Lung vs Heart Power Analysis

Lung vs Heart

This notebook contains code for the power analysis for lung vs heart. Experiment design is a simple one factor.

Data

Data used is from GTEx.

- `filtered_data.rds`: data matrix produced by `process_gtex.Rmd`. For more details please see that notebook. For file size reasons it is not available in this repo, however you can follow the steps outlined in `process_gtex.Rmd` to reproduce the files needed here.
- `GTEx_Analysis_v8_Annotations_SampleAttributesDS.txt`: mappings of the sample IDs to the tissue group and tissue subtype among others. It appears this map contains annotations for not just the samples that are RNAseq but also for all the other types of sequencing, such as RIP-Seq and ChIP-Seq. For our purposes we ignore those samples since they will not be in the data matrix.
- `BI-ECM-proteome-genelist-short.xlsx`: initial genelist of interest
- `BI-Reactome-results-ECM-gene-list.csv`: pathways of interest

```
data <- readRDS("filtered_data.rds")
proteome_xlsx <- read.xlsx("BI-ECM-proteome-gene-list-short.xlsx", 1)
filtgenedf <- filter(proteome_xlsx, !is.na(Ensembl)) %>% select(Gene, Ensembl) %>%
    distinct
genelist <- filtgenedf$Ensembl
genenames <- filtgenedf$Gene
tissue_map <- read.delim("GTEx_Analysis_v8_Annotations_SampleAttributesDS.txt")

# NOTEs on realizing tissue_map has mappings for not just RNAseq data UHOH, we
# might have made an error with selection? :OOOOO filter(tissue_map,
# SMTSD=='Heart - Left Ventricle') %>% .$SMTS %>% unique reveals everything is
# indeed heart

# > select(data, filter(tissue_map, SMTSD=='Heart - Left Ventricle')$SAMPID)
# Error: Can't subset columns that don't exist.  x Columns
# `GTEX-111FC-0826-SM-CYKRT`, `GTEX-11ZTT-5001-SM-AN3YD`,
# `GTEX-1211K-5001-SM-AN3X5`, `GTEX-12BJ1-5001-SM-AN3X8`,
# `GTEX-12WSG-5001-SM-AN3X9`, etc. don't exist.  Run `rlang::last_error()` to see
# where the error occurred.  filter(tissue_map,
# SAMPID=='GTEX-111FC-0826-SM-CYKRT') shows that it's RIPseq under the column
# SMGEBTCHT. Looking at the sample attributes data dictionary (see GTEx portal)
# it indicates type of genotype or expression batch/tech used to analyze DNA/RNA.
# Also looking further several samples from tissue map for heat/lung/skin are not
# present in data matrix, they are all ones with different sequencing tech

# comprehensive pathway ECM difficult protein to analyze in proteomics, because
```

```
# of insolubility only in past few years have been able to make soluble and
# analyze
```

Pathway parsing

```
pathway_csv <- read.csv("BI-Reactome-results-ECM-gene-list.csv")
pathway_genes <- pathway_csv$Submitted.entities.found %>% as.character %>% strsplit(";") %>%
    unlist %>% unique
all(pathway_genes %in% proteome_xlsx$Gene)  # not all of the genes in the pathway entities are in the p
```

```
## [1] FALSE
```

NOTES:

- Combine because significant overlap between the 2

Genelist questions

- Many of the genes in the proteome excel sheet don't have ensembl ids associated with them. Why is that? Where did this data come from? c(1, 2, 3, 3, 4, 5), c(NA, NA, 104, NA, NA, 81), c(140, 58, 88, 309, 197, 11)
- There are 599 pathways, which only translates into 213 genes, not all of which are found in the proteome gene list.

