Pathways of Distinction Analysis of Liver Cancer Data:

Genetic Differences Between Males and Females

by

Erik Jon Olson

A Thesis Presented in Partial Fulfillment
of the Requirements for the Degree
Master of Science

Approved April 2021 by the
Graduate Supervisory Committee:

Kenneth Buetow, Chair
Melissa Wilson
Reed Cartwright

ARIZONA STATE UNIVERSITY

May 2021

ABSTRACT

The Pathways of Distinction Analysis (PoDA) program calculates relationships between a given group of genes contained within a pathway, and a disease state. It was used here to investigate liver cancer, and to explore how genetic variability may contribute to the different rates of development of the disease in males and females. The goal of the study was to identify germline variation that differs by sex in hepatocellular carcinoma. Using the program, multiple pathways and genes were identified to have significant differences in their relationship to liver cancer in males and females. In animal studies, the genes which were identified using the PoDA analysis have been shown to impact liver cancer, often with different results for males and females. While these genes are often the focus in animal models, they are absent from current Genome Wide Association Studies (GWAS) catalogs for humans. By working to bridge the results of animal studies and human studies, the results help to identify the causes of liver cancer, and more specifically, the reason the disease affects males at much higher rates. The differences in pathways identified to be significant for the two sexes indicate the germline variance may play sex-specific roles in the development of hepatocellular carcinoma. Additionally, these results reinforce the capacity of the PoDA analysis to identify genes that may be missed by more traditional GWAS methods. This study lays the groundwork for further investigations into the identified genes and pathways, and how they behave differently within males and females.

ACKNOWLEDGMENTS

First, I would like to thank Dr. Buetow for his constant guidance and support throughout the process of my Master's research project and thesis. I am particularly appreciative of the fact that he invited me to be a member of his lab and equipped me with the resources necessary for success. I also would like to thank Dr. Wilson for serving as a member of my supervisory committee and introducing me to bioinformatics and coding along with Dr. Buetow. Additionally, my thanks goes to Dr. Cartwright for the insight and expertise he provided as a committee member.

I would also like to thank the members of the Buetow lab. Your willingness to answer questions and help with implementing and running models is greatly appreciated. In particular, I would like to recognize and thank Yih-Kuang Lu, who assisted me when I was learning the PoDA program and provided continued support throughout the research process.

Finally, I would like to recognize and thank my past professors and advisors at Arizona State University. Without your dedication to the field of education and willingness to pass along your knowledge and expertise, the work embodied in this paper would not have been possible.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

Cancer continues to be one of the leading causes of death in the United States and throughout the world. For decades, the disease has thwarted the best efforts of scientists and doctors to find and develop cures. While significant progress has been made, cancer continues to devastate countless lives. One of the reasons advances in cancer research comes so slowly is the disease is challenging in that it can take many forms, and often has unique aspects for each individual who contracts it. (Liu et al., 2015) For this reason, numerous approaches must be taken to study the various types of cancer and identify the root causes as well as practical solutions.

One of the most devastating forms of cancer is hepatocellular carcinoma, the most common form of liver cancer. It is the third leading cause of cancer related death worldwide, affecting roughly 800,000 individuals worldwide as of 2012 with incidences rising since then. (Bertuccio et al., 2017) (Baecker et al. 2018) Interestingly, this cancer affects males at a much higher rate than females. (Petrick et al., 2016) Roughly 30,000 of the 43,000 afflicted individuals in the United States will be males, while only 13,000 will be females. The gap between males and females is even more pronounced in other regions of the world. (Petrick et al., 2016) Despite significant research involving liver cancer, there is noy yet a clear understanding of the etiologic basis of these differences. The lack of sex-specific studies makes it difficult to determine which genetic factors may be contributing to the significantly different rates of liver cancer development between males and females. (Magi et al., 2010)

Liver cancer rates are influenced by a number of factors. Risk factors, including HBV, obesity, type 2 diabetes, and alcohol consumption are well known. However,

inherited genetic factors can also lead to liver cancer, and these factors are not as well understood. (Su et al., 2013) Specifically, there needs to be a better understanding of how genetics contributes to the gap in the development of liver cancer between males and females. Investigation of these relationships can be accomplished through various methods of genetic analysis.

RATIONAL

Analyzing genome-wide germline variation allows for the identification of the molecular components that underpin a disease. This variation that is associated with the disease will be overrepresented in the individuals that have the disease compared to the general population. These individuals with the disease comprise the case group, while the general population represents the controls. The variation may be aggregated in biologic processes that result in the disease, and different individuals may have variation in different components of the process that is only seen when examined at the level of the process as a whole. This variation may have different effects when in the context of other variation. This may be due to epistasis effects, or the context of other risk factors. The effects of the variation may be influenced by the sex of an individual. Different patterns of variation and different biologic processes, pathways, may therefore be observed to be associated with disease in males and females.

ANALYTIC APPROCHES

Genetic analysis is performed in multiple ways. One of the most common approaches is through Genome-wide Association Studies, or GWAS. The process determines if a specific variant, usually a single-nucleotide polymorphism (SNP) is associated with a specific trait within a population. (Korte & Farlow, 2013) The biologic

processes associated with these variants can be assessed using Gene Set Enrichment Analysis (GSEA). GSEA determines whether or not a group of significant results is overrepresented within specific gene networks. (Hung et al., 2012) While traditional GWAS analytic strategies are beneficial to understanding the relationship between genetics and disease, they have significant drawbacks that have not always been successfully addressed in the past.

The Pathways of Distinction Analysis (PoDA) was initially published in a 2011 paper to extend traditional GWAS analysis and has been used to investigate genetic susceptibility to liver and breast cancer. (Braun & Buetow, 2011) The process has also been utilized with diseases other than cancer, such as nonalcoholic fatty liver disease (Chen et al., 2013) PoDA uses concepts related to as traditional GWAS studies but focuses on multi-SNP analysis permitting the study if interactions as opposed to the assumptions of independence of single-SNP analysis. (Braun & Buetow, 2011) Results are achieved by investigating groups of genes collected within biologic pathways. PoDA is capable of identifying the pathways that show significant distinctions between cases and controls. This is accomplished by using pathway-to-gene, and gene-to-SNP associations. These pathways provide the SNP sets used, and are collected from a variety of databases, including Kegg and Biocarta. By incorporating pathway analysis, PoDA attempts to solve some of the issues that occur when using traditional GWAS approaches, namely the epistatic interaction of variation inherent in complex traits. Liver cancer is a complex disease, one which cannot be explained by single-SNP analysis alone. (Liu, 2020) There are multi-gene effects interacting with risk factos which contribute to the specific disease phenotype, and therefore the genetic analysis must employ a multi-gene

approach in order to fully capture the complexity of the disease. An additional benefit to PoDA is that it may detect epistatic associations not visible to traditional GWAS approaches. While any single gene mutation may not be significant, interactions of variation of the genes within a certain pathway may be. The benefit to the pathway system is it incorporates the idea that multiple interacting genes are involved in single biologic processes, and the disruption to any one of these genes can result in the disruption of the process. PoDA is able to capitalize on this concept while more traditional methods cannot. The ability to integrate the effect of multiple variant loci and identify gene interactions is an important component to understanding a disease and makes PoDA an attractive approach to explore complex phenotypes such as sex differences in cancer incidence.

The primary goal of this study was to explore the role of germline variation in sex differences in liver cancer susceptibility. While the specific variations are not different between males and females, they may have different effects on susceptibility due to the environment they are contained in. This was investigated by identifying significantly associated pathways and the genes driving the pathways' significance, for both males and females diagnosed with liver cancer. The significance was determined by comparing the SNPs from a group of individuals who have liver cancer to a group who does not. The groups were further divided into subsets of male and female participants, and significant pathways were compared between the two groups. Overall, this division improved on past studies which did not investigate the role of biologic sex as a modifying factor. Additionally, this study provides insight into the genetic component of the development

of liver cancer and yields a basis for further analysis into possible methods for identifying, managing, and treating the disease.

CHAPTER 2

METHODS

PoDA. The PoDA program assesses if a sample is more closely related to the group of controls or the group of cases, given a collection of SNPs representing genes within a biologic pathway. This is measured through the distance statistic below. After this distance is calculated for each sample, the distribution of this statistic is compared between the two groups, using the nonparametric Wilcoxon rank sum statistic. The process is completed for all pathway-based SNP sets, which enables pathways which are closest to either the cases or controls to be found. The distance is calculated based on the mean of its distance to the controls minus the mean of distances to the cases. (Braun et al. 2009) (Braun & Buetow, 2011)

$$D_{Y,i} = |y_i - f_i| - |y_i - g_i| \quad (1)$$

The above equation calculates the individual's distance score for a particular locus. The variables are as follows: D = distances score, y = individual, i= locus, gi and fi = minor allele frequencies of F and G samples. A positive value for D indicates that the individual is closer to the minor allele frequency for group G than group F. In this study, group G indicates those individuals with cancer, while F indicates those without cancer.

$$S_Y = \frac{E(D_{Y,i})}{\sqrt{Var(D_{Y,i})/l}} \quad (2)$$

The above equation calculates the distance of individual across a set of loci. The variables are as follows: S_y = distance score for an individual y, $E(D_{y,i})$= mean of $D_{(y,i)}$ across the loci under consideration, l = the group of loci of interest. A positive value for

Sy indicates that the individual is closer to group G than group F for the set of loci within the group.

$$W_P = \sum_{Y \epsilon case} R_{Y,P} - \frac{n_{case}(n_{case}+1)}{2} \quad (3)$$

The above equation calculates the of similarity of cases to cases and controls to controls. It is the equation for determining which members of the predefined groups are closest to the other members of the group, for all individuals using a set of loci. This is achieved using a Wilcoxon rank sum statistic, which non-parametrically determines this feature. $R_{y,p}$ is the rank of $S_{y,p}$ relative to the other samples for the selected pathway.

The Wilcoxon statistic can be affected by two variables in each pathway: the number of genes in the pathway and the number of SNPs per gene. These are corrected using a normalized Pathway Distinction score and is achieved by a resampling procedure. (Braun & Buetow, 2011) This process involves permuting labels for the cases and controls and recalculating the Wilcoxon statistic. The new statistic acts as a null hypothesis and can be used to normalize the original score.

$$DS_P = \frac{W_P - E(W_P^*)}{SD(W_P^*)} \quad (4)$$

The above equation represents the normalized Wilcoxon score. It is the equation for correction of the Wilcoxon statistic to account for biases. Variables for the equation are as follows: DS_p = normalized Pathway Distinction Score, W_p = Wilcoxon statistic for the given pathway, W^*_p = Wilcoxon statistic using permuted labels during recalculation. For the recalculation, 1000 permutations are used.

The second score generated by the PoDA program is the odds ratio. It is achieved by taking the relative distance from the two groups for a sample and using it to calculate

the odds of disease. This is done using a logistic regression model. (Braun & Buetow, 2011) PoDA is an open-source R program available for download at https://github.com/Buetow-Lab/PoDA.

For genes which have multiple SNPs, the SNP with the highest significance for each gene was selected. This significance was calculated using a Chi-squared test of independence.

Networks. At the core of the PoDA program is the use of pathways, or groups of related genes. These pathways come from pathway databases, such as Kegg and Biocarta, or from published literature. For this analysis only selected pathways were used, termed networks. Pathways may be broad, loosely defined, or consist of thousands of genes. A subset of pathways, called networks, are smaller and have more targeted biologic processes related to the genes contained within them. This allows for more focused analysis and improved computational time. There are a total of 2540 pathways which are available for use within the PoDA analysis, and 1290 of which are considered networks. The network list used for this analysis is available at https://github.com/Buetow-Lab/PoDA.

Datasets. The dataset in the experiment comes from previously a published case-control study of Korean individuals and includes 973 individuals, of which 704 are males and 269 are females. (Clifford et al., 2010) (Braun & Buetow, 2011) Of the males, 299 were cases and 405 were controls. Of the females, 87 were cases and 182 were controls.

These individuals were genotyped using the Affymetrix 6.0 platform. (Clifford et al., 2010) The platform detects over 900,000 SNPS. (Clifford et al., 2010)

Graphing Results. The results for the PoDA analysis were quality assessed by plotting using R version Rx64 3.2.0 available at https://www.r-project.org/, and the plotting script (PlotSvals.R) available at https://github.com/Buetow-Lab/PoDA. The script uses primitives, which is part of the core release of R. The networks that returned expected distributions of results were retained for further analysis, while the pathways with discontinuous distributions were removed. The unusual distributions could be attributed to a number of different aspects within the data, but overall indicates that the network does not follow the assumptions made within the PoDA program.

Significance. After an initial screening through the graphing process, significant pathways were selected. The criteria to determine significance was as follows: The network was multiple comparisons adjusted significant in both distance score and odds ratio for either males or females, but non-significant for the other sex. The criteria were chosen in order to identify the pathways whose pattern of germline variation differed in case-control contrasts in males and females. The *p* value cutoff for distance score was < 0.05, and < 3.94×10^{-5} for the odds ratio. The odds ratio cutoff was a result of 0.05, adjusted for the number of networks, using a Bonferroni correction.

Overlap. Pathways which shared greater than 60% overlap in genes with another significant pathway were removed from further analysis. This step was taken in order to

improve computation times while maintaining the representation of major biologic processes within the data. When overlap occurred, the smaller pathway containing the overlap was removed in order to preserve the largest number of genes possible for further analysis.

Control Comparisons. The male and female control groups were compared to each other using the same PoDA process employed to compare cases to controls. The purpose of this comparison was to determine the extent to which the significant pathways were a result of germline variation differences associated with sex differences, unrelated to liver cancer.

Step-down. The significant pathways underwent the process of feature selection, using a step-down reduction procedure. This process involved removing each gene from the pathway and recalculating the PoDA scores. The gene whose removal resulted in the highest remaining significance of the pathway, as measured by odds ratio, was removed from the pathway. The process was repeated until only three genes (the minimum definition of a pathway) remained in the network. Through this method, the genes with the largest contribution to a pathway's significance were discovered. This step-down process retains genes allelic interactions differentiating case and control status. This is in contrast to step-up processes, where genes are added according to the identification of those that generate the highest level of significance.

Subsets. The results of the step-down process were a list of genes, ordered by the gene whose removal resulted in the greatest increase or least decrease of the overall significance of the pathway. Of particular interest were groups of genes that showed specific characteristics. The first subset of genes produced the most significant odds ratio. This group represented genes from a pathway that had the greatest differences between liver cancer cases and controls. The next group of genes was those that produced the same significance as the original pathway but resulted in a lower significance than the original pathway when any additional gene was removed. These genes represent the minimum number of genes necessary in order to generate the same significance as the original pathway. Every pathway had a set of genes that produced a maximum significance, but not every pathway had a minimum gene subset greater than three genes. This was due to the fact that the step-down process did not always end in a result that had less significance than the original pathway. For those pathways which did not meet the criteria for the minimum gene subset, the last three genes from the step-down process were used.

Combined Gene Lists. Because of the overlap observed in the step-down lists of genes, new gene sets were created from the results of the step-down processes. The gene sets were generated by combining the genes that made up the maximum significance and minimum significance gene set lists. This resulted in two new pathways for each of the male and female datasets. The new gene sets underwent the same step-down and resampling process in order to determine the degree to which the genes they contained impacted their significance.

Identify SNPs. The individual, independent significance of SNPs within genes of particular interest were analyzed, using a Chi-Squared test of independence to determine if an association between genotype and phenotype existed. The three genotypes were homozygous, heterozygous, and homozygous alternative. Significance was based on a *p* value of < 0.05, adjusted using a Bonferroni correction based on the number of SNPs within the gene of interest.

Databases. The pathways used in analysis were pulled from online databases and included: Biocarta, Reactome, KEGG, and Nci/Nature. The pathways are constantly updated in order to represent the current best understanding of the genes which interact with each other to perform biologic processes.

CHAPTER 3

RESULTS

The PoDA program was used on the complete dataset, as well as a male only and female only dataset. Below is a sample representation of the results. Lists of all significant results can be found in the Appendix. As indicated, there were pathways of significance that appeared in all, none, or some combination of the datasets. Significance was determined via a < 0.05 *p* value for distance score. *P* values for the odds ratio values were adjusted using a Bonferroni correction.

Pathway Names	DS	p(DS)	O.R	q(O.R)
respiratory electron transport, atp synthesis by chemiosmotic coupling, and heat production by uncoupling	8.360027	0	3.205931	1.53E-33
respiratory electron transport(reactome)	8.286362	0	3.22177	3.94E-33
retrograde endocannabinoid signaling(kegg)	4.623409	0.02	2.89522	3.07E-28
formation of the beta-catenin:tcf transactivating complex(reactome)	4.164194	0.005	1.664608	4.53E-14
transcriptional regulation by tp53(reactome)	4.105537	0.04	2.830578	3.08E-33
antigen processing and presentation(kegg)	3.925168	0.014	1.48451	6.86E-13
signaling in immune system(nci/nature)	3.462283	0.039	2.078045	1.41E-24

Figure 1. Sample PoDA Results

An example of the results generated by the PoDA program. From left to right the columns represent: The pathway name, the distance score, the *p* value of the distance score as calculated by resampling, the odds ratio, and the adjusted *p* value for the odds ratio. The pathway name includes the database from which it was retrieved.

Pathways that showed significance in both scores were then graphed. Graphs indicated whether the pathway showed results consistent with assumptions made by the PoDA program, or if they deviated from those assumptions. Below is a graph which follows standard distributions patterns, as well as one in which the normal distribution is not seen.

