Gene Set Analyses Under a Bayesian Partial Factor Regression Model

May 27, 2016

Abstract

In recent years, there has been a drive to study the genetic etiology of diseases. Many statistical models, centered around regression techniques, have been proposed in order to find genes or groups of genes that affect diseases; however, these models often ignore many biological facts, such as the effect of the interdependence of genes and the role genes play in various biological pathways. Additionally, many of these models, such as genome wide association studies, may take as long as several months to complete and have problems with predictor selection. In this paper, we propose the use of a Bayesian partial factor regression model in order to address the covariance structure of the genes and gene sets, aid in variable selection and decrease computation time. Through the use of this model, we find a decrease in computational expense, reproduce genetic results in two separate real data examples, and suggest a novel biological interpretation that provides unique insight into the genetic etiology of many diseases.

Introduction

With the decreasing cost of sequencing genomic data, it comes as no surprise that the demand for statistical methods that accurately model our biological understanding of genetics has increased. Namely, gene set analyses have become progressively important due to the biological understanding that complex phenotypes are jointly influenced by multiple genes [3]. As the goal of many biological studies is to identify groups of genes with a common biological function that explain a given phenotype, many statistical models have arisen that attempt to quantify the relationship between these predefined gene sets and the phenotype. In the literature, these are often referred to as deriving a "molecular signature." There are many problems with deriving a molecular signature, such as its instability, thus motivating further statistical research in developing more complete methods of analyzing gene sets [2]. Additionally, many current gene set approaches do not consider the interdependence of pathways because they only test one gene at a time, such as gene set analysis (GSA) [4].

Recent works have shown that using predefined gene sets often improves both statistical power and interpretability of the model [1, 5, 6]. Many of these works incorporate the use of Bayesian factor models, which allow for analyses in a reduced dimension setting yet retaining the ability to make inferences on the original predictors. In this paper, we implement a Bayesian partial factor regression model that identifies functionally relevant enriched gene sets and genes in cancer data and we discuss this model's statistical and biological relevance.

In section 2, we evaluate the ability of the partial factor regression model to identify genes and pathways associated with two types of cancer: melanoma and colon. In section 3, we discuss the statistical relevance of the partial factor model in gene set analyses and the novel biological interpretability of the partial factor model. We also comment on the limitations of the partial factor regression model and describe potential avenues for future work. In section 4, we outline specifically the partial factor regression model.

Results

Partial Factor Method Overview

We begin with a summary of the Bayesian partial factor model (PFRM). First, it is important to recognize that this model is best understood to be a combination of two models: a linear regression model for the response variable, Y, and a marginal model for the predictor variables, X. We can model Y as:

$$Y = X\beta + \epsilon, \epsilon \sim MVN_n(0, \sigma^2 I_n). \tag{1}$$

Here, Y is an n-dimensional vector of test statistics (p-values may be used as well) derived from genetic signatures from n genes. X is an $n \times p$ incidence matrix, where n is the number of genes and p is the number of pathways being considered, with entries of 1 if the gene is in the pathway or 0 otherwise. β is a p-dimensional vector of effect sizes for the pathways being analyzed. ϵ is an n-dimensional vector of idiosyncratic noise with a variance component of σ^2 modeled by an n-dimensional multivariate normal distribution with mean μ and variance Σ: $MVN_n(\mu, \Sigma)$. For our marginal model for the predictor variables, we have:

$$X^T = BF + \nu, \nu \sim MVN_n(0, \Psi I_n), \tag{2}$$

where X remains the $n \times p$ incidence matrix of genes in the pathways; B is a $p \times k$ dimensional matrix of the p pathways in the k factors, commonly known as the factor loadings matrix; and F is a $k \times n$ dimensional matrix of the n genes in the k factors. By combining Equations 1 and 2, we see that the regression model can be written as

$$Y = F\theta + \epsilon, \tag{3}$$

where θ is a $1 \times k$ vector of effect sizes for the factors.

There are many advantages to using a Bayesian partial factor model in gene set analyses, both from a statistical and a biological viewpoint. From a statistical standpoint, there is a unique equation relating the lower dimensional coefficient parameter to the original high dimensional coefficient parameter, which allows for biological interpretations on original predictors. Additionally, through implementing principal component analysis (PCA), the dimensionality on X is reduced, thus resulting in a decrease in computational expense and no collinearity. See Appendix A to see the computational efficiency of this model. Another important consequence to underscore is that when using a latent factor regression model, y_i is conditionally independent of x_i given f_i.

From a biological viewpoint, we find a novel interpretation of the partial factor model in a genetics setting. One way in which the factors can be viewed are as meta-pathways, meaning groups of gene sets that appear to be functionally related to the phenotype, or response. Looking at Equations 2 and 3, we can see that there are several interpretations B and F can take. B, the factor loadings matrix with dimensions $p \times k$, can be interpreted as the strength of pathway membership in the meta-pathways or the probability of pathway membership in the meta-pathways. F, the factor matrix with dimensions $k \times n$, can be interpreted as the strength of gene membership in the meta-pathways or the probability of gene membership in the meta-pathways. In this way, the partial factor model provides a novel avenue where it can be seen which collection of pathways and genes most influence the phenotype. This unique interpretation may allow for unprecedented insight into many biological problems, especially in genetics.

We implemented our partial factor model on two real data sets that have well known genetic signatures and established pathways: a $BRAF^{V600}$-mutant melanoma study and a colon-APC loss cancer study. Each data set contained genetic signatures of the enriched genes from each study. The test statistic derived from the genetic signatures was treated as the response variable in each analysis. All gene sets were downloaded from the MSigDB website from the C2 KEGG and C2 REACTOME sets. For each example, a binary design matrix based on these genetic signatures was created, where a value of 1 was given if the gene appeared in the gene set and 0 otherwise. The final predictor matrix was completed when all of the pathways containing no enriched genes and the enriched genes that were not in any pathways were removed. This matrix was then scaled so that the genes in each pathway were centered.

The gene set and number of factors used in the analysis were selected so that theta was identifiable, meaning that it converged to its posterior distribution. This was determined by looking at the trace plots of theta. If more than one gene set and factor combination resulted in posterior estimates on theta, then the one with the largest number of pathways in the gene set and the largest number of factors was used. In this paper, the significant genes and pathways were determined subjectively, due to time constraints [26]. In the following subsections, we will describe each example in more detail in addition to the model in context and subsequent results.

