11.3 Creating a count matrix

11.3.1 Count matrices

At this point in the process, we have aligned each sample of raw RNA-seq reads to the reference genome using HISAT2 to acquire a SAM alignment file. As you saw, the SAM file is very large and contains a lot of information that we do not need. All we are interested in right now is the number of reads that were counted per gene. To do this, we need to convert all of our SAM alignment files to a single count matrix.

Count matrices are a condensed way to store gene expression information across multiple samples. You already learned about matrices in week 4, so there is nothing too new and scary here! In a count matrix, the rows are genes and the columns are samples, with the values of the matrix being the number of RNA-seq reads that were counted for each gene in that sample. For example, if a column labelled "Sample 2" has a value of "89" for a row labelled "Gene 5", this would mean that 89 reads were counted for the "Gene 5" gene for "Sample 2".

	Sample 1	Sample 2	Sample 3	Sample 4
Gene 1	0	5	7	3
Gene 2	50	52	4	0
Gene 3	0	1	0	0
Gene 4	2	0	0	4
Gene 5	172	89	108	110

The gene expression information is representative of all of the cells from a particular sample. This means that these genes are expressed at the tissue or sample level. So if counts are being compared across samples or groups of samples, the differences measured can only tell you what's happening to that area of the tissue, not the individual cells within the tissue. This is a great way to compare healthy and diseased tissue, or diseased tissue that has been treated with a variety of drugs. Looking at the gene expression response to a drug can help determine its impact on a diseased tissue, or figure out what needs to be changed about the treatment.

The data stored in count matrices cannot be taken as the complete truth ("ground truth") of gene expression in a sample for a variety of reasons. Because RNA-sequencing doesn't sequence every single RNA molecule in a cell, low abundance transcripts can be missed. This means that low-expression genes can be counted as having zero expression.

Technical factors that differ between RNA-seq experiments (a.k.a "batch effects") can also be strong in RNA-seq data, with different protocols potentially impacting gene expression values (e.g. if one protocol causes more tissue damage than another). This can cause differences in gene expression values due to the experimental "batch" rather than a result of the biological phenotype. This is why it is helpful to have multiple samples representing the same phenotype, so that we can be more confident in which differences are real, and which may be due to batch effect alone. Large differences in reads that are consistent across these "replicates" typically represent a true biological difference and can be very informative.

11.3.2 Converting alignment files to counts

A SAM or BAM alignment file can be converted to a count matrix with a variety of tools. At this point, you can imagine that we have run the HISAT2 alignment on all six FASTQ files, each corresponding to a different sample. This would have created six individual SAM files, that we can refer to as follows:

16N_hisat2.sam, 18N_hisat2.sam, 19N_hisat2.sam, 16T_hisat2.sam, 18T_hisat2.sam, and 19T_hisat2.sam. Each of these will represent a different column in our count matrix.

The count matrix can be created using a tool called FeatureCounts. We will not be running FeatureCounts in this course, but just like with HISAT2, we will walk through how it is used. This piece of software takes all of the SAM files as input at once and creates the multi-sample matrix. The only other piece of information you need is the GTF file, which will tell FeatureCounts exactly what genes the reads in the SAM file correspond to. Remember, the SAM file recorded genome contigs and coordinates, so the GTF file is now required to convert these coordinates to the appropriate gene symbols. It's very important to use a GTF file from the same source as the reference genome file genome file that you indexed with HISAT2.

The command to convert the SAM files into a count matrix will look something like this, where "-d" indicates the length of the RNA-seq reads, "-a" indicates the GTF file, "-o" indicates the name of the output TSV file, and the arguments without tags are the different SAM files:

```
featureCounts -d 38 -a genome_annotation.gtf \

-o countMatrixIntermediate.tsv \

16N_hisat2.sam 18N_hisat2.sam 19N_hisat2.sam \

16T_hisat2.sam 18T_hisat2.sam 19T_hisat2.sam
```

The output is an intermediate file that contains the information you will need for your count matrix. Let's take a peak at the file below, using "less". Here you can see the first few lines:

```
# Program:featureCounts v2.0.3; Command:"featureCounts" "-d" "38" "-M"
"-a" "genome_annotation.gtf" "-o" "all.hisat2.counts.tsv"
"16N_hisat2.sam" "18N_hisat2.sam" "19N_hisat2.sam" "16T_hisat2.sam"
"18T_hisat2.sam" "19T_hisat2.sam"

Geneid  Chr     Start   End     Strand  Length  16N    18N   19N   16T
18T    19T

DDX11L1 chr1;chr1;chr1  11874;12613;13221       12227;12721;14409
+;+;+   1652    2       0       0       2       3       1

WASH7P  chr1;chr1;chr1;chr1;chr1;chr1;chr1;chr1;chr1;chr1;chr1
14362;14970;15796;16607;16858;17233;17606;17915;18268;24738;29321
14829;

15038;15947;16765;17055;17368;17742;18061;18366;24891;29370       -;-
;-;-;-;-;-;-;-;-;-   1769    121     124     107     158     204
201

MIR6859-1       chr1;chr1;chr1  17369;17369;17409
17436;17391;17431       -;-;-   68      1       1       0       0
2       2

MIR1302-2HG     chr1;chr1;chr1  29926;30564;30976
30039;30667;31295       +;+;+   538     0       1       0       0
0       1
```

```
MIR1302-2       chr1;chr1       30366;30438     30503;30458     +;+
138     0       0       0       0       0       0

FAM138A chr1;chr1;chr1  34611;35277;35721       35174;35481;36081
-;-;-   1130    0       0       0       2       0       0

OR4F5   chr1    69091   70008   +       918     0       0       0
0       0       0
```

Here we have a TSV file that not only has the counts for each sample, but additional information in additional columns that specify strandedness (i.e. positive or negative strand), gene length, etc. The very top line is commented out (i.e. beginning with "#"), and is simply the command-line argument that we specified for FeatureCounts. Below this are the column names (16N to 19T). All that we are interested in at this point are the columns that contain (1) the gene symbol and (2) the counts for each sample.

The first column contains the gene symbols, and then columns two to six contain these extra (and in our case, unnecessary) pieces of information for each gene. Following this are the columns with count information that are easily recognizable by their column names which will be labelled with their corresponding SAM file name (e.g. "16N_hisat2.sam"). We have six samples, so these columns are seven to twelve, inclusive. These columns can be extracted using the "cut" function directly on the command line.

```
cut -f1,7-12 countMatrixIntermediate.tsv > countMatrix.tsv
```

In this case, we created a new file from the FeatureCounts output, and called it countMatrix.tsv. This file is now ready to be analyzed in R!