Lung vs heart

RNAseqsamplesize (Zhao et al. 2018) was used to do the analysis. With expected fold change between groups = 2, FDR set = 0.01 and number of samples = 5, power was computed to be 0.029 for the selected genes of interest which is poor. Some gene IDs pulled from Biomart are not in the dataset. Dispersion was also computed as 0.3663078.

```
heart_left_vent <- filter(tissue_map, SMTSD == "Heart - Left Ventricle")$SAMPID %>%
    as.character
lung <- filter(tissue_map, SMTSD == "Lung")$SAMPID %>% as.character
lung_heart <- select(data, Name, Description, any_of(c(heart_left_vent, lung)))
lh_datamat <- data.matrix(lung_heart[, 3:length(colnames(lung_heart))])
rownames(lh_datamat) <- lung_heart$Name

# estimate gene read count and dispersion distribution
distribution <- est_count_dispersion(lh_datamat, group = colnames(lh_datamat) %in%
    lung %>% as.numeric)
```

```
## Disp = 0.36631 , BCV = 0.6052
```

```
# power estimation 6 selectedGenes were not found in distributionObject,
# discarded
power_n5 <- est_power_distribution(n = 5, f = 0.01, rho = 2, distributionObject = distribution,
    selectedGenes = genelist, storeProcess = TRUE)
mean(power_n5$power)  # 0.02885
```

```
## [1] 0.02885254
```

```
power_n10 <- est_power_distribution(n = 30, f = 0.05, rho = 2, distributionObject = distribution,
    selectedGenes = genelist, storeProcess = TRUE)
mean(power_n10$power)
```

```
## [1] 0.825588
```

From changing around FDR (`fdr` parameter in `est_power_curve()`) and coverage (`lambda0` parameter in `est_power_curve()`) as well as making an optimization plot, it appears that around 35 samples are needed with an average coverage > 80 reads/gene depending on desired FDR in order to achieve >80% power. To be precise, 35 samples with 80 coverage and FDR=0.05 will give 80% power (equation is `est_power(n=35, lambda0=80, phi0=distribution$common.dispersion, f=0.05, m=56200,m1=158)`= 0.8). You could probably get away with a few less than 35 samples, as coverage will likely exceed 80 if sequencing is closer to 30M reads.

Notes

It should be noted that increasing sample size increases power much more than increasing the coverage after a certain depth (approx. 10M reads/sample) (Liu, Zhou, and White 2014). This can be seen in the optimization plots. From `est_power()`, `m=` the number of genes for testing, i.e. the number of genes in the full data matrix, while `m1=` the expected number of prognostic genes, so the number of genes we are interested in.

Relationship between coverage and sequencing depth is relatively simple: if `N` is the number of reads, `L` is the average read length, and `G` is the genome (or transcriptome) size, then coverage will be `N x L / G`. RNAseq experiment coverage is also significantly affected by the relative expression of the genes of interest, but on average 30M 75bp reads can reach 90% of transcriptome coverage (Wang et al. 2011).

