IBS574 Project Write-up

Kameryn McCarty

April 15, 2014

**Functions of genes differentially expressed in cancers highly associated with Type 2 Diabetes**

Introduction:

Cancer and Type 2 Diabetes (T2D) are both growing health problems, in the U.S. and worldwide. The incidence of both diseases being diagnosed in the same person is much higher than would be expected by chance and the presence of one disease makes the treatment of the other increasingly difficult. Cancer and diabetes share many risk factors such as age, obesity, physical activity, and diet among others. There is evidence to suggest that metabolic dysfunction found in diabetic patients can contribute to the progression of cancerous tumors by providing an environment high in insulin and glucose (hyperinsulinemia and hyperglycemia)1. In models of metabolic dysfunction specifically, tumor cells appear to be able to overcome the insulin resistant phenotype through a Wnt signaling pathway and it is predicted that the cells can then take advantage of the high circulating levels of insulin and glucose in the blood to facilitate their growth and energy needs2.

Individuals with T2D are observed to have a greatly increased risk for cancers of the liver, pancreas, and endometrium, as well as colon, breast and bladder to a lesser extent1. Approximately 24 cancers have been described to have increased incidence in individuals with T2D3. To date, lung cancer has not been found to be associated with diabetes, and prostate cancer appears to be less common in diabetic men1,3. If metabolic dysfunction in diabetes is what enhances tumor development and progression, why aren’t all tissues equally susceptible to tumors? In my project, I aimed to examine differential gene expression patterns in several cancers that diabetic patients are at a greater risk for developing and compare them to the expression patterns of the less common or unassociated cancers in order to identify genes commonly dysregulated in the high risk cancers. I then performed Gene Ontology (GO) and pathway annotation and overrepresentation analysis on the resulting list of genes to examine if these genes were enriched for functions or pathways also dysregulated in diabetic patients, such as glucose metabolism or inflammation.

Project Workflow:

To begin, I downloaded normalized RNA-seq data for five cancers (liver hepatocellular carcinoma, LIHC; uterine corpus endometrial carcinoma, UCEC; lung adenocarcinoma, LUAD; and prostate adenocarcinoma, PRAD; pancreatic adenocarcinoma, PAAD) from The Cancer Genome Atlas (TCGA). This included a file for each cancer containing a table of normalized read counts for 20,531 genes for each sample included. Using python scripts, I was able to exact from this table the sample descriptions and classify each sample as normal (1) or tumor (2), which is necessary for differential expression analysis, based on the barcode system TCGA uses to name their samples. I employed the edgeR package in R/Bioconductor to identify genes differentially expressed (DE) between normal and tumor samples separately for each cancer type. I read in my table of read counts and the sample group classification and then ran the classic analysis, which involved calculating common and tagwise dispersions of the data and performing a fisher’s exact test for each gene. I had the program return to me all of the genes sorted by p-value for differential expression. The analysis also returned log of fold change (FC), log of counts per million (CPM), and FDR values for each gene.

Once I had files for each cancer, I examined the quality of the data by generating plots of fold change versus gene expression (logFC vs. logCPM) and coloring data points based on FDR (above or below 0.1) (Figure 1). From this, I could tell that the PAAD dataset did not appear to have enough statistical power to call differential expression appropriately, since it only had two normal samples. As a result, I elected to exclude the PAAD data from further analyses. Because there were many small p-values in the remaining cancer sets, I used a more conservative Bonferroni correction to calculate my cutoff threshold for significant differential expression (0.05/20531 = 2.44e-06). Using basic R functions, I created a data frame of the p-values of all the tested genes for the four cancers. Next I created logic vectors that indicated if each gene was significant based on my p-value cutoff for the T2D-risk cancers and the non-risk cancers. For the risk cancers, TRUE indicated that the p-value was significant for a gene in both the LIHC and UCEC cancers and FALSE meant that the gene was not significant in both cancers. Similarly for the non-risk cancers, TRUE indicated that a gene was significant in both LUAD and PRAD cancers and FALSE meant it is not significant in both. I created a two by two matrix using these vectors and performed a Fisher’s exact test on the TRUE/FALSE counts. This test indicated that I had significantly different numbers of genes in each category than expected, specifically there were fewer genes significant in both T2D-associated cancers and not in the non-associated cancers than would be expected and vice versa. I made another data frame from the TRUE/FALSE vectors and used that to return the genes that were commonly differentially expressed in the T2D-risk cancers but not in the non-risk cancers, referred to as “T2D-associated”. I also obtained a list of genes DE in both non-risk cancers but not the T2D-risk cancers, termed “Non-associated”. I repeated this process again to find genes that were DE in both risk-cancers but not in either non-risk cancer and still found that the number of genes was significantly different from expected using the Fisher’s exact test. I returned these genes to a new file and distinguish them from the other list of genes by referring to them as “T2D-associated only”.

