Midterm

Brian Lee
COP5859: Semantic Web Programming

Dr. Ravi Shankar
Department of Engineering
Florida Atlantic University

April 4, 2019

Please refer to the folder ’Exam Documents’ in the Files section. I have uploaded there a chapter from the book: "Genetics of Obesity’ by S.F.A. Grant. It documents all the Genome-Wide Association Studies as pertinent to obesity. Please study the chapter and make a list of the genes and gene products discussed.

Part 1. Conduct a GO Term enrichment study and determine the biological pathways, molecular functions, and cellular components that are involved in obesity. Summarize into a 4-5 page report - 15 points.

Part 2. Obesity and cardiovascular diseases may overlap. Use your research topic as representative of cardiovascular diseases. Find ways to determine common pathways involved in both. Read more on such pathways. Develop insight on possible reasons for this overlap. Summarize into a 4-5 page report - 15 points.

For both parts, report format is the same as that used in your first assignment submission. For part 2, there is no correct answer. It is possible you may not find any overlap. However, good due diligence is expected, both at the level of the determination of the overlap (smart use of the tools) and and possible explanation for overlap/non-overlap (review of relevant literature).

1 Part 1

1.1 Background

The central dogma of biology is that the essentials of life start at the level of DNA, DNA sequences are read and transcribed into mRNA, and mRNA is then translated into protein. Proteins are the main structural and functional macromolecules of cells that allow a cell to grow, develop, and proliferate. While there are exceptions to these processes, such as mRNA being the final product of DNA transcription, protein products and the genes that encode them are the main focus in the investigation possible pathways and mechanisms of disease.

When describing genes and their relation to disease, there is often a reference to genetic variants. Genetic variants are instances of a particular gene that differ from one individual to another. These differences are caused by mutations in the DNA. The different forms of a gene in genetic variants are sometimes referred to as *alleles.* When talking about significant genetic variants, researchers often refer to the *minor allele frequency.* The minor allele frequency is the frequency at which the second most common allele occurs in a population as compared to the wild type. When the frequency is $> 5\%$, then that allelic mutation is considered a *common variant* (Pierce, 2017).

Genetic variants arise from various mutations, most often from point mutations. Point mutations are single base substitutions that can have varying effects. In a *silent* mutation, the changed base has no effect on the amino acid incorporated into the protein, because it occurs in the wobble-base position of the codon, and codes for the same amino acid. In a *nonsense* mutation, there are major repercussions, because the mutation causes a stop codon to be incorporated in the middle of the protein chain which stops translation. In a *missense* mutation, the point mutation can cause an amino acid substitution that has similar chemical and physical properties to the original amino acid, and so there is little effect. However, if a chemically different amino acid is introduced, this can have a major effect on the protein's structure and hinder or completely inhibit the protein's function.

Alleles that carry these point mutations are called *single nucleotide variants.* If these single nucleotide variants occur in gametes and are inherited by offspring, they begin to integrate into the population as they are passed from generation to generation. If they become common enough, to the point where these variants occur in $> 1\%$ of the population, then they are referred to as *single nucleotide polymorphisms* (SNPs). It is at this point that the mutation is usually common enough to gain attention, and it's potential role in causing disease is investigated (Pierce, 2017).

There are two main methods toward identifying the gene(s) responsible for a particular disease or phenotype: linkage analysis and association. Historically, finding disease causing genes has been done through genetic linkage analysis studies. Linkage analysis is based on the principle of identifying linked genes. Before independent assortment of chromatids occurs in meiosis, crossing over between chromosomes occurs which exchanges fragments between duplicated chromosomes. This causes genes in the genome to be seen as inherited independently of most other genes in the genome. In an oversimplified example, this means that having brown eyes doesn't make you any more likely to also have brown hair. However, if the genes for brown hair and brown eyes are located closely together on a chromosome, then they will likely be conserved together no matter how much genetic recombination occurs. Therefore, having brown eyes would make you more likely to also have brown hair, thus the genes are said to be "linked".

Before the use of next-generation DNA sequencing, scientists relied on polymorphic genetic markers to look for the presence of particular sequences. These markers can be easily screened and cross referenced with a familiar pedigree. Instead of looking for a linkage between two sequences or between two phenotypes, like in the example above, researchers would look at the linkage between a specific genetic marker and a phenotype (the disease). If a specific genetic marker is found that is conserved for diseased family members, then there is a linkage between that genetic marker and the disease. Therefore, it can be hypothesized that the gene responsible for that phenotype is located in close proximity in the chromosome to the marker. Since the location of the marker is known, further studies can be conducted to narrow down and eventually identify the gene responsible for the disease (Oikkonen, n.d.).