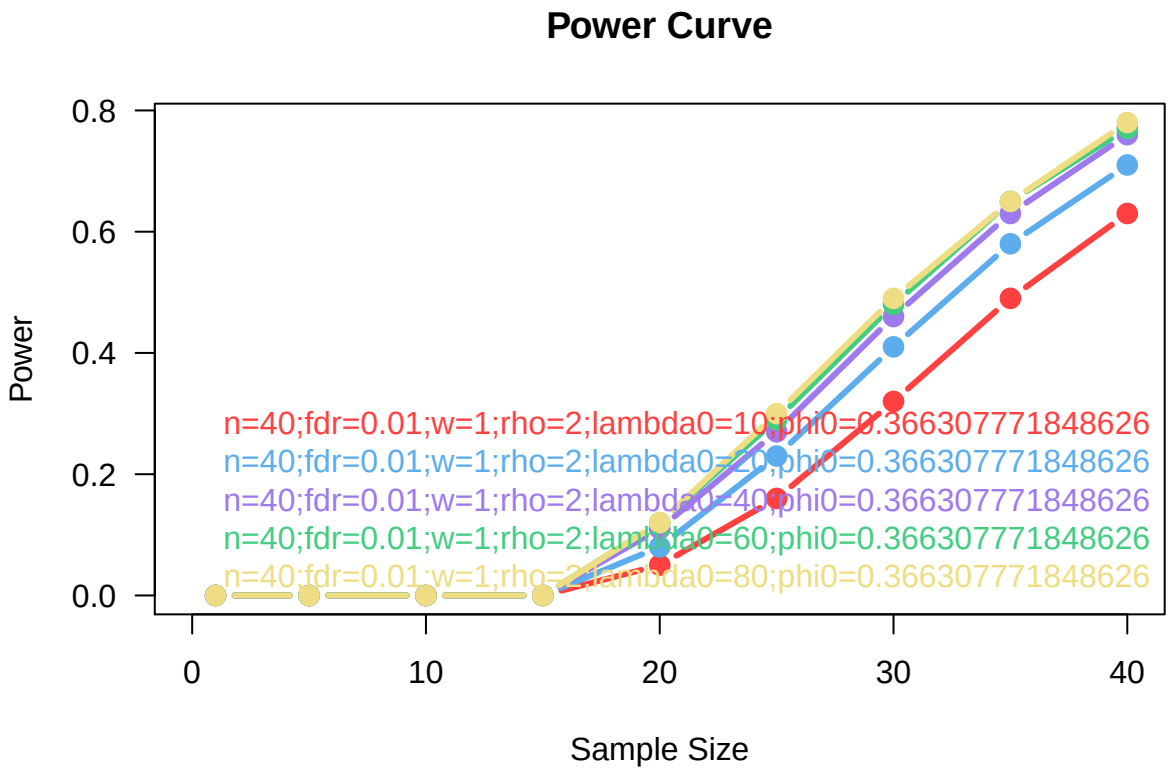
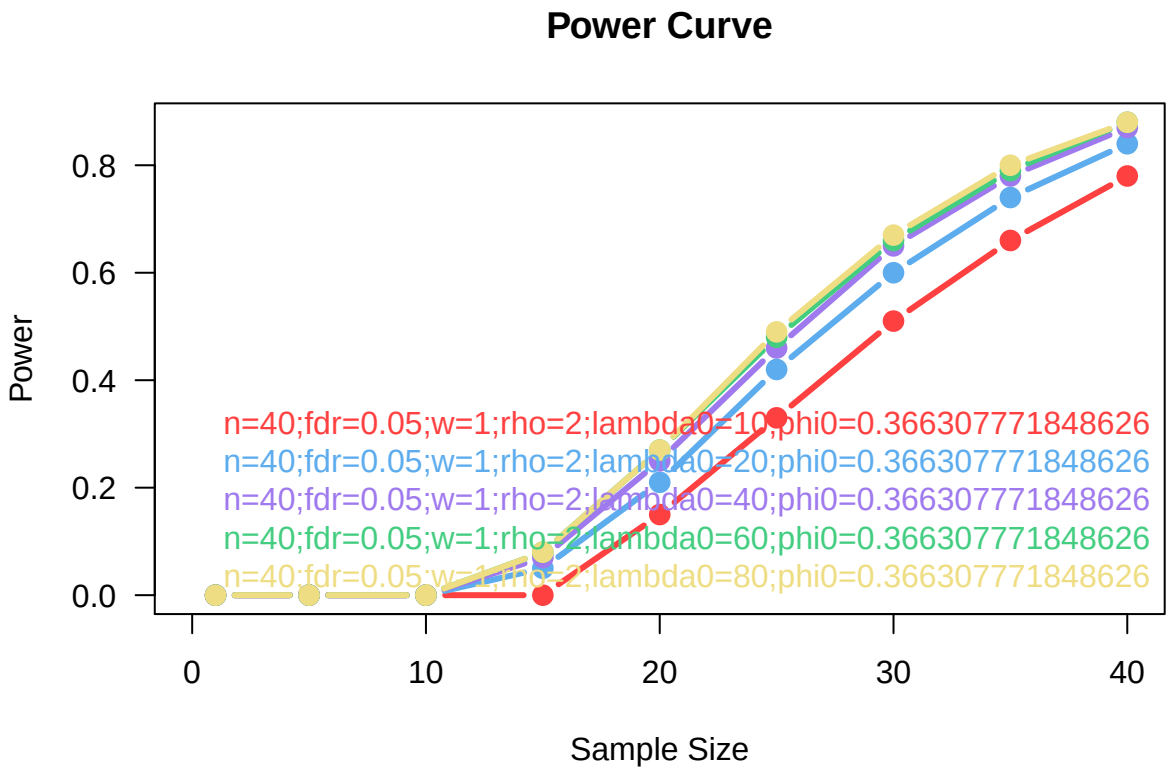


