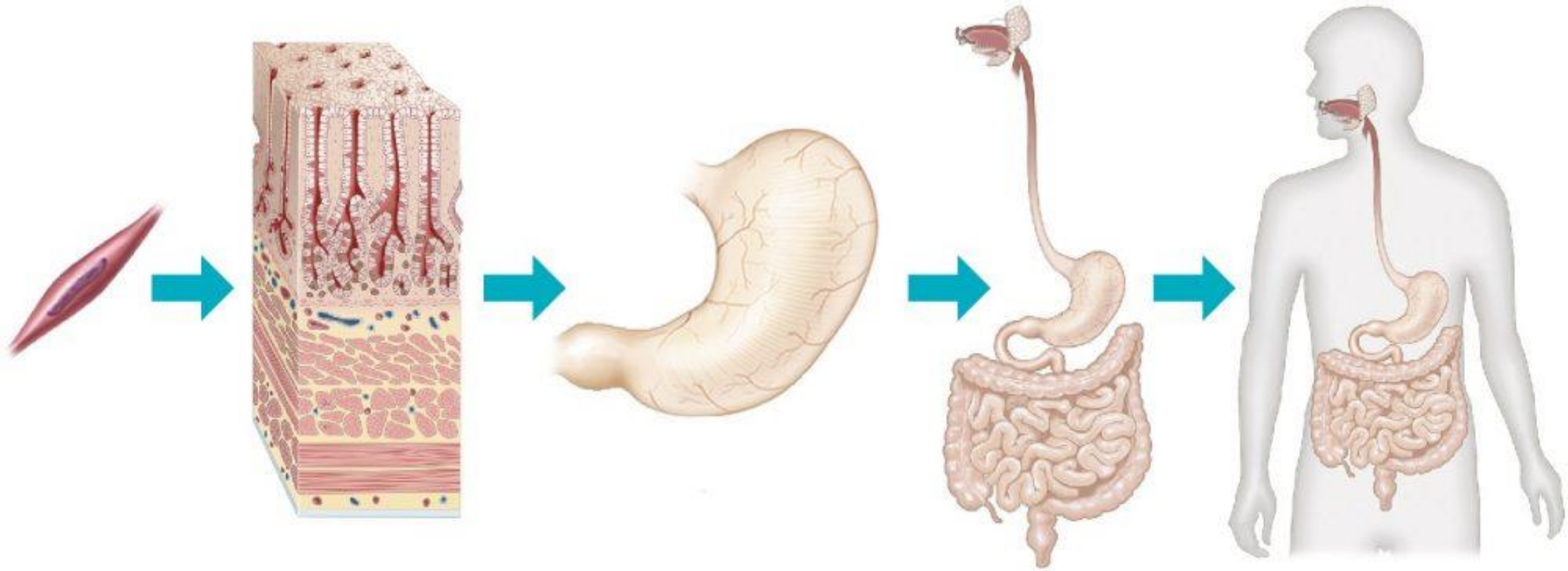