This method works well for highly penetrative binary traits, but most common diseases are *quantitative*, meaning that they are controlled by multiple genes. Because of the advent of whole genome sequencing and its subsequent dramatic price decrease, the modern method of finding disease has been shifted toward genetic association studies, or, more specifically, genome wide association studies (GWAS). These types of studies are much better for assessing quantitative diseases. They use statistical analysis to look for direct associations between traits and genetic loci. They do not rely on familial pedigree, but instead use large populations of case and control and are much higher resolution than linkage based studies (Oikkonen, n.d.).

1.2 Methods

The following set of genes from both pre-GWAS and GWAS obesity studies was obtained from (Grant, 2014):

```
LEP , LEPR , MC4R , POMC , PCSK1 , SNRPN , NDN , CCDC28B , ARL6 ,
BBS1 , ALMS1 , AGRP , BDNF , SDC1 , SDC3 , SIM1 , CARTPT , UCP1 ,
UCP3 , GHRL , GHSR , PYY , PPARG , PPARGC1A , NR0B2 , ENPP1 , ADRB2 ,
 ADRB3 , ADCY3 , CADM2 , KCTD15 , LRP1B , MAP2K5 , MTCH2 , MTIF3 ,
NRXN3 , NUDT3 , PRKD1 , PTBP2 , RPL27A , SLC39AB , TMEM150A ,
TNNI3K , ZNF608 , ETV5 , FAIM2 , FANCL , FLJ3577 , FTO , GIPR ,
GNPDA2 , GRPC5B , LINGO2 , NEGR1 , SEC16B , SH2B1 , TFAP2B , TMEM18
, HS6ST3 , KCNMA1 , MAF , OLFM4 , PACS1 , PRKCH , RMST , TNKS , MSRA
, ZZZ3 , ADAMTS9 , CPEB4 , DNM3 , PIGC , GRB14 , HOXC13 , ITPR2 ,
SSPN , LY96 , LYPLAL1 , NFE2L3 , NISCH , STAB1 , RSPO3 , TBX15 ,
WARS2
```

There are many Gene Ontology (GO) enrichment services available, both web and local based. The R library *clusterProfiler* was used to perform GO enrichment because of it's ability to produce easy to read enrichment bar graphs (Yu, Wang, Han, & He, 2012).

```
> library(org.Hs.eg.db)
> library(clusterProfiler)
> library(enrichplot)
> OB_Genes <- readLines('ob_gene_list.txt')
> OB_entrez_temp <- bitr(OB_Genes, fromType="SYMBOL",
+ toType="ENTREZID", OrgDb=org.Hs.eg.db)
> OB_entrez <- OB_entrez_temp$ENTREZID
> OB_BP_GO <- enrichGO(gene=OB_entrez, universe=Human_Genes,
+ OrgDb=org.Hs.eg.db, ont="BP", pAdjustMethod="bonferroni",
+ pvalueCutoff=0.03, qvalueCutoff=0.03)
> barplot(OB_BP_GO)

> OB_MF_GO <- enrichGO(gene=OB_entrez, universe=Human_Genes,
+ OrgDb=org.Hs.eg.db, ont="MF", pAdjustMethod="bonferroni",
+ pvalueCutoff=0.03, qvalueCutoff=0.03)
> barplot(OB_MF_GO)

> OB_CC_GO <- enrichGO(gene=OB_entrez, universe=Human_Genes,
+ OrgDb=org.Hs.eg.db, ont="CC", pAdjustMethod="bonferroni",
+ pvalueCutoff=0.03, qvalueCutoff=0.03)
> barplot(OB_CC_GO)
```