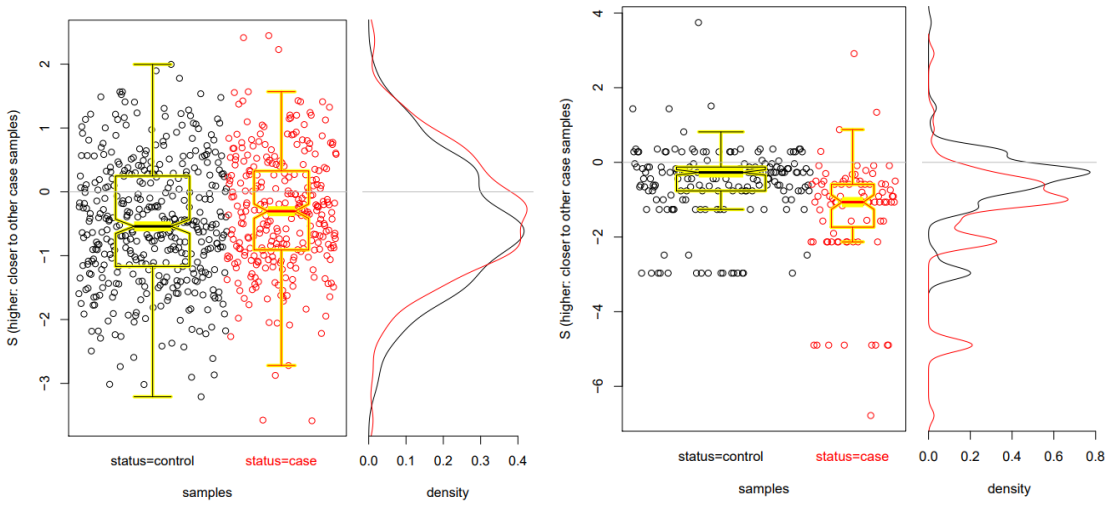

Figure 2. Distribution Graphs

The graph on the left indicates a distribution that is considered normal. The graph on the right indicates a distribution that is abnormal, as the spikes indicate it is breaking with the assumptions made by the PoDA program. A pathway which resulted in a graph similar to that on the right would be removed from further analysis, while pathways with results similar to the graph on the left would be selected for continued analysis.

Pathways significant in both distance scores and odds ratios for any of the three datasets were then compared in order to determine overlap. Pathways with more than 60% of shared genes were removed, as calculated by the number of genes in the pathway. The pathway that was smaller in terms of overall gene count was removed.

The final selection of pathways was based on the pathways that had significance in both distance score and odds ratio, for either the male or female group, as well as being non-significant in these categories for the opposite sex. This selection resulted in 10 pathways for males and 12 pathways for females. The scores for each pathway across each dataset are presented in the table below. In most instances, the male scores were

most similar to the combined dataset scores due to the larger number of male genomes compared to females in the combined dataset. These pathways represent biologic processes, and the genes which control them, influential to liver cancer for one sex, but not the other. This eliminated pathways that may be useful in predicting cancer for both genders, which while important are not the topic of this study and have neem previously published for this dataset. (Braun & Buetow, 2011) (Lu et al., 2018)

Table 1.

Male significant Pathways

Pathway Names	DS	p(DS)	O.R	q(O.R)	Male DS	Male p(DS	Male O.R	Male q(O.	Female D	Female p(	Female O.	Female q(
antigen processing and presentation(kegg)	3.925168	0.014	1.48451	6.86E-13	3.13012	0.016	1.569151	2.73E-11	0.953365	0.267	1.524476	0.000229
chaperones modulate interferon signaling pathway(biocarta)	3.431543	0.005	1.54141	8.87E-10	2.669804	0.005	1.554896	5.98E-08	0.476337	0.377	1.569899	0.001022
interferon gamma signaling(reactome)	3.071326	0.024	1.699004	4.15E-17	2.641183	0.017	1.801122	1.52E-14	-0.94859	0.129	1.660684	7.40E-05
digestion(reactome)	0.889149	0.275	1.293229	0.00015	2.47182	0.006	1.488229	1.45E-06	1.445519	0.087	1.651982	0.000148
tnfr2 signaling pathway(biocarta)	1.186691	0.137	1.255164	0.000366	2.289308	0.017	1.422892	3.26E-06	-1.39597	0.071	1.042428	0.749002
il3-mediated signaling events(nci/nature)	2.026405	0.052	1.406553	3.71E-07	2.124242	0.033	1.452754	3.1E-06	-1.29628	0.067	1.194589	0.171042
calnexin/calreticulin cycle(reactome)	1.760564	0.077	1.382968	4.06E-06	2.047567	0.023	1.447218	2.2E-06	-0.82584	0.172	1.168718	0.227194
il23-mediated signaling events(nci/nature)	2.345129	0.024	1.501357	6.43E-10	2.03547	0.043	1.550908	7.46E-08	-0.45469	0.242	1.430978	0.006972
erythropoietin mediated neuroprotection through nf-kb(biocarta)	2.168784	0.025	1.345878	4.33E-06	1.831919	0.033	1.387407	1.87E-05	0.142644	0.498	1.261377	0.072941
sumoylation by ranbp2 regulates transcriptional repression(nci/nature)	1.653505	0.085	1.400469	2.99E-06	1.78372	0.042	1.430137	7.93E-06	0.091301	0.549	1.466187	0.005799

The above table indicates the selected pathways for the male dataset. Green highlighted cells indicate a significant value while red highlighted cells indicate a non-significant cell. The columns are labeled in a similar manner to the example above, with the first four indicating scores for the complete dataset, the next four indicating values for the male dataset and the final four representing values for the female dataset. Each pathway in this group has significant values for all scores from the male dataset, and non-significant values for all scores from the female dataset. The scores from the combined dataset are listed, but not considered when deciding if a dataset was selected for further analysis.

Table 2.

Female Significant Pathways

Pathway Names	DS	p(DS)	O.R	q(O.R)	Male DS	Male p(DS	Male O.R	Male q(O.	Female DS	Female p(	Female O.	Female q(
1- and 2-methylnaphthalene degradation(kegg)	1.512969	0.104	1.160618	0.001391	-0.02985	0.425	1.11998	0.100712	2.251776	0.009	1.609675	2.52E-05
metabolism of xenobiotics by cytochrome p450(kegg)	1.094594	0.254	1.134418	3.95E-05	-0.07244	0.333	1.122889	0.001554	1.986714	0.049	1.469141	4.94E-06
g0 and early g1(reactome)	0.068956	0.581	1.200339	0.010723	-0.9361	0.128	1.181708	0.032959	1.977526	0.025	1.890637	1.58E-05
aminosugars metabolism(kegg)	-0.83478	0.149	1.208055	0.00402	0.119785	0.544	1.344876	0.000161	1.964259	0.023	2.196258	1.31E-06
internalization of erbb1(nci/nature)	-0.65606	0.174	1.278889	0.000538	-1.2746	0.06	1.28615	0.001329	1.95047	0.039	2.346698	1.18E-08
cbl mediated ligand-induced downregulation of egf receptors pathway(biocarta)	0.509519	0.363	1.253483	0.00057	-0.49906	0.265	1.208809	0.009273	1.949206	0.017	1.800297	9.14E-06
elastic fibre formation(reactome)	-2.3073	0.006	1.129995	0.073002	-0.42443	0.217	1.364016	0.00011	1.891625	0.038	2.819741	7.49E-09
negative regulation of fgfr3 signaling(reactome)	0.462186	0.427	1.235748	0.002264	0.301403	0.469	1.272862	0.003011	1.883919	0.028	1.995265	3.63E-06
paxillin-dependent events mediated by a4b1(nci/nature)	1.028864	0.219	1.367114	1.55E-06	-0.40837	0.256	1.295193	0.001769	1.761643	0.047	2.198993	2.59E-07
adp-ribosylation factor(biocarta)	0.639638	0.398	1.270425	0.000214	0.874082	0.257	1.33954	0.000314	1.743569	0.049	1.929792	7.18E-06
role of egf receptor transactivation by gpcrs in cardiac hypertrophy(biocarta)	0.368614	0.502	1.372964	3.97E-06	-0.50966	0.231	1.323058	0.00052	1.736496	0.042	2.399661	1.24E-07
mets affect on macrophage differentiation(biocarta)	0.366838	0.439	1.250436	0.001622	0.099271	0.563	1.295017	0.00105	1.678908	0.049	1.873562	4.31E-06

The above table indicates the selected pathways for the female dataset. Green highlighted cells indicate a significant value while red highlighted cells indicate a non-significant cell. The columns are labeled in a similar manner to the example above, with the first four indicating scores for the complete dataset, the next four indicating values for the female dataset, and the final four representing values for the male dataset. Each pathway in this group has significant values for all scores from the female dataset, and non-significant values for all scores from the male dataset. The scores from the combined dataset are listed, but not considered when deciding if a dataset was selected for further analysis.

The step-down process, as described above, was completed for each of the pathways with all three datasets. Below is a sample pathway, highlighting the results and indicating the manner in which odds ratios change based on the genes remaining in the pathway. Subsets of genes were selected based on the genes that generated the most

significant odds ratios, as well as the minimum number of genes within a pathway that could be used to generate a odds ratio as significant as the original pathway. The cut off points are highlighted in the example. The subsets of genes also underwent the step-down process. This allowed the genes with the largest impact on the subsets to be highlighted. An example is shown below.

Chr	Gene ID	Gene Symb	Gene Name	DS	p(D.S)	OR	p(O.R)
	---			2.168784445	0.025	1.345877828	4.33E-06
6	6648	SOD2	superoxide dismutase 2	2.654674846	0	1.359085328	1.29E-06
14	3091	HIF1A	hypoxia inducible factor 1 subunit alpha	2.931877618	0	1.360545822	6.08E-07
6	1026	CDKN1A	cyclin dependent kinase inhibitor 1A	2.925933184	0	1.358276386	4.86E-07
7	2056	EPO	erythropoietin	2.912492262	0.012	1.349446401	4.02E-07
8	10987	COPS5	COP9 signalosome subunit 5	2.957081246	0	1.321442315	3.83E-07
1	405	ARNT	aryl hydrocarbon receptor nuclear translocator	2.718555023	0.012	1.315460793	4.95E-07
1	3725	JUN	Jun proto-oncogene, AP-1 transcription factor subunit	2.733660704	0.024	1.189591926	1.44E-06
9	3717	JAK2	Janus kinase 2	2.032074472	0.024	1.16315092	9.07E-06
16	1387	CREBBP	CREB binding protein	NA	0	NA	NA
22	2033	EP300	E1A binding protein p300	NA	0	NA	NA
4	4790	NFKB1	nuclear factor kappa B subunit 1	NA	0	NA	NA

Figure 3. Pathway Step-down

An example of a pathway that has undergone the step-down process. Highlighted in purple is the cutoff point where the pathway becomes the most significant. In this case, when gene 10987 is removed, the remaining genes result in the most significant odds ratio possible. Highlighted in red is the cutoff point where the remaining genes have the same or greater significance than the original pathway, but the removal of any additional gene results in a lesser significance than the original pathway. The three most significant genes have no scores, as the step-down program requires four or more genes to function.

Chr	Gene ID	Gene Symbol	Gene Name	DS	pDS	OR	pOR
4	4790	NFKB1	nuclear factor kappa B subunit 1	2.669703772	0.012	1.215489823	2.00E-05
1	3725	JUN	Jun proto-oncogene, AP-1 transcription factor subunit	2.960520708	0.012	1.240422666	8.73E-07
9	3717	JAK2	Janus kinase 2	2.487583106	0.012	1.256262669	1.27E-05
16	1387	CREBBP	CREB binding protein	2.109453817	0.012	1.277108687	8.16E-06
22	2033	EP300	E1A binding protein p300	2.263942845	0.012	1.309027582	6.47E-07
1	405	ARNT	aryl hydrocarbon receptor nuclear translocator	2.718555023	0.012	1.315460793	4.95E-07
	0			2.957081246	0.012	1.321442315	3.83E-07

Figure 4. Most Significant Gene Subset

An example of a Most Significant gene subset. The figure shows the genes which make up the most significant gene subset of the pathway above after undergoing the additional step-down process. The genes are sorted by largest pOR, an indication that those genes at the top most impact the significance after their removal. The original score for the subset is highlighted in purple below.

Chr	Gene ID	Gene Symbol	Gene Name	DS	pDS	OR	pOR
22	2033	EP300	E1A binding protein p300	1.874595448	0.072	1.04700565	0.003677054
4	4790	NFKB1	nuclear factor kappa B subunit 1	2.321010974	0.012	1.050974981	0.000102531
16	1387	CREBBP	CREB binding protein	1.995319044	0.048	1.146400197	2.25E-05
9	3717	JAK2	Janus kinase 2	2.032074472	0.036	1.16315092	9.07E-06
	0			2.733660704	0.024	1.189591926	1.44E-06

Figure 5. Minimum Gene Set

An example of a minimum gene subset. The genes which make up the minimum gene subset of the pathway above after undergoing the additional step-down process. The genes are sorted by largest pOR, an indication that those genes at the top of the list most impacted the significance after their removal. The original score for the subset is highlighted in red below the genes listed.

The PoDA program was also used to compare the male control group to the female control group. The purpose of the comparison was to identify pathways whose variation was associated with biologic sex, independent of disease statues. Below is a sample of the results, with all significant results appearing in the Appendix.

Pathway Names	DS	p(DS)	O.R	q(O.R)
synthesis of pips at the golgi membrane(reactome)	7.429426828	0	9.332414423	8.02E-21
integrin signaling pathway(biocarta)	7.347556922	0.002	3.782494786	4.96E-27
rho gtpases activate paks(reactome)	6.28289503	0.002	3.155915083	6.90E-24
rho gtpases activate cit(reactome)	7.078735778	0.002	3.057645864	2.82E-22
rho gtpases activate rocks(reactome)	6.98634906	0.002	2.849587477	2.95E-21
rho cell motility signaling pathway(biocarta)	7.106539557	0.003	3.468608686	1.86E-25
rac1 cell motility signaling pathway(biocarta)	6.637605346	0.004	3.363884284	3.85E-25
pkc-catalyzed phosphorylation of inhibitory phosphoprotein of myosin phosphatase(biocarta)	6.308222335	0.004	3.634558216	5.40E-25
dcc mediated attractive signaling(reactome)	6.277872766	0.005	4.205601686	3.26E-25
rho gtpases activate pkns(reactome)	6.623621277	0.005	3.448700105	6.86E-25
tnfr2 signaling pathway(biocarta)	2.44642503	0.005	1.520316691	4.60E-06
thrombin signaling and protease-activated receptors(biocarta)	6.940067243	0.006	3.974030712	2.28E-27
ccr3 signaling in eosinophils(biocarta)	5.418632653	0.006	2.739737618	8.50E-20
pi metabolism(reactome)	6.832756855	0.007	23.94586132	3.81E-27
sema3a pak dependent axon repulsion(reactome)	2.416396242	0.008	1.765040055	1.22E-09

Figure 6. Control-Control Comparison

Sample results of the PoDA program run on female controls compared to male controls. These results follow the same format as past figures. The results indicate there is minimal overlap with the significant pathways identified for males and females, seen in Tables 1 and 2.

Genes from the overall pathways of significance for males and females were compared and results indicated there was very little overlap between the pathways. Between the 248 genes belonging to the male pathways (duplicate genes only counted once) and 302 genes for female pathways, only 17 overlapped. The step-down process resulted in gene subsets found 133 genes for male in the most significant subsets and 163 genes for females. There was an overlap of only five genes between these groups. The minimum gene list contained 50 genes for males and 62 genes for females and these groups contained an overlap of a single gene.

Figure 7. Overlap of significant pathways between sexes.

Diagram of the overlap between the significant pathways and their various subsets. A total of 17 genes were shared between all significant male and female pathways, five genes (CHIA, CHIT1, GRB2, JUN and SCH1) shared in the most significant subsets, and one gene (CHIA) shared in the minimum subsets.

Of particular importance were the genes present across multiple significant pathways. For males, four genes (JAK2, NKKB1, IFNG, and NFKBIA) were present three or more times across the selected pathways. Of these genes, two (JAK2 and NFKB1) were present three or more times in the maximum significance subsets and one (JAK2) that was present three or more times in the minimum subsets. For females, following the same criteria, there were six genes (CBL, GRB2, EGFR, EGF, HRAS, and SRC) repeated in the pathways, two (CBL and GRB2) that remained in the maximum significance subset and one (CBL) that remained in the minimum subsets. These genes can be seen in Figures 8 and 9, which show the genes that were repeated across multiple significant pathways. Of particular interest were those genes which remained in multiple pathways at the maximum significance cutoff. Of these genes that were often repeated in the gene subsets, JAK2 was not found in any female significant pathways, and NF-KB

was found in only one. Neither of cbl proto-oncogene nor epidermal growth factor were found in any of the male significant pathways.

Gene ID	Gene Name	Occurances- Significant Pathways	Occurances - Max gene Subset	Occurances - Minimum Gene Set
3717	Janus kinase 2	5	5	5
4790	nuclear factor kappa B subunit 1	4	3	
3458	interferon gamma	4		
4792	NFKB inhibitor alpha	3		

Figure 8. Male Repeated Genes

Figure indicating the genes that were repeated more than three times in the male significant pathways. The figure lists the gene ID, the gene name, the times it occurred across all male significant pathways, the times it occurred in the maximum significance subset, and the times it occurred in the minimum subset.