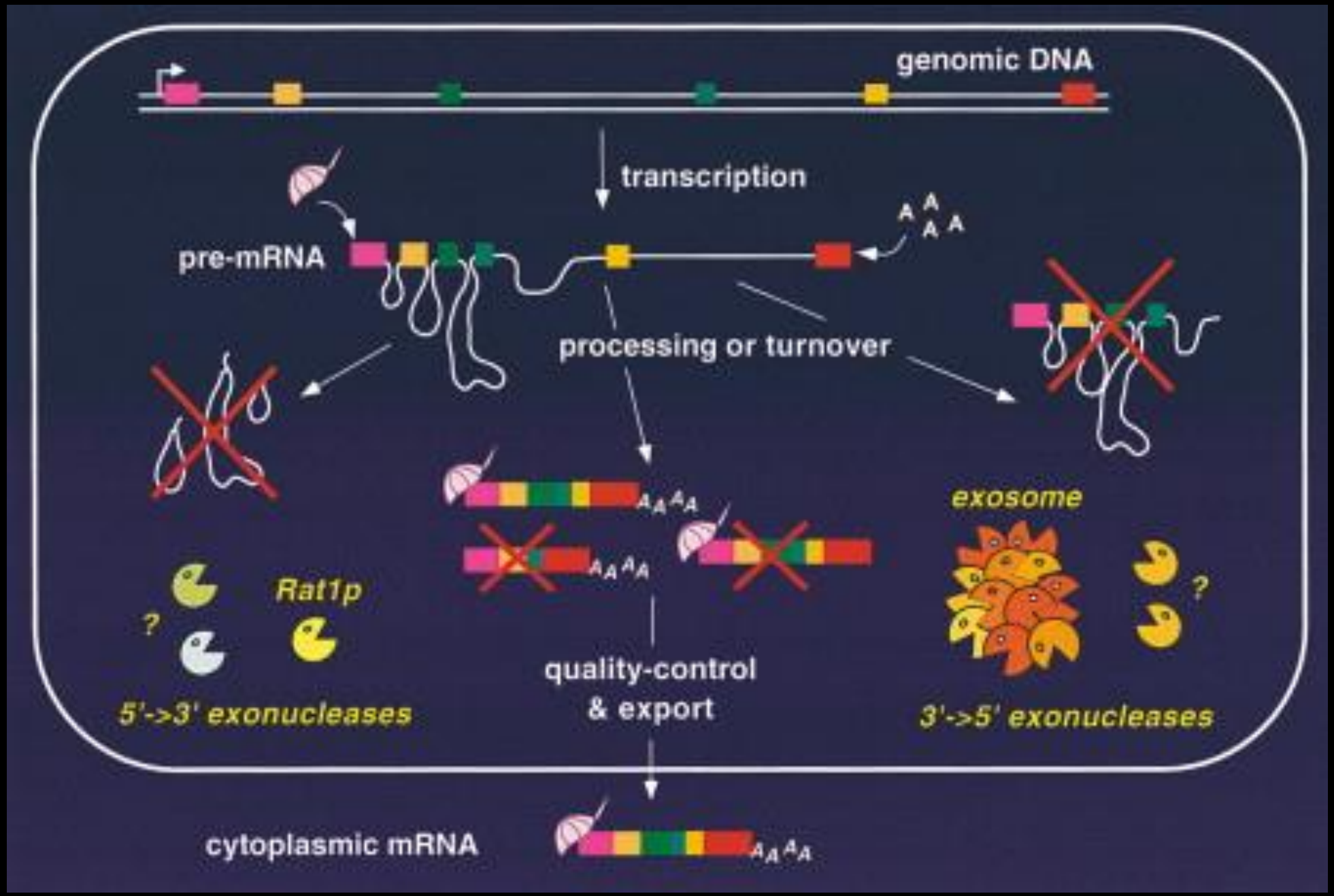
RNA-seq and Genomics File Formats

