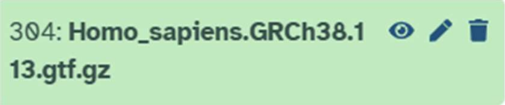
Week 5 – Building count matrix and metadata files

A count matrix file is a .csv file that talks about how many sequencing reads (or counts) map to each gene (or transcript) across different samples.

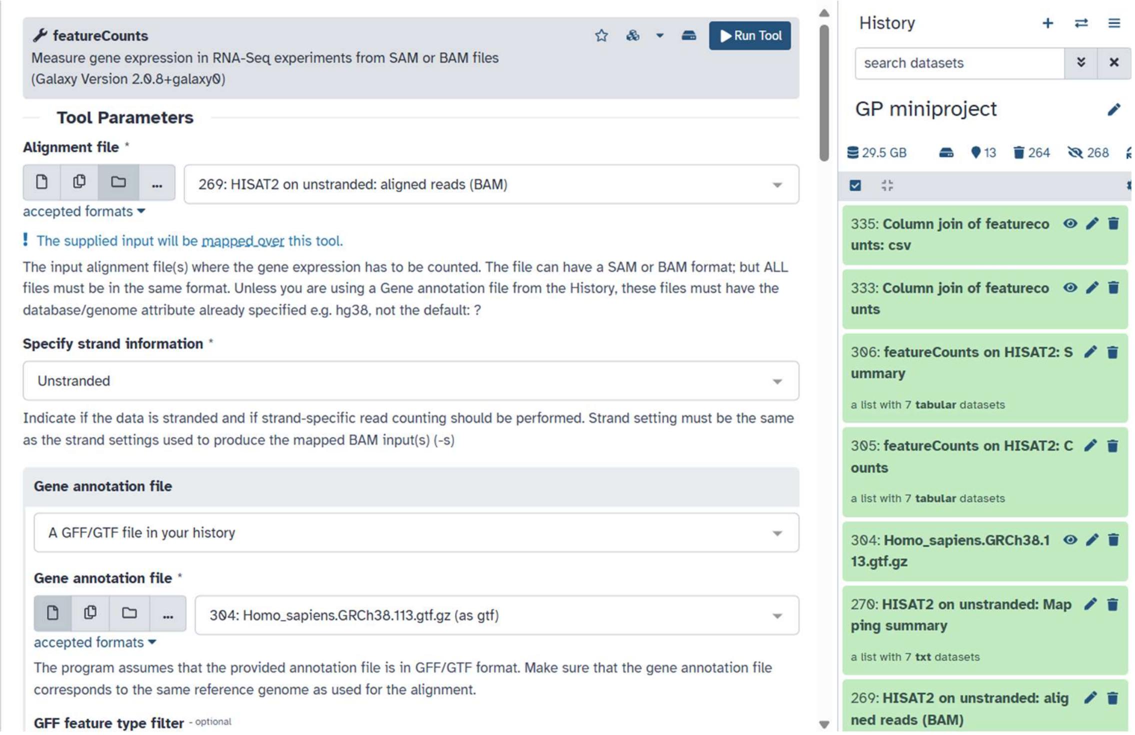
To build a counts.csv file, first the GTF file of the human reference transcriptome is downloaded. A GTF (Gene Transfer Format) file is a tab-delimited text file used to store information about gene structure and annotations, primarily focusing on gene-centric features. This will help us extract the number of transcripts of different genes across samples.

This was downloaded and loaded onto the galaxy server. Following this, the 'featureCounts' tool was used to assemble the feature counts matrix. Since it was done individually for each sample, the Column join option was used to join all the columns and the file was converted into a .csv file (GP.csv) for further analysis in R.



304: Homo_sapiens.GRCh38.113.gtf.gz

Loading .gtf file



featureCounts
Measure gene expression in RNA-Seq experiments from SAM or BAM files
(Galaxy Version 2.0.8+galaxy0)

Run Tool

Tool Parameters

Alignment file *
269: HISAT2 on unstranded: aligned reads (BAM)
accepted formats

! The supplied input will be mapped over this tool.

The input alignment file(s) where the gene expression has to be counted. The file can have a SAM or BAM format; but ALL files must be in the same format. Unless you are using a Gene annotation file from the History, these files must have the database/genome attribute already specified e.g. hg38, not the default: ?

Specify strand information *
Unstranded

Indicate if the data is stranded and if strand-specific read counting should be performed. Strand setting must be the same as the strand settings used to produce the mapped BAM input(s) (-s)

Gene annotation file
A GFF/GTF file in your history

Gene annotation file *
304: Homo_sapiens.GRCh38.113.gtf.gz (as gtf)
accepted formats

The program assumes that the provided annotation file is in GFF/GTF format. Make sure that the gene annotation file corresponds to the same reference genome as used for the alignment.