Gene ID	Gene Name	Occurances- Significant Pathways	Occurances - Max Gene Subsets	Occurances - Minimum Gene Set
867	Cbl proto-oncogene	4	4	4
2885	growth factor receptor bound protein 2	4	4	
1956	epidermal growth factor receptor	3		
1950	epidermal growth factor	3		
3265	HRas proto-oncogene, GTPase	3		
6714	SRC proto-oncogene, non-receptor tyrosine kinase	3		

Figure 9. Female Repeated Genes

Figure indicating the genes that were repeated more than three times in the female significant pathways. The figure lists the gene ID, the gene name, the times it occurred across all female significant pathways, the times it occurred in the maximum significance subset, and the times it occurred in the minimum subset.

The genes identified from the step-down process were analyzed using gene ontology. This process compares the genes to known pathways and attempts to determine if there are pathways where the genes are overrepresented. This process can aid in grouping the genes by identifying characteristics and biologic processes that may tie them together. The results of this process are seen below in Figures 10 - 15. Gene ontology

highlights possible biologic processes these genes may be impacting, and the different biologic functions impacted for males versus females.

Based on the gene ontology results, it is clear there are many overlapping pathways between males and females. Absent from the female list are cytokine related pathways, while interestingly, these same pathways appear multiple times in the male results. The first cytokine pathway for females is the 137th most significant *p* value, far below those same pathways for males. While this process can be useful for uncovering some general underlying characteristics, in many cases it simply provides broad categories of biologic activity. For example, the pathway 'response to stimulus' consists of 8502 genes. This is far too large a set to provide any meaningful insight into the manner in which the genes interact within it, and too general a process to have a meaningful impact on the understanding of liver cancer. As such, the gene ontology search was filtered. In the following figures, 12 and 13, the same gene ontology process is complete, but pathways larger than 1000 genes are removed. The filtering process was again repeated, removing pathways with greater than 100 genes, seen in Figures 14 and 15. This resulted in many interferon pathways appearing significant for males, with the same pathways absent for females.

GO biological process complete	Homo sapiens (REF) #	upload_1 (Hierarchy) NEW! #	expected	Fold Enrichment	+/-	raw P value	FDR
cellular response to cytokine stimulus	1036	51	8.05	6.34	+	1.32E-26	2.10E-22
response to cytokine	1125	52	8.74	5.95	+	6.20E-26	4.92E-22
response to organic substance	3006	80	23.35	3.43	+	9.40E-26	4.97E-22
cytokine-mediated signaling pathway	701	41	5.45	7.53	+	7.53E-24	2.99E-20
cellular response to organic substance	2350	68	18.26	3.72	+	3.82E-23	1.21E-19
cellular response to chemical stimulus	2924	74	22.72	3.26	+	4.31E-22	1.14E-18
response to chemical	4392	90	34.12	2.64	+	1.07E-21	2.43E-18
immune response	1951	56	15.16	3.69	+	1.98E-18	3.93E-15
response to stimulus	8502	121	66.05	1.83	+	2.31E-18	4.07E-15
cell surface receptor signaling pathway	2510	63	19.50	3.23	+	4.18E-18	6.63E-15

Figure 10. Male Most Significant Genes Gene Ontology Analysis

Data generated by gene ontology. The input genes were all genes that appeared in the male significant pathways. Pathways are sorted by most significant *p* value. From left to right, the columns are: pathway name, number of genes in pathway, number of input genes present in pathway, expected number of input genes in pathway, increase enrichment, *p* value, and the false discovery rate.

GO biological process complete	Homo sapiens (REF) #	upload_1 (Hierarchy) NEW! #	expected	Fold Enrichment	+/-	raw P value	FDR
response to stimulus	8502	148	82.98	1.78	+	2.61E-20	4.14E-16
regulation of response to stimulus	4275	97	41.72	2.32	+	9.22E-18	7.31E-14
cellular response to stimulus	6793	126	66.30	1.90	+	1.24E-17	6.57E-14
anatomical structure morphogenesis	2180	65	21.28	3.06	+	1.13E-16	4.48E-13
cellular response to chemical stimulus	2924	76	28.54	2.66	+	1.73E-16	5.48E-13
regulation of signal transduction	3068	77	29.94	2.57	+	6.74E-16	1.78E-12
extracellular matrix organization	363	27	3.54	7.62	+	1.12E-15	2.54E-12
extracellular structure organization	364	27	3.55	7.60	+	1.19E-15	2.36E-12
response to endogenous stimulus	1417	50	13.83	3.62	+	2.05E-15	3.61E-12
cellular response to organic substance	2350	65	22.94	2.83	+	4.07E-15	6.45E-12

Figure 11. Female Most Significant Genes Gene Ontology Analysis

Data generated by gene ontology. The input genes were all genes that appeared in the female significant pathways. Pathways are sorted by most significant *p* value. From

left to right, the columns are pathway name, number of genes in pathway, number of input genes present in pathway, expected number of input genes in pathway, increase enrichment, *p* value, and the false discovery rate.

Pathway(Male)	Genes	#	Expected	+/-	Fold	Raw p-value	FDR
cytokine-mediated signaling pathway (GO:0019221)	701	51	5.96	+	8.56	4.29E-32	1.70E-28
interferon-gamma-mediated signaling pathway (GO:0060333)	72	22	0.61	+	35.96	5.58E-26	1.11E-22
cellular response to interferon-gamma (GO:0071346)	163	26	1.39	+	18.77	3.39E-24	5.98E-21
response to interferon-gamma (GO:0034341)	185	26	1.57	+	16.54	6.20E-23	8.94E-20
response to organonitrogen compound (GO:0010243)	997	38	8.47	+	4.49	9.91E-15	5.42E-12
regulation of cytokine production (GO:0001817)	739	32	6.28	+	5.1	6.52E-14	3.34E-11
positive regulation of cytokine production (GO:0001819)	453	25	3.85	+	6.49	3.15E-13	1.56E-10
innate immune response (GO:0045087)	843	32	7.16	+	4.47	1.97E-12	8.67E-10
type I interferon signaling pathway (GO:0060337)	68	12	0.58	+	20.77	3.21E-12	1.34E-09
cellular response to type I interferon (GO:0071357)	68	12	0.58	+	20.77	3.21E-12	1.30E-09

Figure 12. Male Most Significant Pathways less than 1000 genes

Data generated using gene ontology. Follows the same procedures and format as Figure 10. Pathways with greater than 1000 genes were removed.

Pathway(Female)	Genes	#	Expected	+/-	Fold	Raw p-value	FDR
cellular response to growth factor stimulus (GO:0071363)	500	34	5.41	+	6.28	5.37E-17	8.52E-14
enzyme linked receptor protein signaling pathway (GO:0007167	730	40	7.9	+	5.06	8.24E-17	1.19E-13
response to growth factor (GO:0070848)	528	34	5.72	+	5.95	2.50E-16	2.48E-13
animal organ morphogenesis (GO:0009887)	971	44	10.51	+	4.18	1.46E-15	1.22E-12
extracellular matrix organization (GO:0030198)	363	28	3.93	+	7.12	1.99E-15	1.44E-12
extracellular structure organization (GO:0043062)	364	28	3.94	+	7.1	2.12E-15	1.35E-12
transmembrane receptor protein tyrosine kinase signaling path	524	30	5.67	+	5.29	3.01E-13	1.29E-10
response to tumor necrosis factor (GO:0034612)	275	16	2.34	+	6.85	3.61E-09	7.85E-07
regulation of small molecule metabolic process (GO:0062012)	454	20	3.86	+	5.18	3.65E-09	7.82E-07
response to molecule of bacterial origin (GO:0002237)	338	17	2.87	+	5.92	9.12E-09	1.90E-06

Figure 13. Female Most Significant pathways less than 1000 genes.

Data generated using gene ontology. Follows the same procedures and format as Figure 10. Pathways with greater than 1000 genes were removed.

Pathway(Male)	Genes	#	Expected	+/-	Fold	Raw p-val	FDR
interferon-gamma-mediated signaling pathway (GO:0060333)	72	22	0.61	+	35.96	5.58E-26	1.26E-22
type I interferon signaling pathway (GO:0060337)	68	12	0.58	+	20.77	3.21E-12	1.49E-09
cellular response to type I interferon (GO:0071357)	69	12	0.59	+	20.47	3.74E-12	1.64E-09
response to type I interferon (GO:0034340)	75	12	0.64	+	18.83	8.97E-12	3.54E-09
positive regulation of interferon-gamma production (GO:0032729)	67	10	0.57	+	17.57	9.39E-10	2.60E-07
regulation of leukocyte mediated cytotoxicity (GO:0001910)	79	10	0.67	+	14.9	3.99E-09	8.99E-07
polysaccharide digestion (GO:0044245)	6	5	0.05	+	98.07	1.78E-08	3.60E-06
regulation of T cell mediated immunity (GO:0002709)	78	9	0.66	+	13.58	5.08E-08	8.91E-06
antigen processing and presentation of endogenous peptide antigen (GO:0002483)	19	6	0.16	+	37.16	5.09E-08	8.82E-06
positive regulation of leukocyte mediated cytotoxicity (GO:0001912)	55	8	0.47	+	17.12	5.65E-08	9.28E-06

Figure 14. Male Most Significant pathways less than 100 genes.

Data generated using gene ontology. Follows the same procedures and format as Figure 10. Pathways with greater than 100 genes were removed.

Pathway(Female)	Genes	#	Expected	+/-	Fold	Raw p-val	FDR
integrin-mediated signaling pathway (GO:0007229)	98	12	1.06	+	11.31	2.37E-09	3.37E-07
negative regulation of epidermal growth factor receptor signaling pathway (GO:0042059)	48	9	0.52	+	17.32	9.10E-09	1.16E-06
embryonic eye morphogenesis (GO:0048048)	34	8	0.37	+	21.73	1.32E-08	1.64E-06
negative regulation of ERBB signaling pathway (GO:1901185)	53	9	0.57	+	15.68	1.96E-08	2.27E-06
gland morphogenesis (GO:0022612)	99	11	1.07	+	10.26	2.80E-08	3.11E-06
cell adhesion mediated by integrin (GO:0033627)	25	7	0.27	+	25.86	3.95E-08	4.26E-06
positive regulation of DNA replication (GO:0045740)	40	8	0.43	+	18.47	3.99E-08	4.28E-06
fibroblast growth factor receptor signaling pathway (GO:0008543)	81	10	0.88	+	11.4	4.90E-08	5.19E-06
salivary gland morphogenesis (GO:0007435)	27	7	0.29	+	23.94	6.20E-08	6.43E-06
regulation of epidermal growth factor receptor signaling pathway (GO:0042058)	87	10	0.94	+	10.62	9.07E-08	9.11E-06

Figure 15. Female Most Significant pathways less than 100 genes.

Data generated using gene ontology. Follows the same procedures and format as Figure 10. Pathways with greater than 100 genes were removed.

Cross-pathway Analysis. After the step-down process, new male and female gene sets were created by combining the gene subsets obtained from the individual pathway step-down analysis. These sets contained either all the genes that comprised the most significant subsets for each pathway, or the minimum gene set for the pathways. The new pathways can be found in the supplemental materials. The step-down process was performed on these new sets in order to identify the genes with the greatest and least impact Below is an example of these results. Full pathways are provided in the Appendix.

Gene ID	Gene Name	DS	pDS	OR	pOR
---		0.31276	NA	1.479274	1.99E-06
7412	vascular cell adhesion molecule 1	0.504059	0.52	1.514224	4.54E-07
8807	interleukin 18 receptor accessory protein	0.621723	0.325	1.523247	3.07E-07
3725	Jun proto-oncogene, AP-1 transcription factor subunit	0.797151	0.365	1.543349	1.49E-07
4684	neural cell adhesion molecule 1	0.921162	0.208	1.55881	7.75E-08
405	aryl hydrocarbon receptor nuclear translocator	1.019027	0.241	1.572824	4.44E-08
3106	major histocompatibility complex, class I, B	1.163071	0.275	1.596929	1.81E-08
3118	major histocompatibility complex, class II, DQ alpha 2	1.275602	0.17	1.609927	1.07E-08
3553	interleukin 1 beta	1.414105	0.188	1.619504	6.66E-09
3811	killer cell immunoglobulin like receptor, three Ig domains and lor	1.479288	0.12	1.629808	4.39E-09
8844	kinase suppressor of ras 1	1.558535	0.084	1.641366	2.82E-09
4049	lymphotoxin alpha	1.47318	0.12	1.646601	2.43E-09
3459	interferon gamma receptor 1	1.497027	0.06	1.660547	1.46E-09
3394	interferon regulatory factor 8	1.668241	0.094	1.662159	9.57E-10
3308	HSPA4	1.755223	0.133	1.671228	6.39E-10
3938	lactase	1.798126	0.097	1.67302	5.64E-10
5295	phosphoinositide-3-kinase regulatory subunit 1	1.795509	0.08	1.671544	5.41E-10
11253	mannosidase alpha class 1B member 1	1.795509	0.036	1.671544	5.41E-10
119548	pancreatic lipase related protein 3	1.795509	0.072	1.671544	5.41E-10
6476	sucrase-isomaltase	1.801217	0.06	1.670344	5.28E-10
10083	USH1 protein network component harmonin	1.801217	0.012	1.670344	5.28E-10
149233	interleukin 23 receptor	1.801217	0.06	1.670344	5.28E-10
23225	nucleoporin 210	1.801217	0.131	1.670344	5.28E-10
23321	tripartite motif containing 2	1.801217	0.097	1.670344	5.28E-10
27159	chitinase acidic	1.801217	0.068	1.670344	5.28E-10
55757	UDP-glucose glycoprotein glucosyltransferase 2	1.801217	0.036	1.670344	5.28E-10
91445	ring finger protein 185	1.801217	0.072	1.670344	5.28E-10
4261	class II major histocompatibility complex transactivator	1.842044	0.06	1.668597	5.44E-10
5905	Ran GTPase activating protein 1	1.815576	0.012	1.659578	7.02E-10
3115	major histocompatibility complex, class II, DP beta 1	1.947119	0.048	1.64908	9.66E-10
821	calnexin	1.99626	0.095	1.620267	2.12E-09
7726	tripartite motif containing 26	1.977472	0.049	1.592136	5.44E-09
7185	TNF receptor associated factor 1	2.322042	0.032	1.564788	1.51E-08
6890	transporter 1, ATP binding cassette subfamily B member	2.259258	0.012	1.565975	1.13E-08
4214	mitogen-activated protein kinase kinase kinase 1	2.240474	0.036	1.56318	9.23E-09
4790	nuclear factor kappa B subunit 1	2.211333	0.036	1.558165	7.42E-09
3675	integrin subunit alpha 3	2.061388	0	1.550549	6.35E-09
3717	Janus kinase 2	1.989754	0.036	1.546997	6.11E-09
3635	inositol polyphosphate-5-phosphatase D	1.983863	0.107	1.540003	6.45E-09
4843	nitric oxide synthase 2	2.08786	0.036	1.526943	7.81E-09
7133	TNF receptor superfamily member 1B	2.203296	0.032	1.515771	1.18E-08
7124	tumor necrosis factor	1.973713	0.036	1.513169	1.94E-08
2033	E1A binding protein p300	1.988841	0.048	1.478608	7.89E-08
9759	histone deacetylase 4	1.953127	0.06	1.449367	1.90E-07
2984	guanylate cyclase 2C	1.604856	0.06	1.43054	1.61E-07
3117	major histocompatibility complex, class II, DQ alpha 1	1.522123	0.118	1.348883	6.10E-07
5407	pancreatic lipase related protein 1	1.248029	0.25	1.331163	9.05E-07
3122	major histocompatibility complex, class II, DR alpha	0.081371	0.577	1.30093	2.26E-05
3113	major histocompatibility complex, class II, DP alpha 1	NA	NA	NA	NA
3663	interferon regulatory factor 5	NA	NA	NA	NA
815	calcium/calmodulin dependent protein kinase II alpha	NA	NA	NA	NA

Figure 16. Male Minimum Pathway Step-down

The above chart represents the step-down results for the new pathway generated by using significant pathway subsets. Highlighted in purple is the point at which the pathway becomes most significant. In red is the point at which the remaining genes are as significant as the original pathway, and the removal of any additional gene will result in it being less significant. For the two largest pathways, the pathways composed of the maximum significant subsets, the pDS score was not calculated due to limited computing process power and time constraints.

Individual SNP Association Analysis. Finally, the SNP used in calculations involving the JAK2 gene were analyzed. Based on a Chi-squared test of independence, only one SNP was significant for the male dataset when correcting for the number of SNPs per gene. Below are the calculations for each SNP in the male dataset. The SNP used in the PoDA calculations was 'rs2149556' for the male and combined dataset. A different SNP was used for the female dataset, based on the one that had the largest test statistic.