Levels of organization




genomic DNA
transcription
pre-mRNA
processing or turnover
exosome
Rat1p
?
?
quality-control
& export
5'->3' exonucleases
3'->5' exonucleases
cytoplasmic mRNA

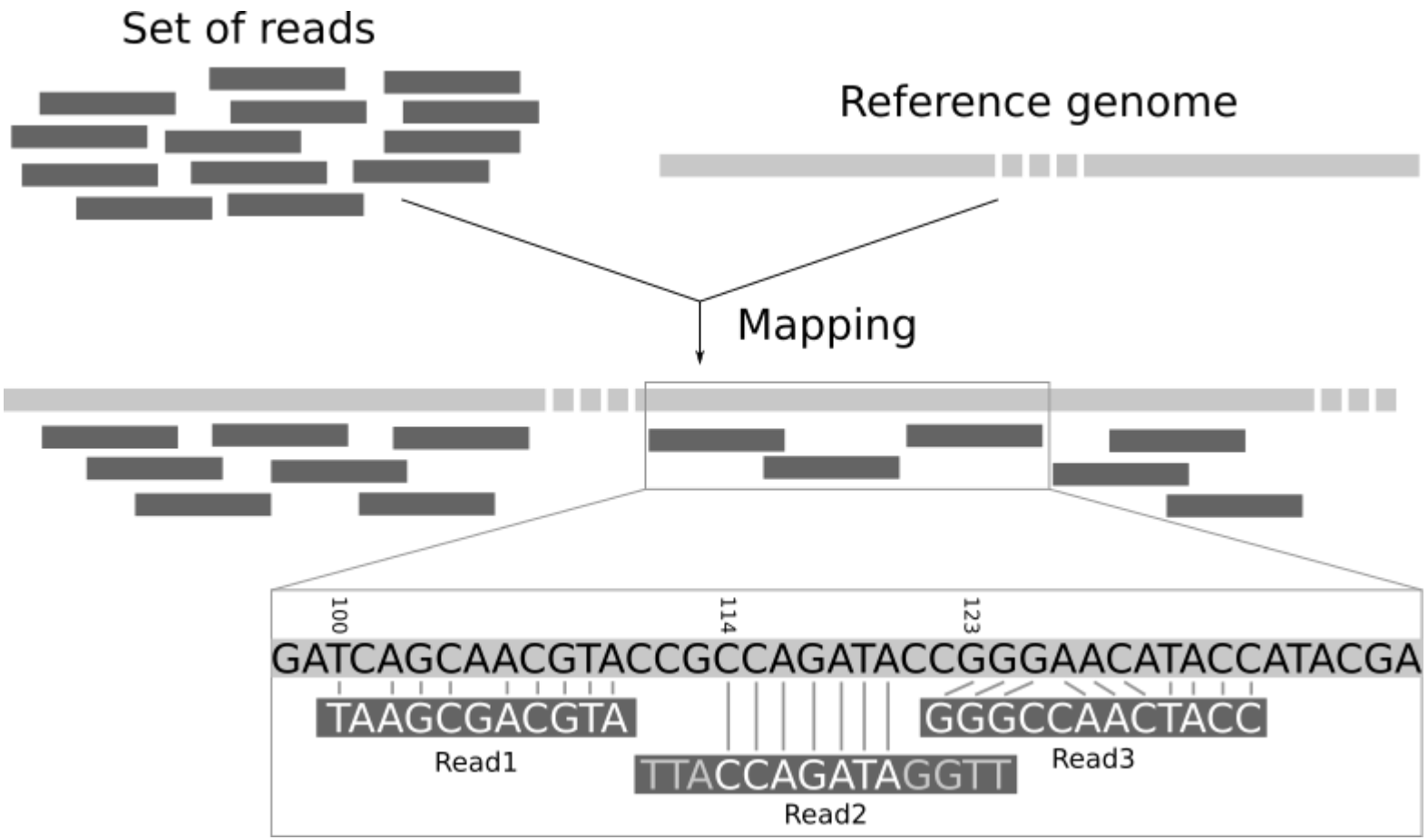
RNA-seq

- Isolate a tissue or single cell
- Extract total RNA
- Use the poly-A tail to enrich for mRNA
- Convert the mRNA to cDNA
- Sequence the cDNA on a next-generation sequencer (Illumina)
- The reads from the sequencer are returned as fastq files.

Quality Control

- Use FASTQC to check fastq files for quality
- Use Trimmomatic to trim
  - Remove adapter sequences
  - Remove low quality reads or parts of reads
- Recheck with FASTQC
- Interpreting FASTQC results: https://hbctraining.github.io/Intro-to-rnaseq-hpc-salmon/lessons/qc_fastqc_assessment.html

Read Mapping (many options)



Normalizing Read Counts

- Longer genes will have more reads because they occupy more of the genome
- Samples with more reads will have higher read counts
- Normalization
  - RPKM: The number of reads normalized by the total read number and length of each transcript
  - FPKM: Takes into account that reads can be paired but otherwise the same is RPKM
  - TPM: Like RPKM/FPKM, but normalized to ensure that the average value is constant across samples (TPM is probably the best to use because it ensures the normalized values sum to the same total value across samples)

