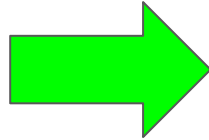


Introduction to RNA-seq

Overview & Alignment

Why do RNA-seq?

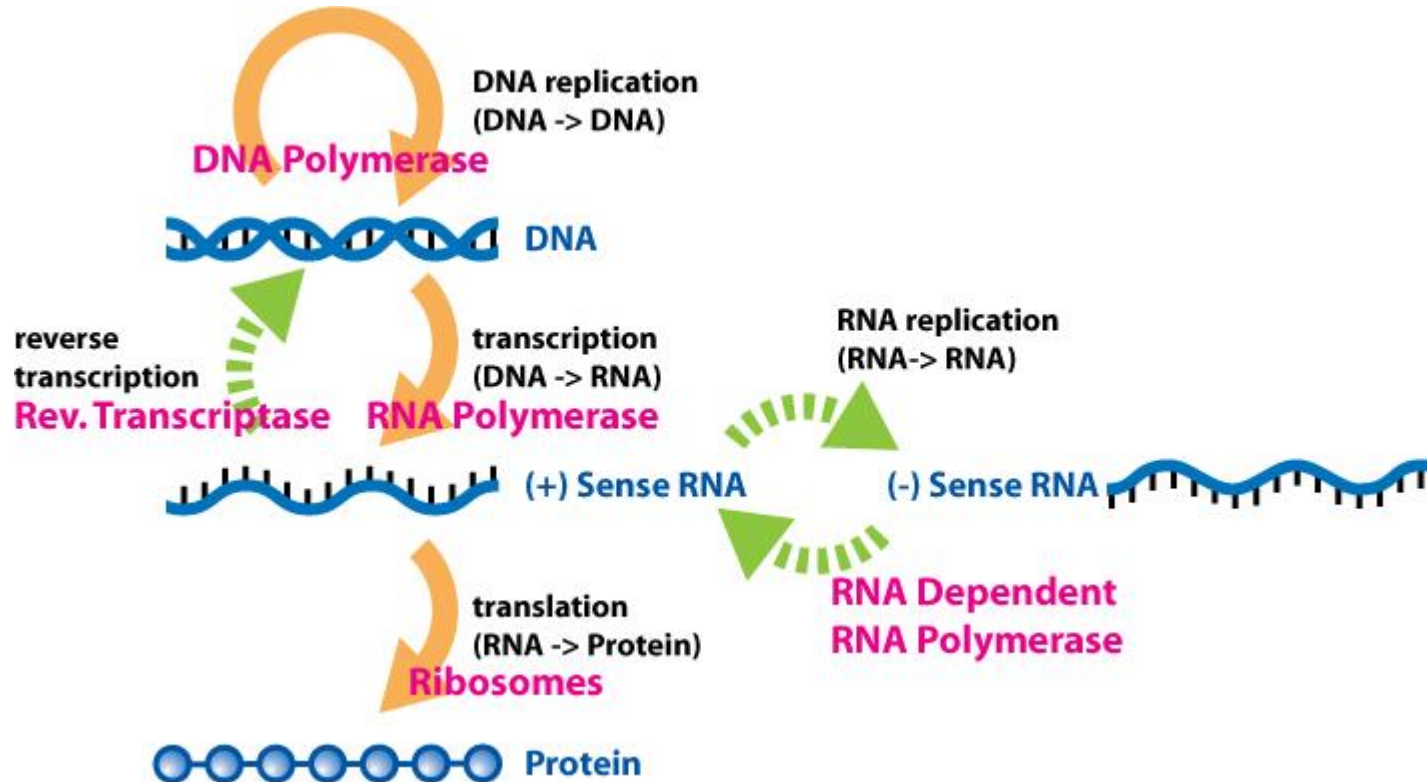
- Mapping from genotype-to-phenotype
- Annotation of genes and transcripts
- Tissue biology
- Gene regulatory networks