SNP	rs1011900	rs1049165	rs1081514	rs1081515	rs1081516	rs1097491	rs1097494	rs1179365	rs1233966	rs1234089	rs1327493	rs1571437	rs1576271	
Statistic	8.5556	0.013	7.226	5.209	5.692	6.56	4.397	5.347	4.836	5.707	0	6.244	7.504	
P-value	0.013873	0.9935	0.0269	0.0739	0.058	0.0376	0.1109	0.069	0.089	0.057	1	0.04406	0.0234	
Adjusted Significace	0.001852	0.001852	0.001852	0.001852	0.001852	0.001852	0.001852	0.001852	0.001852	0.001852	0.001852	0.001852	0.001852	
Significant	No	No	No	No	No	No	No	No	No	No	No	No	No	
SNP	rs1742581	rs2031904	rs2104685	rs2149556	rs4372063	rs6476934	rs7031456	rs7034539	rs7034878	rs7847294	rs7851969	rs7859390	rs913594	rs1742563
Statistic	2.471	8.774	5.133	12.821	10.523	0	0.185	6.971	11.222	5.701	0.605	7.047	5.524	6.489
P-value	0.2906	0.01243	0.0768	0.001644	0.00518	1	0.9116	0.0306	0.0036	0.0578	0.7389	0.029	0.0631	0.0389
Adjusted Significace	0.001852	0.001852	0.001852	0.001852	0.001852	0.001852	0.001852	0.001852	0.001852	0.001852	0.001852	0.001852	0.001852	0.001852
Significant	No	No	No	Yes	No	No	No	No	No	No	No	No	No	No

Figure 17. Jak2 Chi-Squared Results

The figure indicates Chi-squared test results for each SNP of the JAK2 gene. The statistic row indicates the Chi-squared test statistic associated with the SNP. The *p* value indicates the corresponding *p* value for that SNP. The adjusted significance indicates the

cutoff for significance when using a Bonferroni correction. The significance indicates whether or not the SNP was significant for the male dataset.

CHAPTER 4

DISCUSSION

The use of the PoDA program identified pathways of significance related to liver cancer. When used on datasets grouped by sex, it had the added benefit of identifying pathways that may be significant for one group but not the other. The difference in which pathways are significant for each group helps identify sex-specific etiologic differences that could explain differences in male and female incidence. A similar experiment could be performed using GWAS in order to identify individual genes and SNPs, but the PoDA program is able to detect more subtle impacts by individual genes and their interactions, through the process of grouping them into pathways. (Braun & Buetow, 2011)

Previous studies used PoDA to identify significant pathways in the same data set but did not look at sex-specific differences. (Braun & Buetow, 2011) The current research uncovered a variety of pathways, including those seen in previous work as well as novel ones. Below is a sample of the two sets of results, highlighting similarities and differences. This comparison is important for two reasons. First, it acts as a check on the PoDA program, ensuring it works as intended by finding similar results to past uses of the program. Second, it also acts as an update previous work and applies a known technique to a known dataset, with the benefit of optimizing and improving the selection process.

Table 3.

Comparison of Results from past studies

Pathway	Source	Length	DS_P	$p(DS_P)$	O.R.	q(O.R.)
Cell adhesion molecules (CAMs)	Kegg	86	1.57	9.09e-03	1.66	3.56e-13
ErbB signaling pathway	Kegg	76	1.45	3.45e-02	1.61	2.59e-10
Signaling events mediated by Stem cell factor receptor (c-Kit)	NCI-Nature	40	2.35	5.45e-03	1.58	7.31e-10
Neurotrophic factor-mediated Trk receptor signaling	NCI-Nature	50	1.60	2.36e-02	1.55	2.49e-08
Lissencephaly gene (LIS1) in neuronal migration and development	NCI-Nature	21	2.02	7.27e-03	1.52	1.44e-07
Angiopoietin receptor Tie2-mediated signaling	NCI-Nature	40	2.36	1.36e-02	1.51	5.77e-08
Reelin signaling pathway	NCI-Nature	28	1.62	5.45e-03	1.46	7.35e-08
Syndecan-4-mediated signaling events	NCI-Nature	27	1.74	1.64e-02	1.46	1.19e-06
Galactose metabolism	Kegg	19	1.65	2.27e-02	1.44	5.01e-06
Vibrio cholerae infection	Kegg	35	1.84	2.64e-02	1.43	6.67e-07
Paxillin-independent events mediated by a4b1 and a4b7	NCI-Nature	19	2.14	1.00e-02	1.40	6.67e-07
Antigen processing and presentation	Kegg	34	3.26	1.36e-02	1.40	3.71e-08
Corticosteroids and Cardioprotection	BioCarta	21	1.98	3.55e-02	1.39	1.24e-05
Lissencephaly gene (Lis1) in neuronal migration and development	BioCarta	15	1.60	1.36e-02	1.37	2.52e-05
IL12 signaling mediated by STAT4	NCI-Nature	25	1.93	4.55e-02	1.37	1.58e-05
Biosynthesis of unsaturated fatty acids	Kegg	13	1.76	1.64e-02	1.36	6.44e-05
Growth hormone signaling pathway	BioCarta	18	1.75	3.18e-02	1.36	7.46e-05
Canonical Wnt signaling pathway	NCI-Nature	28	1.92	4.73e-02	1.35	9.36e-06
NO2-dependent IL-12 pathway in NK cells	BioCarta	8	1.82	2.73e-03	1.32	5.83e-05
Signaling events mediated by HDAC Class III	NCI-Nature	19	2.12	3.91e-02	1.32	4.19e-05
Removal of aminoterminal propeptides from γ-carboxylated proteins	Reactome	7	3.12	5.45e-03	1.29	8.46e-05
Aminophosphonate metabolism	Kegg	13	1.91	3.36e-02	1.26	8.17e-04
Antigen processing and presentation	BioCarta	6	2.61	1.82e-03	1.22	3.36e-05
Classical complement pathway	BioCarta	12	2.27	1.55e-02	1.19	1.67e-04
Chylomicron-mediated lipid transport	Reactome	7	1.94	3.27e-02	1.16	1.49e-02

Pathway Names	DS	p(DS)	O.R	q(O.R)
respiratory electron transport, atp synthesis by chemiosmotic coupling, and heat productio	8.360027317	0	3.205930689	1.52577E-33
respiratory electron transport(reactome)	8.286362477	0	3.221769821	3.93547E-33
retrograde endocannabinoid signaling(kegg)	4.623409391	0.02	2.895220157	3.0683E-28
formation of the beta-catenin:tcf transactivating complex(reactome)	4.164193541	0.005	1.664608068	4.53322E-14
transcriptional regulation by tp53(reactome)	4.105536795	0.04	2.830577956	3.08173E-33
antigen processing and presentation(kegg)	3.925168424	0.014	1.484509815	6.8571E-13
signaling in immune system(nci/nature)	3.462283309	0.039	2.078044773	1.41437E-24
chaperones modulate interferon signaling pathway(biocarta)	3.431543262	0.005	1.541409601	8.86674E-10
glycosphingolipid biosynthesis - neo-lactoseries(kegg)	3.155144319	0.021	2.037506004	7.01243E-21
blood group systems biosynthesis(reactome)	3.130129105	0.003	1.513169983	1.9681E-09
pre-notch expression and processing(reactome)	3.08644809	0.013	1.808219952	2.6202E-16
interferon gamma signaling(reactome)	3.071325782	0.024	1.699004061	4.15418E-17
positive epigenetic regulation of rrna expression(reactome)	3.053183808	0.02	1.558995498	2.4602E-10
signaling by erythropoietin(reactome)	2.992968959	0.006	1.608596735	1.58887E-11
activation of hox genes during differentiation(reactome)	2.983660448	0.017	1.652926135	1.10896E-12
interleukin-3, interleukin-5 and gm-csf signaling(reactome)	2.849417273	0.014	1.708752918	2.03833E-12
tpo signaling pathway(biocarta)	2.775687367	0.013	1.558593081	1.1725E-09
calcineurin-regulated nfat-dependent transcription in lymphocytes(nci/nature)	2.761941602	0.027	1.665372317	1.60007E-11
antigen processing and presentation(biocarta)	2.72088214	0.005	1.292334922	4.19113E-06
rna polymerase i transcription(reactome)	2.631727176	0.022	1.505421302	1.5734E-09
insertion of tail-anchored proteins into the endoplasmic reticulum membrane(reactome)	2.62475205	0.017	1.470014789	1.49593E-08
il5-mediated signaling events(nci/nature)	2.57775443	0.011	1.380966826	5.58637E-07
t cell receptor signaling pathway(biocarta)	2.576016787	0.03	1.702589128	9.41958E-14
influenza a(kegg)	2.553655017	0.032	1.641751538	2.02835E-14
epigenetic regulation of gene expression(reactome)	2.523239805	0.042	1.689726052	6.33644E-14

The top table represents the published results from the first use of the PoDA program (Braun & Buetow, 2011), while the second table represents the current results for the dataset. For the current results, only the 25 most significant pathways are listed.

The first table represents the PoDA results from the original 2011 study. The second table represents the top 25 significant pathways in the combined dataset for the current study. Specific pathways are present in both tables, while others are absent in one or the other. There are multiple reasons for the results. First, in the top chart, pathways which overlap by more than 60% were removed. This process was also performed in the current research, but not for the results of the combined dataset. Second, many pathways were updated after the first study was completed. This includes combining pathways, eliminating them, or changing the number of genes within a given pathway. Finally, the second table represents only the 25 most significant pathways. Additional pathways were significant, but the combined pathways were not the subject of this study, so they did not undergo additional analysis. Overall, this work built on the original study, applying new techniques in order to investigate liver cancer. Most significantly, it focused on how genetic variation contributes to liver cancer differently in males and females.

Multiple pathways were identified to be significant for only either males or females. These pathways can be used to determine the biologic relevance of the PoDA process. For example, some of the significant pathways for males were antigen processing and presentation, digestion, and interferon gamma signaling. In contrast, the significant pathways for females included elastic fibre formation, aminosugars metabolism and negative regulation of fgfr3 signaling. All indicate a variety of biologic processes that may impact cancer development rates and highlight stark differences

between the two sexes. However, in both the male and female datasets, it is extremely difficult to identify a clear trend that relates all of the significant pathways to one another. The use of gene ontology addresses this problem. The gene ontology process looks for genes that are overrepresented in pathways, based on the data. By examining the subsets of genes in the significant pathways with gene ontology, the difference between overrepresented genes between males and females is identified. A sample of the two sets is presented in Figure 10 and 11. As shown in the figures, it appears both male and female gene sets were overrepresented in pathways that related in some way to cells responding to their environment. However, the results for males specifically included pathways that related to cytokines. At the same significant levels, the same type of results were absent from the female data. Three of the top five most significant results for males related to cytokines. For females, the first pathway relating to cytokines appears at the 137 most significant result. The trend continues when pathways greater than 1000 genes are removed. This step was employed in order to provide more specific results since extremely large pathways can be poorly defined and provide limited information about the specific biological processes contained within them. Continuing with this process, pathways with greater than 100 genes were removed which continued to show clear differences in the pathways that were represented by the significant genes in males and females. Results seen in Figures 14 and 15 indicate that interferon related pathways are present for males, but those same pathways are absent for females. This points to the key biologic factors that may be impacting the rates at which males and females develop hepatocellular carcinoma. Past work with liver cancer supports this result and suggests genetic risk factors for hepatocellular carcinoma include mutations to loci related to

immune response. (Clifford et al., 2010) Further research would be needed to determine any possible connection between the female pathways.

While all the genes within significant pathways potentially have value in terms of identifying key factors relating to liver cancer, the analysis also identified genes that were overrepresented in significant pathways. The overrepresented genes were present in many of the significant pathways, indicating an importance in relation to the development of liver cancer. Additionally, the overrepresented genes were different for males and females. This discovery is significant because it indicates that, at both the pathway and gene level, there are different factors which are significant for males and females. There were not simply the same significant genes divided into different pathways for the two sexes. The genes also maintained significance throughout the step-down process, and contributed significantly towards the observed odds ratio, as indicated by the step-down of the gene subsets. This result indicates which genes are driving the significance of the pathways and holds true for multiple pathways.

Each of the genes identified has a unique relationship to cancer, and many have been shown to have different impacts on males and females in the literature. For example, some of the repeated genes include janus kinase 2 and NFKB for males, and cbl proto-oncogene and epidermal growth factor for females. A 2017 paper identified that JAK2 knockout male mice did not develop liver cancer at the expected rates when exposed to the carcinogen DEN. (Shi et al., 2017) The results, however, could not be replicated in female mice. Additionally, a 2016 paper indicated a JAK2 deletion and siRNA knockout were protected against oxidative DNA damage and led to a delay in cancer development. (Themanns, 2016) The study only used male mice and therefore results were not provided

for any females. Finally, a 2019 paper indicated a knockout of JAK2 in hepatocytes led to a variety of diseases in the mice, including hepatocellular carcinoma. (Corbit, 2019) Interestingly, where there was an additional knockout of JAK2 in adipocytes, the same diseases were prevented. (Corbit, 2019) This may explain, in part, the reason the PoDA results indicate mutations to the JAK2 gene increase risk, while certain past papers indicate a JAK2 knockout in mice is protective. It is a complex gene, and the context of its mutation can determine its effect related to liver cancer. Therefore, further investigations are needed in order to determine the effect of SNPs and characterize the gene's relation to hepatocellular carcinoma. Clearly, there is a complexity present related to the cells in which the mutations occur, the overall impact of mutations on the organism, and possibly the sex of the organism that develops the mutations. Once again, the 2019 study did not use any female mice, so it is difficult to know if the knockouts would have had the same effect in females. Overall, the role of JAK2 was identified as inflammation, and it is believed this is the root of liver cancer as it is related to the JAK2 gene. NF-KB has also been identified as a regulator of inflammation, and similarly only appeared significant in the male dataset. (Luedde & Schwabe, 2011) (Wilson et al., 2015) A possible explanation for this phenomenon was identified in a 2008 paper, which indicated that NF-KB may be inhibited by estrogen. (Sun & Karin, 2008) This could explain why the gene appears significant for only one sex.

As for the genes that appeared significant only in the female dataset, cbl proto-oncogene has been reported as a negative regulator of T-cell activity. The same paper indicated that x-linked miRNAs were overexpressed in female t-cells, and that cbl may be the target of these miRNAs. (Clocchiatti et al., 2016) While the paper primarily focused

on autoimmune disorders, this finding highlights a potential rationale for the gene's significance only in females. Finally, Grb2 has been found to be overexpressed in many breast cancer cell lines. (Daly et al., 1994) While this is not directly related to its role in liver cancer, it is interesting that it is found in a cancer that primarily affects females. Additionally, this gene has been the target of therapeutic studies, and research findings suggest that it may be upregulated in hepatocellular carcinoma tissues. (Ly et al., 2020) Further research around this specific gene may reveal the reason it does not appear significant in the development of liver cancer for males but is significant for females.

The genes identified here are largely absent from the GWAS catalog of genes associated with hepatocellular carcinoma. (Lu et al., 2018) Despite this fact, there is a wide volume of literature that describes genes, specifically NF-KB and JAK2, as relating to liver cancer in mouse models. Clearly these genes play an important role in the development of liver cancer. Additionally, the use of only male models, with either a lack of female models or an inability to replicate results using females, indicates the genes may be part of the driving force that leads to a higher rate of liver cancer in males. Due to necessary limits placed on human research, most progress in this area comes either from bioinformatics or model studies. Both groups exhibit differing results regarding these genes and their relationship to liver cancer. The results presented here act as a bridge between the two, matching genetic data to results observed in animal models.

The combined pathway results, generated by combining subsets developed in the step-down process, give valuable insight into the relationship between key genes. Additionally, the results aid in indicating the genes most likely to be associated with the development of hepatocellular carcinoma. The Jak2 gene and NFKB gene were seen to

be removed late in the step-down process, illustrating that they are highly significant in the development of hepatocellular carcinoma in males. Additionally, the PoDA program provides the SNP which is most significant for each gene. For example, the most significant SNP for JAK2 was ‘rs2149556’. Further investigation of this SNP can now be performed in order to determine if it plays a particularly key role in the development of liver cancer in males, as the data indicated. In addition, it could be investigated as to why this SNP does not play a similar role for females, as a different SNP was more significant for that dataset.

Additionally, the conformation of gene significance by animal models indicates PoDA provides novel insights. It is able to identify genes that may have been overlooked through the use of traditional GWAS studies. The reason may be due, in part, to the fact that the individuals were segregated by sex in this analysis. While the genes and pathways were significant for one sex, this was not always true for the other. Therefore, the signals from the data may be canceling each other out, thus limiting the ability of researchers to identify significant genes when the sexes are pooled. Furthermore, the PoDA process purposely focuses on groups of genes and their interactions. This is a feature missing from more traditional GWAS systems. While these genes may not be significant individually, grouping them within pathways enables the program to find statistical significance that may otherwise be missed. The step-down process is then able to identify the genes which drive the significance of a specific pathway and further shed biological insight on the genes playing a role in the cancer development process.

These differences are important to identify because they point to sex-specific mechanisms of etiology which could explain why males develop liver cancer at much

higher rates than females. Variations in these could be used to identify the susceptible individual who would be monitored for early detection, leading to better outcomes for patients. Additionally, if different biologic processes are responsible for the development of cancer, it may be necessary to develop different treatments for males and females. It is important to be clear that males and females share the same genetic variance, they do not have significantly different sets of SNPs between them. These SNPs are contributing differently to the development of liver cancer, based on the results presented here.

The identification of the genes specific to males or females may also point to issues in previous liver cancer research. Some mutations may appear insignificant when males and females are grouped together but may become relevant when the two populations are split. Additionally, this fact calls into questions studies using participants from only one sex, whether it be animal or human. While such studies are still valuable, the results may not be applicable across sexes. For this reason, studies should include populations of both males and females, and provide comparisons across these groups. Although conducting studies in this manner may be more time consuming and costly, the results will be more applicable to the real world. Further work related to PoDA could work to incorporate a comparisons of distance scores between males and females. This may improve upon the method of selecting pathways based on significance in each sex. By comparing the actual values, the difference in significance for each pathway could be further quantified, and this may yield additional insight in how germline variation contributes to liver cancer differently for each sex.

