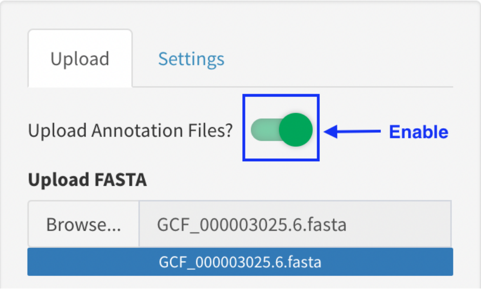
1. Alignment files
    a. Upload
        i. Enable upload annotation files

        ii. FASTA file formats: .fna, .fa, .fasta (note: **.gz is not supported**)

        iii. Gene annotation file formats: .gtf, .gff, .gff3, .gtf.gz, .gff.gz, .gff3.gz

        iv. Wait for each file to finish uploading completely before uploading the next or proceeding.
        (note: Especially for FASTA files which can be very big)

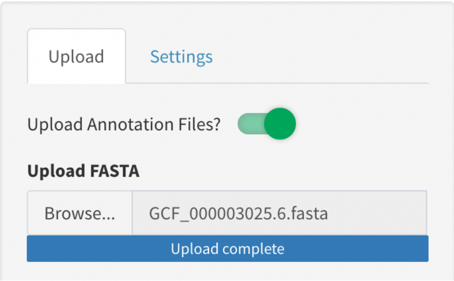


Upload Annotation Files? ● ← Enable

Upload FASTA

Browse... GCF_000003025.6.fasta

GCF_000003025.6.fasta



Upload Annotation Files? ●

Upload FASTA

Browse... GCF_000003025.6.fasta

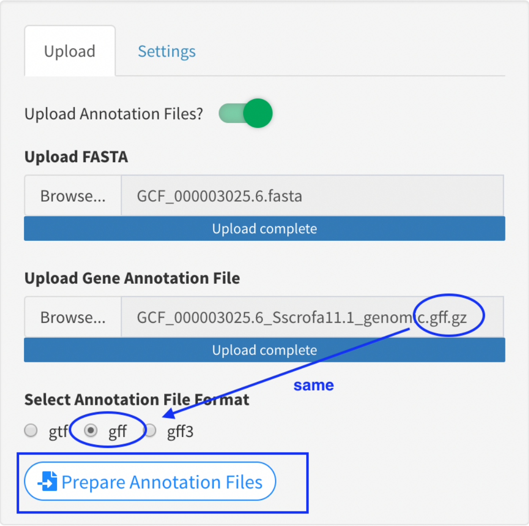
Upload complete

Fig 4.1: Upload incomplete

Fig 4.2: Upload complete

        v. Select correct file format for gene annotation file uploaded
        (note: .gtf.gz → select "gtf" / .gff.gz → select "gff" / .gff3.gz → select "gff3")

        vi. Click 'Prepare Annotation Files' button once done



Upload Annotation Files? ●

Upload FASTA

Browse... GCF_000003025.6.fasta

Upload complete

Upload Gene Annotation File

Browse... GCF_000003025.6_Sscrofa11.1_genomic.gff.gz

Upload complete

same

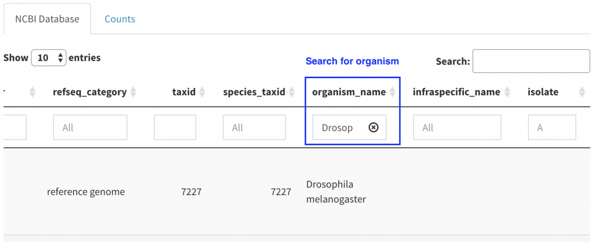
Select Annotation File Format

○ gtf ● gff ○ gff3

Fig 4.3: Prepare annotation files

b. Download from NCBI

i. In the NCBI tab, search the organism name and find the ncbi accession number.