Melanoma Example

Microarray gene expression data was obtained on $BRAF^{V600}$-mutant melanoma metastatic samples from NCBI's Gene Expression Omnibus. After preprocessing steps to ensure data quality, there was a total of 68 samples and 11,657 genes. In order to find the enriched genes for MAPK pathway addiction, a genetic signature was derived using the logistic version of the Bayesian approximate kernel regression (BAKR-logit) model, which yielded 68 enriched genes. We used the KEGG gene set, freely available from the MSigDB website, in this analysis. The number of pathways and genes that were used, after removing the pathways that contained no enriched genes and the enriched genes that were not in any pathways, were 61 and 29, respectfully. The number of factors, or meta-pathways, was set equal to 2. We report the pathways and genes with posterior means greater than 0.3 and 1 as the elements that most influenced melanoma. In both meta-pathways, 6 genes and 8 gene sets were found to be most influential. PFRM replicated findings of 2 genes known to be associated with melanoma, *MYC* and *CCND1*, and found several pathways related to various cancers among the most influential pathways.

The estimated factors and factor loadings that are above our threshold are provided in Tables 1 and 2, respectfully. See Appendix B for a complete list of values for each pathway and gene in each meta-pathway. The most interesting result was that the melanoma pathway and the MAPK pathway did not have large posterior values. The melanoma pathway had posterior values of 0.183 (SD = 0.113) and -0.006 (SD = 0.081) for the first and second meta-pathways, respectfully. The MAPK pathway had posterior values of 0.017 (SD = 0.075) and -0.019 (SD = 0.075) for the first and second meta-pathways, respectfully. Additionally, Figures 1 and 2 have been provided that depict the posterior means for the strength of association of the genes and pathways in the melanoma meta-pathways. Each figure provides a visual image of how PFRM is able to select variables as there are clear groups of association.

Table 1: The genes recognized as having the largest posterior value in the melanoma example.

Meta-Pathway	Genes	Posterior Value
1	MYC	6.953
	CCND1	5.809
	HLA-DMB	3.215
	HLA-DMA	3.215
	ACTN1	1.031
	VEGFB	1.028
2	HLA - DMA	5.759
	HLA - DMB	5.758
	VEGFB	1.566
	ACTN1	1.565
	POLR3G	1.150
	IRAK1	1.150

Table 2: The pathways recognized as having the largest posterior value in the melanoma example.

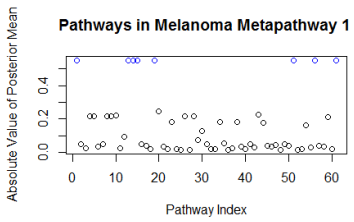
Meta-Pathway	Pathways	Posterior Value
1	Acute Myeloid Leukemia	0.553
	Colorectal Cancer	0.553
	Endometrial Cancer	0.553
	Chronic Myeloid Leukemia	0.553
	Thyroid Cancer	0.553
	Small Cell Lung Cancer	0.553
	WNT Signaling Pathway	0.553
	Cell Cycle	0.553
2	Type 1 Diabetes Mellitus	0.493
	Autoimmune Thyroid Disease	0.493
	Intestinal Immune Netowrk for IGA Production	0.493
	Asthma	0.493
	Graft vs. Host Disease	0.493
	Antigen Processing and Presentation	0.493
	Allograft Rejection	0.493
	Viral Myocarditis	0.448

One of the motivating factors on which gene set was used and the number of factors chosen was due to the proper mixing of θ. It is important that θ converges because that shows the model reached convergence at the posterior distribution. As shown in Figure 6 in Appendix B, θ converged. The effect size of the first meta-pathway was 0.215 (SD = 0.300) and the effect size of the second meta-pathway was -0.128 (SD = 0.260).

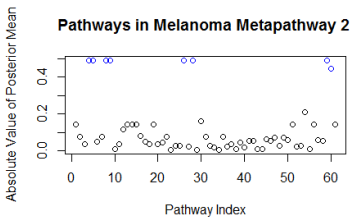
Melanoma is common cancer in the first world, especially in regions with high rates of Caucasian individuals [36]. While it is known that exposure to UV radiation from the sun is a major risk factor, several genetic mutations have been recognized for their role in the progression of melanoma, which makes this a valuable disease to use for testing the effectiveness of PFRM [36]. In the first meta-pathway, two genes previously recognized as important to the etiology of melanoma were found to be influential: *MYC* and *CCND1*. *MYC* is an oncogene that has been linked to many types of cancers, specifically melanoma [39].

MYC is widely regarded as a major player in melanoma due to the detrimental results of its overexpression, namely a continuous and high rate of cell growth and the regulation of angiogenic factors [33] [39]. *CCND1*, also an oncogene, is known to play an important role in the cell cycle, where its function results in the cell's entry into the S phase [38]. Any abnormalities during the S phase of the cell cycle may cause numerous negative consequences for the cell, as it is in the S phase where DNA is replicated and mutations to the DNA are more likely to occur. The replication of these genes are important in that it demonstrates the ability of PFRM to find correct results.

The novel genes found in the first meta-pathway were *HLA-DMA*, *HLA-DMB*, *ACTN1* and *VEGFB*. Both *HLA-DMA* and *HLA-DMB* are part of a class of immune system proteins called major histocompatability complexes (MHC). More specifically, they are part of the MHC class II proteins that help stimulate the immune system to recognize foreign antigens outside of the cell [29, 30]. Until recently, there has not been a significant number of studies that have attempted to link the immune system to cancer, but immuno-oncology is becoming a more prevalent research topic and has even shown promising results in increasing survival for patients with melanoma [32, 40]. Seeing *HLA-DMA* and *HLA-DMB* among these results may provide evidence that further research should be done on the role these genes play in melanoma, which may lead to actual clinical targets for treatment.

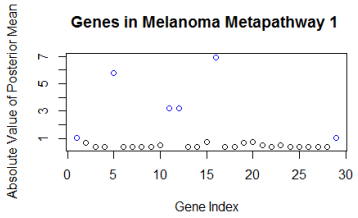


(a) Posterior Means for Pathways in Meta-pathway 1

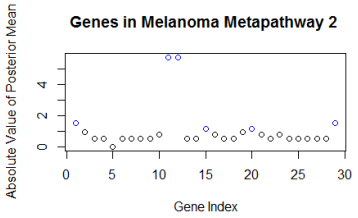


(b) Posterior Means for Pathways in Meta-pathway 2

Figure 1: Graphs of the absolute value of the posterior means for the pathways in each melanoma meta-pathway. The pathways reported as significant are in blue.