While this study has made progress determining sex-specific patterns of variation the work is far from over. The complexity of cancer slows progress and makes

generalizations exceedingly difficult. A variety of techniques are required to capture all the interactions occurring within the disease. This includes the use of analysis such as this one, traditional techniques, animal studies, and approaches not yet imagined.

REFERENCES

Baecker, A., Liu, X., La Vecchia, C., & Zhang, Z. F. (2018). Worldwide incident hepatocellular carcinoma cases attributable to major risk factors. European journal of cancer prevention: the official journal of the European Cancer Prevention Organisation (ECP), 27(3), 205.

Bertuccio, P., Turati, F., Carioli, G., Rodriguez, T., La Vecchia, C., Malvezzi, M., & Negri, E. (2017). Global trends and predictions in hepatocellular carcinoma mortality. Journal of Hepatology, 67(2), 302-309.

Braun, R., & Buetow, K. (2011). Pathways of distinction analysis: a new technique for multi–SNP analysis of GWAS data. PLoS Genet, 7(6), e1002101.

Braun, R., Rowe, W., Schaefer, C., Zhang, J., & Buetow, K. (2009). Needles in the haystack: identifying individuals present in pooled genomic data. PLoS Genet, 5(10), e1000668.

Chen, Q. R., Braun, R., Hu, Y., Yan, C., Brunt, E. M., Meerzaman, D., ... & Buetow, K. (2013). Multi-SNP analysis of GWAS data identifies pathways associated with nonalcoholic fatty liver disease. PloS one, 8(7), e65982.

Clifford, R. J., Zhang, J., Meerzaman, D. M., Lyu, M. S., Hu, Y., Cultraro, C. M., ... & Buetow, K. H. (2010). Genetic variations at loci involved in the immune response are risk factors for hepatocellular carcinoma. Hepatology, 52(6), 2034-2043.

Clocchiatti, A., Cora, E., Zhang, Y., & Dotto, G. P. (2016). Sexual dimorphism in cancer. Nature Reviews Cancer, 16(5), 330-339.

Corbit, K. C., Wilson, C. G., Lowe, D., Tran, J. L., Vera, N. B., Clasquin, M., ... & Weiss, E. J. (2019). Adipocyte JAK2 mediates spontaneous metabolic liver disease and hepatocellular carcinoma. JCI insight, 4(17).

Daly, R. J., Binder, M. D., & Sutherland, R. L. (1994). Overexpression of the Grb2 gene in human breast cancer cell lines. Oncogene, 9(9), 2723-2727.

Hung, J. H., Yang, T. H., Hu, Z., Weng, Z., & DeLisi, C. (2012). Gene set enrichment analysis: performance evaluation and usage guidelines. Briefings in bioinformatics, 13(3), 281-291.

Korte, A., & Farlow, A. (2013). The advantages and limitations of trait analysis with GWAS: a review. Plant methods, 9(1), 1-9.

Liu, C. Y., Chen, K. F., & Chen, P. J. (2015). Treatment of liver cancer. Cold Spring Harbor perspectives in medicine, 5(9), a021535.

Liu, M., Liu, X., Liu, S., Xiao, F., Guo, E., Qin, X., ... & Zeng, X. (2020). Big Data-Based Identification of Multi-Gene Prognostic Signatures in Liver Cancer. Frontiers in oncology, 10, 847.

Lu, Y. K., Brill, J. M., Aghili, A., & Buetow, K. H. (2018). Pathway analysis provides insight into the genetic susceptibility to hepatocellular carcinoma and insight into immuno-therapy treatment response. Hepatoma Research, 4.

Luedde, T., & Schwabe, R. F. (2011). NF-κB in the liver—linking injury, fibrosis and hepatocellular carcinoma. Nature reviews Gastroenterology & hepatology, 8(2), 108-118.

Lv, J., Zhang, S., Wu, H., Lu, J., Lu, Y., Wang, F., ... & Yin, Z. (2020). Deubiquitinase PSMD14 enhances hepatocellular carcinoma growth and metastasis by stabilizing GRB2. Cancer letters, 469, 22-34.

Magi, R., Lindgren, C. M., & Morris, A. P. (2010). Meta-analysis of sex-specific genome-wide association studies. Genetic epidemiology, 34(8), 846-853.

Petrick, J. L., Braunlin, M., Laversanne, M., Valery, P. C., Bray, F., & McGlynn, K. A. (2016). International trends in liver cancer incidence, overall and by histologic subtype, 1978–2007. International journal of cancer, 139(7), 1534-1545.

Shi, S. Y., Luk, C. T., Schroer, S. A., Kim, M. J., Dodington, D. W., Sivasubramaniyam, T., ... & Woo, M. (2017). Janus Kinase 2 (JAK2) dissociates hepatosteatosis from hepatocellular carcinoma in mice. Journal of Biological Chemistry, 292(9), 3789-3799.

Su, C. H., Lin, Y., & Cai, L. (2013). Genetic factors, viral infection, other factors and liver cancer: an update on current progress. Asian Pacific Journal of Cancer Prevention, 14(9), 4953-4960.

Sun, B., & Karin, M. (2008). NF-κB signaling, liver disease and hepatoprotective agents. Oncogene, 27(48), 6228-6244.

Themanns, M., Mueller, K. M., Kessler, S. M., Golob-Schwarzl, N., Mohr, T., Kaltenecker, D., ... & Moriggl, R. (2016). Hepatic deletion of Janus kinase 2 counteracts oxidative stress in mice. Scientific reports, 6(1), 1-13.

Wilson, C. L., Jurk, D., Fullard, N., Banks, P., Page, A., Luli, S., ... & Mann, D. A. (2015). NFκB1 is a suppressor of neutrophil-driven hepatocellular carcinoma. Nature communications, 6(1), 1-15.

APPENDIX A

PODA RESULTS COMBINED DATASET

Pathway Names	DS	p(DS)	O.R	q(O.R)
respiratory electron transport, atp synthesis by chemiosmotic coupling, and heat production by uncoupling proteins.(reactome)	8.360027317	0	3.205930689	1.53E-33
respiratory electron transport(reactome)	8.286362477	0	3.221769821	3.94E-33
retrograde endocannabinoid signaling(kegg)	4.623409391	0.02	2.895220157	3.07E-28
formation of the beta-catenin:tcf transactivating complex(reactome)	4.164193541	0.005	1.664608068	4.53E-14
transcriptional regulation by tp53(reactome)	4.105536795	0.04	2.830577956	3.08E-33
antigen processing and presentation(kegg)	3.925168424	0.014	1.484509815	6.86E-13
signaling in immune system(nci/nature)	3.462283309	0.039	2.078044773	1.41E-24
chaperones modulate interferon signaling pathway(biocarta)	3.431543262	0.005	1.541409601	8.87E-10
glycosphingolipid biosynthesis - neo-lactoseries(kegg)	3.155144319	0.021	2.037506004	7.01E-21
blood group systems biosynthesis(reactome)	3.130129105	0.003	1.513169983	1.97E-09
pre-notch expression and processing(reactome)	3.08644809	0.013	1.808219952	2.62E-16
interferon gamma signaling(reactome)	3.071325782	0.024	1.699004061	4.15E-17
positive epigenetic regulation of rrna expression(reactome)	3.053183808	0.02	1.558995498	2.46E-10
signaling by erythropoietin(reactome)	2.992968959	0.006	1.608596735	1.59E-11
activation of hox genes during differentiation(reactome)	2.983660448	0.017	1.652926135	1.11E-12
interleukin-3, interleukin-5 and gm-csf signaling(reactome)	2.849417273	0.014	1.708752918	2.04E-12
tpo signaling pathway(biocarta)	2.775687367	0.013	1.558593081	1.17E-09
calcineurin-regulated nfat-dependent transcription in lymphocytes(nci/nature)	2.761941602	0.027	1.665372317	1.60E-11
antigen processing and presentation(biocarta)	2.72088214	0.005	1.292334922	4.19E-06
rna polymerase i transcription(reactome)	2.631727176	0.022	1.505421302	1.57E-09
insertion of tail-anchored proteins into the endoplasmic reticulum membrane(reactome)	2.62475205	0.017	1.470014789	1.50E-08
il5-mediated signaling events(nci/nature)	2.57775443	0.011	1.380966826	5.59E-07
t cell receptor signaling pathway(biocarta)	2.576016787	0.03	1.702589128	9.42E-14
influenza a(kegg)	2.553655017	0.032	1.641751538	2.03E-14
epigenetic regulation of gene expression(reactome)	2.523239805	0.042	1.689726052	6.34E-14
complement cascade(reactome)	2.514505076	0.036	1.631366938	1.66E-11
downstream signaling of activated fgfr3(reactome)	2.466434489	0.018	1.453952151	8.78E-08
glycosphingolipid biosynthesis - lacto and neolacto series(kegg)	2.412243435	0.016	1.516469167	6.81E-10
classical complement pathway(biocarta)	2.394279061	0.018	1.416660884	7.01E-08
gmcsf-mediated signaling events(nci/nature)	2.374310769	0.012	1.485057936	1.04E-08
il23-mediated signaling events(nci/nature)	2.345129484	0.024	1.501357097	6.43E-10
angiopoietin receptor tie2-mediated signaling(nci/nature)	2.337908358	0.028	1.642361208	1.39E-11
signaling events mediated by stem cell factor receptor (c-kit)(nci/nature)	2.3107894	0.041	1.792974096	6.94E-15
costimulation by the cd28 family(reactome)	2.308921519	0.034	1.574432791	4.14E-14
regulation of retinoblastoma protein(nci/nature)	2.304649195	0.039	1.706489418	7.14E-13
downstream signaling of activated fgfr2(reactome)	2.242301208	0.037	1.462603578	7.04E-08
rna polymerase i promoter clearance(reactome)	2.230735606	0.039	1.512364014	1.46E-09
erythropoietin mediated neuroprotection through nf-kb(biocarta)	2.168784445	0.025	1.345877828	4.33E-06
axon guidance(kegg)	2.150444107	0.041	1.679117453	9.42E-14
lectin induced complement pathway(biocarta)	2.106913846	0.021	1.429280103	1.84E-07
phospholipase c signaling pathway(biocarta)	2.093782222	0.017	1.296987416	3.82E-06
acetylation and deacetylation of rela in nucleus(biocarta)	2.068024584	0.026	1.333561795	1.07E-05
eicosanoid metabolism(biocarta)	2.055670861	0.041	1.462208341	3.35E-08
il12 signaling mediated by stat4(nci/nature)	1.985659551	0.044	1.434162931	4.63E-08
bmp receptor signaling(nci/nature)	1.978282218	0.043	1.523240685	2.84E-09
initial triggering of complement(reactome)	1.917362413	0.046	1.359452454	4.90E-06
platelet adhesion to exposed collagen(reactome)	1.868260205	0.046	1.403083531	4.66E-07
cadmium induces dna synthesis and proliferation in macrophages(biocarta)	1.75426054	0.046	1.315061706	7.63E-06
toll like receptor 7/8 (tlr7/8) cascade(reactome)	-1.174714265	0.047	1.413635323	1.57E-06
signaling by tgf-beta family members(reactome)	-1.196438803	0.048	1.329306964	3.18E-05
cilium assembly(reactome)	-1.288969992	0.025	1.568009743	3.41E-10
circadian entrainment(kegg)	-1.41489471	0.039	1.653538258	1.98E-11
er to golgi anterograde transport(reactome)	-1.427918441	0.024	1.455043573	3.83E-07
mitotic prometaphase(reactome)	-1.486687926	0.024	1.470125704	1.04E-07
assembly of collagen fibrils and other multimeric structures(reactome)	-1.65276857	0.026	1.329796257	3.26E-06
intra-golgi and retrograde golgi-to-er traffic(reactome)	-1.775746416	0.008	1.515780942	3.64E-08
regulation of tp53 activity(reactome)	-1.781665818	0.009	1.40250879	9.45E-07
transport of inorganic cations/anions and amino acids/oligopeptides(reactome)	-1.964634787	0.013	1.47752467	3.03E-08
degradation of the extracellular matrix(reactome)	-1.991790553	0.007	1.590753744	2.06E-12
morphine addiction(kegg)	-1.995105163	0.014	1.38734219	2.80E-06

APPENDIX B

PODA RESULTS MALE DATASET

Pathway Names	DS	p(DS)	O.R	q(O.R)
respiratory electron transport, atp synthesis by chemiosmotic coupling, and heat production by uncoupling proteins.(reactome)	7.726435954	0	3.310546443	2.09E-27
respiratory electron transport(reactome)	7.436885266	0	3.1324932	3.23E-26
retrograde endocannabinoid signaling(kegg)	3.607341727	0.011	3.075503477	1.11E-24
calcineurin-regulated nfat-dependent transcription in lymphocytes(nci/nature)	3.150813885	0.013	1.906913906	1.69E-13
antigen processing and presentation(kegg)	3.130120097	0.016	1.569151464	2.73E-11
signaling in immune system(nci/nature)	2.781814298	0.037	2.334584526	2.56E-22
lissencephaly gene (lis1) in neuronal migration and development(biocarta)	2.756005338	0.006	1.649677448	5.40E-10
chaperones modulate interferon signaling pathway(biocarta)	2.669804441	0.005	1.554895699	5.98E-08
il2 signaling events mediated by stat5(nci/nature)	2.658654621	0.012	1.686785288	1.30E-09
interleukin-3, interleukin-5 and gm-csf signaling(reactome)	2.648829894	0.013	1.780279875	2.15E-11
interferon gamma signaling(reactome)	2.641183139	0.017	1.801121777	1.52E-14
digestion(reactome)	2.471820189	0.006	1.488229208	1.45E-06
regulation of spermatogenesis by crem(biocarta)	2.423645726	0.013	1.52322935	4.01E-08
ecm proteoglycans(reactome)	2.344470121	0.04	2.136571902	6.37E-19
glycosphingolipid biosynthesis - neo-lactoseries(kegg)	2.342186108	0.039	2.162329428	2.94E-18
glycerophospholipid metabolism(kegg)	2.319445947	0.031	1.97349549	6.37E-13
tnfr2 signaling pathway(biocarta)	2.28930788	0.017	1.422892295	3.26E-06
antigen processing and presentation(biocarta)	2.217289101	0.011	1.364020661	2.10E-05
gmcsf-mediated signaling events(nci/nature)	2.165567936	0.021	1.50637025	3.87E-07
il3-mediated signaling events(nci/nature)	2.124242381	0.033	1.452754451	3.10E-06
glycosphingolipid biosynthesis - lacto and neolacto series(kegg)	2.11857634	0.018	1.587609858	9.08E-09
blood group systems biosynthesis(reactome)	2.098269575	0.023	1.615145203	4.64E-09
calnexin/calreticulin cycle(reactome)	2.047567145	0.023	1.447218456	2.20E-06
ether lipid metabolism(kegg)	2.045402404	0.034	1.707832432	3.41E-09
il23-mediated signaling events(nci/nature)	2.035469758	0.043	1.550907557	7.46E-08
vegf ligand-receptor interactions(reactome)	2.031857794	0.015	1.390599888	9.21E-06
il5-mediated signaling events(nci/nature)	1.962658519	0.03	1.424504653	7.19E-06
signaling events mediated by stem cell factor receptor (c-kit)(nci/nature)	1.950406288	0.043	1.946931158	9.40E-14
signaling by erythropoietin(reactome)	1.919195719	0.042	1.576728447	2.09E-08
repression of pain sensation by the transcriptional regulator dream(biocarta)	1.895083492	0.034	1.541594632	5.52E-07
axon guidance(kegg)	1.888162047	0.048	1.857792351	5.80E-14
classical complement pathway(biocarta)	1.881676774	0.032	1.381118408	1.48E-05
erythropoietin mediated neuroprotection through nf-kb(biocarta)	1.831919242	0.033	1.387406812	1.87E-05
alpha-linolenic acid metabolism(kegg)	1.787394166	0.042	1.418841284	5.34E-06
sumoylation by ranbp2 regulates transcriptional repression(nci/nature)	1.783719652	0.042	1.430136564	7.93E-06
neuroregulin receptor degredation protein-1 controls erbb3 receptor recycling(biocarta)	1.514449258	0.048	1.308991874	1.52E-05
membrane trafficking(reactome)	-1.161208471	0.019	2.974281349	1.12E-28
transport of inorganic cations/anions and amino acids/oligopeptides(reactome)	-1.238754239	0.042	1.668068778	5.11E-10
regulation of lipid metabolism by peroxisome proliferator-activated receptor alpha (pparalpha)(reactome)	-1.298541061	0.039	1.612884165	1.27E-08
fatty acid metabolism(reactome)	-1.418240966	0.028	1.643617653	6.14E-10
semaphorin interactions(reactome)	-1.523629978	0.039	1.427265573	4.13E-06
gabaergic synapse(kegg)	-1.62389507	0.033	1.495698591	2.80E-07
calcium signaling pathway(kegg)	-1.901272548	0.017	1.72336393	1.98E-10
cilium assembly(reactome)	-2.019307938	0.004	1.694604461	3.74E-09
regulation of tp53 activity(reactome)	-2.341054336	0.002	1.461895426	5.88E-06
opposing roles of aif in apoptosis and cell survival(biocarta)	-3.018199726	0.011	0.693852354	3.90E-08

