TDM

Jeffrey A. Thompson

Quantitative Biomedical Sciences Program, Geisel School of Medicine at Dartmouth College

March 14, 2016

1 Training Distribution Matching

To perform the TDM transformation you need to have a reference dataset and a target dataset. The reference dataset should be from microarray expression experiments and the target dataset should be from RNA-seq. The target dataset will be transformed to have similar characteristics to the reference.

As an example, the TDM package contains some sample data. These data can be loaded as follows:

```
data(meta)
data(tcga)
```

The data are loaded into variables **meta** and **tcga**. Here is a summary of their characteristics:

```
summary(as.vector(as.matrix(meta)))

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   4.776   6.804   7.774   8.006   8.966  14.880

summary(as.vector(as.matrix(tcga)))

##      Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
##       0.0     184.8     613.6    2143.0    1702.0 2066000.0
```

If we simply scaled the TCGA data to be in the same range, the distribution would be quite different:

```
load_it("scales")
tcga_vec = rescale(as.vector(as.matrix(tcga)), to=c(min(meta), max(meta)))
summary(tcga_vec)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   4.776   4.776   4.779   4.786   4.784  14.880
```

One might try log transforming the RNA-seq data, but this also is unsatisfactory:

```
load_it("data.table")
tcga_log = log_transform_p1(data.table(cbind(gene=rownames(tcga), tcga)))
summary(as.vector(data.matrix(tcga_log[,2:ncol(tcga_log),with=F])))

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.000   7.537   9.263   8.999  10.730  20.980
```

If we TDM transform the data, the results appear to be much improved:

```
tcga_tdm = tdm_transform(ref_data = data.table(cbind(gene=rownames(meta), meta)),
target_data = data.table(cbind(gene=rownames(tcga), tcga)))
summary(as.vector(data.matrix(tcga_tdm[,2:ncol(tcga_tdm),with=F])))

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   4.776   5.766   7.338   7.448   8.793  14.880
```

Finally, here is a plot comparing the distributions of the reference data, the scaled data, the log transformed data, and the TDM transformed data:

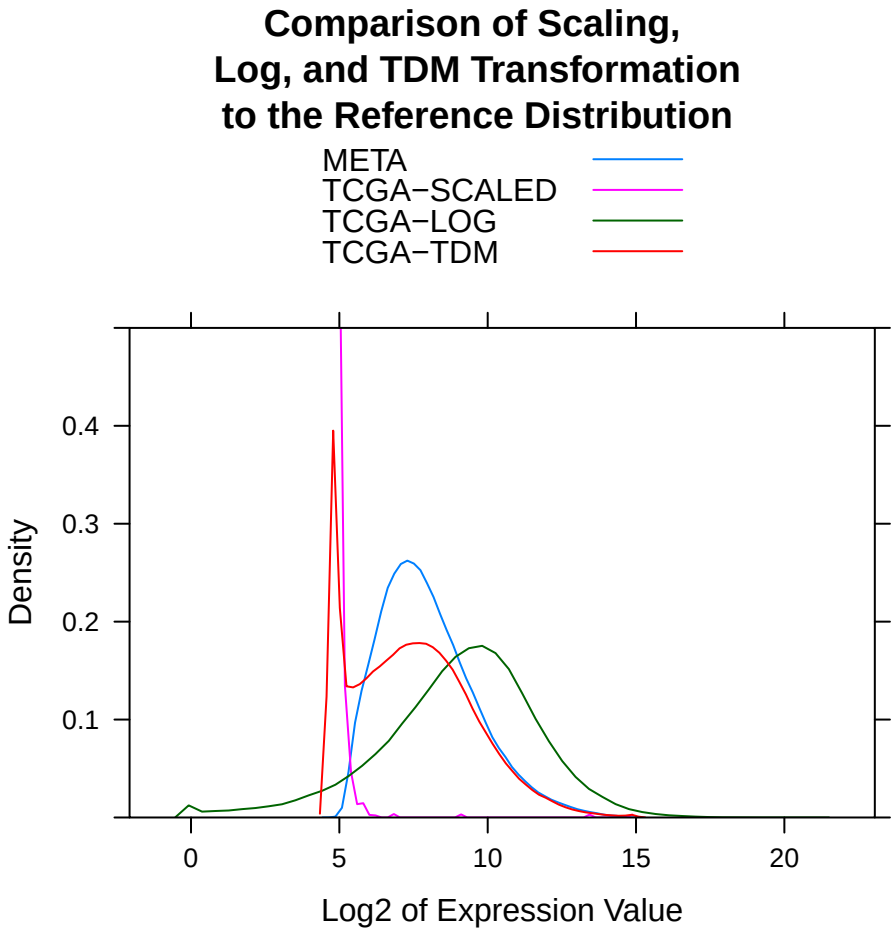

Figure 1: TDM brings the sample RNA-seq data closest to the reference distribution. Log2 transformation creates a left-skewed distribution that is not typical of microarray data, making comparison between the datasets difficult, while simple scaling creates a right-skewed distribution.