Finally, I used these three gene lists to examine abundances of gene ontology terms. For this I used the web-based sites for PANTHER Classification System and DAVID Functional Annotation tool and input the Entrez IDs of each list. For PANTHER, I performed functional classification and statistical overrepresentation tests and in DAVID, I used the functional annotation tool to examine pathways and gene ontology terms annotated to the genes.

Results and Discussion

From the DE analysis with edgeR, I identified thousands of differentially expressed genes in each cancer (Table 1). There were 1182 significant DE genes that were common to LIHC and UCEC but not to both LUAD and PRAD cancers and 2556 genes common to LUAD/PRAD but not LIHC/UCEC cancers. A smaller number of genes (316) were common to LIHC/UCEC but not LUAD or PRAD individually (figure 2). Overall the distribution of GO terms within the biological process, molecular function, and cellular component categories as annotated by PANTHER appeared similar across the three gene lists (figure 3). Some detectable differences were visible when examining the distribution of genes in biological pathways in PANTHER. In particular, pathways related to Ras signaling and inflammation mediated by chemokine and cytokine signaling were significantly enriched in the smaller T2D-risk cancer gene list (figure 4).

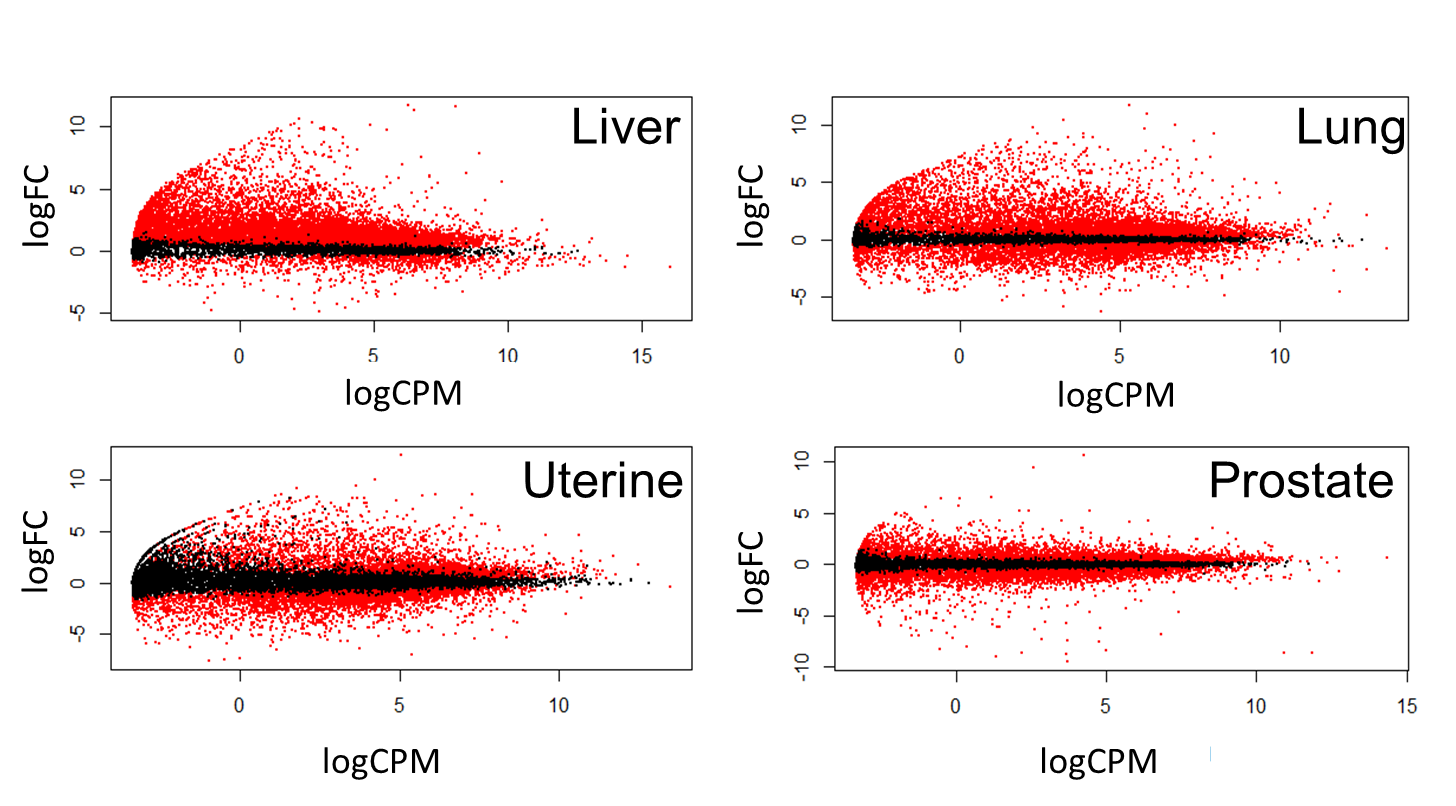
For the Non-associated DE genes, the list was enriched for terms and pathways related to metabolism and biogenesis of biological macromolecules, pathways which are upregulated in cancer cells generally. The T2D-associated DE genes generated terms related to cell cycle, energy production, and neurodegenerative pathways. The more narrowed T2D-associated only list also returned terms associated with energy production, inflammation, and neurodegenerative diseases. Diabetic patients have altered levels of insulin and glucose, which are intimately related to energy production by the mitochondria of cells. Energy production would therefore be expected to be disrupted in diabetic patients. Cancer cells also require a lot of energy to proliferate and might also be expected to have altered energy production. T2D is associated with increased cellular inflammation and is also considered a risk factor for cognitive dysfunction such as is found in Alzheimer’s, Huntington’s, and Parkinson’s diseases. Interestingly, both of the T2D-associated gene lists returned pathways annotated by Reactome, a database of biological pathways, related to diabetes and integration of energy metabolism.