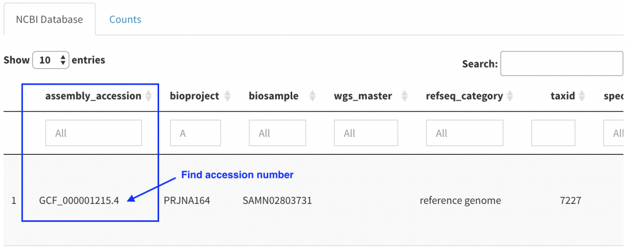

NCBI Database | Counts

Show 10 entries

Search for organism

Search:

	refseq_category	taxid	species_taxid	organism_name	infraspecific_name	isolate
	All		All	Drosop	All	A
	reference genome	7227	7227	Drosophila melanogaster		


NCBI Database | Counts

Show 10 entries

Search:

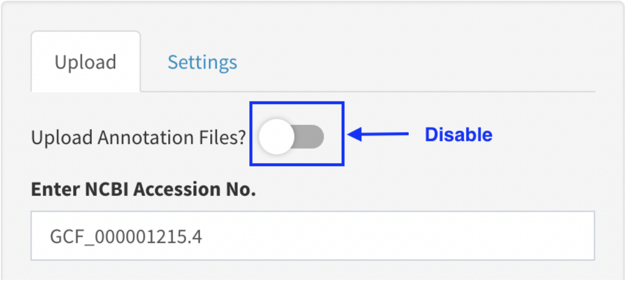
	assembly_accession	bioproject	biosample	wgs_master	refseq_category	taxid	spec
	All	A	All	All	All		All
1	GCF_000001215.4	PRJNA164	SAMN02803731		reference genome	7227	

Find accession number

Fig 4.4: Search organism

Fig 4.5: Search accession no.

ii. Enter ncbi accession number in the left panel


Upload | Settings

Upload Annotation Files?

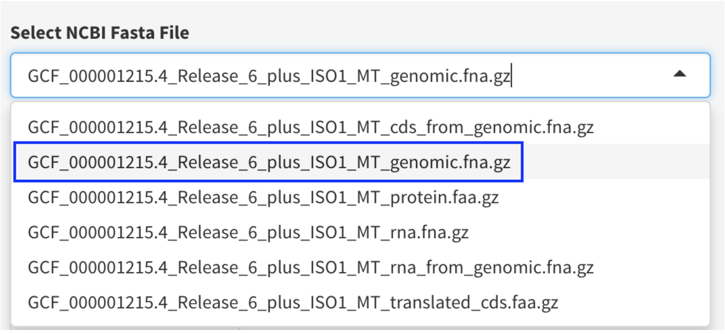
Disable

Enter NCBI Accession No.

GCF_000001215.4

Fig 4.5: Enter accession no.

iii. FASTA file: From the dropdown, select the option “XXX_genomic.fna.gz”

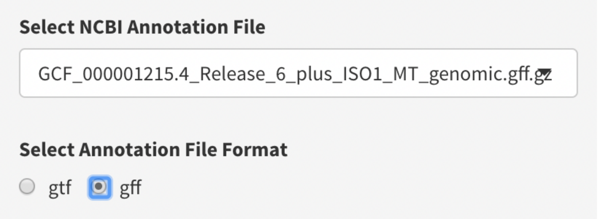

Select NCBI Fasta File

GCF_000001215.4_Release_6_plus_ISO1_MT_genomic.fna.gz

- GCF_000001215.4_Release_6_plus_ISO1_MT_cds_from_genomic.fna.gz
- GCF_000001215.4_Release_6_plus_ISO1_MT_genomic.fna.gz
- GCF_000001215.4_Release_6_plus_ISO1_MT_protein.faa.gz
- GCF_000001215.4_Release_6_plus_ISO1_MT_rna.fna.gz
- GCF_000001215.4_Release_6_plus_ISO1_MT_rna_from_genomic.fna.gz
- GCF_000001215.4_Release_6_plus_ISO1_MT_translated_cds.faa.gz