1.3 Results

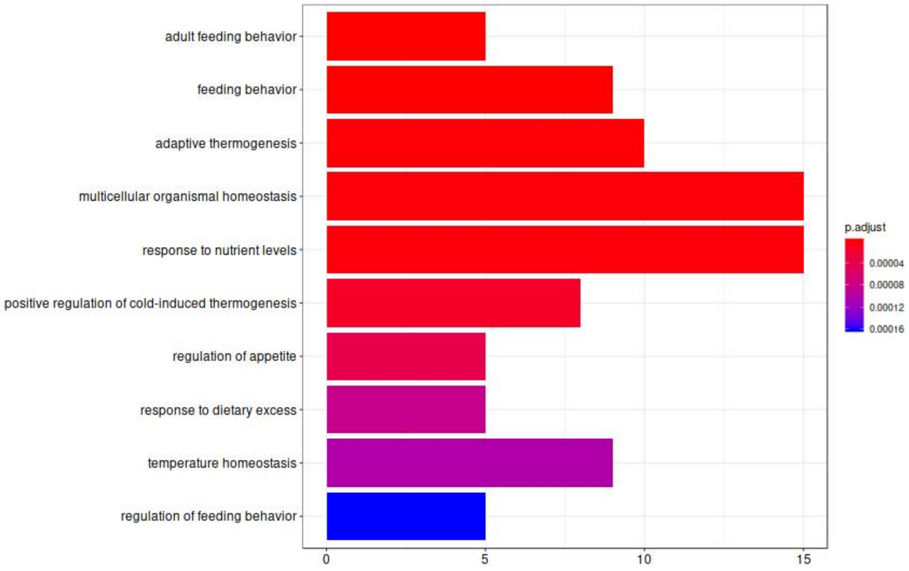

Figure 1: Obesity Biological Process GO Enrichment

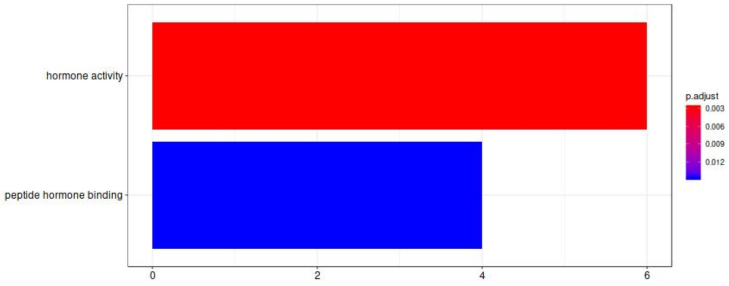

Figure 2: Obesity Molecular Function GO Enrichment

There was no significant enrichment terms for the Cellular Component Ontology.

1.4 Discussion

In addition to using the clusterProfiler, results were obtained from web-based enrichment services from PantherDB and Princeton's GOTermFinder. All results were nearly identical, with slight variations in p-values. Note, that we are not performing Gene Set Enrichment Analysis during these runs, but rather Over Representation Analysis as outlined by (Boyle et al., 2004), where enrichment is determined by hypergeometric distribution followed by p-value adjustment by Bonferroni correction. The difference in p-values was due to the number of genes defined in the database being used.

The results from the enrichment for biological processes correlated well with what one might expect from looking at obesity related processes. The enriched molecular functions of hormone activity and binding are also not suprising. The lack of significant cellular components is due to the fact that the cellular component type of gene/gene product is not relevant to it's role obesity.

1.5 Conclusion

There are many tools available to look at which GO terms are enriched for a set of genes. This allows a quick method for viewing what types of characteristics are shared and signficicant for the gene set. The different tools available generally give the same output, it is the quality of the input that really determines the reliability of the results.

2 Part 2

2.1 Background

The relation between obesity and cardiovascular diseases is a continuing area of research and investiagtion. There has been less research however, on the effect of obesity on cardiac repiar after surgery. One study has looked specifically at the role of diet-induced obesity on inflammation and remodelling after myocardial infarction in mice(Thakker et al., 2006).

The reason for using diet-induced obesity mice is that normally obese mice are induced to be obese by creating mutations in the leptin gene which alters the immune system. The authors conclude that this may interfere with myocardial repair, and, since most obesity is a result of high-calorie diet combined with sedentary lifestyle, it would not accurately reflect the majority of cases. The authors observed that adverse remodeling did occur in obese mice in response to myocardial infarction. We can look at biological pathways of genes involved with obesity and myocardial repair to see if the authors were justified in being concerned with genetically altered mice in their experiment.

