Gene Expression Profiles

BIOF2014

Data

Gene expression profiles have been generated by RNA sequencing across different normal tissues and tumour samples. These datasets are available from the GTEx and TCGA projects.

The RNA sequencing reads have been aligned, and the expression of each gene has been quantified and summarized as “transcripts per million” (TPM) values. TPM values can be log-transformed.

The resulting TPM expression matrices have been subsetted to only include genes in the Cancer Gene Census.

Download `expr_gtex_tpm_ccg.rds` and `expr_pancan_ccg.rds` from Moodle.

Checking if data is normally distributed

While there are statistical tests such as the Shapiro–Wilk test to test for normality, such tests do not work well for large samples.

A well-accepted way of checking for normality is by inspecting the quantile-quantile plots against theoretical quantiles from the normal distribution.

Questions

0. Has the expression data from GTEx and TCGA been log-transformed?
1. Log transform any data that has not been log-transformed. Does the distributions of the expression data from GTEx vs. TCGA look similar? What could have caused any differences in distributions?
2. Is the expression data as a whole normally distributed?
3. Is the expression profile for each sample across genes normally distributed?

4. Is the expression level of each gene across samples normally distributed?

5. Are the mean expression levels for each gene normally distributed? You can sample mean gene expression levels by bootstrapping using the following function.

```r
# @param x  data matrix (features by samples)
# @param B  number of bootstrap samples
sample_means <- function (x, B=100) {
    boots <- matrix(
        unlist(lapply(1:B,
            function(b) {
                idx <- sample.int(ncol(x), replace=TRUE);
                rowMeans(x[, idx])
            }
        )),
        nrow = nrow(x)
    );
    rownames(boots) <- rownames(x);
    boots
}
```

6. Implement a statistical model in Stan to compare the difference in expression level in normal vs. tumour tissues for a single gene.