     392     91      344
YDL216C chrIV   68997   70319   -       1323    179     127     203     215     134     490
```

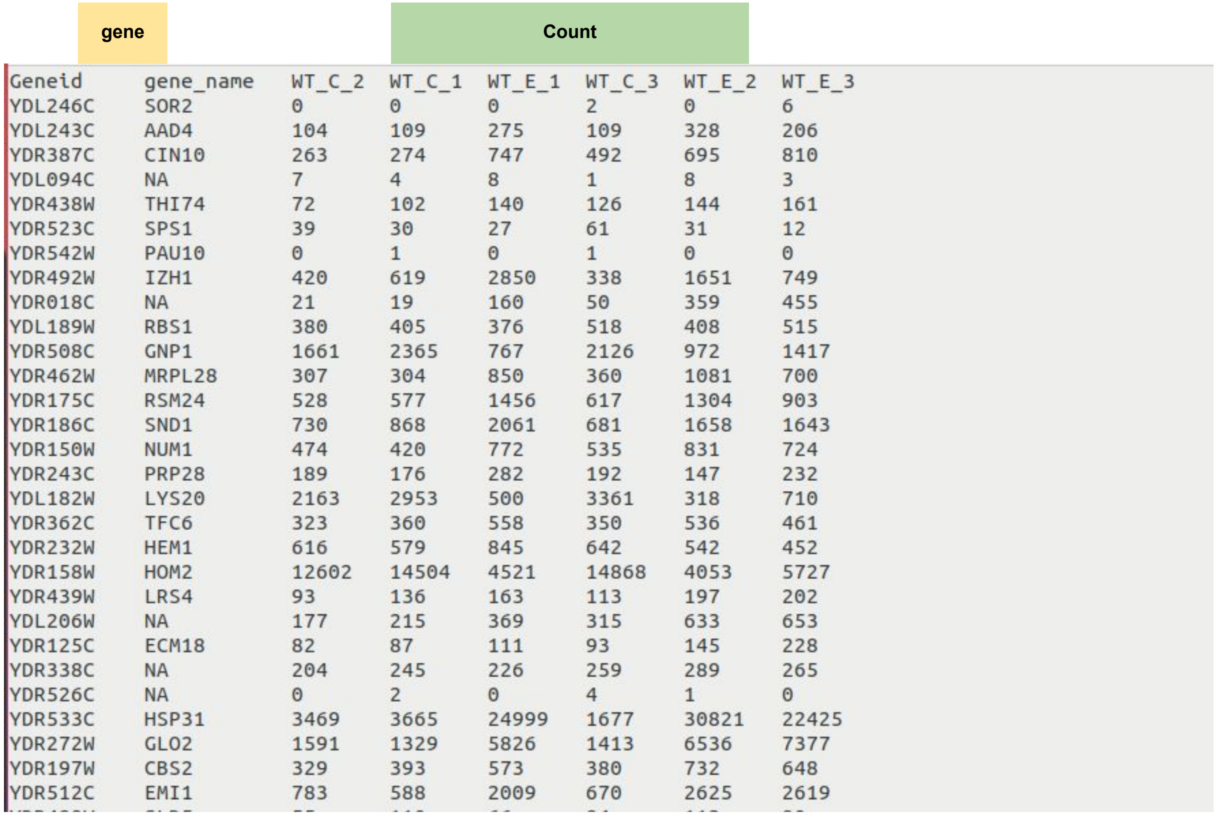
featureCounts output

Output: Raw counts

These are the “raw” counts will be used in statistical programs downstream for differential gene expression.