GFF feature type filter - optional

History

search datasets

GP miniproject

29.5 GB 13 264 268

335: Column join of featurecounts: csv

333: Column join of featurecounts

306: featureCounts on HISAT2: Summary
a list with 7 tabular datasets

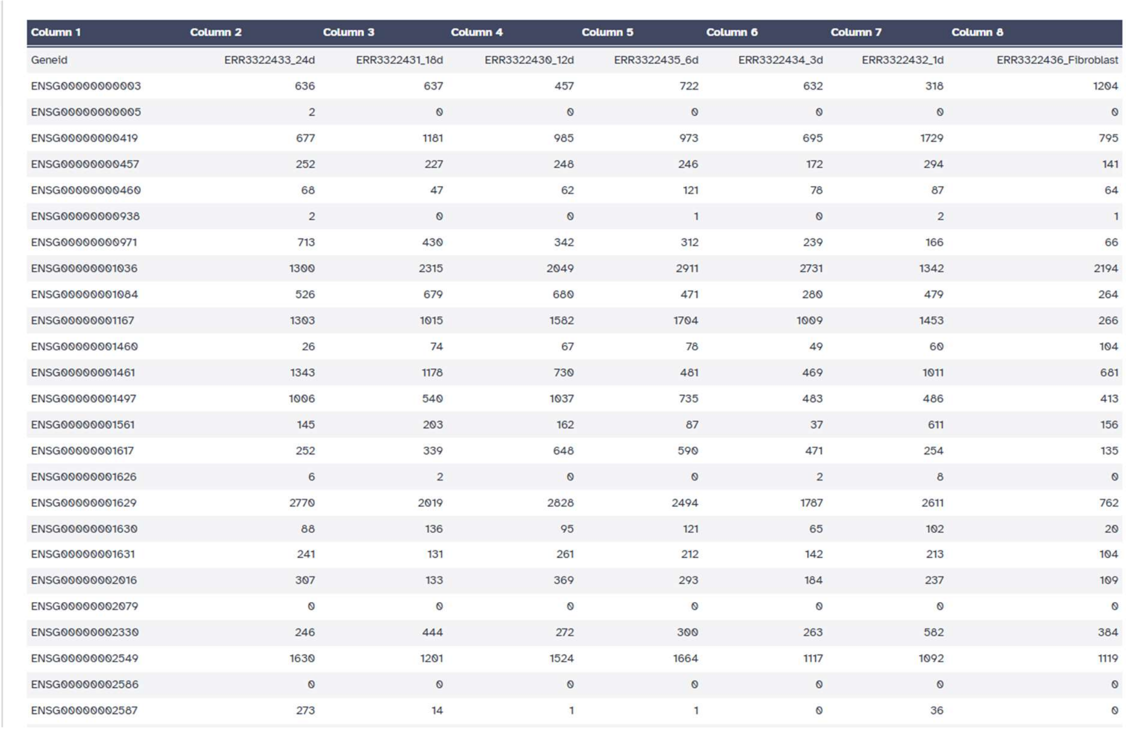
305: featureCounts on HISAT2: Counts
a list with 7 tabular datasets

304: Homo_sapiens.GRCh38.113.gtf.gz

270: HISAT2 on unstranded: Mapping summary
a list with 7 txt datasets

269: HISAT2 on unstranded: aligned reads (BAM)

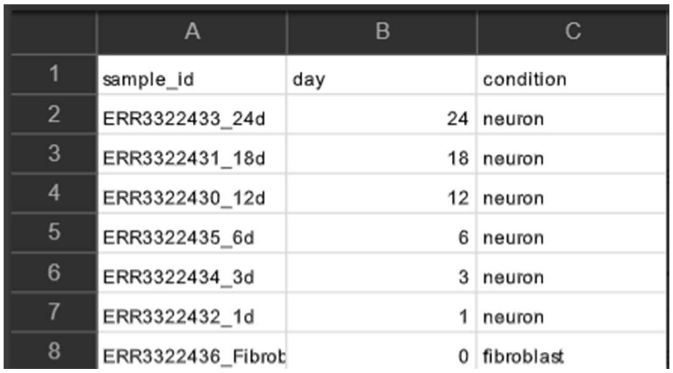
Creating count matrix



Column 1	Column 2	Column 3	Column 4	Column 5	Column 6	Column 7	Column 8
Geneid	ERR3322433_24d	ERR3322431_18d	ERR3322430_12d	ERR3322435_6d	ERR3322434_3d	ERR3322432_1d	ERR3322436_Fibroblast
ENSG00000000003	636	637	457	722	632	318	1204
ENSG00000000005	2	0	0	0	0	0	0
ENSG00000000419	677	1181	985	973	695	1729	795
ENSG00000000457	252	227	248	246	172	294	141
ENSG00000000460	68	47	62	121	78	87	64
ENSG00000000938	2	0	0	1	0	2	1
ENSG00000000971	713	430	342	312	239	166	66
ENSG00000001036	1300	2315	2049	2911	2731	1342	2194
ENSG00000001084	526	679	680	471	280	479	264
ENSG00000001167	1303	1015	1582	1704	1009	1453	266
ENSG00000001460	26	74	67	78	49	60	104
ENSG00000001461	1343	1178	730	481	469	1011	681
ENSG00000001497	1006	540	1037	735	483	486	413
ENSG00000001561	145	203	162	87	37	611	156
ENSG00000001617	252	339	648	590	471	254	135
ENSG00000001626	6	2	0	0	2	8	0
ENSG00000001629	2770	2019	2828	2494	1787	2611	762
ENSG00000001630	88	136	95	121	65	102	20
ENSG00000001631	241	131	261	212	142	213	104
ENSG00000002016	307	133	369	293	184	237	109
ENSG00000002079	0	0	0	0	0	0	0
ENSG00000002330	246	444	272	300	263	582	384
ENSG00000002549	1630	1201	1524	1664	1117	1092	1119
ENSG00000002586	0	0	0	0	0	0	0
ENSG00000002587	273	14	1	1	0	36	0

Final feature counts matrix after joining columns

The metadata file was created as follows:



	A	B	C
1	sample_id	day	condition
2	ERR3322433_24d	24	neuron
3	ERR3322431_18d	18	neuron
4	ERR3322430_12d	12	neuron
5	ERR3322435_6d	6	neuron
6	ERR3322434_3d	3	neuron
7	ERR3322432_1d	1	neuron
8	ERR3322436_Fibrob	0	fibroblast