(a) Posterior Means for Genes in Meta-pathway 1



(b) Posterior Means for Genes in Meta-pathway 2

Figure 2: Graphs of the absolute value of the posterior means for the genes in each melanoma meta-pathway. The pathways reported as significant are in blue.

ACTN1 encodes for a protein in the cytoplasm that is necessary for proper cell matrix adhesions and organization of the cytoskeleton [18, 28]. When *ACTN1* is phosphorylated, it can interact with *tyrosine-protein kinase Src*, which is a proto-oncogene that plays a key role in cancer through its determination of many fundamental cellular processes, such as cell growth, differentiation, and specialized cell signals [34]. Finding *ACTN1* among these results appears to make sense from a biological standpoint and suggests that further research should be done to fully realize the role *ACTN1* plays in melanoma.

The last significant gene was *VEGFB*, which is an important component of endothelial cell growth and survival during vasculogenesis and angiogenesis [8]. Angiogenesis, or the growth of new capillaries from

blood vessels, is especially important for the growth of cancerous cells, as without access to sufficient oxygen and nutrients from the blood, the cancerous cells cannot not survive [7]. Additionally, *VEGFB* can influence the tumor microenvironment by changing the way in which the host organism responds to the tumor [27]. While *VEGFB* has not been specifically linked to melanoma, its role has been explored in other types of cancer, such as breast cancer [27]. Further research should be completed to explore the function of *VEGFB* in melanoma.

In the second meta-pathway, four of the six genes overlapped with genes from the first meta-pathway. These genes were *HLA-DMA*, *HLA-DMB*, *VEGFB* and *ACTN1*. It is interesting to note that *HLA-DMA*, *HLA-DMB* and *VEGFB* were reproduced in this second meta-pathway, especially since there is limited knowledge about their role in melanoma. This provides further motivation to explore the roles of these three genes in melanoma and should be a focus of future research. The last two genes found to be influential in the second meta-pathway were *POLR3G* and *IRAK1*. *POLR3G* is a gene that codes for one type of RNA polymerase III, which transcribes genes that control the cell cycle and growth. Moreover, *MYC* binds to the *POLR3G* promoter, thus suggesting that *POLR3G* may play a role in the progression of melanoma due to this relationship with *MYC* [35]. *IRAK1* has previously been linked with melanoma through its effects on activating proteins involved in cell division and survival, most notably the p38 MAPK pathway [37] [19]. Reproducing results in which *MYC* and *IRAK1* are important in melanoma is important in that it provides evidence that our model is identifying genes truly associated with melanoma and not only false positives.

While PFRM recognized several well-known genes as most important in the meta-pathways, PFRM failed to identify two key pathways in melanoma: the melanoma pathway and the MAPK signaling pathway. This is unexpected, as *MYC* plays a central role in both pathways. One reason PFRM may not have uncovered these pathways is because *MYC* has a stronger representation in other pathways considered, such as those found to be most prevalent in the first meta-pathway: acute myeloid lukemia, colorectal cancer, endometrial cancer, chronic myeloid leukemia, thyroid cancer, small cell lung cancer and WNT signaling pathways. While we did not necessarily expect to see these specific cancer pathways, we are not surprised as many cancers share common features [7]. Additionally, the cell cycle pathway, which encodes key regulators of mitosis checkpoints, was found to be significant. We expected to see this pathway as it is well known that unregulated cell division contributes to cancer.

In the second meta-pathway, we found that the pathways with the highest posterior means were associated with the immune system: type 1 diabetes mellitus, autoimmune thyroid disease, intestinal immune network for IGA production, asthma, graft vs. host disease, antigen processing and presentation, allograft rejection, and viral myocarditis pathways. While not initially intuitive, these pathways coincide with the most significant genes found in the second meta-pathway. The prevalence of pathways related to the immune system in the second meta-pathway suggests that the immune system contributes to melanoma. This provides a new area for clinical intervention, as the immune system can be targeted to help combat cell abnormalities. Further research should be done on the intersection of the immune system and melanoma.

PFRM has demonstrated its strength in identifying several genes that are known to contribute to melanoma; however, there is significant ambiguity at the ability of PFRM to recognize pathways that affect melanoma. We examine this problem from a statistical perspective in the discussion.

Colon Cancer Example

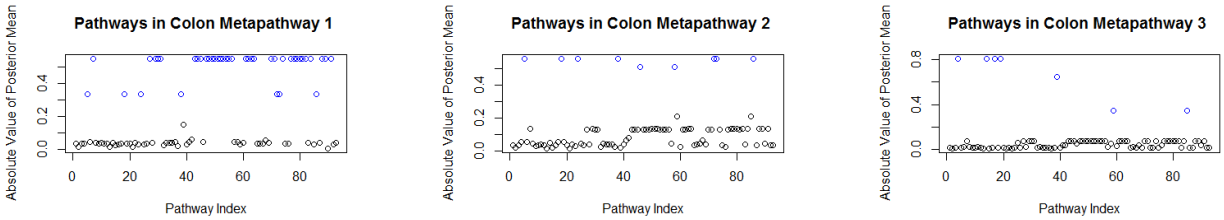
The levels of three gene expression data of matched colon cancer tumor samples was taken from The Cancer Genome Atlas (TCGA). Since this data had already been preprocessed and mapped to the gene level, we labeled each sample by the presence of the *APC* gene and derived the genetic signature using the BAKR-logit model. 49 genes were found to be enriched. The combination of the KEGG and REACTOME gene sets was used in this analysis. 93 pathways and 21 genes were included after removing the pathways that contained no enriched genes and the enriched genes that were not in any pathways. The number of factors, or meta-pathways, was set equal to 3. We report the pathways and genes with posterior means larger than 0.3 and 2, respectfully, as the most influential elements (see Appendix C for a complete list of

pathways and genes in each meta-pathway). In the first meta-pathway, 42 pathways and 2 genes were found to be most influential while in the second meta-pathway, 9 pathways and 2 genes were found to be most significant. In the third meta-pathway, 7 pathways and 1 gene were above the threshold. PFRM identified the driving gene of colon cancer, *WNT2*, and several potential new pathways that may play a role in colon cancer.