Fig 4.6: NCBI fasta file

iv. Gene annotation file: From the dropdown, select the options “XXX_genomic.gtf.gz” (gene annotation format: “gtf’) or “XXX_genomic.gff.gz” (gene annotation format: “gff’)



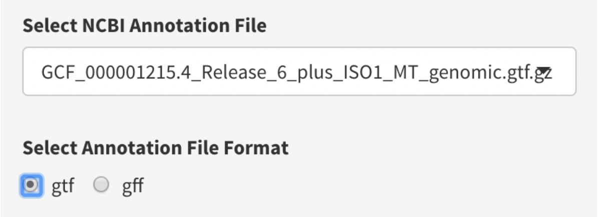
Select NCBI Annotation File

GCF_000001215.4_Release_6_plus_ISO1_MT_genomic.gff.gz

Select Annotation File Format

○ gtf ● gff

Fig 4.7: Find gff file



Select NCBI Annotation File

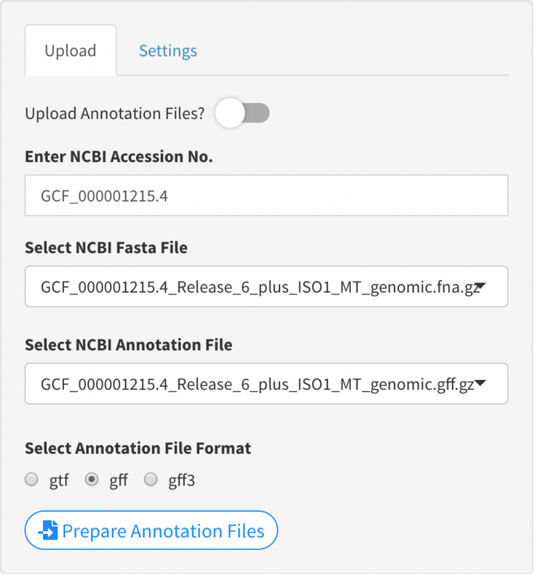
GCF_000001215.4_Release_6_plus_ISO1_MT_genomic.gtf.gz

Select Annotation File Format

● gtf ○ gff

Fig 4.8: Find gtf file

v. Click ‘Prepare Annotation Files’ button once done



Upload Annotation Files? ○

Enter NCBI Accession No.

GCF_000001215.4

Select NCBI Fasta File

GCF_000001215.4_Release_6_plus_ISO1_MT_genomic.fna.gz

Select NCBI Annotation File

GCF_000001215.4_Release_6_plus_ISO1_MT_genomic.gff.gz

Select Annotation File Format

○ gtf ● gff ○ gff3

Fig 4.9: Prepare annotation files

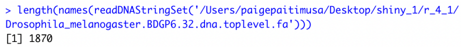
c. Sources of FASTA & gene annotation files

i. Ensembl (however files from ensembl take a long time to download so Ensembl download functionality is not supported in this shiny application)

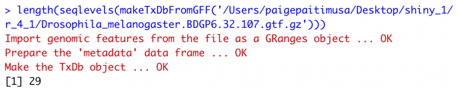
ii. NCBI

d. Importance of ‘Prepare Annotation’ functionality

i. Some chromosomes in fasta file not found in gtf file (note: Often when downloading _genomic.fna.gz from NCBI / dna.primary_assembly.fa.gz or dna.toplevel.fa.gz from Ensembl)



```
> length(names(readDNAStringSet('/Users/paigepaitimusa/Desktop/shiny_1/r_4_1/
Drosophila_melanogaster.BDGP6.32.dna.toplevel.fa')))
[1] 1870
```