Count Matrix

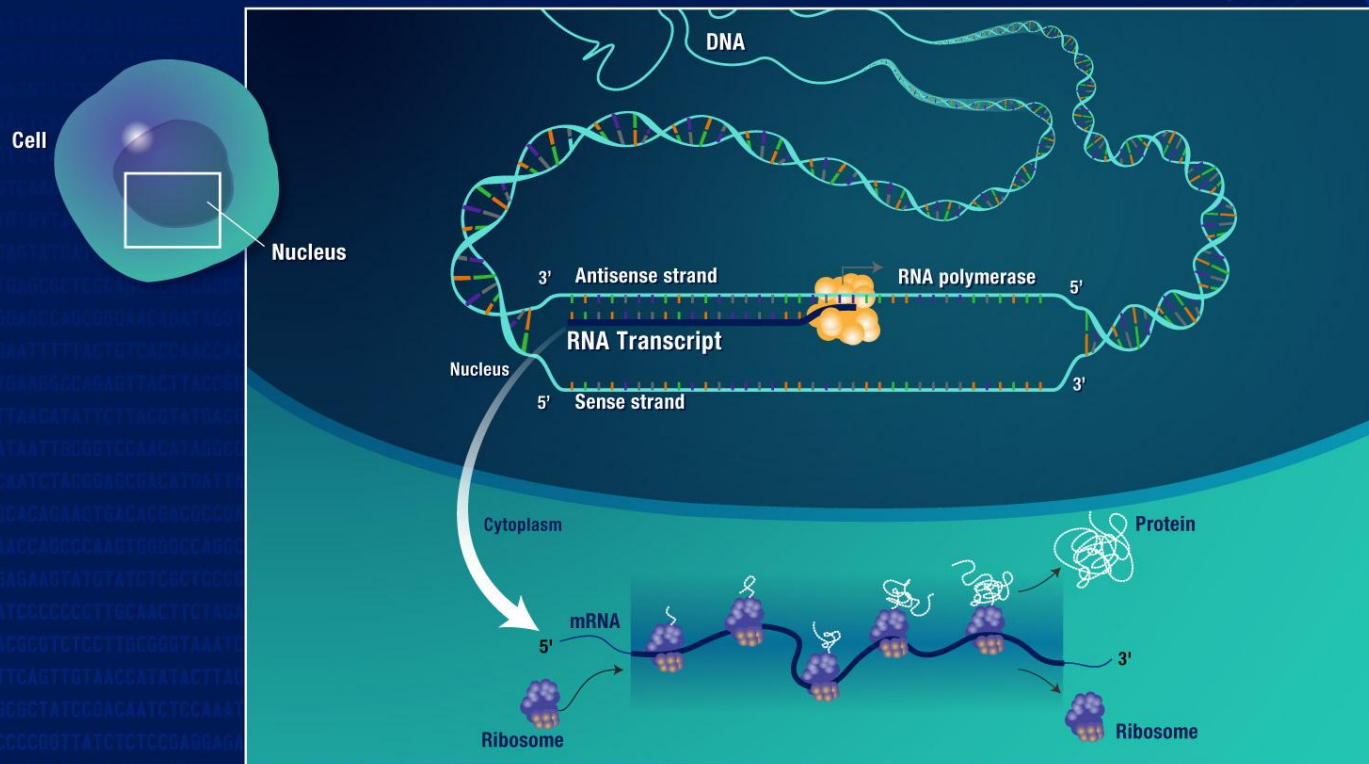
	WT1	treatment1	...
gene1	50	15	
gene2	20	87	
...			