Statistical Analysis

Each column is a sample

Each row is a gene

GENE ID	KD.2	KD.3	OE.1	OE.2	OE.3	IR.1	IR.2	IR.3
1/2-SBSRNA4	57	41	64	55	38	45	31	39
A1BG	71	40	100	81	41	77	58	40
A1BG-AS1	256	177	220	189	107	213	172	126
A1CF	0	1	1	0	0	0	0	0
A2LD1	146	81	138	125	52	91	80	50
A2M	10	9	2	5	2	9	8	4
A2ML1	3	2	6	5	2	2	1	0
A2MP1	0	0	2	1	3	0	2	1
A4GALT	56	37	107	118	65	49	52	37
A4GNT	0	0	0	0	1	0	0	0
AA06	0	0	0	0	0	0	0	0
AAA1	0	0	1	0	0	0	0	0
AAAS	2288	1363	1753	1727	835	1672	1389	1121
AACS	1586	923	951	967	484	938	771	635
AACSP1	1	1	3	0	1	1	1	3
AADAC	0	0	0	0	0	0	0	0
AADACL2	0	0	0	0	0	0	0	0
AADACL3	0	0	0	0	0	0	0	0
AADACL4	0	0	1	1	0	0	0	0
AADAT	856	539	593	576	359	567	521	416
AAGAB	4648	2550	2648	2356	1481	3265	2790	2118
AAK1	2310	1384	1869	1602	980	1675	1614	1108
AAMP	5198	3081	3179	3137	1721	4061	3304	2623
AANAT	7	7	12	12	4	6	2	7
AARS	5570	3323	4782	4580	2473	3953	3339	2666

Types of Files Involved

- Fastq
  - Reads
- Fasta
  - Genome
- GFF, GTF
  - Genome Annotation
- SAM, BAM
  - Aligned Reads
- Output files
  - Tables of RPKM, FPKM, or TPM

```
--->gzip -cd L2I_S1_L001_R1_001.fastq.gz | head
@M00805:5:000000000-A0VLL:1:1101:16473:1320 1:N:0:1
NTTGTCATCAGCTGAAGATGAAATAGGATGTAATCAGACGACACAGGAAGCAGATTTTGCTAAT
TTGGAACTAGGTCAGCTGAAGATCCTGTGAGCGAAGTTCCGGCAGTGTCACAGCAC
+
#55<<?BBDBDDDDDDFFFFFFHHHHHFHHAFHHHHHHHHHHBHHHHHFFFHHHHHHHHDGDGHC
AFHFHHHHHHFGHDDHFBFHDFFHFFFHHHFFA=@BEEEED)@<B?BE3==?EEEE
@M00805:5:000000000-A0VLL:1:1101:15023:1321 1:N:0:1
NAGAAATCACAGACATACAAAGCAGTCTGTGTCCTTAGGTCCTGAGCAGCCTCCAGCACATTCT
AGCATCTGCCGTCACATTGTTCTGCACACACCGTCCTTGTCACTGCAGAAGACAGA
+
#55???BBDDEDDDDDGGGGGGIIIIIIIIIIIIIIIIIHIHIIHFGHHHIIIIIIIIIHIIII
HHHHHHHHHHHHHHHHHHHHHHHHHGGFGEGGGGGGGGGGGGGGGEGGGGCEGG>
@M00805:5:000000000-A0VLL:1:1101:14046:1321 1:N:0:1
NTTTCGTGGAAGTGGGTTACCTGACAGTGTGCACGCCCCCAGCAGGTTCACAATATTCTCGTGG
ACATGAGTGCCTCTCTTTCAGAGCTGTCTGCTTTTTCTGTCAAAGAAAGGAGCATT
```

FASTQ quality scores

- Modern sequencers use Phred+33 for their quality scores

- What do quality scores mean? — HTS2018 1.0 documentation (duke.edu)

>NG_008679.1:5001-38170 Homo sapiens paired box 6 (PAX6)