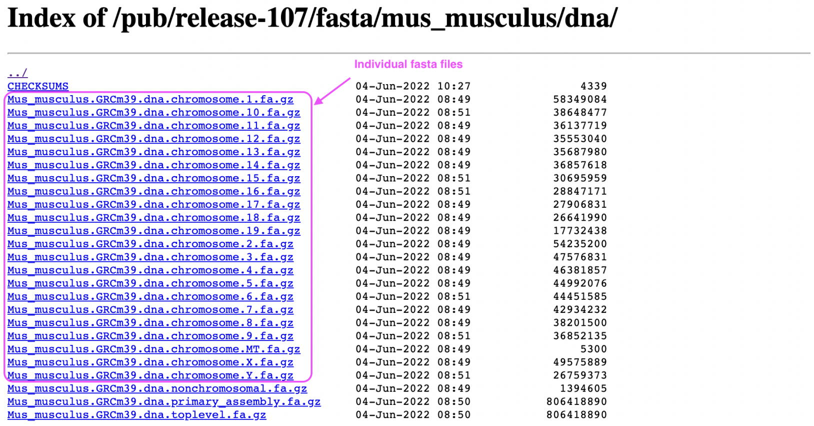


```
> length(seqlevels(makeTxDbFromGFF('/Users/paigepaitimusa/Desktop/shiny_1/
r_4_1/Drosophila_melanogaster.BDGP6.32.107.gtf.gz')))
Import genomic features from the file as a GRanges object ... OK
Prepare the 'metadata' data frame ... OK
Make the TxDb object ... OK
[1] 29
```

Fig 4.10: Fasta file from Ensembl

Fig 4.11: GTF file from Ensembl

ii. Some chromosomes in gtf file not found in FASTA file (note: When downloading individual chromosome fasta files from Ensembl)



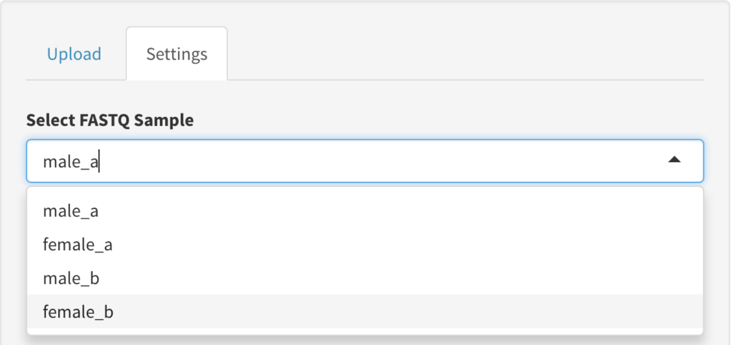
Index of /pub/release-107/fasta/mus_musculus/dna/

```
../
CHECKSUMS                                          04-Jun-2022 10:27                4339
Mus_musculus.GRCm39.dna.chromosome.1.fa.gz         04-Jun-2022 08:49            58349084
Mus_musculus.GRCm39.dna.chromosome.10.fa.gz        04-Jun-2022 08:51            38648477
Mus_musculus.GRCm39.dna.chromosome.11.fa.gz        04-Jun-2022 08:49            36137719
Mus_musculus.GRCm39.dna.chromosome.12.fa.gz        04-Jun-2022 08:49            35553040
Mus_musculus.GRCm39.dna.chromosome.13.fa.gz        04-Jun-2022 08:49            35687980
Mus_musculus.GRCm39.dna.chromosome.14.fa.gz        04-Jun-2022 08:49            36857618
Mus_musculus.GRCm39.dna.chromosome.15.fa.gz        04-Jun-2022 08:51            30695959
Mus_musculus.GRCm39.dna.chromosome.16.fa.gz        04-Jun-2022 08:51            28847171
Mus_musculus.GRCm39.dna.chromosome.17.fa.gz        04-Jun-2022 08:49            27906831