Figure 1: Red line is FDR=0.01, coverage=10. Blue line is FDR=0.05, coverage=5. Purple line is FDR=0.01, coverage=10. Green line is FDR=0.05, coverage=10. Yellow line is FDR=0.05, coverage=20.

```
coverage10_fdr5 <- est_power_curve(n=40,f=0.05,rho=2,lambda0=10,phi=distribution$common.dispersion, m=5
coverage20_fdr5 <- est_power_curve(n=40,f=0.05,rho=2,lambda0=20,phi=distribution$common.dispersion, m =
coverage40_fdr5 <- est_power_curve(n=40,f=0.05,rho=2,lambda0=40,phi=distribution$common.dispersion, m=5
coverage60_fdr5 <- est_power_curve(n=40,f=0.05,rho=2,lambda0=60,phi=distribution$common.dispersion, m =
coverage80_fdr5 <- est_power_curve(n=40,f=0.05,rho=2,lambda0=80,phi=distribution$common.dispersion, m=5
plot_power_curve(list(coverage10_fdr5,coverage20_fdr5,coverage40_fdr5,coverage60_fdr5,coverage80_fdr5))
```




```
est_power(n = 8, lambda0 = 20, phi0 = 0.07154, f = 0.05, m = 52000, m1 = 71)
```

```
## [1] 0.51
```

```
power35 <- est_power_distribution(n = 35, f = 0.05, rho = 2, distributionObject = distribution,
    selectedGenes = genelist, storeProcess = TRUE)
```

```
## Warning in selecteGeneByName(selectedGenes, distributionObject): 6 selectedGenes were not found in di
```

```
mean(power35$power)
```

```
## [1] 0.8712377
```

This experiment doesn't really suffer from too many lowly expressed genes, it does seem like a few (LGALS1, LAMA1, PHPT1) will have coverage less than 50.

```
gene_readcounts <- distribution$pseudo.counts.mean[which(names(distribution$pseudo.counts.mean) %in% ge
gene_dispersions <- distribution$tagwise.dispersion[which(names(distribution$pseudo.counts.mean) %in% g
mean(gene_readcounts)
```

```
## [1] 10975.04
```

```
min(gene_readcounts)
```

```
## [1] 3.129038
```