44

APPENDIX C

PODA RESULTS FEMALE DATASET

Pathway Names	DS	p(DS)	O.R	q(O.R)
respiratory electron transport(reactome)	4.339749467	0	3.525940055	5.09E-12
respiratory electron transport, atp synthesis by chemiosmotic coupling, and heat production by uncoupling proteins.(reactome)	3.852150503	0	3.721787129	6.19E-12
rna polymerase i transcription(reactome)	3.125493062	0.004	2.369189628	2.84E-09
negative epigenetic regulation of rrna expression(reactome)	3.124673817	0.003	2.509395752	1.64E-09
rna polymerase i promoter clearance(reactome)	3.100511985	0.011	2.388029738	2.14E-09
pre-notch expression and processing(reactome)	2.983315586	0.004	3.490228907	1.62E-11
base-excision repair, ap site formation(reactome)	2.852099262	0.004	1.929879235	3.48E-06
retrograde endocannabinoid signaling(kegg)	2.782033847	0.008	4.973115847	1.14E-14
depyrimidination(reactome)	2.6716891	0.002	1.872227206	5.98E-06
lissencephaly gene (lis1) in neuronal migration and development(nci/nature)	2.592705742	0.009	2.682676456	3.97E-09
gene silencing by rna(reactome)	2.483531708	0.012	2.865311468	4.07E-10
chemical carcinogenesis(kegg)	2.363479792	0.019	2.013999253	1.22E-07
epigenetic regulation of gene expression(reactome)	2.334099912	0.022	2.578393193	2.45E-10
signaling by ptk6(reactome)	2.326238649	0.028	3.109575963	2.96E-10
class a/1 (rhodopsin-like receptors)(reactome)	2.310616934	0.02	5.848536917	2.41E-15
gpcr downstream signalling(reactome)	2.278136901	0.015	14.96263882	1.94E-12
glycerolipid metabolism(kegg)	2.276831966	0.016	3.513503178	1.82E-11
positive epigenetic regulation of rrna expression(reactome)	2.252961678	0.023	2.187883374	4.78E-08
1- and 2-methylnaphthalene degradation(kegg)	2.251776368	0.009	1.609674924	2.52E-05
formation of the beta-catenin:tcf transactivating complex(reactome)	2.206822369	0.018	2.094408717	3.31E-08
platelet adhesion to exposed collagen(reactome)	2.196629782	0.018	2.056128132	5.50E-07
activation of hox genes during differentiation(reactome)	2.177071216	0.034	2.412812747	4.02E-09
erk/mapk targets(reactome)	2.117039017	0.013	2.041005876	1.89E-06
htlv-i infection(kegg)	2.112208615	0.032	2.883507387	2.58E-11
transmission across chemical synapses(reactome)	2.07262153	0.049	12.07613085	1.56E-14
glycosphingolipid biosynthesis - neo-lactoseries(kegg)	2.047784299	0.039	3.073662094	1.20E-12
sphingolipid metabolism(reactome)	2.025292335	0.04	3.025094569	3.09E-11
cellular senescence(reactome)	2.019226621	0.048	3.775633963	2.62E-12
metabolism of xenobiotics by cytochrome p450(kegg)	1.986713893	0.049	1.469140528	4.94E-06
bacterial invasion of epithelial cells(kegg)	1.98497475	0.034	2.814327404	8.30E-10
role of mef2d in t-cell apoptosis(biocarta)	1.980653133	0.028	2.189869152	1.50E-07
g0 and early g1(reactome)	1.977525521	0.025	1.890636788	1.58E-05
aminosugars metabolism(kegg)	1.964259381	0.023	2.196257665	1.31E-06
internalization of erbb1(nci/nature)	1.950470446	0.039	2.346698462	1.18E-08
cbl mediated ligand-induced downregulation of egf receptors pathway(biocarta)	1.949206234	0.017	1.800297249	9.14E-06
calcineurin-regulated nfat-dependent transcription in lymphocytes(nci/nature)	1.904412742	0.027	2.380660657	3.94E-08
elastic fibre formation(reactome)	1.891624933	0.038	2.819741017	7.49E-09
negative regulation of fgfr3 signaling(reactome)	1.883918762	0.028	1.995264535	3.63E-06
mitotic prophase(reactome)	1.87310375	0.048	2.993612794	1.03E-09
reelin signaling pathway(nci/nature)	1.855963604	0.047	2.311566091	3.19E-08
trans-golgi network vesicle budding(reactome)	1.833150467	0.043	3.125972421	4.30E-11
fgfr3 ligand binding and activation(reactome)	1.793736758	0.036	1.809369056	2.16E-05
p73 transcription factor network(nci/nature)	1.775445258	0.049	3.060375377	5.72E-10
paxillin-dependent events mediated by a4b1(nci/nature)	1.761643395	0.047	2.198992961	2.59E-07
adp-ribosylation factor(biocarta)	1.743568887	0.049	1.929791828	7.18E-06
role of egf receptor transactivation by gpcrs in cardiac hypertrophy(biocarta)	1.73649617	0.042	2.39966131	1.24E-07
tgf-beta receptor signaling activates smads(reactome)	1.734115138	0.049	1.97532038	3.69E-06
mets affect on macrophage differentiation(biocarta)	1.678908104	0.049	1.873562491	4.31E-06
paxillin-independent events mediated by a4b1 and a4b7(nci/nature)	1.641899353	0.048	2.041092821	7.54E-07
mitotic metaphase and anaphase(reactome)	-1.471566771	0.035	2.0422147	3.02E-06
cdc42 signaling events(nci/nature)	-1.493800486	0.047	1.832658724	5.56E-06
cell-cell junction organization(reactome)	-1.621575075	0.028	1.824765408	1.26E-05
class i mhc mediated antigen processing & presentation(reactome)	-1.887778817	0.013	2.787570765	6.34E-11
metabolism of steroids(reactome)	-1.99506062	0.022	2.034660012	8.23E-07
metabolism of amino acids and derivatives(reactome)	-2.645159418	0.001	2.788030643	6.06E-10
fatty acid biosynthesis(kegg)	-4.53635869	0	0.588429221	1.48E-05
fructose metabolism(reactome)	-6.219412732	0	0.415455001	4.29E-10

APPENDIX D

CONTROL-CONTROL COMPARISONS

Pathway Names	DS	p(DS)	O.R	q(O.R)
synthesis of pips at the golgi membrane(reactome)	7.429426828	0	9.332414423	8.02E-21
integrin signaling pathway(biocarta)	7.347556922	0.002	3.782494786	4.96E-27
rho gtpases activate paks(reactome)	6.28289503	0.002	3.155915083	6.90E-24
rho gtpases activate cit(reactome)	7.078735778	0.002	3.057645864	2.82E-22
rho gtpases activate rocks(reactome)	6.98634906	0.002	2.849587477	2.95E-21
rho cell motility signaling pathway(biocarta)	7.106539557	0.003	3.468608686	1.86E-25
rac1 cell motility signaling pathway(biocarta)	6.637605346	0.004	3.363884284	3.85E-25
pkc-catalyzed phosphorylation of inhibitory phosphoprotein of myosin phosphatase(biocarta)	6.308222335	0.004	3.634558216	5.40E-25
dcc mediated attractive signaling(reactome)	6.277872766	0.005	4.205601686	3.26E-25
rho gtpases activate pkns(reactome)	6.623621277	0.005	3.448700105	6.86E-25
tnfr2 signaling pathway(biocarta)	2.44642503	0.005	1.520316691	4.60E-06
thrombin signaling and protease-activated receptors(biocarta)	6.940067243	0.006	3.974030712	2.28E-27
ccr3 signaling in eosinophils(biocarta)	5.418632653	0.006	2.739737618	8.50E-20
pi metabolism(reactome)	6.832756855	0.007	23.94586132	3.81E-27
sema3a pak dependent axon repulsion(reactome)	2.416396242	0.008	1.765040055	1.22E-09
calcineurin activates nfat(reactome)	2.216136388	0.008	1.530682227	4.72E-06
glycosaminoglycan metabolism(reactome)	-2.530075522	0.009	1.585056263	2.15E-06
ikk complex recruitment mediated by rip1(reactome)	2.58087282	0.009	1.521614872	2.18E-06
aflatoxin activation and detoxification(reactome)	2.151352953	0.01	1.609408128	1.26E-06
metabolism of amino acids and derivatives(reactome)	-2.098809373	0.011	2.160522358	4.46E-12
foxo-mediated transcription of oxidative stress, metabolic and neuronal genes(reactome)	2.582473589	0.011	1.890469969	1.12E-10
clec7a (dectin-1) induces nfat activation(reactome)	2.157565949	0.011	1.615028282	2.04E-07
ticam1, rip1-mediated ikk complex recruitment(reactome)	2.341870137	0.012	1.47233086	5.82E-06
phospholipid metabolism(reactome)	6.629960312	0.013	19.89729012	3.91E-30
sema3a-plexin repulsion signaling by inhibiting integrin adhesion(reactome)	2.347540483	0.013	1.619859202	1.79E-08
huntington's disease(kegg)	2.080159493	0.013	1.630000785	2.53E-08
regulation of plk1 activity at g2/m transition(reactome)	5.193899474	0.014	2.962948963	6.64E-22
hormone ligand-binding receptors(reactome)	2.067543745	0.014	1.457427411	5.37E-06
vascular smooth muscle contraction(kegg)	5.512808288	0.015	4.410587051	5.53E-30
netrin-1 signaling(reactome)	5.454310194	0.015	4.763016356	9.80E-28
free fatty acids regulate insulin secretion(reactome)	2.2146824	0.015	1.283830404	1.52E-05
senescence-associated secretory phenotype (sasp)(reactome)	2.747809387	0.016	1.817546965	1.31E-08
multi-step regulation of transcription by pitx2(biocarta)	2.157308543	0.016	1.663009037	1.79E-07
validated transcriptional targets of ap1 family members fra1 and fra2(nci/nature)	2.315843337	0.016	1.642597382	1.36E-06
metabolism of lipids(reactome)	5.920089331	0.017	34.47171948	1.18E-27
gap junction trafficking(reactome)	2.095161475	0.018	1.577192244	2.09E-06
mrna decay by 5' to 3' exoribonuclease(reactome)	2.152258095	0.018	1.514011213	1.73E-05
pink/parkin mediated mitophagy(reactome)	1.928102954	0.019	1.507292055	3.65E-06
ap-1 transcription factor network(nci/nature)	2.590602285	0.02	2.014673784	1.25E-11
ddx58/ifih1-mediated induction of interferon-alpha/beta(reactome)	2.634529406	0.021	1.792693091	1.92E-10
igf1 pathway(nci/nature)	2.396615867	0.022	1.977561551	1.19E-10
mitophagy(reactome)	2.265099518	0.022	1.778994178	1.61E-08
tumor suppressor arf inhibits ribosomal biogenesis(biocarta)	2.06611628	0.022	1.625208494	3.18E-07
met promotes cell motility(reactome)	2.125873733	0.023	2.09069724	2.46E-12
signaling mediated by p38-alpha and p38-beta(nci/nature)	2.054829998	0.023	1.87264122	7.08E-11
gap junction trafficking and regulation(reactome)	2.192586049	0.023	1.644653447	3.77E-07
ketone body metabolism(reactome)	1.945193622	0.023	1.470181056	5.49E-06
mitotic g2-g2/m phases(reactome)	4.704839694	0.024	3.453528633	8.38E-25
beta-alanine metabolism(kegg)	1.998396165	0.024	1.666707963	1.34E-07
toll like receptor 7/8 (tlr7/8) cascade(reactome)	2.425274874	0.025	2.034865883	2.51E-12
regulation of cortical dendrite branching(reactome)	1.808689868	0.025	1.445631648	4.87E-07
toll like receptor tlr6:tlr2 cascade(reactome)	2.270200809	0.026	2.107519375	5.27E-13
myd88:mal(tirap) cascade initiated on plasma membrane(reactome)	2.577098259	0.027	2.107519375	5.27E-13
nephrin family interactions(reactome)	2.134962647	0.027	1.83604315	2.14E-10
butanoate metabolism(kegg)	1.8697368	0.027	1.702856162	3.75E-08
trp channels(reactome)	1.96574489	0.029	1.724345125	1.81E-08
rip-mediated nfkb activation via zbp1(reactome)	1.93368259	0.03	1.472387705	1.25E-05
myd88 dependent cascade initiated on endosome(reactome)	2.300791423	0.031	2.034865883	2.51E-12
sumoylation of transcription cofactors(reactome)	2.129483844	0.031	1.870041571	7.31E-10
calcineurin-regulated nfat-dependent transcription in lymphocytes(nci/nature)	2.024124683	0.031	1.82911596	1.03E-08
crmps in sema3a signaling(reactome)	1.769531624	0.031	1.634283776	1.72E-08
digestion and absorption(reactome)	1.979809219	0.031	1.51782486	1.08E-05
g2/m transition(reactome)	4.359876391	0.033	3.147457637	2.02E-23
regulation of nuclear beta catenin signaling and target gene transcription(nci/nature)	2.086923213	0.033	2.103138799	3.88E-13
regulation of tp53 activity through acetylation(reactome)	1.805012437	0.033	1.547166727	1.35E-06
post-translational modification: synthesis of gpi-anchored proteins(reactome)	2.120388379	0.034	2.280608633	2.46E-14
met activates rap1 and rac1(reactome)	1.794821797	0.034	1.670866549	2.40E-06
digestion(reactome)	1.807788231	0.034	1.478404604	1.45E-05
circadian entrainment(kegg)	2.264716411	0.035	2.730684178	3.79E-20
regulation of nuclear smad2/3 signaling(nci/nature)	2.604563917	0.035	2.306267139	1.65E-14
toll like receptor 4 (tlr4) cascade(reactome)	2.179600496	0.036	2.265921405	6.30E-16
toll like receptor tlr1:tlr2 cascade(reactome)	2.230755378	0.036	2.116804636	4.29E-13
traf6 mediated induction of nfkb and map kinases upon tlr7/8 or 9 activation(reactome)	2.089283636	0.036	2.034865883	2.51E-12
toll like receptor 9 (tlr9) cascade(reactome)	2.434052554	0.036	2.002984572	2.57E-12
amyotrophic lateral sclerosis (als)(kegg)	1.992279457	0.036	1.875461567	4.14E-10
mecp2 regulates neuronal receptors and channels(reactome)	1.780449967	0.036	1.669608319	4.90E-08
platelet sensitization by ldl(reactome)	1.844828351	0.036	1.498969916	1.57E-05
toll like receptor 2 (tlr2) cascade(reactome)	2.133749226	0.037	2.116804636	4.29E-13
alcoholism(kegg)	1.974575769	0.039	1.997983678	2.39E-11
toll like receptor 10 (tlr10) cascade(reactome)	2.024138033	0.04	1.970833976	7.15E-12
toll like receptor 5 (tlr5) cascade(reactome)	2.02270726	0.04	1.970833976	7.15E-12
macroautophagy(reactome)	2.200395158	0.04	2.055991158	7.17E-12
signal dependent regulation of myogenesis by corepressor mitr(biocarta)	1.575858844	0.041	1.393854428	1.70E-05
long-term potentiation(kegg)	1.958217172	0.042	2.455108411	2.74E-16
myd88 cascade initiated on plasma membrane(reactome)	1.948765691	0.042	1.970833976	7.15E-12
assembly and cell surface presentation of nmda receptors(reactome)	1.811066724	0.042	1.854015814	5.10E-10
smooth muscle contraction(reactome)	1.735351139	0.043	1.84358459	7.23E-10
meiotic recombination(reactome)	1.945428845	0.043	1.670604151	3.47E-07
synthesis of ketone bodies(reactome)	1.621300591	0.043	1.453529825	4.70E-06
glutamatergic synapse(kegg)	1.863913566	0.044	2.518718284	5.75E-20
foxm1 transcription factor network(nci/nature)	1.830032584	0.044	1.783595234	3.81E-08
branched-chain amino acid catabolism(reactome)	1.769810942	0.044	1.618067184	9.46E-07
traf6 mediated nf-kb activation(reactome)	1.763154441	0.044	1.477594545	4.24E-06
cell division(biocarta)	1.674728315	0.044	1.493402255	7.77E-06
notch1 intracellular domain regulates transcription(reactome)	1.8269416	0.046	1.95473046	9.05E-11
downstream signaling in naive cd8+ t cells(nci/nature)	1.855943338	0.047	1.665416788	3.15E-08
adp-ribosylation factor(biocarta)	1.888561423	0.047	1.50638275	5.69E-06
mapk family signaling cascades(reactome)	-1.601557155	0.048	2.410731963	4.02E-15
measles(kegg)	2.033991472	0.048	1.641129601	1.28E-08
bacterial invasion of epithelial cells(kegg)	1.816293684	0.049	2.206886797	7.10E-13
nf-kappa b signaling pathway(kegg)	1.913291606	0.049	1.864458581	2.25E-10
amoebiasis(kegg)	1.838122518	0.049	1.886423262	3.19E-10