Based on these results, it would appear that the two cancers that are more commonly diagnosed in T2D patients exhibit altered expression of genes related to pathways dysregulated in diabetes. This could be a consequence of the cancer samples being taken from individuals who also had T2D. Alternatively, these pathways might be normally disrupted in these particular cancer types and thus more susceptible to disruption by diabetes or greatly exacerbated by the condition. The fact that the non-associated gene list did not exhibit similar gene ontology or pathway terms gives support to this being a nonrandom result; however, there were more genes in the non-associated list, which may have hidden gene similarities with the T2D-associated gene list. More cancers will need to be examined to confirm if this pattern is true for other cancers associated with T2D. One limitation of my analysis is that the direction of gene expression change was not taken into account. Given the pathways that were identified, it would be interesting to note which pathways were up or down-regulated specifically.

Tables and Figures

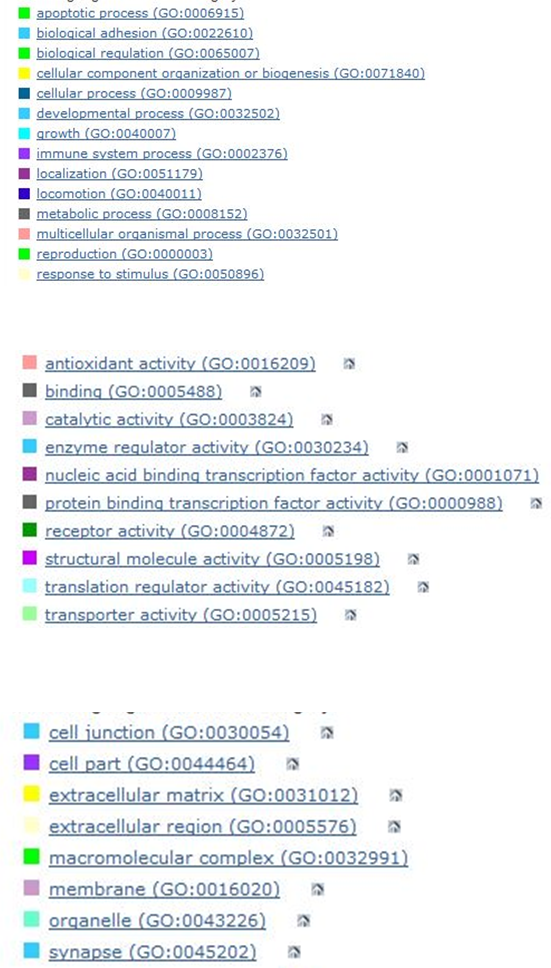
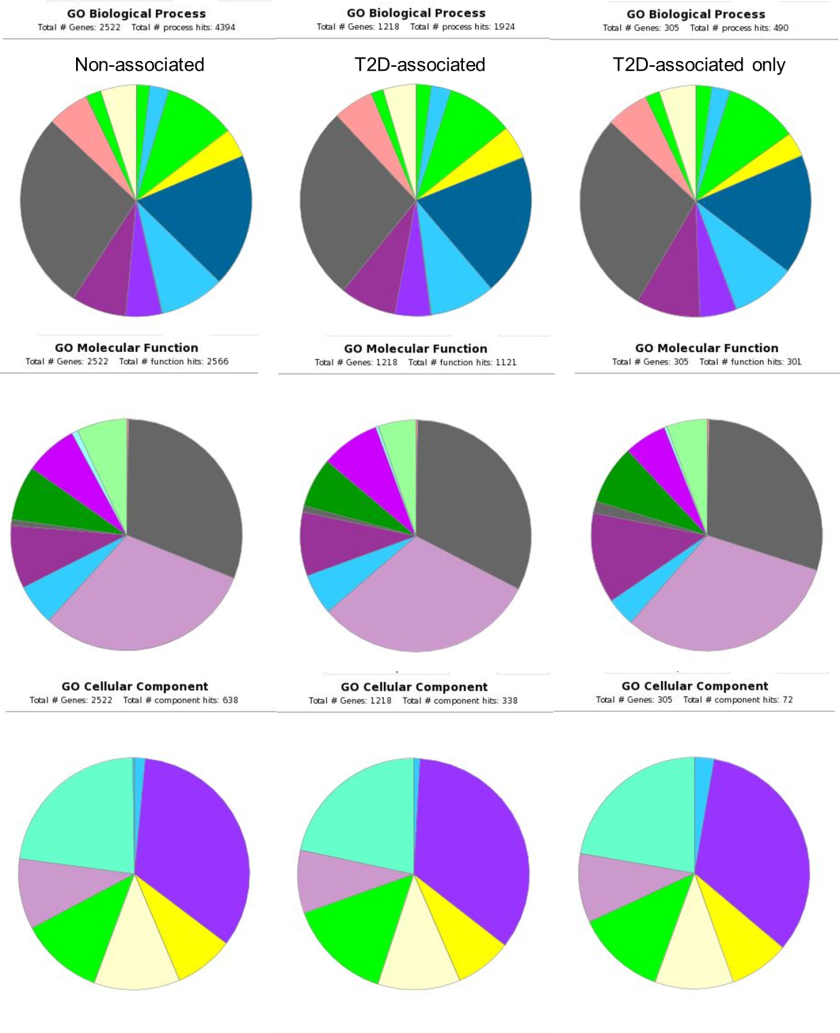
**Table 1**. The cancer types examined and the number of samples and differentially expressed genes per cancer.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Cancer Type** | **Abbreviation** | **T2D associated risk?** | **Tumor samples** | **Normal samples** | **Number of Genes tested** | **DE Genes** |
| Liver hepatocellular carcinoma | LIHC | Yes | 147 | 50 | 20531 | 10887 |
| Lung adenocarcinoma | LUAD | No | 490 | 58 | 20531 | 9556 |
| Prostate adenocarcinoma | PRAD | No | 278 | 50 | 20531 | 5447 |
| Uterine corpus endometrial carcinoma | UCEC | Yes | 370 | 11 | 20531 | 3179 |
| Pancreatic adenocarcinoma | PAAD | Yes | 56 | 2 | n/a | n/a |