Central dogma



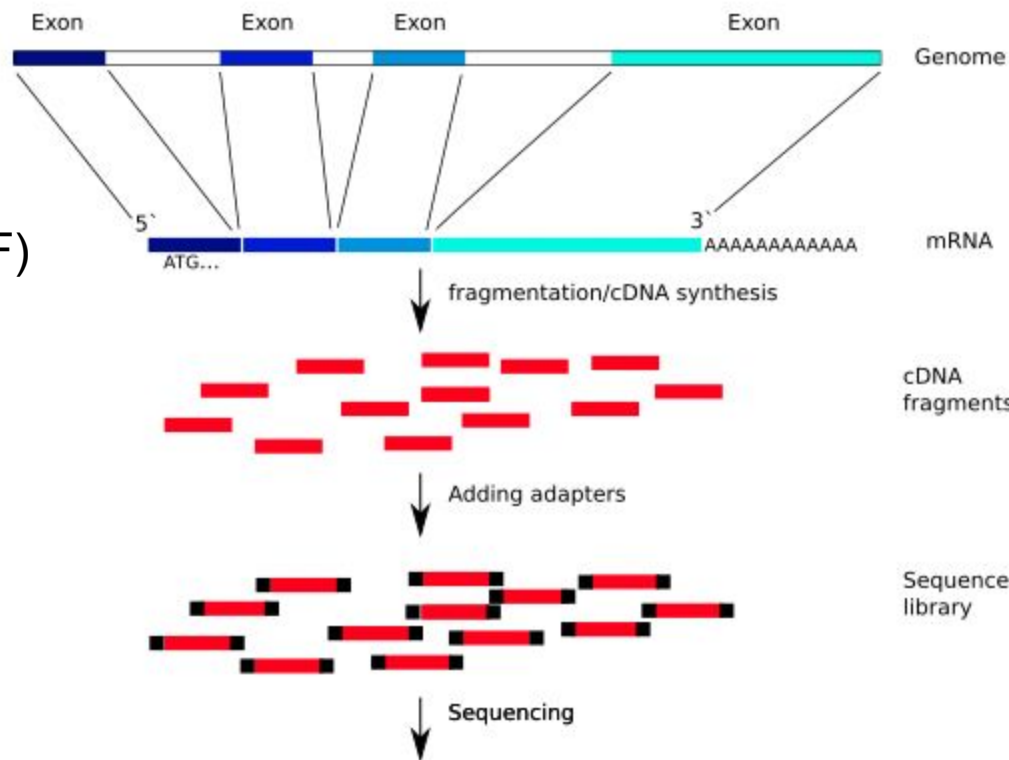
Transcriptome

NHGRI FACT SHEETS
genome.gov



Reference (FASTA)

Annotation (GTF)



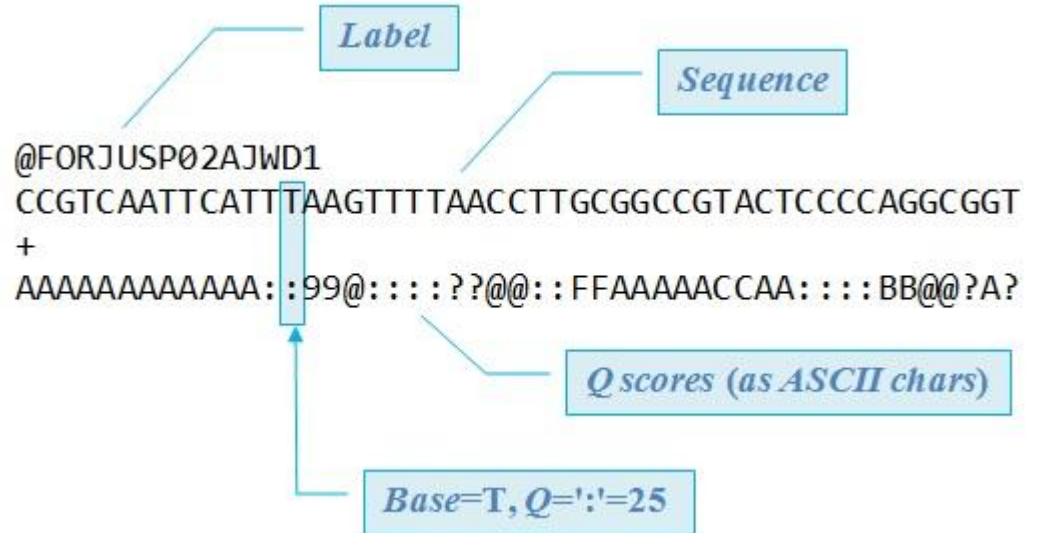
```
@HWI-ST100R:6:73:941:1973#0/1
AGTCCCGTCGAAGCTCGATTTCATCAGTTTATATATTTAT
+
!''*(((****))888++)(888).1**CCCCFF>>>>
@HWI-ST100R:6:73:941:1974#0/1
TCACCCGTCGAAAAAAATTCATCAGTTTATATATAAAA
+
!''*(((****))888+CCCCCCCC>>>>CCFF>FFF
```

```
@HWI-ST100R:6:73:941:1973#0/2
AGTCCCGTCGAAGCTCGATTTCATCAGTTTATATATTTAT
+
!''*(((****))888++)(888).1**CCCCFF>>>>
@HWI-ST100R:6:73:941:1974#0/2
TCACCCGTCGAAAAAAATTCATCAGTTTATATATAAAA
+
!''*(((****))888+CCCCCCCC>>>>CCFF>FFF
```