APPENDIX E

NEW PATHWAYS

MALE MOST SIGNIFICANT		
GENE ID	GENE SYMBOL	GENE NAME
248	ALPI	ALKALINE PHOSPHATASE, INTESTINAL
405	ARNT	ARYL HYDROCARBON RECEPTOR NUCLEAR TRANSLOCATOR
598	BCL2L1	BCL2 LIKE 1
815	CAMK2A	CALCIUM/CALMODULIN DEPENDENT PROTEIN KINASE II ALPHA
816	CAMK2B	CALCIUM/CALMODULIN DEPENDENT PROTEIN KINASE II BETA
821	CANX	CALNEXIN
916	CD3E	CD3E MOLECULE
925	CD8A	CD8A MOLECULE
972	CD74	CD74 MOLECULE
1051	CEBPB	CCAAT ENHANCER BINDING PROTEIN BETA
1118	CHIT1	CHITINASE 1
1385	CREB1	CAMP RESPONSIVE ELEMENT BINDING PROTEIN 1
1439	CSF2RB	COLONY STIMULATING FACTOR 2 RECEPTOR SUBUNIT BETA
1508	CTSB	CATHEPSIN B
1520	CTSS	CATHEPSIN S
2033	EP300	E1A BINDING PROTEIN P300
2633	GBP1	GUANYLATE BINDING PROTEIN 1
2635	GBP3	GUANYLATE BINDING PROTEIN 3
2885	GRB2	GROWTH FACTOR RECEPTOR BOUND PROTEIN 2
2984	GUCY2C	GUANYLATE CYCLASE 2C
3065	HDAC1	HISTONE DEACETYLASE 1
3105	HLA-A	MAJOR HISTOCOMPATIBILITY COMPLEX, CLASS I, A
3106	HLA-B	MAJOR HISTOCOMPATIBILITY COMPLEX, CLASS I, B
3108	HLA-DMA	MAJOR HISTOCOMPATIBILITY COMPLEX, CLASS II, DM ALPHA
3111	HLA-DOA	MAJOR HISTOCOMPATIBILITY COMPLEX, CLASS II, DO ALPHA
3112	HLA-DOB	MAJOR HISTOCOMPATIBILITY COMPLEX, CLASS II, DO BETA
3113	HLA-DPA1	MAJOR HISTOCOMPATIBILITY COMPLEX, CLASS II, DP ALPHA 1
3115	HLA-DPB1	MAJOR HISTOCOMPATIBILITY COMPLEX, CLASS II, DP BETA 1
3117	HLA-DQA1	MAJOR HISTOCOMPATIBILITY COMPLEX, CLASS II, DQ ALPHA 1
3118	HLA-DQA2	MAJOR HISTOCOMPATIBILITY COMPLEX, CLASS II, DQ ALPHA 2
3122	HLA-DRA	MAJOR HISTOCOMPATIBILITY COMPLEX, CLASS II, DR ALPHA
3134	HLA-F	MAJOR HISTOCOMPATIBILITY COMPLEX, CLASS I, F
3306	HSPA2	HEAT SHOCK PROTEIN FAMILY A (HSP70) MEMBER 2
3308	HSPA4	HSPA4
3309	HSPA5	HEAT SHOCK PROTEIN FAMILY A (HSP70) MEMBER 5
3312	HSPA8	HEAT SHOCK PROTEIN FAMILY A (HSP70) MEMBER 8
3394	IRF8	INTERFERON REGULATORY FACTOR 8
3459	IFNGR1	INTERFERON GAMMA RECEPTOR 1

3553	IL1B	INTERLEUKIN 1 BETA
3562	IL3	INTERLEUKIN 3
3594	IL12RB1	INTERLEUKIN 12 RECEPTOR SUBUNIT BETA 1
3605	IL17A	INTERLEUKIN 17A
3635	INPP5D	INOSITOL POLYPHOSPHATE-5-PHOSPHATASE D
3660	IRF2	INTERFERON REGULATORY FACTOR 2
3663	IRF5	INTERFERON REGULATORY FACTOR 5
3664	IRF6	INTERFERON REGULATORY FACTOR 6
3675	ITGA3	INTEGRIN SUBUNIT ALPHA 3
3717	JAK2	JANUS KINASE 2
3725	JUN	JUN PROTO-ONCOGENE, AP-1 TRANSCRIPTION FACTOR SUBUNIT
3804	KIR2DL3	KILLER CELL IMMUNOGLOBULIN LIKE RECEPTOR, TWO IG DOMAINS AND LONG CYTOPLASMIC TAIL 3
3805	KIR2DL4	KILLER CELL IMMUNOGLOBULIN LIKE RECEPTOR, TWO IG DOMAINS AND LONG CYTOPLASMIC TAIL 4
3811	KIR3DL1	KILLER CELL IMMUNOGLOBULIN LIKE RECEPTOR, THREE IG DOMAINS AND LONG CYTOPLASMIC TAIL 1
3821	KLRC1	KILLER CELL LECTIN LIKE RECEPTOR C1
3938	LCT	LACTASE
4049	LTA	LYMPHOTOXIN ALPHA
4214	MAP3K1	MITOGEN-ACTIVATED PROTEIN KINASE KINASE KINASE 1
4261	CIITA	CLASS II MAJOR HISTOCOMPATIBILITY COMPLEX TRANSACTIVATOR
4281	MID1	MIDLINE 1
4283	CXCL9	C-X-C MOTIF CHEMOKINE LIGAND 9
4353	MPO	MYELOPEROXIDASE
4684	NCAM1	NEURAL CELL ADHESION MOLECULE 1
4790	NFKB1	NUCLEAR FACTOR KAPPA B SUBUNIT 1
4800	NFYA	NUCLEAR TRANSCRIPTION FACTOR Y SUBUNIT ALPHA
4801	NFYB	NUCLEAR TRANSCRIPTION FACTOR Y SUBUNIT BETA
4843	NOS2	NITRIC OXIDE SYNTHASE 2
4938	OAS1	2'-5'-OLIGOADENYLATE SYNTHETASE 1
5290	PIK3CA	PHOSPHATIDYLINOSITOL-4,5-BISPHOSPHATE 3-KINASE CATALYTIC SUBUNIT ALPHA
5295	PIK3R1	PHOSPHOINOSITIDE-3-KINASE REGULATORY SUBUNIT 1
5407	PNLIPRP1	PANCREATIC LIPASE RELATED PROTEIN 1
5408	PNLIPRP2	PANCREATIC LIPASE RELATED PROTEIN 2 (GENE/PSEUDOGENE)
5641	LGMN	LEGUMAIN
5721	PSME2	PROTEASOME ACTIVATOR SUBUNIT 2
5770	PTPN1	PROTEIN TYROSINE PHOSPHATASE NON-RECEPTOR TYPE 1
5903	RANBP2	RAN BINDING PROTEIN 2
5905	RANGAP1	RAN GTPASE ACTIVATING PROTEIN 1
6048	RNF5	RING FINGER PROTEIN 5
6347	CCL2	C-C MOTIF CHEMOKINE LIGAND 2
6400	SEL1L	SEL1L ADAPTOR SUBUNIT OF ERAD E3 UBIQUITIN LIGASE

6464	SHC1	SHC ADAPTOR PROTEIN 1
6476	SI	SUCRASE-ISOMALTASE
6672	SP100	SP100 NUCLEAR ANTIGEN
6774	STAT3	SIGNAL TRANSDUCER AND ACTIVATOR OF TRANSCRIPTION 3
6775	STAT4	SIGNAL TRANSDUCER AND ACTIVATOR OF TRANSCRIPTION 4
6776	STAT5A	SIGNAL TRANSDUCER AND ACTIVATOR OF TRANSCRIPTION 5A
6890	TAP1	TRANSPORTER 1, ATP BINDING CASSETTE SUBFAMILY B MEMBER
6892	TAPBP	TAP BINDING PROTEIN
7124	TNF	TUMOR NECROSIS FACTOR
7133	TNFRSF1B	TNF RECEPTOR SUPERFAMILY MEMBER 1B
7185	TRAF1	TNF RECEPTOR ASSOCIATED FACTOR 1
7297	TYK2	TYROSINE KINASE 2
7311	UBA52	UBIQUITIN A-52 RESIDUE RIBOSOMAL PROTEIN FUSION PRODUCT 1
7412	VCAM1	VASCULAR CELL ADHESION MOLECULE 1
7532	YWHAG	TYROSINE 3-MONOOXYGENASE/TRYPTOPHAN 5-MONOOXYGENASE ACTIVATION PROTEIN GAMMA
7534	YWHAZ	TYROSINE 3-MONOOXYGENASE/TRYPTOPHAN 5-MONOOXYGENASE ACTIVATION PROTEIN ZETA
7726	TRIM26	TRIPARTITE MOTIF CONTAINING 26
7844	RNF103	RING FINGER PROTEIN 103
8021	NUP214	NUCLEOPORIN 214
8638	OASL	2'-5'-OLIGOADENYLATE SYNTHETASE LIKE
8807	IL18RAP	INTERLEUKIN 18 RECEPTOR ACCESSORY PROTEIN
8809	IL18R1	INTERLEUKIN 18 RECEPTOR 1
8844	KSR1	KINASE SUPPRESSOR OF RAS 1
8972	MGAM	MALTASE-GLUCOAMYLASE
9020	MAP3K14	MITOGEN-ACTIVATED PROTEIN KINASE KINASE KINASE 14
9695	EDEM1	ER DEGRADATION ENHANCING ALPHA-MANNOSIDASE LIKE PROTEIN 1
9759	HDAC4	HISTONE DEACETYLASE 4
9830	TRIM14	TRIPARTITE MOTIF CONTAINING 14
9972	NUP153	NUCLEOPORIN 153
10083	USH1C	USH1 PROTEIN NETWORK COMPONENT HARMONIN
10256	CNKSR1	CONNECTOR ENHANCER OF KINASE SUPPRESSOR OF RAS 1
10346	TRIM22	TRIPARTITE MOTIF CONTAINING 22
10612	TRIM3	TRIPARTITE MOTIF CONTAINING 3
10956	OS9	OS9 ENDOPLASMIC RETICULUM LECTIN
10987	COPS5	COP9 SIGNALOSOME SUBUNIT 5
11009	IL24	INTERLEUKIN 24
11181	TREH	TREHALASE
11253	MAN1B1	MANNOSIDASE ALPHA CLASS 1B MEMBER 1
23193	GANAB	GLUCOSIDASE II ALPHA SUBUNIT
23225	NUP210	NUCLEOPORIN 210

23321	TRIM2	TRIPARTITE MOTIF CONTAINING 2
23650	TRIM29	TRIPARTITE MOTIF CONTAINING 29
27159	CHIA	CHITINASE ACIDIC
51009	DERL2	DERLIN 2
55741	EDEM2	ER DEGRADATION ENHANCING ALPHA-MANNOSIDASE LIKE PROTEIN 2
55757	UGGT2	UDP-GLUCOSE GLYCOPROTEIN GLUCOSYLTRANSFERASE 2
56886	UGGT1	UDP-GLUCOSE GLYCOPROTEIN GLUCOSYLTRANSFERASE 1
79097	TRIM48	TRIPARTITE MOTIF CONTAINING 48
80263	TRIM45	TRIPARTITE MOTIF CONTAINING 45
80314	EPC1	ENHANCER OF POLYCOMB HOMOLOG 1
91445	RNF185	RING FINGER PROTEIN 185
115362	GBP5	GUANYLATE BINDING PROTEIN 5
117854	TRIM6	TRIPARTITE MOTIF CONTAINING 6
119548	PNLIPRP3	PANCREATIC LIPASE RELATED PROTEIN 3
149233	IL23R	INTERLEUKIN 23 RECEPTOR

MALE MINIMUM		
GENE ID	GENE SYMBOL	GENE NAME
405	ARNT	ARYL HYDROCARBON RECEPTOR NUCLEAR TRANSLOCATOR
815	CAMK2A	CALCIUM/CALMODULIN DEPENDENT PROTEIN KINASE II ALPHA
821	CANX	CALNEXIN
2033	EP300	E1A BINDING PROTEIN P300
2984	GUCY2C	GUANYLATE CYCLASE 2C
3106	HLA-B	MAJOR HISTOCOMPATIBILITY COMPLEX, CLASS I, B
3113	HLA-DPA1	MAJOR HISTOCOMPATIBILITY COMPLEX, CLASS II, DP ALPHA 1
3115	HLA-DPB1	MAJOR HISTOCOMPATIBILITY COMPLEX, CLASS II, DP BETA 1
3117	HLA-DQA1	MAJOR HISTOCOMPATIBILITY COMPLEX, CLASS II, DQ ALPHA 1
3118	HLA-DQA2	MAJOR HISTOCOMPATIBILITY COMPLEX, CLASS II, DQ ALPHA 2
3122	HLA-DRA	MAJOR HISTOCOMPATIBILITY COMPLEX, CLASS II, DR ALPHA
3308	HSPA4	HSPA4
3394	IRF8	INTERFERON REGULATORY FACTOR 8
3459	IFNGR1	INTERFERON GAMMA RECEPTOR 1
3553	IL1B	INTERLEUKIN 1 BETA
3635	INPP5D	INOSITOL POLYPHOSPHATE-5-PHOSPHATASE D
3663	IRF5	INTERFERON REGULATORY FACTOR 5
3675	ITGA3	INTEGRIN SUBUNIT ALPHA 3
3717	JAK2	JANUS KINASE 2
3725	JUN	JUN PROTO-ONCOGENE, AP-1 TRANSCRIPTION FACTOR SUBUNIT
3811	KIR3DL1	KILLER CELL IMMUNOGLOBULIN LIKE RECEPTOR, THREE IG DOMAINS AND LONG CYTOPLASMIC TAIL 1
3938	LCT	LACTASE

4049	LTA	LYMPHOTOXIN ALPHA
4214	MAP3K1	MITOGEN-ACTIVATED PROTEIN KINASE KINASE KINASE 1
4261	CIITA	CLASS II MAJOR HISTOCOMPATIBILITY COMPLEX TRANSACTIVATOR
4684	NCAM1	NEURAL CELL ADHESION MOLECULE 1
4790	NFKB1	NUCLEAR FACTOR KAPPA B SUBUNIT 1
4843	NOS2	NITRIC OXIDE SYNTHASE 2
5295	PIK3R1	PHOSPHOINOSITIDE-3-KINASE REGULATORY SUBUNIT 1
5407	PNLIPRP1	PANCREATIC LIPASE RELATED PROTEIN 1
5905	RANGAP1	RAN GTPASE ACTIVATING PROTEIN 1
6476	SI	SUCRASE-ISOMALTASE
6890	TAP1	TRANSPORTER 1, ATP BINDING CASSETTE SUBFAMILY B MEMBER
7124	TNF	TUMOR NECROSIS FACTOR
7133	TNFRSF1B	TNF RECEPTOR SUPERFAMILY MEMBER 1B
7185	TRAF1	TNF RECEPTOR ASSOCIATED FACTOR 1
7412	VCAM1	VASCULAR CELL ADHESION MOLECULE 1
7726	TRIM26	TRIPARTITE MOTIF CONTAINING 26
8807	IL18RAP	INTERLEUKIN 18 RECEPTOR ACCESSORY PROTEIN
8844	KSR1	KINASE SUPPRESSOR OF RAS 1
9759	HDAC4	HISTONE DEACETYLASE 4
10083	USH1C	USH1 PROTEIN NETWORK COMPONENT HARMONIN
11253	MAN1B1	MANNOSIDASE ALPHA CLASS 1B MEMBER 1
23225	NUP210	NUCLEOPORIN 210
23321	TRIM2	TRIPARTITE MOTIF CONTAINING 2
27159	CHIA	CHITINASE ACIDIC
55757	UGGT2	UDP-GLUCOSE GLYCOPROTEIN GLUCOSYLTRANSFERASE 2
91445	RNF185	RING FINGER PROTEIN 185
119548	PNLIPRP3	PANCREATIC LIPASE RELATED PROTEIN 3
149233	IL23R	INTERLEUKIN 23 RECEPTOR

FEMALE MAX SIGNIFICANCE		
GENE ID	GENE SYMBOL	GENE NAME
125	ADH1B	ALCOHOL DEHYDROGENASE 1B (CLASS I), BETA POLYPEPTIDE
126	ADH1C	ALCOHOL DEHYDROGENASE 1C (CLASS I), GAMMA POLYPEPTIDE
128	ADH5	ALCOHOL DEHYDROGENASE 5 (CLASS III), CHI POLYPEPTIDE
131	ADH7	ALCOHOL DEHYDROGENASE 7 (CLASS IV), MU OR SIGMA POLYPEPTIDE
186	AGTR2	ANGIOTENSIN II RECEPTOR TYPE 2
220	ALDH1A3	ALDEHYDE DEHYDROGENASE 1 FAMILY MEMBER A3
222	ALDH3B2	ALDEHYDE DEHYDROGENASE 3 FAMILY MEMBER B2
273	AMPH	AMPHIPHYSIN
387	RHOA	RAS HOMOLOG FAMILY MEMBER A
393	ARHGAP4	RHO GTPASE ACTIVATING PROTEIN 4