Mus_musculus.GRCm39.dna.chromosome.18.fa.gz        04-Jun-2022 08:49            26641990
Mus_musculus.GRCm39.dna.chromosome.19.fa.gz        04-Jun-2022 08:49            17732438
Mus_musculus.GRCm39.dna.chromosome.2.fa.gz         04-Jun-2022 08:49            54235200
Mus_musculus.GRCm39.dna.chromosome.3.fa.gz         04-Jun-2022 08:49            47576831
Mus_musculus.GRCm39.dna.chromosome.4.fa.gz         04-Jun-2022 08:49            46381857
Mus_musculus.GRCm39.dna.chromosome.5.fa.gz         04-Jun-2022 08:49            44992076
Mus_musculus.GRCm39.dna.chromosome.6.fa.gz         04-Jun-2022 08:51            44451585
Mus_musculus.GRCm39.dna.chromosome.7.fa.gz         04-Jun-2022 08:49            42934232
Mus_musculus.GRCm39.dna.chromosome.8.fa.gz         04-Jun-2022 08:49            38201500
Mus_musculus.GRCm39.dna.chromosome.9.fa.gz         04-Jun-2022 08:51            36852135
Mus_musculus.GRCm39.dna.chromosome.MT.fa.gz        04-Jun-2022 08:49                5300
Mus_musculus.GRCm39.dna.chromosome.X.fa.gz         04-Jun-2022 08:49            49575889
Mus_musculus.GRCm39.dna.chromosome.Y.fa.gz         04-Jun-2022 08:51            26759373
Mus_musculus.GRCm39.dna.nonchromosomal.fa.gz       04-Jun-2022 08:49             1394605
Mus_musculus.GRCm39.dna.primary_assembly.fa.gz     04-Jun-2022 08:50           806418890
Mus_musculus.GRCm39.dna.toplevel.fa.gz             04-Jun-2022 08:50           806418890
```