ACCCTCTTTTCTTATCATTGACATTTAAACTCTGGGGCAGGTCCTCGCGTAGAACGCGGCTGTCAGATCT

GCCACTTCCCCTGCCGAGCGGCGGTGAGAAGTGTGGGAACCGGCGCTGCCAGGCTCACCTGCCTCCCCGC

CCTCCGCTCCCAGGTAACCGCCCGGGCTCCGGCCCCGGCCCGGCTCGGGGCCCGCGGGGCCTCTCCGCTG

CCAGCGACTGCTGTCCCCAAATCAAAGCCCGCCCCAAGTGGCCCCGGGGCTTGATTTTTGCTTTTAAAAG

GAGGCATACAAAGATGGAAGCGAGTTACTGAGGGAGGGATAGGAAGGGGGGTGGAGGAGGGACTTGTCTT

TGCCGAGTGTGCTCTTCTGCAAAAGTAGCAAAATGTTCCACTCCTAAGAGTGGACTTCCAGTCCGGCCCT

GAGCTGGGAGTAGGGGGCGGGAGTCTGCTGCTGCTGTCTGCTAAAGCCACTCGCGACCGCGAAAAATGCA

GGAGGTGGGGACGCACTTTGCATCCAGACCTCCTCTGCATCGCAGTTCACGACATCCACGCTTGGGAAAG

TCCGTACCCGCGCCTGGAGCGCTTAAAGACACCCTGCCGCGGGTCGGGCGAGGTGCAGCAGAAGTTTCCC

GCGGTTGCAAAGTGCAGATGGCTGGACCGCAACAAAGTCTAGAGATGGGGTTCGTTTCTCAGAAAGACGC

>sp|P11274|BCR_HUMAN Breakpoint cluster region protein OS=Homo sapiens
OX=9606 GN=BCR PE=1 SV=2
MVDPVGFAEAWKAQFPDSEPPRMELRSVGDIEQELERCKASIRRLEQEVNQERFRMIYLQ
TLLAKEKKSYDRQRWGFRRAAQAPDGASEPRASASRPQPAPADGADPPPAEEPEARPDGE
GSPGKARPGTARRPGAAASGERDDRGPPASVAALRSNFERIRKGHGQPGADAEKPFYVNV
EFHHERGLVKVNDKEVSDRISSLGSQAMQMERKKSQHGAGSSVGDASRPPYRGRSSESSC
GVDGDYEDAELNPRFLKDNLIDANGGSRPPWPPLEYQPYQSIYVGGMMEGEGKGPLLRSQ
STSEQEKRLTWPRRSYSPRSFEDCGGGYTPDCSSNENLTSSEEDFSSGQSSRVSPSPTTY
RMFRDKSRSPSQNSQQSFDSSSPPTPQCHKRHRHCPVVVSEATIVGVRKTGQIWPNDGEG
AFHGDADGSFGTPPGYGCAADRAEEQRRHQDGLPYIDDSPSSSPHLSSKGRGSRDALVSG
ALESTKASELDLEKGLEMRKWVLSGILASEETYLSHLEALLLPMKPLKAAATTSQPVLTS
QQIETIFFKVPELYEIHKEFYDGLFPRVQQWSHQQRVGDLFQKLASQLGVYRAFVDNYGV
AMEMAEKCCQANAQFAEISENLRARSNKDAKDPTTKNSLETLLYKPVDRVTRSTLVLHDL
LKHTPASHPDHPLLQDALRISQNFLSSINEEITPRRQSMTVKKGEHRQLLKDSFMVELVE
GARKLRHVFLFTDLLLCTKLKKQSGGKTQQYDCKWYIPLTDLSFQMVDELEAVPNIPLVP
DEELDALKIKISQIKNDIQREKRANKGSKATERLKKKLSEQESLLLLMSPSMAFRVHSRN
GKSYTFLISSDYERAEWRENIREQQKKCFRSFSLTSVELQMLTNSCVKLQTVHSIPLTIN
KEDDESPGLYGFLNVIVHSATGFKQSSNLYCTLEVDSFGYFVNKAKTRVYRDTAEPNWNE
EFEIELEGSQTLRILCYEKCYNKTKIPKEDGESTDRLMGKGQVQLDPQALQDRDWQRTVI
AMNGIEVKLSVKFNSREFSLKRMPSRKQTGVFGVKIAVVTKRERSKVPYIVRQCVEEIER
RGMEEVGIYRVSGVATDIQALKAAFDVNNKDVSVMMSEMDVNAIAGTLKLYFRELPEPLF
TDEFYPNFAEGIALSDPVAKESCMLNLLLSLPEANLLTFLFLLDHLKRVAEKEAVNKMSL
HNLATVFGPTLLRPSEKESKLPANPSQPITMTDSWSLEVMSQVQVLLYFLQLEAIPAPDS
KRQSILFSTEV

```
0  ##gff-version 3.2.1
1  ##sequence-region ctg123 1 1497228
2  ctg123 . gene            1000  9000  .  +  .  ID=gene00001;Name=EDEN