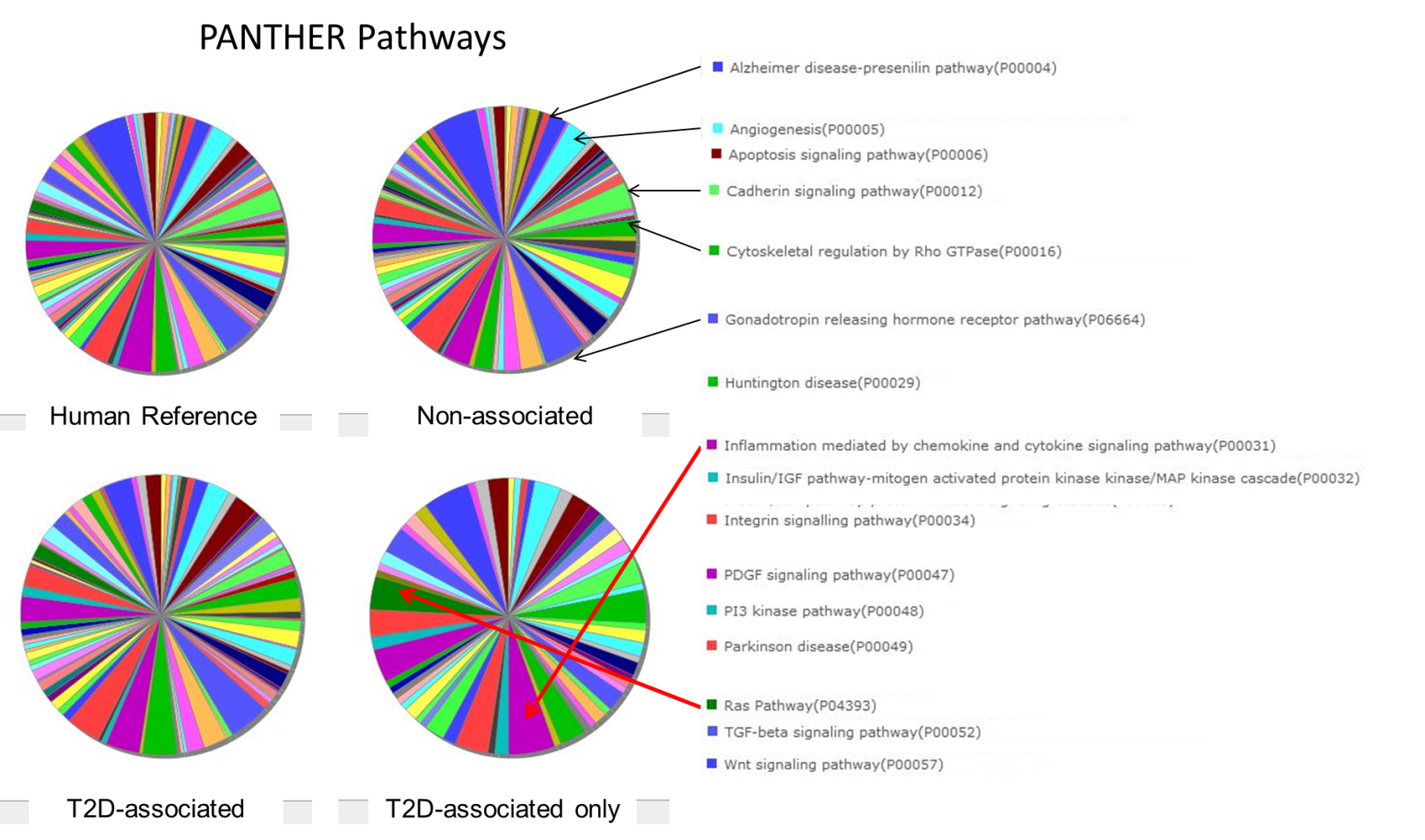


**Figure 1.** Plots of fold change versus gene expression for each cancer. Black dots represent genes with FDR > 0.1 and red dots have FDR < 0.1.

**Figure 2**. Venn diagram demonstrating the overlap of differentially expressed genes shared amongst the four different cancer types. The overlap of the orange and green ovals, minus the 700 genes common to all cancers, represents the 1182 genes in the T2D-associated list. The overlap of the purple and red ovals, minus the 700 common genes, contains the 2556 genes in the non-associated gene list. Finally, the T2D-associated only gene list contains the 316 genes common to the orange and green ovals only.



**Figure 3**. All three gene lists appeared to have similar distributions of GO terms in the biological process, molecular function, and cellular component categories.



**Figure 4.** Graphical representation of PANTHER-annotated pathways associated with the genes from the three gene lists generated, along with the human reference for the entire genome. Part of the legend is included for prevalent pathways, especially those that are enriched in the T2D-associated only gene list. Many pathways were excluded from the legend to conserve space. Black arrows illustrate where the terms are found on the pie chart. Red arrows highlight enriched pathways in the T2D-associated only pie chart.

References:

1. Giovannucci, E. *et al*. (2010) Diabetes and Cancer: A Consensus Report. CA CANCER J CLIN 60, 207–221.
2. Hirabayashi, S., Baranski, T.J, and Ross L. Cagan. (2013) Transformed Drosophila Cells Evade Diet-Mediated Insulin Resistance through Wingless Signaling. Cell 154, 664–675.
3. Handelsman, Y. et al. (2013) Diabetes and cancer—an AACE/ACE consensus statement. Endocrine Practice 19(4), 675-693.
4. Robinson, M.D., McCarthy, D.J, and Gordon K. Smyth. (2010) edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics 26(1), 139-140.
5. Huaiyu , M. *et al*. (2013) [Large-scale gene function analysis with the PANTHER classification system.](http://www.nature.com/nprot/journal/v8/n8/full/nprot.2013.092.html) *Nature Protocols 8, 1551 – 1566.*
6. Huang DW, Sherman BT, Lempicki RA. (*2009)* Systematic and integrative analysis of large gene lists using DAVID Bioinformatics Resources. *Nature Protoc. 4(1):44-57.*
7. Huang DW, Sherman BT, Lempicki RA. (*2009)* Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Res. 37(1):1-13.*