The estimated factors and factor loadings that are above our threshold are provided in Tables 3 and 4, respectfully. It should be noted that only the first ten pathways for the first meta-pathway are listed in Table 4 due to space constraints. Of special note is the absence of the WNT pathway, which is known to contribute significantly to colon cancer. In the first meta-pathway, its posterior mean was -0.032 (SD = 0.093); in the second, its posterior mean was -0.044 (SD = 0.106); and in the third, its posterior mean was 0.054 (SD = 0.127). Additionally, Figures 3 and 4 show a graph of the posterior values of each pathway and gene in each meta-pathway, thus providing a good visual to see the groupings of elements PFRM identified as important in colon cancer. Just as in the melanoma example, the gene set and the number of factors was chosen so that θ was able to converge to its posterior distribution (Figure 7 in Appendix C). The effect size of theta in the first meta-pathway was -0.172 (SD = 0.401), the second meta-pathway was -0.325 (SD = 0.242) and the third meta-pathway was -0.118 (SD = 0.359).

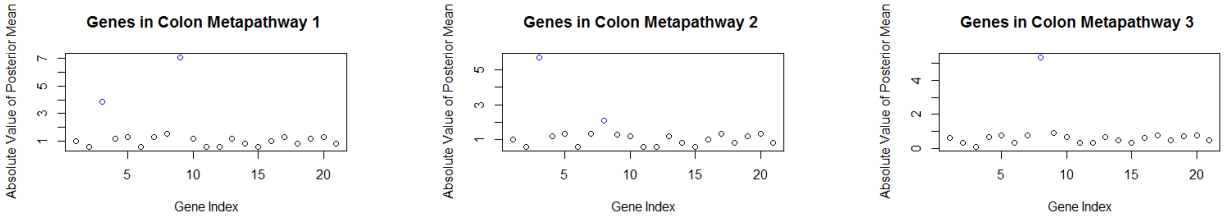
Table 3: The genes recognized as having the largest posterior value in the colon cancer example.

Meta-Pathway	Genes	Posterior Value
1	GNG7	7.090
	ADORA2A	-3.879
2	ADORA2A	3.215
	WNT2	5.759
3	WNT2	5.758



(a) Posterior Means for Pathways in Metapathway 1 (b) Posterior Means for Pathways in Metapathway 2 (c) Posterior Means for Pathways in Metapathway 3

Figure 3: Graphs of the absolute value of the posterior means for the pathways in each colon meta-pathway. The pathways reported as significant are in blue.



(a) Posterior Means for Genes in Metapathway 1 (b) Posterior Means for Genes in Metapathway 2 (c) Posterior Means for Genes in Metapathway 3

Figure 4: Graphs of the absolute value of the posterior means for the genes in each colon cancer meta-pathway. The pathways reported as significant are in blue.

PFRM identified several genes in this data set that have not previously been found linked to colon cancer. In the first meta-pathway, the two genes that were identified as most influential were *GNG7* and *ADORA2A*.

Table 4: The pathways recognized as having the largest posterior value in the colon cancer example.

Meta-Pathway	Pathway	Posterior Value
1	REACTOME_NEUROTRANSMITTER_RECEPTOR_BINDING_AND_DOWNSTREAM_TRANSMISSION_IN_THE_POSTSYNAPTIC_CELL	0.550
	REACTOME_G_ALPHA_Q_SIGNALLING_EVENTS	0.550
	REACTOME_GABA_RECEPTOR_ACTIVATION	0.550
	REACTOME_GLUCAGON_TYPE_LIGAND_RECEPTORS	0.550
	REACTOME_INHIBITION_OF_VOLTAGE_GATED_CA2_CHANNELS_VIA_GBETA_GAMMA_SUBUNITS	0.550
	REACTOME_NEURONAL_SYSTEM	0.550
	REACTOME_POTASSIUM_CHANNELS	0.550
	REACTOME_ADP_SIGNALLING_THROUGH_P2RY1	0.550
	REACTOME_TRANSMISSION_ACROSS_CHEMICAL_SYNAPSES	0.550
	REACTOME_ADP_SIGNALLING_THROUGH_P2RY12	0.550
2	REACTOME_NUCLEOTIDE_LIKE_PURINERGIC_RECEPTORS	0.561
	KEGG_CALCIUM_SIGNALING_PATHWAY	0.561
	REACTOME_CLASS_A1_RHODOPSIN_LIKE_RECEPTORS	0.561
	KEGG_NEUROACTIVE_LIGAND_RECEPTOR_INTERACTION	0.561
	KEGG_VASCULAR_SMOOTH_MUSCLE_CONTRACTION	0.560
	REACTOME_NGF_SIGNALLING_VIA_TRKA_FROM_THE_PLASMA_MEMBRANE	0.560
	REACTOME_SIGNALLING_BY_NGF	0.560
	REACTOME_GPCR_DOWNSTREAM_SIGNALING	0.510
	REACTOME_G_ALPHA_S_SIGNALLING_EVENTS	0.509
3	KEGG_BASAL_CELL_CARCINOMA	0.813
	KEGG_HEDGEHOG_SIGNALING_PATHWAY	0.813
	KEGG_PATHWAYS_IN_CANCER	0.813
	KEGG_MELANOGENESIS	0.813
	REACTOME_CLASS_B_2_SECRETIN_FAMILY_RECEPTORS	0.646
	REACTOME_SIGNALING_BY_GPCR	0.349
	REACTOME_GPCR_LIGAND_BINDING	0.346

GNG7 is part of a group of large G-proteins that may function to suppress tumors and halt cell growth [41]. The role of *GNG7* in colon cancer remains unclear, however it is known to to be downregulated in gastrointestinal cancer, oseophageal cancer, and carcinoma [41, 42]. These three types of cancer are closely related to colon cancer, as the colon is simply the end of the digestive tract and an internal organ. Thus, it is not completely unexpected to find the presence of *GNG7* in this first meta-pathway. Future research should focus on the role of *GNG7* in colon cancer, especially as it is known to contribute to similar cancers. *ADORA2A* functions to increase intracellular cAMP levels, which aids many biological functions, such as blood flow in

the kidneys, heart and cerebrum, a component in the immune system and a contributor to pain regulation [20]. Recent research has shown it to be a potential site of clinical intervention due to its ability to slow the growth of tumors via the immune system. In tumors, there is a high rate of cell death and dying cells release the nucleotide adenosine which binds to *ADORA2A*. When adenosine is bound to *ADORA2A*, T cells of the immune system become T_{Reg} cells, which then stimulates an inflammatory response to the cancerous cells [43]. This is an unexpected, yet potentially insightful, result that warrants further research.