RNA-SEQ: QUANTIFICATION

Counting reads using featureCounts



Geneid	gene_name	WT_C_2	WT_C_1	WT_E_1	WT_C_3	WT_E_2	WT_E_3
YDL246C	SOR2	0	0	0	2	0	6
YDL243C	AAD4	104	109	275	109	328	206
YDR387C	CIN10	263	274	747	492	695	810
YDL094C	NA	7	4	8	1	8	3
YDR438W	THI74	72	102	140	126	144	161
YDR523C	SPS1	39	30	27	61	31	12
YDR542W	PAU10	0	1	0	1	0	0
YDR492W	IZH1	420	619	2850	338	1651	749
YDR018C	NA	21	19	160	50	359	455
YDL189W	RBS1	380	405	376	518	408	515
YDR508C	GNP1	1661	2365	767	2126	972	1417
YDR462W	MRPL28	307	304	850	360	1081	700
YDR175C	RSM24	528	577	1456	617	1304	903
YDR186C	SND1	730	868	2061	681	1658	1643
YDR150W	NUM1	474	420	772	535	831	724
YDR243C	PRP28	189	176	282	192	147	232
YDL182W	LYS20	2163	2953	500	3361	318	710
YDR362C	TFC6	323	360	558	350	536	461
YDR232W	HEM1	616	579	845	642	542	452
YDR158W	HOM2	12602	14504	4521	14868	4053	5727
YDR439W	LRS4	93	136	163	113	197	202
YDL206W	NA	177	215	369	315	633	653
YDR125C	ECM18	82	87	111	93	145	228
YDR338C	NA	204	245	226	259	289	265
YDR526C	NA	0	2	0	4	1	0
YDR533C	HSP31	3469	3665	24999	1677	30821	22425
YDR272W	GLO2	1591	1329	5826	1413	6536	7377
YDR197W	CBS2	329	393	573	380	732	648
YDR512C	EMI1	783	588	2009	670	2625	2619

A table of counts

Don’t need information about the genomic coordinates, length

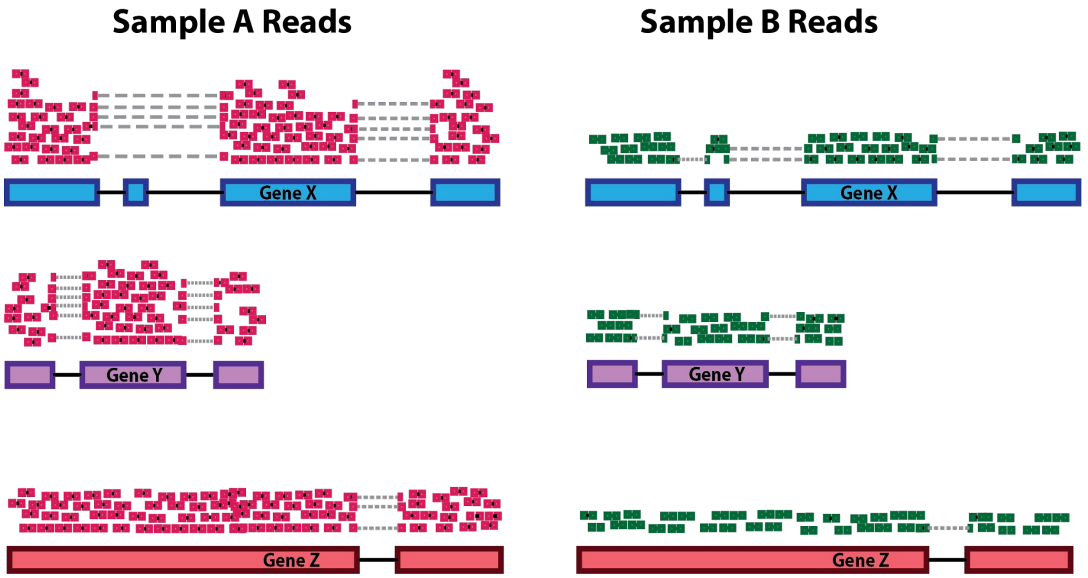
→ Cleaning up the featureCounts matrix

Final output:
A count matrix, with genes as rows and samples are columns

RNA-SEQ: QUANTIFICATION

Normalization

Sequencing depth: Accounting for sequencing depth is necessary for comparison of gene expression between samples.