395	ARHGAP6	RHO GTPASE ACTIVATING PROTEIN 6
650	BMP2	BONE MORPHOGENETIC PROTEIN 2
655	BMP7	BONE MORPHOGENETIC PROTEIN 7
673	BRAF	B-RAF PROTO-ONCOGENE, SERINE/THREONINE KINASE
867	CBL	CBL PROTO-ONCOGENE
868	CBLB	CBL PROTO-ONCOGENE B
898	CCNE1	CYCLIN E1
998	CDC42	CELL DIVISION CYCLE 42
1017	CDK2	CYCLIN DEPENDENT KINASE 2
1109	AKR1C4	ALDO-KETO REDUCTASE FAMILY 1 MEMBER C4
1118	CHIT1	CHITINASE 1
1147	CHUK	COMPONENT OF INHIBITOR OF NUCLEAR FACTOR KAPPA B KINASE COMPLEX
1212	CLTB	CLATHRIN LIGHT CHAIN B
1314	COPA	COPI COAT COMPLEX SUBUNIT ALPHA
1435	CSF1	COLONY STIMULATING FACTOR 1
1486	CTBS	CHITOBIASE
1555	CYP2B6	CYTOCHROME P450 FAMILY 2 SUBFAMILY B MEMBER 6
1558	CYP2C8	CYTOCHROME P450 FAMILY 2 SUBFAMILY C MEMBER 8
1571	CYP2E1	CYTOCHROME P450 FAMILY 2 SUBFAMILY E MEMBER 1
1576	CYP3A4	CYTOCHROME P450 FAMILY 3 SUBFAMILY A MEMBER 4
1605	DAG1	DYSTROGLYCAN 1
1645	AKR1C1	ALDO-KETO REDUCTASE FAMILY 1 MEMBER C1
1793	DOCK1	DEDICATOR OF CYTOKINESIS 1
1859	DYRK1A	DUAL SPECIFICITY TYROSINE PHOSPHORYLATION REGULATED KINASE 1A
1869	E2F1	E2F TRANSCRIPTION FACTOR 1
1874	E2F4	E2F TRANSCRIPTION FACTOR 4
1875	E2F5	E2F TRANSCRIPTION FACTOR 5
1906	EDN1	ENDOTHELIN 1
1909	EDNRA	ENDOTHELIN RECEPTOR TYPE A
1956	EGFR	EPIDERMAL GROWTH FACTOR RECEPTOR
2052	EPHX1	EPOXIDE HYDROLASE 1
2060	EPS15	EPIDERMAL GROWTH FACTOR RECEPTOR PATHWAY SUBSTRATE 15
2113	ETS1	ETS PROTO-ONCOGENE 1, TRANSCRIPTION FACTOR
2114	ETS2	ETS PROTO-ONCOGENE 2, TRANSCRIPTION FACTOR
2192	FBLN1	FIBULIN 1
2199	FBLN2	FIBULIN 2
2200	FBN1	FIBRILLIN 1
2201	FBN2	FIBRILLIN 2
2202	EFEMP1	EGF CONTAINING FIBULIN EXTRACELLULAR MATRIX PROTEIN 1
2249	FGF4	FIBROBLAST GROWTH FACTOR 4

2250	FGF5	FIBROBLAST GROWTH FACTOR 5
2532	ACKR1	ATYPICAL CHEMOKINE RECEPTOR 1 (DUFFY BLOOD GROUP)
2645	GCK	GLUCOKINASE
2673	GFPT1	GLUTAMINE--FRUCTOSE-6-PHOSPHATE TRANSAMINASE 1
2767	GNA11	G PROTEIN SUBUNIT ALPHA 11
2822	GPLD1	GLYCOSYLPHOSPHATIDYLINOSITOL SPECIFIC PHOSPHOLIPASE D1
2885	GRB2	GROWTH FACTOR RECEPTOR BOUND PROTEIN 2
2939	GSTA2	GLUTATHIONE S-TRANSFERASE ALPHA 2
2948	GSTM4	GLUTATHIONE S-TRANSFERASE MU 4
2954	GSTZ1	GLUTATHIONE S-TRANSFERASE ZETA 1
3066	HDAC2	HISTONE DEACETYLASE 2
3073	HEXA	HEXOSAMINIDASE SUBUNIT ALPHA
3074	HEXB	HEXOSAMINIDASE SUBUNIT BETA
3098	HK1	HEXOKINASE 1
3101	HK3	HEXOKINASE 3
3265	HRAS	HRAS PROTO-ONCOGENE, GTPASE
3551	IKBKB	INHIBITOR OF NUCLEAR FACTOR KAPPA B KINASE SUBUNIT BETA
3676	ITGA4	ITGA4
3678	ITGA5	INTEGRIN SUBUNIT ALPHA 5
3683	ITGAL	INTEGRIN SUBUNIT ALPHA L
3688	ITGB1	INTEGRIN SUBUNIT BETA 1
3689	ITGB2	INTEGRIN SUBUNIT BETA 2
3695	ITGB7	INTEGRIN SUBUNIT BETA 7
3696	ITGB8	INTEGRIN SUBUNIT BETA 8
3725	JUN	JUN PROTO-ONCOGENE, AP-1 TRANSCRIPTION FACTOR SUBUNIT
4015	LOX	LYSYL OXIDASE
4017	LOXL2	LYSYL OXIDASE LIKE 2
4052	LTBP1	LATENT TRANSFORMING GROWTH FACTOR BETA BINDING PROTEIN 1
4053	LTBP2	LATENT TRANSFORMING GROWTH FACTOR BETA BINDING PROTEIN 2
4149	MAX	MYC ASSOCIATED FACTOR X
4237	MFAP2	MICROFIBRIL ASSOCIATED PROTEIN 2
4239	MFAP4	MICROFIBRIL ASSOCIATED PROTEIN 4
5111	PCNA	PROLIFERATING CELL NUCLEAR ANTIGEN
5238	PGM3	PHOSPHOGLUCOMUTASE 3
5516	PPP2CB	PROTEIN PHOSPHATASE 2 CATALYTIC SUBUNIT BETA
5518	PPP2R1A	PROTEIN PHOSPHATASE 2 SCAFFOLD SUBUNIT AALPHA
5575	PRKAR1B	PROTEIN KINASE CAMP-DEPENDENT TYPE I REGULATORY SUBUNIT BETA
5576	PRKAR2A	PROTEIN KINASE CAMP-DEPENDENT TYPE II REGULATORY SUBUNIT ALPHA
5577	PRKAR2B	PROTEIN KINASE CAMP-DEPENDENT TYPE II REGULATORY SUBUNIT BETA
5578	PRKCA	PROTEIN KINASE C ALPHA

5579	PRKCB	PROTEIN KINASE C BETA
5604	MAP2K1	MITOGEN-ACTIVATED PROTEIN KINASE KINASE 1
5879	RAC1	RAC FAMILY SMALL GTPASE 1
5933	RBL1	RB TRANSCRIPTIONAL COREPRESSOR LIKE 1
5934	RBL2	RB TRANSCRIPTIONAL COREPRESSOR LIKE 2
6195	RPS6KA1	RIBOSOMAL PROTEIN S6 KINASE A1
6453	ITSN1	INTERSECTIN 1
6456	SH3GL2	SH3 DOMAIN CONTAINING GRB2 LIKE 2, ENDOPHILIN A1
6464	SHC1	SHC ADAPTOR PROTEIN 1
6654	SOS1	SOS RAS/RAC GUANINE NUCLEOTIDE EXCHANGE FACTOR 1
6675	UAP1	UDP-N-ACETYLGLUCOSAMINE PYROPHOSPHORYLASE 1
7040	TGFB1	TRANSFORMING GROWTH FACTOR BETA 1
7043	TGFB3	TRANSFORMING GROWTH FACTOR BETA 3
7153	TOP2A	DNA TOPOISOMERASE II ALPHA
7316	UBC	UBIQUITIN C
7321	UBE2D1	UBIQUITIN CONJUGATING ENZYME E2 D1
7322	UBE2D2	UBIQUITIN CONJUGATING ENZYME E2 D2
7323	UBE2D3	UBIQUITIN CONJUGATING ENZYME E2 D3
7364	UGT2B7	UDP GLUCURONOSYLTRANSFERASE FAMILY 2 MEMBER B7
8038	ADAM12	ADAM METALLOPEPTIDASE DOMAIN 12
8074	FGF23	FIBROBLAST GROWTH FACTOR 23
8076	MFAP5	MICROFIBRIL ASSOCIATED PROTEIN 5
8200	GDF5	GROWTH DIFFERENTIATION FACTOR 5
8260	NAA10	N-ALPHA-ACETYLTRANSFERASE 10, NATA CATALYTIC SUBUNIT
8516	ITGA8	INTEGRIN SUBUNIT ALPHA 8
8569	MKNK1	MAPK INTERACTING SERINE/THREONINE KINASE 1
8644	AKR1C3	ALDO-KETO REDUCTASE FAMILY 1 MEMBER C3
8729	GBF1	GOLGI BREFELDIN A RESISTANT GUANINE NUCLEOTIDE EXCHANGE FACTOR 1
8817	FGF18	FIBROBLAST GROWTH FACTOR 18
8822	FGF17	FIBROBLAST GROWTH FACTOR 17
8853	ASAP2	ARFGAP WITH SH3 DOMAIN, ANKYRIN REPEAT AND PH DOMAIN 2
8898	MTMR2	MYOTUBULARIN RELATED PROTEIN 2
9101	USP8	UBIQUITIN SPECIFIC PEPTIDASE 8
9107	MTMR6	MYOTUBULARIN RELATED PROTEIN 6
9134	CCNE2	CYCLIN E2
9267	CYTH1	CYTOHESIN 1
9446	GSTO1	GLUTATHIONE S-TRANSFERASE OMEGA 1
9564	BCAR1	BCAR1 SCAFFOLD PROTEIN, CAS FAMILY MEMBER
10007	GNPDA1	GLUCOSAMINE-6-PHOSPHATE DEAMINASE 1
10020	GNE	GLUCOSAMINE (UDP-N-ACETYL)-2-EPIMERASE/N-ACETYLMANNOSAMINE KINASE

10253	SPRY2	SPROUTY RTK SIGNALING ANTAGONIST 2
10564	ARFGEF2	ADP RIBOSYLATION FACTOR GUANINE NUCLEOTIDE EXCHANGE FACTOR 2
10617	STAMBP	STAM BINDING PROTEIN
10818	FRS2	FIBROBLAST GROWTH FACTOR RECEPTOR SUBSTRATE 2
10941	UGT2A1	UDP GLUCURONOSYLTRANSFERASE FAMILY 2 MEMBER A1 COMPLEX LOCUS
10945	KDELR1	KDEL ENDOPLASMIC RETICULUM PROTEIN RETENTION RECEPTOR 1
11015	KDELR3	KDEL ENDOPLASMIC RETICULUM PROTEIN RETENTION RECEPTOR 3
23309	SIN3B	SIN3 TRANSCRIPTION REGULATOR FAMILY MEMBER B
25942	SIN3A	SIN3 TRANSCRIPTION REGULATOR FAMILY MEMBER A
26018	LRIG1	LEUCINE RICH REPEATS AND IMMUNOGLOBULIN LIKE DOMAINS 1
27128	CYTH4	CYTOHESIN 4
27159	CHIA	CHITINASE ACIDIC
27294	DHDH	DIHYDRODIOL DEHYDROGENASE
28964	GIT1	GIT ARFGAP 1
28976	ACAD9	ACYL-COA DEHYDROGENASE FAMILY MEMBER 9
29785	CYP2S1	CYTOCHROME P450 FAMILY 2 SUBFAMILY S MEMBER 1
30011	SH3KBP1	SH3 DOMAIN CONTAINING KINASE BINDING PROTEIN 1
50807	ASAP1	ARFGAP WITH SH3 DOMAIN, ANKYRIN REPEAT AND PH DOMAIN 1
54576	UGT1A8	UDP GLUCURONOSYLTRANSFERASE FAMILY 1 MEMBER A8
55738	ARFGAP1	ADP RIBOSYLATION FACTOR GTPASE ACTIVATING PROTEIN 1
57732	ZFYVE28	ZINC FINGER FYVE-TYPE CONTAINING 28
64816	CYP3A43	CYTOCHROME P450 FAMILY 3 SUBFAMILY A MEMBER 43
84034	EMILIN2	ELASTIN MICROFIBRIL INTERFACER 2
84171	LOXL4	LYSYL OXIDASE LIKE 4
84467	FBN3	FIBRILLIN 3
90187	EMILIN3	ELASTIN MICROFIBRIL INTERFACER 3
91750	LIN52	LIN-52 DREAM MUVB CORE COMPLEX COMPONENT
112476	PRRT2	PROLINE RICH TRANSMEMBRANE PROTEIN 2
116985	ARAP1	ARFGAP WITH RHOGAP DOMAIN, ANKYRIN REPEAT AND PH DOMAIN 1
132660	LIN54	LIN-54 DREAM MUVB CORE COMPLEX COMPONENT
140838	NANP	N-ACETYLNEURAMINIC ACID PHOSPHATASE
221357	GSTA5	GLUTATHIONE S-TRANSFERASE ALPHA 5
253558	LCLAT1	LYSOCARDIOLIPIN ACYLTRANSFERASE 1

FEMALE MINIMUM LIST		
GENE ID	GENE SYMBOL	GENE NAME
131	ADH7	ALCOHOL DEHYDROGENASE 7 (CLASS IV), MU OR SIGMA POLYPEPTIDE
395	ARHGAP6	RHO GTPASE ACTIVATING PROTEIN 6
650	BMP2	BONE MORPHOGENETIC PROTEIN 2
673	BRAF	B-RAF PROTO-ONCOGENE, SERINE/THREONINE KINASE

867	CBL	CBL PROTO-ONCOGENE
868	CBLB	CBL PROTO-ONCOGENE B
898	CCNE1	CYCLIN E1
1314	COPA	COPI COAT COMPLEX SUBUNIT ALPHA
1435	CSF1	COLONY STIMULATING FACTOR 1
1486	CTBS	CHITOBIASE
1645	AKR1C1	ALDO-KETO REDUCTASE FAMILY 1 MEMBER C1
1793	DOCK1	DEDICATOR OF CYTOKINESIS 1
1869	E2F1	E2F TRANSCRIPTION FACTOR 1
1906	EDN1	ENDOTHELIN 1
2113	ETS1	ETS PROTO-ONCOGENE 1, TRANSCRIPTION FACTOR
2192	FBLN1	FIBULIN 1
2200	FBN1	FIBRILLIN 1
2201	FBN2	FIBRILLIN 2
2250	FGF5	FIBROBLAST GROWTH FACTOR 5
2673	GFPT1	GLUTAMINE--FRUCTOSE-6-PHOSPHATE TRANSAMINASE 1
2767	GNA11	G PROTEIN SUBUNIT ALPHA 11
2822	GPLD1	GLYCOSYLPHOSPHATIDYLINOSITOL SPECIFIC PHOSPHOLIPASE D1
2885	GRB2	GROWTH FACTOR RECEPTOR BOUND PROTEIN 2
3066	HDAC2	HISTONE DEACETYLASE 2
3098	HK1	HEXOKINASE 1
3551	IKBKB	INHIBITOR OF NUCLEAR FACTOR KAPPA B KINASE SUBUNIT BETA
4053	LTBP2	LATENT TRANSFORMING GROWTH FACTOR BETA BINDING PROTEIN 2
4149	MAX	MYC ASSOCIATED FACTOR X
4237	MFAP2	MICROFIBRIL ASSOCIATED PROTEIN 2
5111	PCNA	PROLIFERATING CELL NUCLEAR ANTIGEN
5518	PPP2R1A	PROTEIN PHOSPHATASE 2 SCAFFOLD SUBUNIT AALPHA
5578	PRKCA	PROTEIN KINASE C ALPHA
5579	PRKCB	PROTEIN KINASE C BETA
5879	RAC1	RAC FAMILY SMALL GTPASE 1
5933	RBL1	RB TRANSCRIPTIONAL COREPRESSOR LIKE 1
5934	RBL2	RB TRANSCRIPTIONAL COREPRESSOR LIKE 2
6195	RPS6KA1	RIBOSOMAL PROTEIN S6 KINASE A1
6453	ITSN1	INTERSECTIN 1
6654	SOS1	SOS RAS/RAC GUANINE NUCLEOTIDE EXCHANGE FACTOR 1
8038	ADAM12	ADAM METALLOPEPTIDASE DOMAIN 12
8074	FGF23	FIBROBLAST GROWTH FACTOR 23
8516	ITGA8	INTEGRIN SUBUNIT ALPHA 8
8729	GBF1	GOLGI BREFELDIN A RESISTANT GUANINE NUCLEOTIDE EXCHANGE FACTOR 1
8822	FGF17	FIBROBLAST GROWTH FACTOR 17

9101	USP8	UBIQUITIN SPECIFIC PEPTIDASE 8
9107	MTMR6	MYOTUBULARIN RELATED PROTEIN 6
9267	CYTH1	CYTOHESIN 1
9446	GSTO1	GLUTATHIONE S-TRANSFERASE OMEGA 1
9564	BCAR1	BCAR1 SCAFFOLD PROTEIN, CAS FAMILY MEMBER
10253	SPRY2	SPROUTY RTK SIGNALING ANTAGONIST 2
10617	STAMBP	STAM BINDING PROTEIN
10818	FRS2	FIBROBLAST GROWTH FACTOR RECEPTOR SUBSTRATE 2
11015	KDELR3	KDEL ENDOPLASMIC RETICULUM PROTEIN RETENTION RECEPTOR 3
25942	SIN3A	SIN3 TRANSCRIPTION REGULATOR FAMILY MEMBER A
27159	CHIA	CHITINASE ACIDIC
28976	ACAD9	ACYL-COA DEHYDROGENASE FAMILY MEMBER 9
57732	ZFYVE28	ZINC FINGER FYVE-TYPE CONTAINING 28
84034	EMILIN2	ELASTIN MICROFIBRIL INTERFACER 2
84467	FBN3	FIBRILLIN 3
116985	ARAP1	ARFGAP WITH RHOGAP DOMAIN, ANKYRIN REPEAT AND PH DOMAIN 1
140838	NANP	N-ACETYLNEURAMINIC ACID PHOSPHATASE
253558	LCLAT1	LYSOCARDIOLIPIN ACYLTRANSFERASE 1