Fig 4.12: Individual fasta files from Ensembl

iii. Ensures that in these 2 instances, chromosomes in fasta file & gtf file are identical to prevent any errors later on when aligning.

2. In the “Settings” tab of the left side panel, select the sample for aligning.



Select FASTQ Sample

male_a

Fig 4.13: Select sample for aligning

3. Configure trim settings
   (note: If spliced alignment is enabled and Rbowtie is chosen as the aligner, SpliceMap will be used which takes much longer than Rhisat2!)

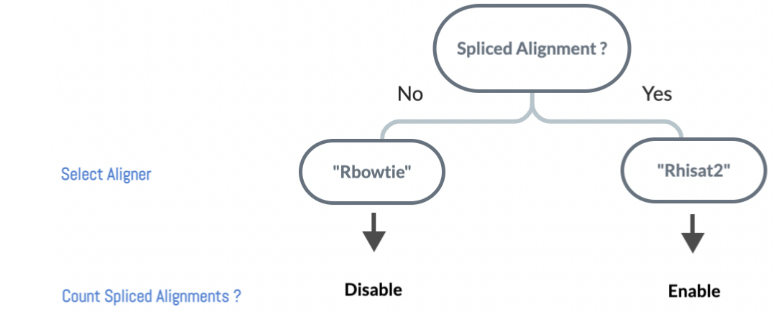


Fig 4.14: Select aligner for spliced alignment

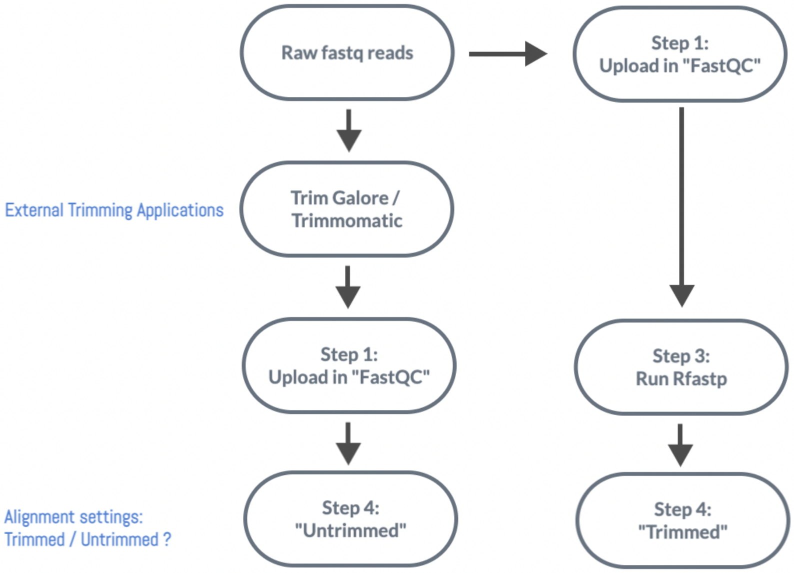
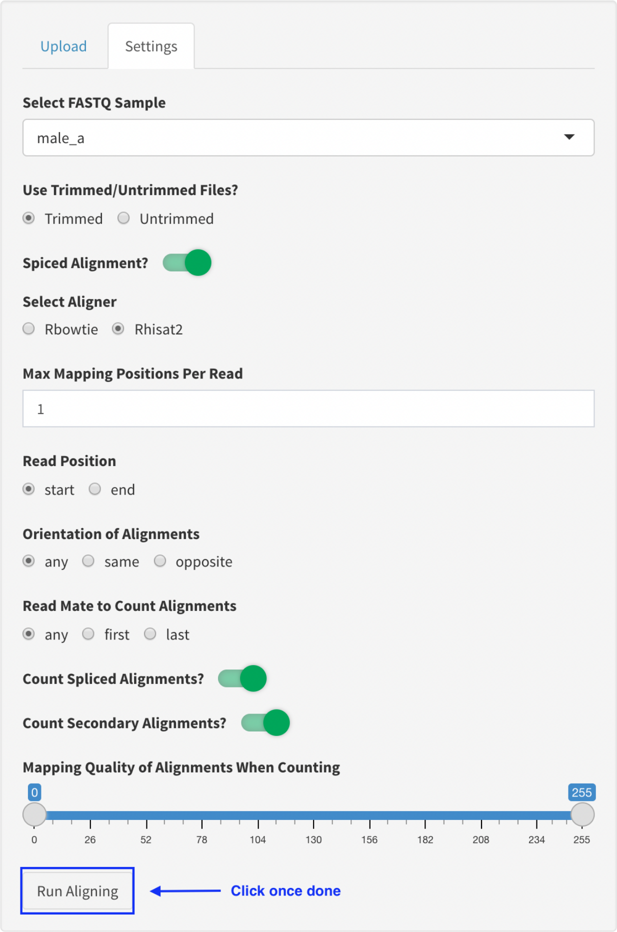


Fig 4.15: Use trimmed / untrimmed fastq files



Select FASTQ Sample

male_a

Use Trimmed/Untrimmed Files?

● Trimmed ○ Untrimmed

Spiced Alignment? on

Select Aligner

○ Rbowtie ● Rhisat2

Max Mapping Positions Per Read

1

Read Position

● start ○ end

Orientation of Alignments

● any ○ same ○ opposite

Read Mate to Count Alignments

● any ○ first ○ last

Count Spliced Alignments? on

Count Secondary Alignments? on

Mapping Quality of Alignments When Counting

0 – 255

Click once done

Fig 4.16: All alignment settings

4. Click "Run Aligning" once done
(note: There'll be 3 modal dialogs which appear: "Aligning", "Generating Counts" and "Quality Reporting")