3  ctg123 . TF_binding_site 1000  1012  .  +  .  ID=tfbs00001;Parent=gene00001
4  ctg123 . mRNA            1050  9000  .  +  .  ID=mRNA00001;Parent=gene00001;Name=EDEN.1
5  ctg123 . mRNA            1050  9000  .  +  .  ID=mRNA00002;Parent=gene00001;Name=EDEN.2
6  ctg123 . mRNA            1300  9000  .  +  .  ID=mRNA00003;Parent=gene00001;Name=EDEN.3
7  ctg123 . exon            1300  1500  .  +  .  ID=exon00001;Parent=mRNA00003
8  ctg123 . exon            1050  1500  .  +  .  ID=exon00002;Parent=mRNA00001,mRNA00002
9  ctg123 . exon            3000  3902  .  +  .  ID=exon00003;Parent=mRNA00001,mRNA00003
10 ctg123 . exon            5000  5500  .  +  .  ID=exon00004;Parent=mRNA00001,mRNA00002,mRNA00003
11 ctg123 . exon            7000  9000  .  +  .  ID=exon00005;Parent=mRNA00001,mRNA00002,mRNA00003
12 ctg123 . CDS             1201  1500  .  +  0  ID=cds00001;Parent=mRNA00001;Name=edenprotein.1
13 ctg123 . CDS             3000  3902  .  +  0  ID=cds00001;Parent=mRNA00001;Name=edenprotein.1
14 ctg123 . CDS             5000  5500  .  +  0  ID=cds00001;Parent=mRNA00001;Name=edenprotein.1
15 ctg123 . CDS             7000  7600  .  +  0  ID=cds00001;Parent=mRNA00001;Name=edenprotein.1
16 ctg123 . CDS             1201  1500  .  +  0  ID=cds00002;Parent=mRNA00002;Name=edenprotein.2
17 ctg123 . CDS             5000  5500  .  +  0  ID=cds00002;Parent=mRNA00002;Name=edenprotein.2
18 ctg123 . CDS             7000  7600  .  +  0  ID=cds00002;Parent=mRNA00002;Name=edenprotein.2
19 ctg123 . CDS             3301  3902  .  +  0  ID=cds00003;Parent=mRNA00003;Name=edenprotein.3
20 ctg123 . CDS             5000  5500  .  +  1  ID=cds00003;Parent=mRNA00003;Name=edenprotein.3
21 ctg123 . CDS             7000  7600  .  +  1  ID=cds00003;Parent=mRNA00003;Name=edenprotein.3
22 ctg123 . CDS             3391  3902  .  +  0  ID=cds00004;Parent=mRNA00003;Name=edenprotein.4
23 ctg123 . CDS             5000  5500  .  +  1  ID=cds00004;Parent=mRNA00003;Name=edenprotein.4
24 ctg123 . CDS             7000  7600  .  +  1  ID=cds00004;Parent=mRNA00003;Name=edenprotein.4
```

Columns:
Seqname
Source
Feature
Start
End
Score
Strand
Frame
Attribute

GFF is derived from GTF

https://m.ensembl.org/info/website/upload/gff.html

SAM (Sequence Alignment Format)