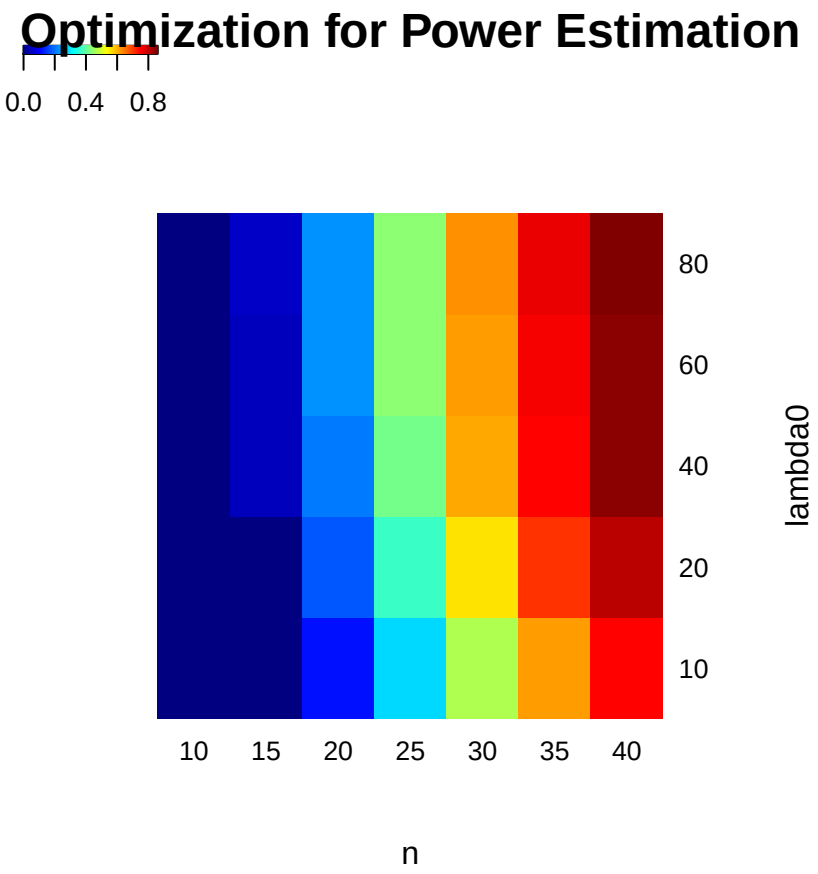


Figure 2: Blue to brown gradient shows power from 0 to 1. Here FDR=0.01.

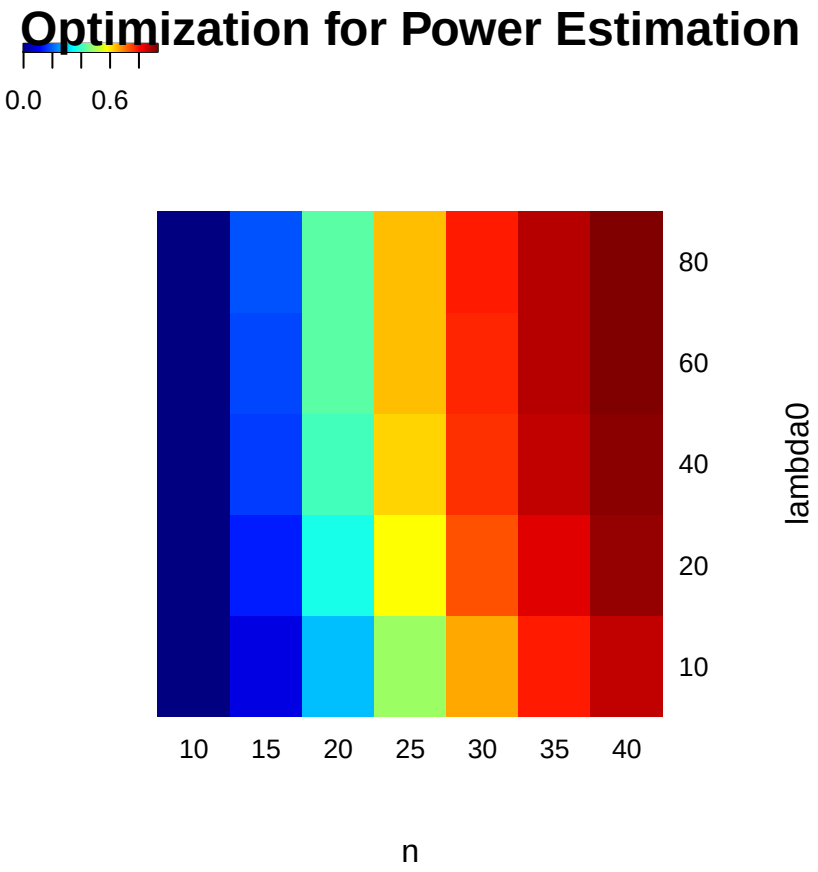
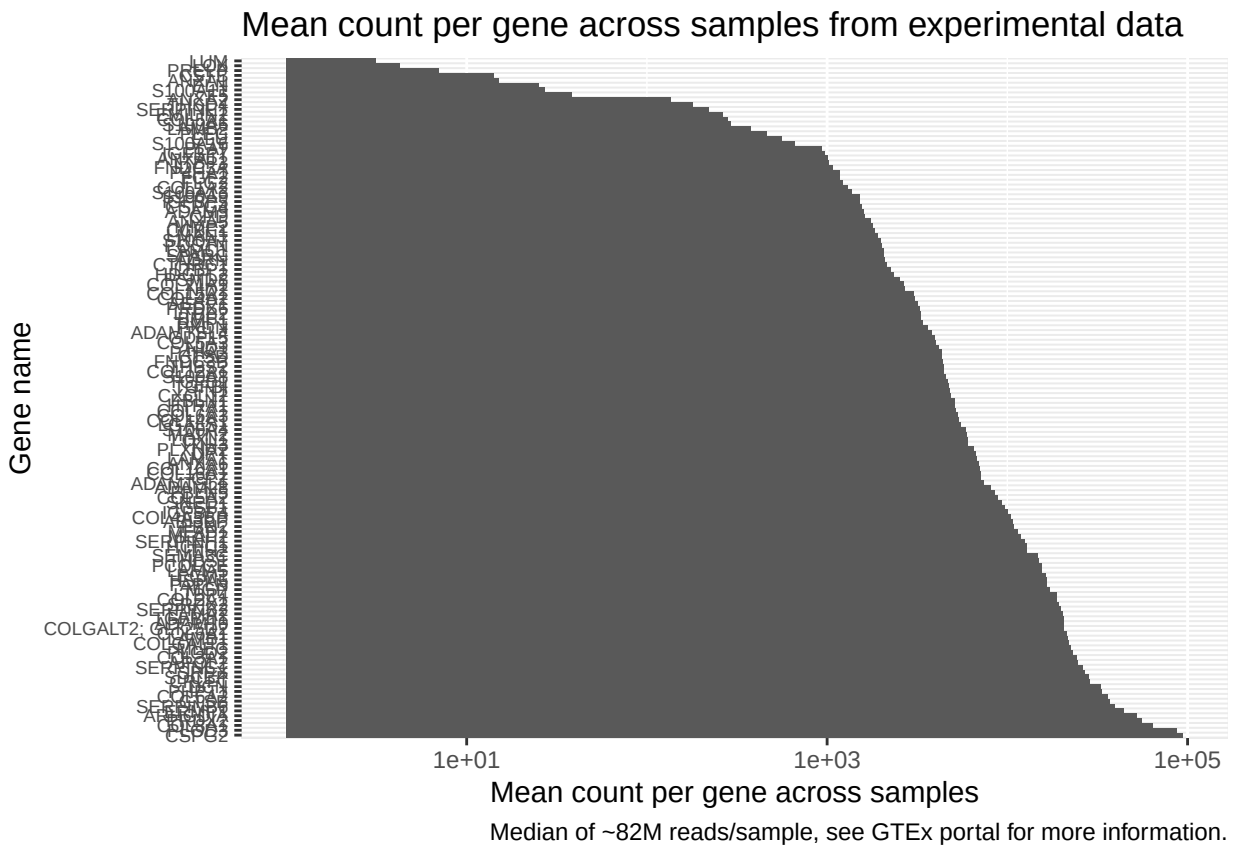


Figure 3: Blue to brown gradient shows power from 0 to 1. Here FDR=0.05.

```
mean(gene_dispersions)
```

```
## [1] 0.3073066
```

```
ggplot(data=data.frame(counts=as.numeric(gene_readcounts),name=genenames[order(match(names(gene_readcou
```



References

Liu, Yuwen, Jie Zhou, and Kevin P. White. 2014. "RNA-seq differential expression studies: More sequence or more replication?" *Bioinformatics* 30 (3). https://doi.org/10.1093/bioinformatics/btt688.

Wang, Ying, Noushin Ghaffari, Charles D. Johnson, Ulisses M. Braga-Neto, Hui Wang, Rui Chen, and Huaijun Zhou. 2011. "Evaluation of the coverage and depth of transcriptome by RNA-Seq in chickens." *BMC Bioinformatics* 12 (SUPPL. 10). https://doi.org/10.1186/1471-2105-12-S10-S5.

Zhao, Shilin, Chung I. Li, Yan Guo, Quanhu Sheng, and Yu Shyr. 2018. "RnaSeqSampleSize: Real data based sample size estimation for RNA sequencing." *BMC Bioinformatics*. https://doi.org/10.1186/s12859-018-2191-5.