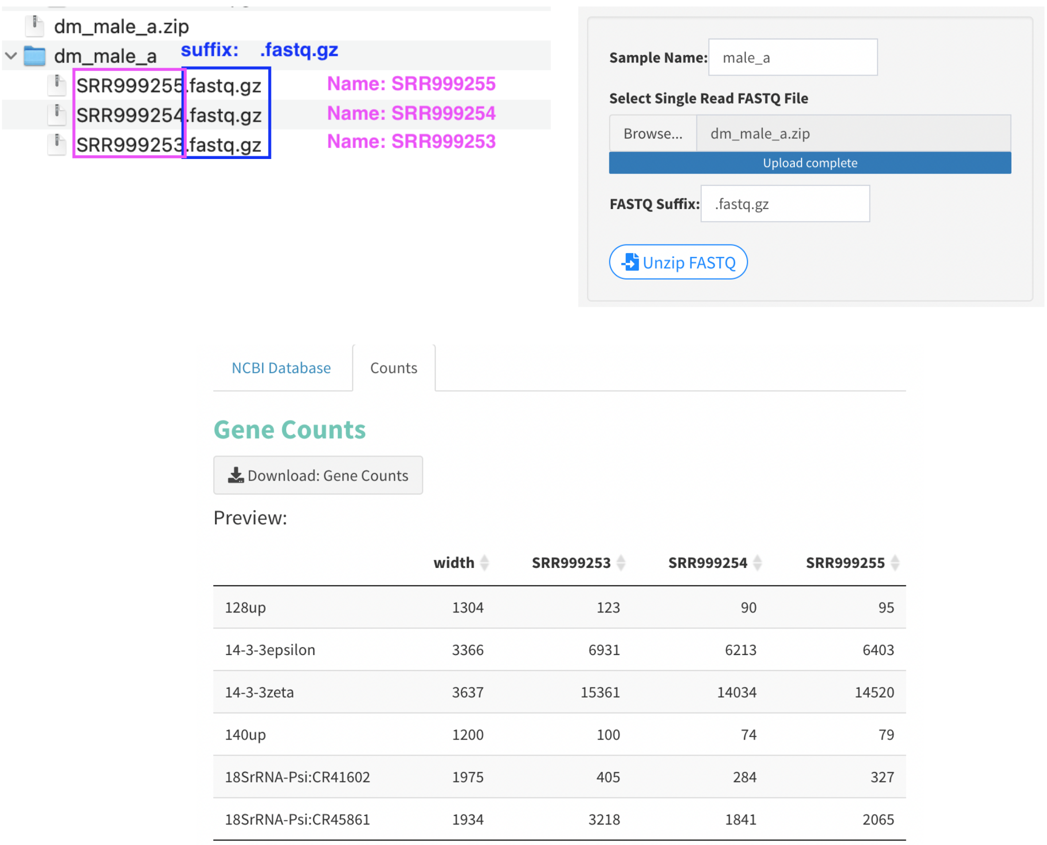
5. In the "Results" tab panel, download the csv files for gene, exon, promoter and junction counts. You'll also see a preview for these files.
(note: Junction counts are only generated when spliced alignment is enabled.)
(note: Column names are identifiers of the fastq files within the sample selected for aligning.)

Example of gene counts:
https://github.com/paigerollex/gene_cloud_omics/blob/main/output_data/male_a_gene_ncbi.csv

Example of exon counts:
https://github.com/paigerollex/gene_cloud_omics/blob/main/output_data/male_a_exon_ncbi.csv

Example of promoter counts:
https://github.com/paigerollex/gene_cloud_omics/blob/main/output_data/male_a_promoter_ncbi.csv

Example of junction counts:
https://github.com/paigerollex/gene_cloud_omics/blob/main/output_data/male_a_junction_ncbi.csv



Sample Name: male_a

Select Single Read FASTQ File: dm_male_a.zip

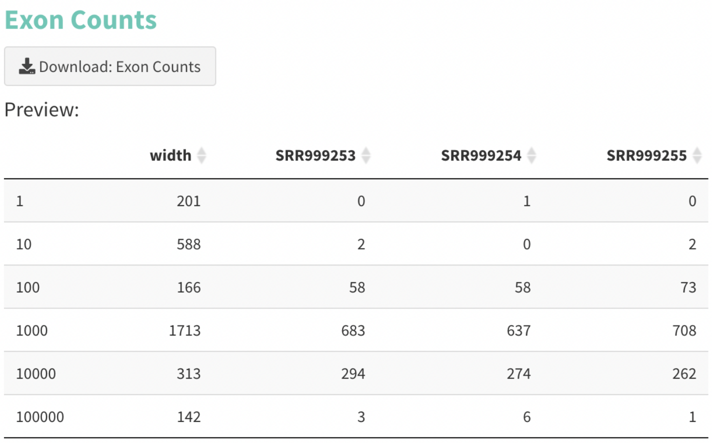
FASTQ Suffix: .fastq.gz

Gene Counts

Preview:

	width	SRR999253	SRR999254	SRR999255
128up	1304	123	90	95
14-3-3epsilon	3366	6931	6213	6403
14-3-3zeta	3637	15361	14034	14520
140up	1200	100	74	79
18SrRNA-Psi:CR41602	1975	405	284	327
18SrRNA-Psi:CR45861	1934	3218	1841	2065