In the second meta-pathway, *ADORA2A* is again found to be influential. However, a more notable gene found to have a large posterior mean was *WNT2*. *WNT2* is widely known for its role in cell proliferation and differentiation in colon cancer [44–46]. It is also the sole influential gene in the third meta-pathway. The fact that we see *WNT2* among our results provides us with evidence that our model successfully identifies genes that relate to colon cancer. This helps to provide evidence that PFRM is able to correctly describe the genetic foundations of diseases.

While PFRM correctly identified *WNT2* as a significant contributing gene in colon cancer, there were several unexpected pathways that were found to be influential in colon cancer in each meta-pathway. Notably, the WNT pathway did not have a posterior mean larger than 0.3 in any of the meta-pathways. This is interesting to note, especially since the *WNT2* gene, which is a major contributor to the WNT pathway, had high posterior means in two of the meta-pathways. Similar to the melanoma example, the *WNT2* gene may play a dominating role in other pathways, which may overshadow the importance of the WNT pathway.

Results found to be larger than 0.3 in the first meta-pathway included many pathways that pertain to metabolism and chemical signaling, such as the GABA receptor activation, the glucagon signaling in metabolic regulation and the potassium channels pathways. Similarly, in the second meta-pathway the pathways with the highest posterior means mostly related to nerve growth factor pathways, nucleotide receptor pathways and signalling pathways. For example, the calcium signaling pathway, the signalling by nerve growth factor pathway and the neuroactive ligand receptor interaction pathway were all found to be influential. While these results were not expected, they are mostly consistent with the known functions of *ADORA2A*. This suggests that *ADORA2A* is a major component in a diverse set of pathways that pertain to many biological functions while it remains absent in known cancer pathways. As there has been limited previous research on the role of *ADORA2A* in cancer, more research should be completed to explore this role.

In the third meta-pathway, there were three pathways that related to cancer: melanogenesis, basal cell carcinoma and the overall pathways in cancer from KEGG. The first two pathways are not necessarily consistent with known pathways in colon cancer, which may suggest that these pathways contain features that are consistent with many types of cancer. Additionally, there were two other pathways with posterior means larger than 0.3 that were related to cell growth and differentiation: the hedgehog signaling pathway and the class B2 secretin family receptors pathway. These two pathways were not expected to be seen, which may show that the *WNT2* gene is a significant contributor to these pathways due to *WNT2*'s role in the cell cycle. Further research should be done into the overlap of these pathways and the role of *WNT2* in each pathway.

While PFRM successfully identified *WNT2* as a contributor to colon cancer, PFRM failed to recognize many of the canonical pathways of colon cancer, such as the WNT pathway. This is a cause for concern, as it is unclear if these pathways are truly connected to colon cancer or are simply false positives. We consider this concern from a statistical viewpoint in the following section, as there are several modeling considerations that need to be addressed.

Discussion

We have proposed a method for gene set analyses that is computationally efficient, is decent at variable selection and provides insightful biological interpretations. The statistical strengths of this model stem from the dimensionality reduction that occurs when finding and using the factors in the regression. While we work in this lower dimensional space, we retain the ability to make inferences on the original parameters.

Additionally, this dimensionality reduction allows for a decrease in computational expense and burden (See Appendix A for further details). Furthermore, the fact that y_i is conditionally independent of x_i given f_i, is non-trivial. Through this fact, we are better able to infer the genetic architecture of the phenotype we are examining. By using the partial factor regression model, instead of a factor regression model, we are better able to select the variables that most contribute to the phenotype.

Biologically, we have found a statistical model with a unique interpretation of collections of genes and pathways. We have examined the idea of meta-pathways, or groups of both genes and pathways that appear to be functionally related to the phenotype beyond biological experiments. This is important because having an unbiased method of relating groups of genes and previously known pathways to phenotypes in new ways may decrease experimental bias and allow for more accurate and precise gene set analyses. In this paper, we have focused on demonstrating the strengths of PFRM using two real life data sets with widely accepted results. We have replicated several of these expected results as well as found new genes and pathways that may influence melanoma and colon cancer.

While PFRM has addressed several known challenges that are common to gene set analyses and can be a valuable tool to many researchers, it is not without its limitations. First, one major assumption is that there is a linear relationship between Y and X. In a genetics context, we cannot necessarily assume that this assumption is met due to our current biological knowledge, especially due to the effects of epistasis. As such, a model that considers and captures non-linear effects is highly desired. An extension of PFRM would be to consider using a kernel function to model these types of relationships between genes and the given phenotype. The use of kernels in statistics to model non-linear processes is not uncommon and its use has become increasingly important in the machine learning community [25]. One way in which kernel models could be incorporated into the PFRM framework is by transforming $F\theta$ from Equation 3 using a kernel, resulting in $Y = K\alpha + \epsilon$, where K is the kernel and α is the coefficients of the kernel. In this way, non-linear effects can be captured.

While using kernels would allow us to consider the non-linear case and still work in a reduced dimension, there are several shortcomings. One drawback of using a kernel model is that the effect size of each explanatory variable is lost because there are no methods that allow the kernel coefficients, α_i, to be related to the original variable coefficients, β. This negatively impacts the biological inferences we can make since we do not have posterior conclusions on the original explanatory variables, only on the factors. Furthermore, there is no way to perform variable selection. This is problematic for several reasons, most notably because it produces additional challenges for subsequent biological research to be completed if there is uncertainty in which predictors influence the response and the extent of this influence. Several of the drawbacks of kernel models have been addressed by the Bayesian approximate kernel regression (BAKR) model.

The main result stemming from BAKR is an analog for the effect size of each explanatory variable when the kernel is shift-invariant [12, 17]. The main link between the coefficients of the original explanatory variables and the coefficients of the factors in the kernel space is known as the basis function. While, this function is very difficult to compute in practice, an approximate basis function is relatively straightforward to compute. Through utilizing this approximate basis function, the relationship between the kernel factor effects, α, and the original variable effects, β can be defined through the feature regression coefficients in the approximate basis function. This allows us to move from the kernel space back to the original space without a loss of information. BAKR combines the empirical factor model's method of mapping back to the original explanatory variables in the context of using a kernel model, thus allowing for non-linearity.