Fastq files

FASTQ format

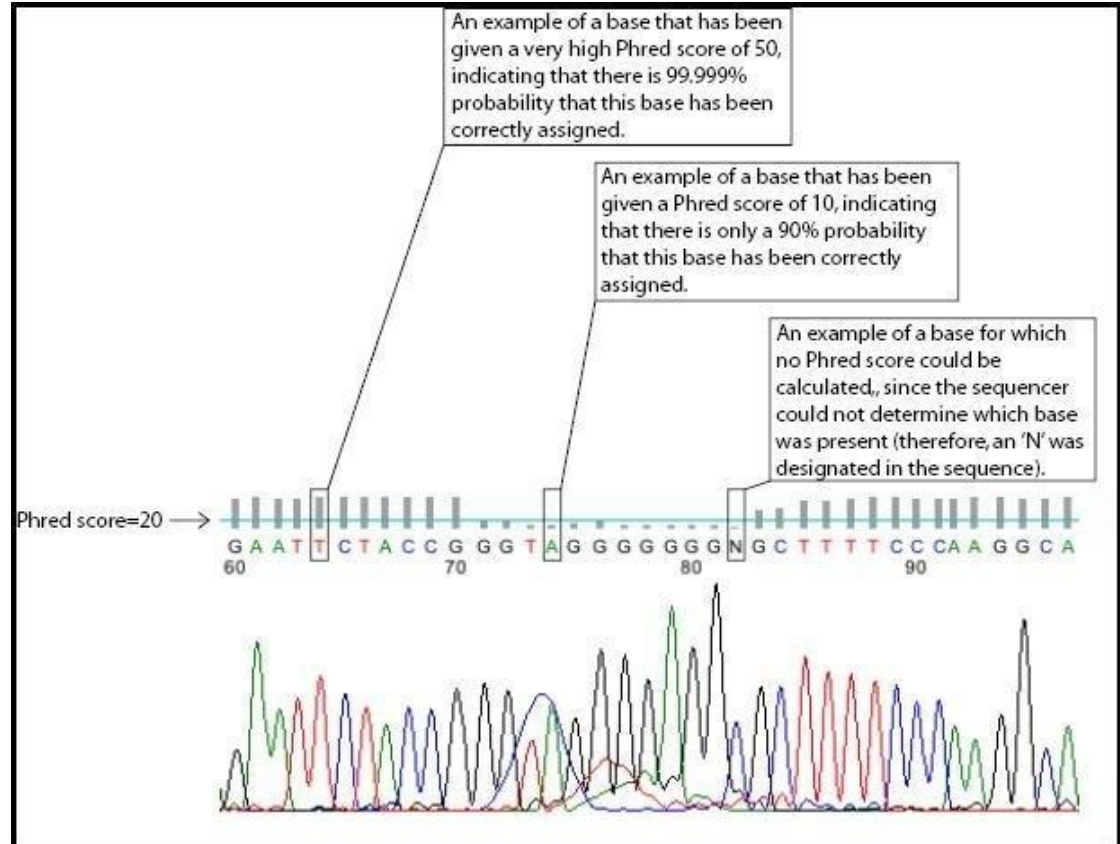
1. Read ID + properties
2. Read sequence
3. (space)
4. Base qualities



Phred quality score

$$Q = -10 \cdot \log_{10}(P)$$

- P is the probability of incorrect base call
- Q is the measure for probability of correct base call