2.2 Methods

Genes related to obesity were obtained from (Grant, 2014), and their biological pathways enriched based on the ReactomeDB using ReactomePA (Yu & He, 2016).

```
> library(org.Hs.eg.db)
> library(reactome.db)
> library(ReactomePA)
> library(enrichplot)
> OB_Genes <- readLines('ob_gene_list.txt')
> OB_entrez_temp <- bitr(OB_Genes, fromType="SYMBOL",
+ toType="ENTREZID", OrgDb=org.Hs.eg.db)
> OB_entrez <- OB_entrez_temp$ENTREZID
> OB_Reactome <- enrichPathway(gene=OB_entrez, pvalueCutoff=0.05,
+ minGSSize=5, readable=T)
> barplot(OB_Reactome)
```

This was then repeated using the KEGG database using clusterProfiler(Yu et al., 2012).

```
> library(org.Hs.eg.db)
> library(clusterProfiler)
> library(enrichplot)
> OB_Genes <- readLines('ob_gene_list.txt')
> OB_entrez <- bitr(OB_Genes, fromType="SYMBOL",
+ toType="ENTREZID", OrgDb=org.Hs.eg.db)
> OB_entrez <- as.numeric(OB_entrez$ENTREZID)
> OB_entrez <- sort(OB_entrez, decreasing=TRUE)
> OB_KEGG <- enrichKEGG(gene=OB_entrez, organism='hsa',
+ minGSSize=5)
> barplot(OB_KEGG)
```

Genes related to post-myocardial infarction cardiac repair were obtained from (Frangogiannis, 2017)(Hudalla & Murphy, 2015)(Jourdan-LeSaux, Zhang, & Lindsey, 2010), and their biological pathways enriched based on the ReactomeDB using ReactomePA (Yu & He, 2016).

```
> library(org.Hs.eg.db)
> library(reactome.db)
> library(ReactomePA)
> library(enrichplot)
> MI_Genes <- readLines('mi_gene_list.txt')
> MI_entrez_temp <- bitr(MI_Genes, fromType="SYMBOL",
+ toType="ENTREZID", OrgDb=org.Hs.eg.db)
> MI_entrez <- MI_entrez_temp$ENTREZID
> MI_Reactome <- enrichPathway(gene=MI_entrez, pvalueCutoff=0.05,
+ minGSSize=5, readable=T)
> barplot(MI_Reactome)
```

This was then repeated using the KEGG database using clusterProfiler(Yu et al., 2012).

```
> library(org.Hs.eg.db)
> library(clusterProfiler)
> library(enrichplot)
> MI_Genes <- readLines('mi_gene_list.txt')
> MI_entrez <- bitr(MI_Genes, fromType="SYMBOL",
+ toType="ENTREZID", OrgDb=org.Hs.eg.db)
> MI_entrez <- as.numeric(MI_entrez$ENTREZID)
> MI_entrez <- sort(MI_entrez, decreasing=TRUE)
> MI_KEGG <- enrichKEGG(gene=MI_entrez, organism='hsa',
+ minGSSize=5)
> barplot(MI_KEGG)
```