While BAKR allows us to move to the original space from the kernel space and allows us to assume non-linearity, PFRM is a hierarchical model that is centered at the Bayesian factor model and conditioned on a fixed number of factors. Fixing the number of factors aids in computational efficiency, however, choosing the number of factors is a serious challenge. If k is chosen to be too large, convergence of θ may not be reached as the model cannot distinguish between each factor. Contrarily, if k is chosen to be too small, inferences may be inaccurate [13]. It would be ideal to allow k to be a variable in a MCMC model; however, this would most likely create numerous computational issues such as increasing the time and computational power necessary

to have the model run. Additionally, the parameters may not mix well due to the variation in the dimensions of k and the starting values for the parameters may be influential in the posterior summaries. This idea of letting the number of factors remain unknown was approached in a paper by Lopes and West in which they utilize a reversible jump MCMC algorithm [15].

Another limitation of PFRM is that it assumes the genes in these pathways taken from MSigDB represent the true membership. While we have great faith in the scientific validity in the experiments that led to the creation of these gene sets, we recognize the fact that these gene sets were found under specific experimental conditions and designs. As such, they are subject to experimental and human error. To our knowledge, there are no methods that explicitly address this issue. While further research should be done to overcome this assumption, it should be noted that the meaning of the factors in PFRM hints at the uncertainty of pathway membership through the interpretation of both B and F: each factor represents a collection of pathways and genes that share common biological aspects with the phenotype. In this way, the factors can be thought of as proposed true meta-pathways for the phenotype.

Lastly, there were situations in which θ did not converge. As such, it is important to recognize that PFRM does not always recognize each factor as being unique. This may be due to the presence of a few genes that are overwhelmingly important in the phenotype, such as MYC and melanoma. These genes strongly influence each factor, which then results in multiple unfixed pathways where gene members are the same or nearly the same. Another reason why θ did not converge may be a result of a lack of variation in the incidence matrix; as such, there is no variation in the factors. These situations will cause theta to be unidentifiable thus preventing posterior inferences on theta to be drawn. Consequently, there must be variation in the incidence matrix of genes and pathways, effectively ensuring that each pathway is unique. Additionally, posterior inferences on theta may not be consistent, as using different sets of pathways, even if each set has significant overlap, may change the results. As such, the MCMC starting value is important for obtaining convergence and posterior summaries.

Further research should be completed on the constraints of this model especially in situations where PFRM is used for inference, as it is unclear as to which cases are completely identifiable and which cases are not identifiable. Another consequence of this identifiably problem is that determining significance becomes an issue. For example, we attempted to use two different imputation methods, the family wise error rate (FWER) and the false discovery rate (FDR), to determine which genes and pathways were significant. However, in nearly all of the imputed data sets, θ was unidentifiable and no conclusive results could be drawn. As such, restricting the parameters of PFRM would allow these significance tests to be used, thus avoiding the case where the thresholds are subjectively chosen. As stated above, due to time constraints we chose the threshold for B and F subjectively. If this was not the case, we would have implemented a bootstrap procedure and gained a better understanding of the distribution of our test statistics.

Methods

For many years, factor models have been studied extensively in several fields of application, notably gene set analyses [1, 3, 5]. They have become popular due to their computational efficiency and implementation, especially under a Bayesian framework. Moreover, when working in fields where high dimensional data is prevalent, statistical methods that can successfully predict and select variables are highly desired. Factor models are especially useful for prediction and variable selection because they can be modified to use sparse priors [9]. In this section, we provide details on both Bayesian factor regression models and the partial factor model.

Factor Regression Models

We begin with a description of a generalized regression model:

$$y = f(x) + \epsilon, \quad \epsilon \sim N(0, \sigma^2 I). \tag{4}$$

Here, y is a vector of responses, $f(x)$ is the link function that relates the linear predictor to the response variable, and ϵ is the idiosyncratic noise vector. In the linear case, $f(x)$ is equal to $X\beta$, where X is an $n \times p$ matrix of predictors and β is a p-dimensional vector of regression coefficients. This model was developed for the case where p < n; however, we are working in the situation where p $\gg$ n, as is true in most gene set analyses. This causes many problems with computation, variable selection, and over-fitting of the model. Moreover, unlike the case where p < n and β can be estimated by its maximum likelihood estimate (MLE): $(X'X)^{-1}X^tY$, when p $\gg$ n, often X^tX is not invertible and multiple MLE's exist.

One way in which these high dimensional problems can be solved is through using Bayesian latent factor models [9]. In these models, the variation in X is separated into two components: latent variables that encode the underlying structure of the predictors and idiosyncratic noise. Specifically,

$$x_i = Bf_i + \nu_i, \quad \nu_i \sim N(\nu_i|0, \Psi^2) \tag{5}$$

$$f_i \sim N(f_i|0, \Delta^2) \tag{6}$$

where x_i is the i^{th} row of X, i=1,...,n, B is the $p \times k$ factor loadings matrix, f_i is the k dimensional vector of latent factors for case i, and ν_i is the idiosyncratic noise vector. There are several constraints on Δ, Ψ and B: Δ and Ψ must be diagonal and B must be a lower triangular matrix. The use of latent variables in the p $\gg$ n setting allows the underlying structure in the predictors to be related directly to the response. Normally, these factors are formed through finding the principal components of X. Since the factors are formed through principal component analysis methods, k $\leq$ n, which means we have reduced our predictor dimensions from p $\gg$ n to k $\leq$ n. Choosing the number of factors, either directly or through learning it from the data, to include in the model is a significant problem that has numerous ramifications if chosen improperly. Traditionally, k is chosen so that both the accuracy of the model fit and the predictive power are maximized. Often, this results in conflicting values of k, where the model fit is unaffected by changing the value of k by one, but the predictive power of the model favors the model with a larger value of k [13].

Under the assumption that the predictors relate directly to the response variable only through the latent factors, we see that:

$$y_i = \theta f_i + \epsilon_i, \quad \epsilon_i \sim N(0, \sigma^2 I) \tag{7}$$

$$cov(X_i, Y_i) = \begin{pmatrix} BB^t + \Psi & V^t \\ V & \xi \end{pmatrix}, \tag{8}$$

$$V = \theta B^t, \tag{9}$$

$$\xi = \sigma^2 + \theta\theta^t. \tag{10}$$

where y_i is the response for case i, θ is a $1 \times k$ row vector of the effect sizes for the factors, ϵ_i is the idiosyncratic noise, and V is the covariation between Y and X. Since the latent variables are derived from the principal components of X, x_i does not directly enter the linear regression and y_i is conditionally independent of x_i given f_i. This is an important fact because it is here that a critical assumption is made: the latent factors provide information about all of the variation in both the predictors and response. A noteworthy consequence of this assumption in gene set analyses is that the variation in the phenotype is most likely not explained solely by the latent factors. Moreover, the incidence matrix is assumed to be error free; however, it can be concluded that the incidence matrix contains experimental and humanistic errors and so assuming that the latent factors completely describes the phenotype may be suboptimal.