- Read mapping will output a SAM file, which is a sequence file that includes alignment information
- SAM files are then analyzed with another program to determine read counts per gene (obviously a GFF file would have to be involved, too)
- BAM files are a binary form of SAM files (essentially compressed to speed up operations)
- SAM/BAM files contain a lot of information, so they are not very intuitive

```
@HD     VN:1.0  SO:coordinate
@SQ     SN:chr20        LN:64444167
@PG     ID:TopHat       VN:2.0.14       CL:/srv/dna_tools/tophat/tophat -N 3 --read-edit-dist 5 --read-rea
lign-edit-dist 2 -i 50 -I 5000 --max-coverage-intron 5000 -M -o out /data/user446/mapping_tophat/index/chr
20 /data/user446/mapping_tophat/L6_18_GTGAAA_L007_R1_001.fastq
HWI-ST1145:74:C101DACXX:7:1102:4284:73714        16      chr20   190930  3       100M    *       0       0
       CCGTGTTTAAAGGTGGATGCGGTCACCTTCCCAGCTAGGCTTAGGGATTCTTAGTTGGCCTAGGAAATCCAGCTAGTCCTGTCTCTCAGTCCCCCCTCT
C    BBDCCDDCCDDDDCDDDDDDCDCCCDBC?DDDDDDDDDDDDDDDCCDCDDDDDDDDDDCCCCEDDDC?DDDDDDDDDDDDDDDDDDDDDBDHFFFFDC@@
    AS:i:-15        XM:i:3  XO:i:0  XG:i:0  MD:Z:55C20C13A9 NM:i:3  NH:i:2  CC:Z:=  CP:i:55352714   HI:i:0
HWI-ST1145:74:C101DACXX:7:1114:2759:41961        16      chr20   193953  50      100M    *       0       0
       TGCTGGATCATCTGGTTAGTGGCTTCTGACTCAGAGGACCTTCGTCCCCTGGGGCAGTGGACCTTCCAGTGATTCCCCTGACATAAGGGGCATGGACGA
G    DCDDDDEDDDDDDCDDDDDDDCCCDDDCDDDDDEEC>DFFFEJJJJJIGJJJJIHGBHHGJIJJJJJJGJJJIJJJJJIHJJJJJJHHHHHFFFFFCCC
    AS:i:-16        XM:i:3  XO:i:0  XG:i:0  MD:Z:60G16T18T3 NM:i:3  NH:i:1
HWI-ST1145:74:C101DACXX:7:1204:14760:4030        16      chr20   270877  50      100M    *       0       0
       GGCTTTATTGGTAAAAAAGGAATAGCAGATTTAATCAGAAATTCCCACCTGGCCCAGCAGCACCAACCAGAAAGAAGGGAAGAAGACAGGAAAAAACCA
C    DDDDDDDDDCCDDDDDDDDDEEEEEEEFFFEFFEGHHHHFGDJJIHJJIJIJJJIIIIGGFJJIHIIIIJJJJJJIGHHFAHGFHJHFGGHFFFDD@BB
    AS:i:-11        XM:i:2  XO:i:0  XG:i:0  MD:Z:0A85G13    NM:i:2  NH:i:1
HWI-ST1145:74:C101DACXX:7:1210:11167:8699        0       chr20   271218  50      50M4700N50M     *       0
       0       GTGGCTCTTCCACAGGAATGTTGAGGATGACATCCATGTCTGGGGTGCACTTGGGTCTCCGAAGCAGAACATCCTCAAATATGACCTCTCG
accepted_hits.sam
```

Here is a detailed account of SAM format: https://samtools.github.io/hts-specs/SAMv1.pdf

In general, the casual user doesn't need to look inside a sam file. Tools such as samtools can be used to manipulate them.

Running Jobs in the Background

- https://linuxize.com/post/how-to-use-linux-screen/

- Use the utility “screen”
  - Just type **`screen`** to create a new screen session
  - Run your script/command in this new session
  - Press **`Ctrl-a`** and then **`d`** to detach from the session
  - Type **`screen -r`** to reattach to the session
  - After you detach, you can log off and your script will keep running
  - Press **`Ctrl-a`** and then **`\`** to terminate all screen sessions
  - See the link above for more details.

By Thursday:

- Download your sequence read file from the SRA and transfer it to the server (this will take a while)

- Bernadette has already downloaded your genome and gff file, which will be necessary for the read mapping