2.3 Results

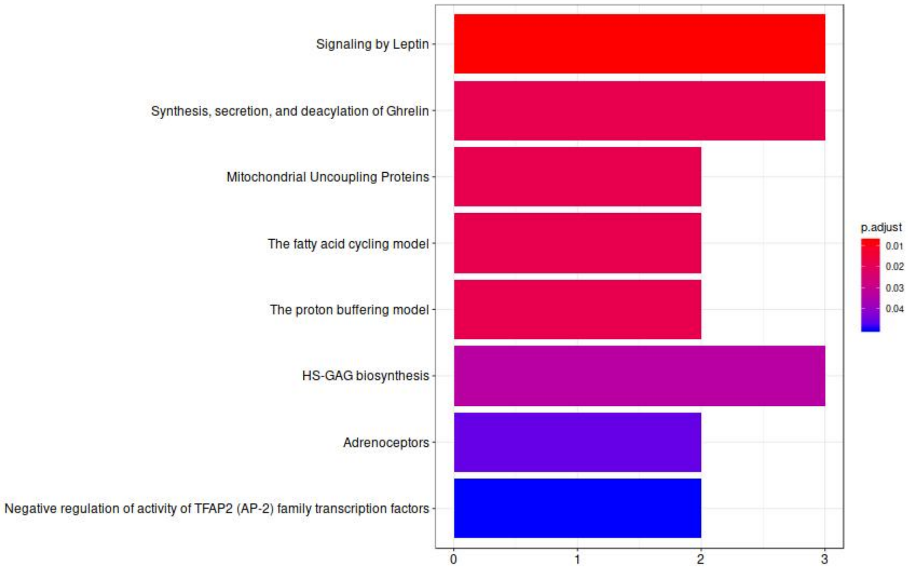

Figure 3: Obesity Reactome Pathway Enrichment

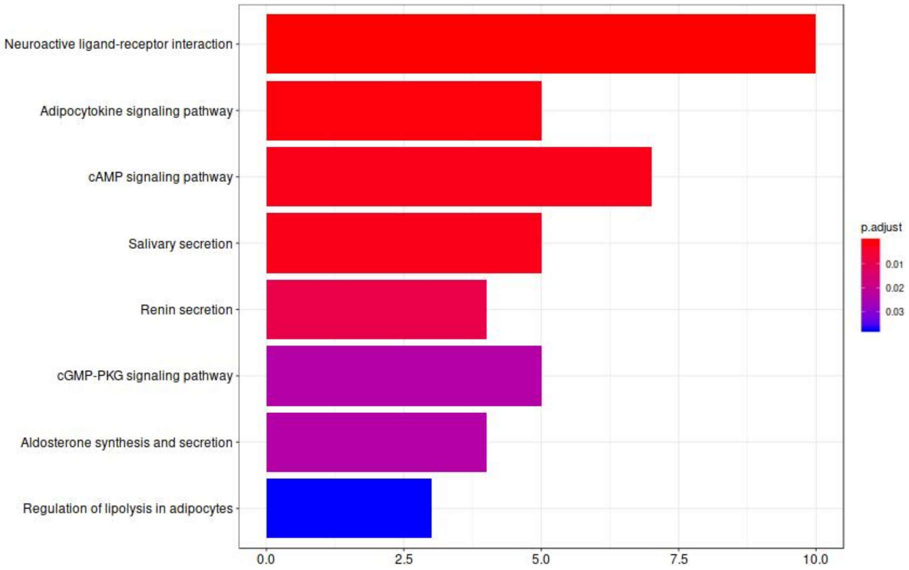

Figure 4: Obesity KEGG Pathway Enrichment

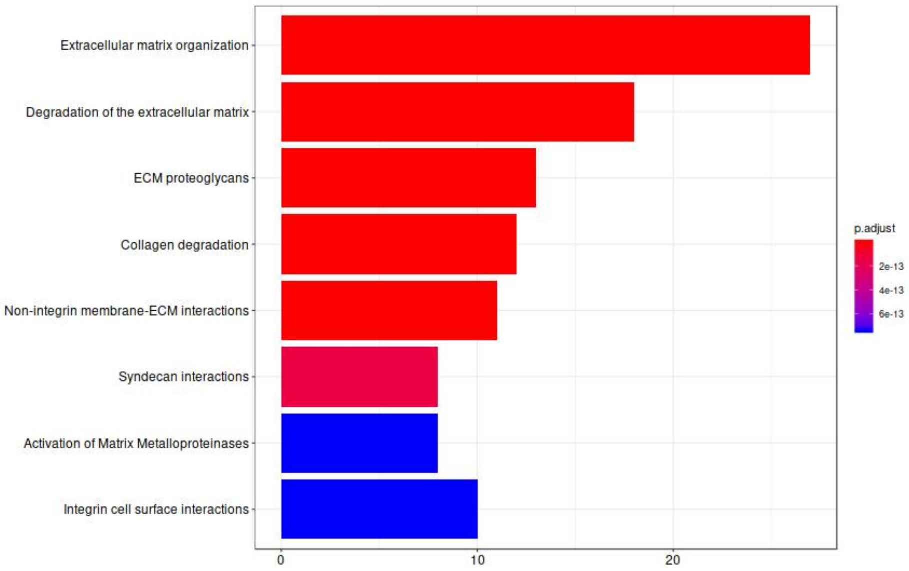

Figure 5: MI Repair Reactome Pathway Enrichment

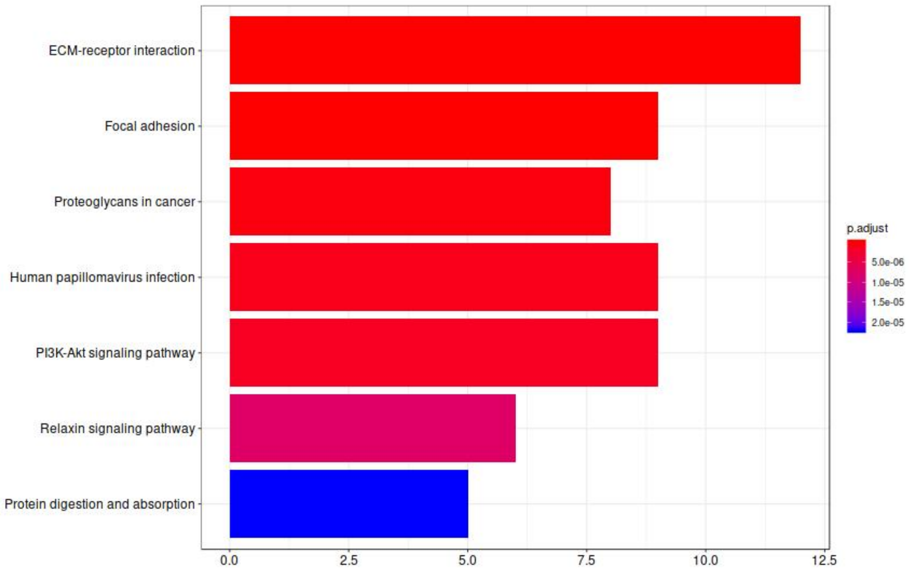

Figure 6: MI Repair KEGG Pathway Enrichment

2.4 Discussion

When comparing pathways of genes involved with obesity and myocardial infarction, there is no commonality in either the Reactome or KEGG databases. This does not mean that obesity does not effect cardiac remodelling, but rather gives evidence that if their is any effect, it is not occuring due to a direct involvement in the same pathway. Rather, other downstream effects of obesity may be contributing to adverse remodelling.

This provides evidence that the authors of the previous study may not have needed to worry about genetically induced obese mice, because the genes would not effect the outcome of the study. This is further supported by the fact that similar results were obtained by other researchers who did use leptin gene altered mice (Greer, Ware, & Lefer, 2006).

2.5 Conclusion

Enrichment studies offer clues toward further research, and may help to guide further hypothesis. The evidence presented here indicates that it may not be worthwhile further investigating the direct role of obesity related genes and their effect on myocardial repair.

References

Boyle, E. I., Weng, S., Gollub, J., Jin, H., Botstein, D., Cherry, J. M., & Sherlock, G. (2004, December 12). GO::TermFinder—open source software for accessing Gene Ontology information and finding significantly enriched Gene Ontology terms associated with a list of genes. *Bioinformatics*, *20*(18), 3710–3715. doi:10.1093/bioinformatics/bth456