A positive aspect of Bayesian factor analysis is that, unlike strict principal component analysis, inferences can be made on the original parameter coefficients as there is a unique equation to relate the k-dimensional regression parameters, θ, to the original p-dimensional parameters, β:

$$\beta = \Psi^{-2}BC\theta \tag{11}$$

$$C^{-1} = \Delta^{-2} + B'\Psi^{-2}B. \tag{12}$$

This property of factor models is very important, especially in biological settings, as it allows for computational efficiency via dimensionality reduction yet it retains the ability to obtain effect sizes on the original predictors.

While Bayesian factor regression models allow for dimensionality reduction and inferences on β, there are numerous issues that need to be addressed. Namely, how to overcome the assumption that the latent factors do not explain all of the variation in the predictors and responses and how to pick the number of factors, k. One way to address several of these issues is by constructing a hierarchical model, centered at the Bayesian factor regression model and conditioned on a fixed number of factors, more commonly known as the Bayesian partial factor model [13]. This model not only addresses these statistical challenges, but also provides a unique and useful biological interpretation in gene set analyses.

Partial Factor Regression Model

The PFRM implements a jointly normal distribution between x_i, y_i, and the latent factors, f_i:

$$Y_i = \theta f_i + [(V - \theta B^t)\Psi^{-1/2}][\Psi^{-1/2}(X_i - Bf_i)] + \epsilon_i. \tag{13}$$

Y_i is the n dimensional vector of responses, θ is the $k \times 1$ vector of effect sizes of the factors, f_i are the k factors, V is the covariance between X and Y, Ψ is the variance of the idiosyncratic noise of the marginal predictor model ($X_i = Bf_i + \nu_i$), B is the $p \times k$ factor loadings matrix, and ϵ_i is the n dimensional idiosyncratic noise vector. The variation of Y_i remains the same between the factor model and PFR, but the change in the expectation of Y_i allows for the covariation between X and Y to change. In a pure factor model, where it is assumed that Y_i is conditionally independent of X_i given f_i and that Y_i linearly depends on the same f_i that encode all of the variation in X_i, V is required to equal θB^t and the relationship between Y_i and X_i can be linear in up to k dimensions. However, in PFRM, V does not need to be restricted to θB^t; Y_i and X_i can now be described in up to p dimensions. An important note is that we can easily return to a factor regression model by setting V equal θB^t.

The prior on V, conditional on θ, B and Ψ is:

$$v_j \sim N(\theta B^t, \omega^2\omega_j^2\psi_j^2), j = 1, ..., p \tag{14}$$

where ω^2 is a global prior variance, w_j^2 is the prior variance for predictor j, and ψ_j^2 is the jth diagonal element of Ψ. This prior implies that the idiosyncratic noise from the latent factors does not affect the regression. Additionally, this prior may be conditioned on Σ_X, which allows for greater flexibility when finding the predictors that are most likely to be predictive of the response. Consequently, the prior for the predictors is less formative than in factor regression models.

Additionally, the PFRM can be adapted to perform variable selection through placing sparse priors on Λ, θ, and B, similar to a spike-and-slap model. This implies that the posterior values of β will be sparse as well. In this way, PFRM places a semi-informative prior on β. This allows the model to account for several cases. First, the case where the response is most strongly and exclusively associated with the least important principal component, commonly known as the least-eigenvalue scenario [21–23], can be acknowledged. Second, for the case where the response depends on predictors that do not have the largest or smallest degree of variation. Third, for the expected cases where the response is most strongly associated with the most important principal component.

By defining $\Lambda = (V - \theta B^t)\Psi^{-1/2}$, the prior on V can be written as $\lambda_j \sim N(0, \omega^2\omega_j^2)$. In this way, a hierarchical model can be constructed that addresses and overcomes the assumption that the latent factors explain the entirety of the variation in both the predictors and the responses. This also centers PFR at the

factor model. Using this parametrization, the complete partial factor model may be defined as

$$X_i|B, f_i, \Psi \sim N(Bf_i, \Psi) \tag{15}$$

$$Y_i|X_i, B, \theta, \Lambda, f_i, \Psi, \sigma^2 \sim N(\theta f_i + \Lambda\Psi^{-1/2}(X_i - Bf_i), \sigma^2) \tag{16}$$

$$\lambda_j \sim N(0, \omega^2\omega_j^2) \tag{17}$$

$$f_i \sim N(0, I_k) \tag{18}$$

$$\theta_h \sim N(0, \tau^2 q_h^2) \tag{19}$$

$$b_{jh} \sim N(0, \tau^2 t_{jh}^2), h = 1, \ldots, k, \tag{20}$$

$$j = 1, \ldots, p. \tag{21}$$

Independent half-Cauchy priors are used for τ and ω as well as for each element in the vectors t, w and q. Using a half-Cauchy prior over t, w and q coincides with the horseshoe priors over B, θ, and Λ, respectfully [24]. The use of a horseshoe-like prior is beneficial in our gene set analysis case due to the sparsity of X. The Strawderman-Berger prior, with density $p(s) \propto s(1 + s^2)^{-3/2}$, is placed on σ and each element in $\Psi^{1/2}$. This model can be implemented using Gibbs Sampling, the details can be found in Appendix A.