Fig 4.17: Gene counts



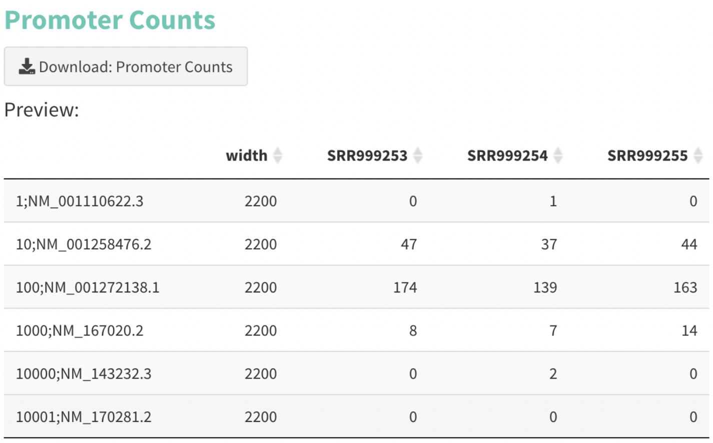
Exon Counts

Download: Exon Counts

Preview:

	width	SRR999253	SRR999254	SRR999255
1	201	0	1	0
10	588	2	0	2
100	166	58	58	73
1000	1713	683	637	708
10000	313	294	274	262
100000	142	3	6	1

Fig 4.18: Exon counts



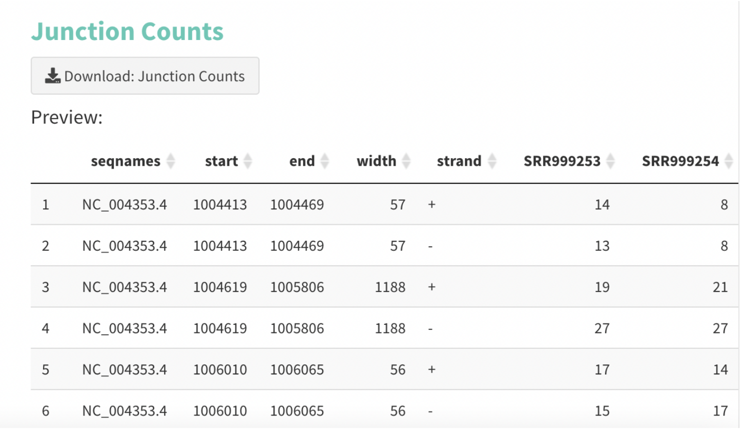
Promoter Counts

Download: Promoter Counts

Preview:

	width	SRR999253	SRR999254	SRR999255
1;NM_001110622.3	2200	0	1	0
10;NM_001258476.2	2200	47	37	44
100;NM_001272138.1	2200	174	139	163
1000;NM_167020.2	2200	8	7	14
10000;NM_143232.3	2200	0	2	0
10001;NM_170281.2	2200	0	0	0

Fig 4.19: Promoter counts



Junction Counts

Download: Junction Counts

Preview:

	seqnames	start	end	width	strand	SRR999253	SRR999254
1	NC_004353.4	1004413	1004469	57	+	14	8
2	NC_004353.4	1004413	1004469	57	-	13	8
3	NC_004353.4	1004619	1005806	1188	+	19	21
4	NC_004353.4	1004619	1005806	1188	-	27	27
5	NC_004353.4	1006010	1006065	56	+	17	14
6	NC_004353.4	1006010	1006065	56	-	15	17

Fig 4.20: Junction counts