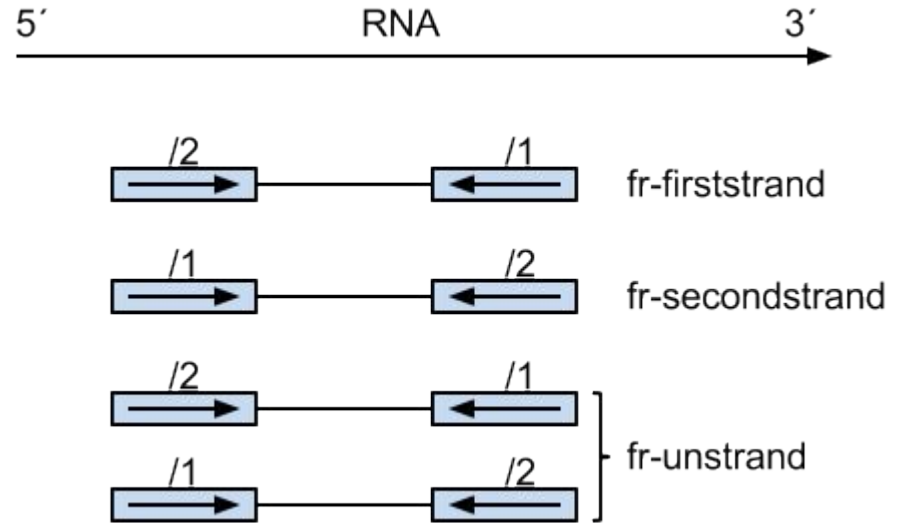
Strandness: single-end vs paired-end

Single-end

- F: sense orientation
- R: antisense orientation

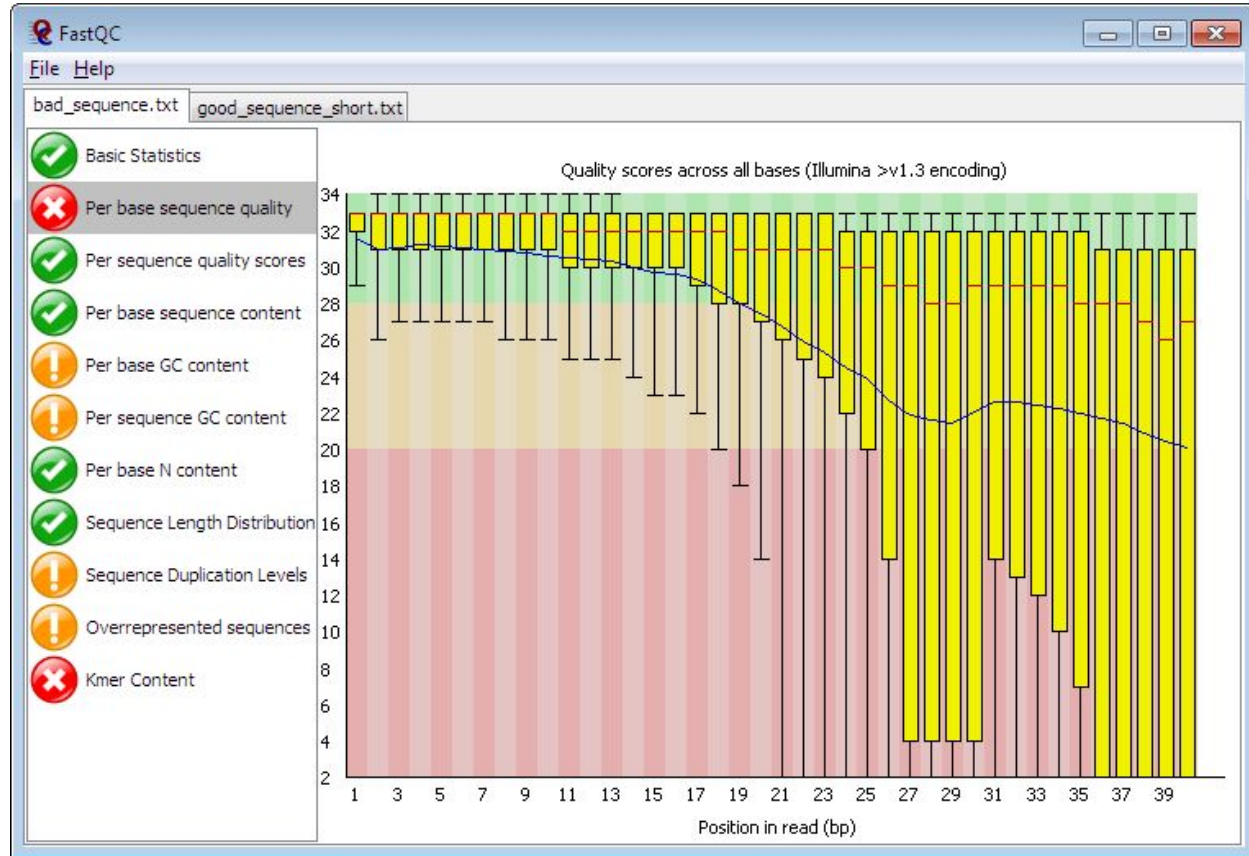
Paired-end

- RF: first read is R, second F
- FR: first read is F, second R



Quality control

- FastQC offers a visual check on raw sequence data
- Per base sequence quality plot
- Distribution of quality scores at each read position across all reads