Frangogiannis, N. G. (2017, May 1). The extracellular matrix in myocardial injury, repair, and remodeling. *Journal of Clinical Investigation*, *127*(5), 1600–1612. doi:10.1172/JCI87491

Grant, S. F. A. (Ed.). (2014). *The genetics of obesity*. New York: Springer. OCLC: ocn857977298.

Greer, J. J. M., Ware, D. P., & Lefer, D. J. (2006, January). Myocardial infarction and heart failure in the db/db diabetic mouse. *American Journal of Physiology. Heart and Circulatory Physiology*, *290*(1), H146–153. doi:10.1152/ajpheart.00583.2005. pmid: 16113078

Hudalla, G. H., & Murphy, W. L. (2015, November 18). *Mimicking the Extracellular Matrix: The Intersection of Matrix Biology and Biomaterials*. Royal Society of Chemistry.

Jourdan-LeSaux, C., Zhang, J., & Lindsey, M. L. (2010, September). Extracellular matrix roles during cardiac repair. *Life Sciences*, *87*(13-14), 391–400. doi:10.1016/j.lfs.2010.07.010

Oikkonen, J. (n.d.). Comparison of genome wide linkage and association analysis methods for a quantitative trait in complex extended pedigrees, 91.

Pierce, B. A. (2017). *Genetics: A conceptual approach* (Sixth edition). New York: W.H. Freeman/Macmillan Learning.

Thakker, G. D., Frangogiannis, N. G., Bujak, M., Zymek, P., Gaubatz, J. W., Reddy, A. K., … Ballantyne, C. M. (2006, November). Effects of diet-induced obesity on inflammation and remodeling after myocardial infarction. *American Journal of Physiology-Heart and Circulatory Physiology*, *291*(5), H2504–H2514. doi:10.1152/ajpheart.00322.2006

Yu, G., & He, Q.-Y. (2016, January 26). ReactomePA: An R/Bioconductor package for reactome pathway analysis and visualization. *Molecular BioSystems*, *12*(2), 477–479. doi:10.1039/C5MB00663E

Yu, G., Wang, L.-G., Han, Y., & He, Q.-Y. (2012, March 28). clusterProfiler: An R Package for Comparing Biological Themes Among Gene Clusters. *OMICS: A Journal of Integrative Biology*, *16*(5), 284–287. doi:10.1089/omi.2011.0118