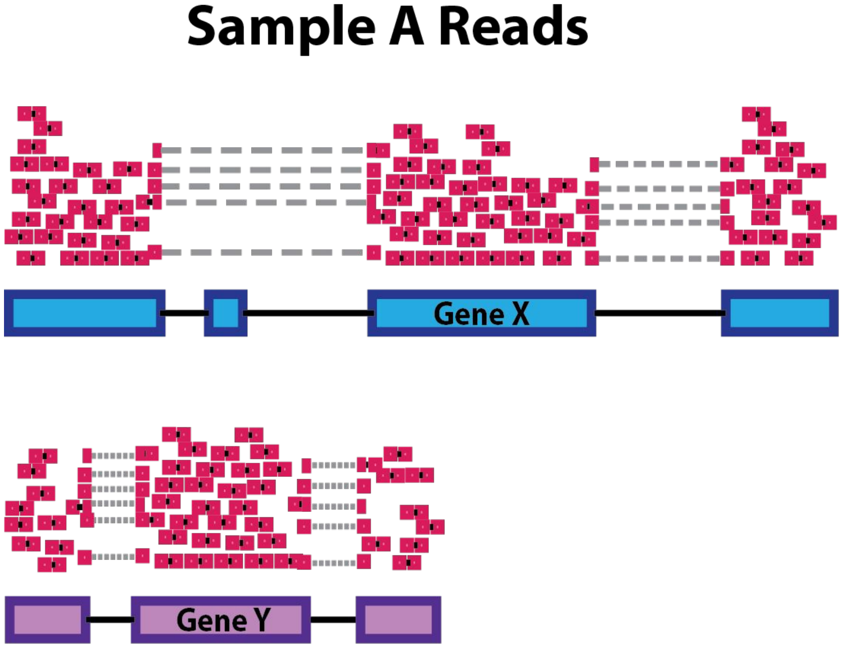


In the example below, each gene appears to have doubled in expression in Sample A relative to Sample B, however this is a consequence of Sample A having double the sequencing depth.

RNA-SEQ: QUANTIFICATION

Normalization

Gene length: Accounting for gene length is necessary for comparing expression between different genes within the same sample.

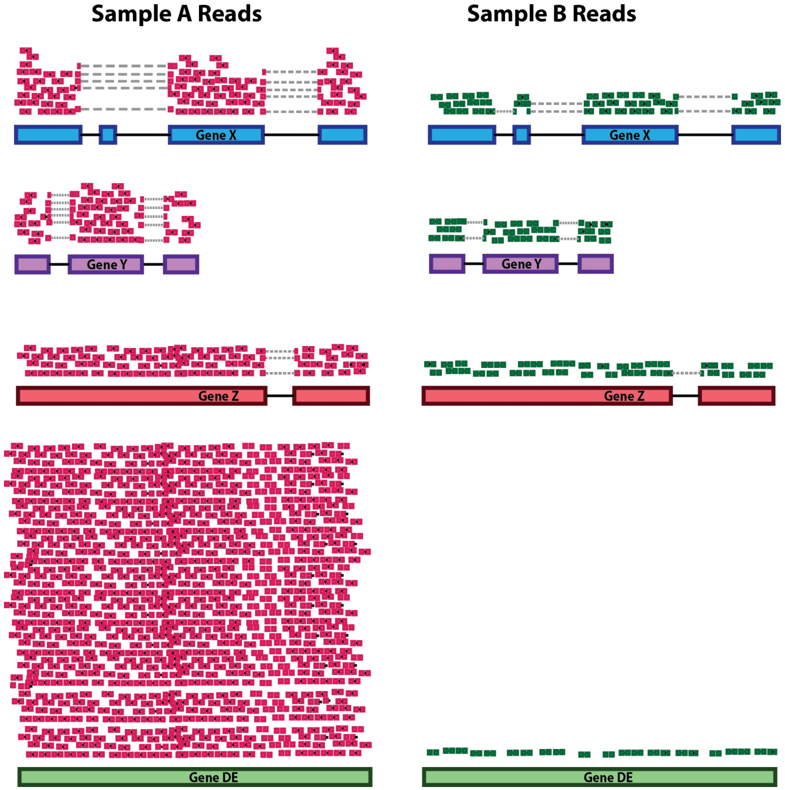


In the example, Gene X and Gene Y have similar levels of expression, but the number of reads mapped to Gene X would be many more than the number mapped to Gene Y because Gene X is longer.

RNA-SEQ: QUANTIFICATION

Normalization

RNA composition: A few highly differentially expressed genes between samples, differences in the number of genes expressed between samples, or presence of contamination can skew some types of normalization methods.


Sample A Reads
Sample B Reads
Gene X
Gene X
Gene Y
Gene Y
Gene Z
Gene Z
Gene DE
Gene DE