Read filtering

[Trimmomatic](#) and [SOAPnuke](#) are commonly used for filtering poor quality reads

- Sequence adapters
- Low quality base rate
- Unknown (N) base rate

Alignment

Types of alignment

Reference-based analysis

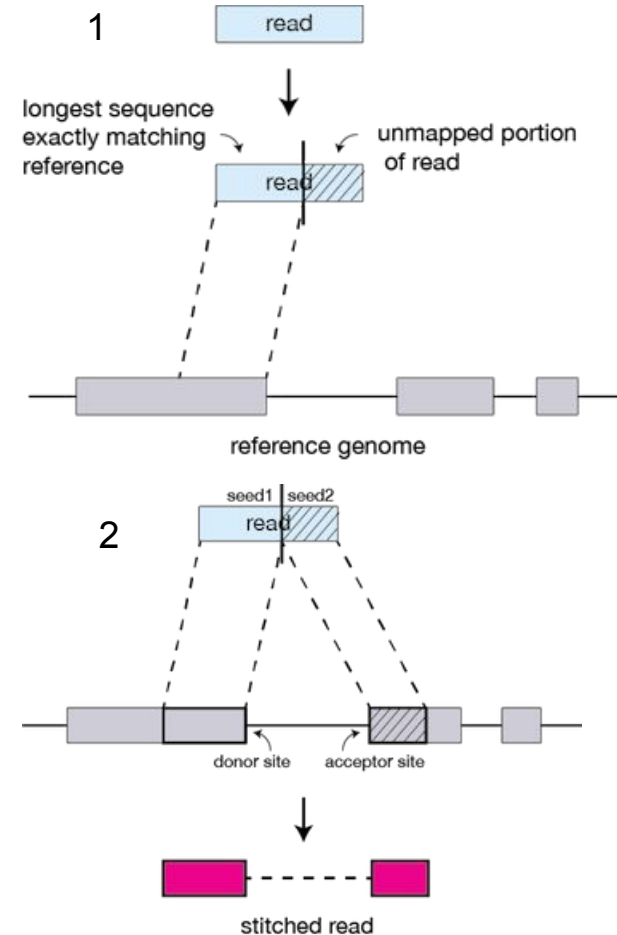
- Mapping to a reference genome
 - Aligners: find the location of reads on the genome
e.g. [STAR](#) and [HISAT2](#)
 - “pseudo-alignment” quantifiers: find transcript counts based on approximate location
e.g. [Salmon](#) and [Kallisto](#)

de-novo assembly

- Building a reference transcriptome
- Contigs resemble RNA transcripts

STAR

1. Seed searching: find longest sequence that matches 1+ locations on reference genome
2. Clustering, stitching, and scoring: seeds are grouped together by anchor and then by score (mismatch, indel, gap)



Sample STAR command

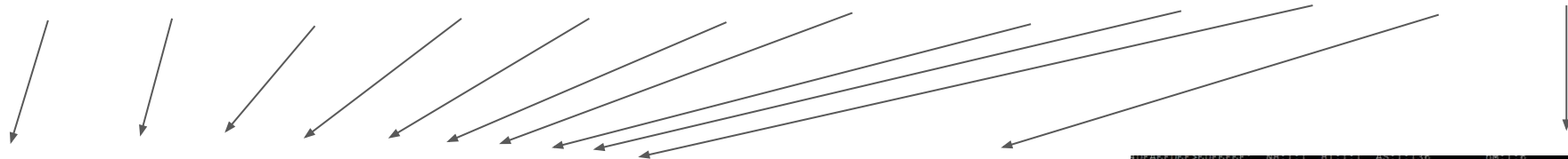
```
STAR --genomeDir genome_index \  
--runThreadN 4 \  
--readFilesIn sample1.fastq \  
--quantMode GeneCounts \  
--outFileNamePrefix STAR/aligned \  
--outSAMattributes Standard \  

```

Sample STAR output

SAM file format

QNAME	FLAG	RNAME	POS	MAPQ	CIGAR	RNEXT	RNEXT	PNEXT	TLEN	SEQ	QUAL
-------	------	-------	-----	------	-------	-------	-------	-------	------	-----	------



RR12095892..92110	16	2	215404541	255	115M1582N35M	* * 0	0	GTAAAGCCACCGGTCACTCGGTAT
RR12095892..92111	0	2	215308026	255	150M *	0	0	CAATCAGGGGCTGGCTTCATATCATCGTGCT
RR12095892..92112	16	2	215375245	255	150M *	0	0	GGGCGGAGAGGACACTTCTCTGGGAGGAGACT
RR12095892..92113	16	11	57818166	255	150M *	0	0	CTCATCTGCTCTCTGCTCTTTCTCTGGTGCT
RR12095892..92114	16	10	88935238	255	129M2694N21M	* * 0	0	GTGGACACATGGAAGCGCCGGCTGCT
RR12095892..92115	16	5	149981484	255	150M *	0	0	ATCCAGAGTGAAGATGTCACTGCACATCTTCCA
RR12095892..92116	16	20	353675	255	150M *	0	0	CCCCCTTCTCACATAACCCCTCTCAAGTTTCCCCAAC
RR12095892..92117	0	15	4461291	255	150M *	0	0	CTTAGTGCTTAATGTGTAACATGAGGAAACTTG
RR12095892..92118	16	17	8382156	255	150M *	0	0	TCCTTTGGGATGGGCACTGCGACCTGTGTACTTGTGCT
RR12095892..92119	0	2	255308	255	150M *	0	0	ATTTTGTATAAAGTGAATCTGGTGGGTACTCCAGCA
RR12095892..92120	16	16	56609030	255	28M205N122M	* * 0	0	GAGTGCACATGCACTCTCTGTCAG
RR12095892..92121	16	9	99218404	255	25148M	* * 0	0	CTTCTCACAACTCTGGGATGCTCACTCTCCG

[illegible]

Summary statistics

- Uniquely mapped reads
- Mismatch rate
- Deletion rate
- Insertion rate

Mapping speed, Million of reads per hour	1688.14
Number of input reads	45017142
Average input read length	150
UNIQUE READS:	
Uniquely mapped reads number	43729271
Uniquely mapped reads %	97.14%
Average mapped length	149.50
Number of splices: Total	23787911
Number of splices: Annotated (sjdb)	23610473
Number of splices: GT/AG	23580970
Number of splices: GC/AG	134444
Number of splices: AT/AC	17558
Number of splices: Non-canonical	54939
Mismatch rate per base, %	0.39%
Deletion rate per base	0.01%
Deletion average length	1.85
Insertion rate per base	0.01%
Insertion average length	1.76
MULTI-MAPPING READS:	
Number of reads mapped to multiple loci	1005151
% of reads mapped to multiple loci	2.23%
Number of reads mapped to too many loci	8267
% of reads mapped to too many loci	0.02%
UNMAPPED READS:	
Number of reads unmapped: too many mismatches	0
% of reads unmapped: too many mismatches	0.00%
Number of reads unmapped: too short	268027
% of reads unmapped: too short	0.60%
Number of reads unmapped: other	6426
% of reads unmapped: other	0.01%

Gene counts

- STAR can generate gene counts
 - Have to strip genes and left-most counts for each aligned FASTQ file and then merge into count matrix
- The more general approach is to use featureCounts from Subread
 - SAM or BAM files as input

ENSG00000225972	16	9	7
ENSG00000225630	66	27	40
ENSG00000237973	1810	677	1174
ENSG00000278791	0	27	0
ENSG00000229344	13	2	11
ENSG00000240409	0	0	0
ENSG00000248527	22723	8427	14296
ENSG00000198744	32	21	11
ENSG00000268663	0	0	0
ENSG00000284662	0	0	0
ENSG00000229376	0	0	0

Other resources

- RNA-seqlopedia
<https://rnaseq.uoregon.edu>
- Gene ontology reference
<http://geneontology.org>