Biological Interpretation Revisited

Perhaps one of the most important consequences of using the partial factor regression model in gene set analyses is its biological interpretation. Returning to the two main equations for the partial factor model:

$$Y = F\theta + \epsilon \tag{1}$$

$$X = BF + \nu \tag{2}$$

we can see that there are insightful interpretations for θ, B, and F. Recall that in a genetics context, the factors can be viewed as meta-pathways, defined as collections of the most influential pathways with respect to the phenotype, thus implying that θ can be interpreted as the effect of each meta-pathway on the response variable. This is important for many genetics studies because these meta-pathways may show which gene sets work together in specific biological contexts, which will help stimulate further research into which pathways can and should be targetable via therapy. B, the factor loadings matrix with dimensions $p \times k$, can be interpreted as the strength of pathway membership in the meta-pathways or the probability of pathway membership in the meta-pathways. B provides valuable information on which gene sets should be further tested under controlled lab conditions. F, the factor matrix with dimensions $k \times n$, can be interpreted as the strength of gene membership in the meta-pathways or the probability of gene membership in the meta-pathways. Similar to the practical usage of B, F draws out the most important genes among all the pathways in each meta-pathway. In this way, the information gained from F can be used to further research on the clinical significance of these genes and the molecular contribution of these genes to the phenotype being studied. In this way, the partial factor model provides a novel and useful interpretation in which to see how collections of pathways and genes most influence any given phenotype.

Acknowledgements

KM would like to acknowledge and thank SM and LC for guidance, advice and comments. KM would also like to thank PRH for specific details on R code for the partial factor regression model.

Appendix A: Computational Implementation and Efficiency

Computational Implementation

We utilize MCMC methods, specifically a Gibbs sampler with one Metropolis Hastings update step, in order to sample from the joint posterior distribution. This is taken directly from Hahn, Carvalho and Mukherjee, 2013 [13]. Throughout this section, a dash to the right of the conditioning bar should be read as "everything else."

1. Sample the latent factors, $(F|-)$, using the joint normal distribution of f_i, X_i and Y_i. Draw $f_i \sim N(\mu_i, S)$, where

$$\mu_i = (B^t\theta^t)\Sigma_{X,Y}^{-1}(X_i^tY_i)^t$$
$$S = I_k - (B^t\theta^t)\Sigma_{X,Y}^{-1}(B^t\theta^t)^t.$$

2. Sample variance components. We update all of the variance components using the same general form. We have random variables $r_l, l = 1, ..., m$, where $r_l \sim N(0, s^2)$ and $\pi(s) \propto s^{2a-1}(1+s^2)^{-(a+1)/2}$. When a = 1/2, this form reduces to a half-Cauchy distribution and is a horseshoe prior on r_l. When a = 1, this form is a Strawdeman-Berger prior on r_l. Define $\eta = 1/s^2$. Sample s by performing the following steps:

I. Draw $(u|\eta) \sim Uniform(0, (1+\eta)^{-(a+1/2)})$
II. Draw $(\eta|r, u) \sim Gamma((m+1)/2, \Sigma_{l=1}^m r_i^2/2)$, restricted to be below $u^{-(1/(a+1/2))} - 1$.
III. Set $s = \eta^{-1/2}$.

(a) Sample $(\Psi|-)$. Let m = n and j=1,...,p. Define $r_l(j) = X_{jl} - b_j f_l$.
(b) Sample $(\sigma|-)$. Let m = n. Define $r_l = Y_l - \theta f_l - \Lambda\Psi^{-1/2}(X_l - Bf_l)$.
(c) Sample $(\omega|-)$. Let $\tilde{\Lambda}$ be the vector of nonzero elements of Λ and $\tilde{w}$ be the corresponding elements of w. Then, m is the length of $\tilde{\Lambda}$ and $r_l(j) = \tilde{\lambda}/w_l$.
(d) Sample $(w|-)$. For each w_j, j = 1,...,p, m = 1 and $r = \lambda_j/\omega$.
(e) Sample $(\tau|-)$. Let $\tilde{B}$ represent the vector of nonzero elements of B, θ and $\tilde{t}$ be the corresponding vector of t, q. Then, m is the length of $\tilde{B}$ and $r_l = \tilde{b}/t_l$.
(f) Sample $(t|-)$. For each $t_{j,h}$, j=1,...,p, h=1,...,k, m = 1 and $r = b_{jg}/\tau$.
(g) Sample $(q|-)$. For each q_h, h = 1,...,k, m = 1 and $r = \theta_h/\tau$.
3. Sample the residual regression coefficients, $(\Lambda|-)$. Let $Y_i^* = Y_i - \theta f_i$. and $X_i^* = \Psi^{-1/2}(X_i - Bf_i)$. For each j=1,...,p, let $\tilde{Y_i} = Y_i^* - \Lambda_{-j}X_{-j}^*$ and $\tilde{X_{ij}} = X_{ij}^*$. Draw $\lambda_j \sim N(\mu, s)$ where $\mu = s\tilde{X}\tilde{Y}/\sigma^2$ and $s = (\tilde{X}\tilde{X^{-1}}/\sigma^2 + \omega^{-2}w_j^{-2})^{-1}$. Set $\lambda_j = 0$ with probability

$$\frac{(1-\alpha_\lambda)\phi(0|\mu, s)}{(1-\alpha_\lambda)\phi(0|\mu, s) + \alpha_\lambda\phi(0|0, \omega^2 w_j^2)}$$

$\phi(-|m, s)$ is the normal density function.
4. Sample the factor regression coefficients, $(\theta|-)$. Let $Y_i^* = Y_i - \Lambda\Psi^{-1/2}(X_i - Bf_i)$. For each h = 1,...,k, let $\tilde{Y_i} = Y_i^* - \theta_{-h}f_{-h,i}$. Draw $\theta_h \sim N(\mu, s)$, where $\mu = sf_h\tilde{Y}/\sigma^2$ and $s = (f_h^tf_h/\sigma^2 + \tau^{-2}q_h^{-2})^{-1}$. Set $\theta_h = 0$ with probability

$$\frac{(1-\alpha_\theta)\phi(0|\mu, s)}{(1-\alpha_\theta)\phi(0|\mu, s) + \alpha_\theta\phi(0|0, \tau^2 q_h^2)}$$

5. Sample the factor loadings, $(B|-)$. This is the Metropolis-Hasting update. The proposal distribution is:

$$\pi(B|X) = \frac{f(X|B)\pi(B)}{\int f(X|B)\pi(B)dB}.$$

We sample from this distribution by setting $\tilde{X_{ji}} = X_{ji} - b_{j,-h} f_{-h,i}$ for h=1,...,k and j=1,...,p. Then draw $b_{jh} \sim N(\mu, s)$ where $\mu = s f_h \tilde{X}_j^t / \phi_j^2$ and $s = (f_h^t f_h / \phi_j^2 + \tau^{-2} t_{jh}^{-2})^{-1}$. The rejection probability is:

$$min\left(1, \frac{\Pi_{i=1}^n \phi(Y_i | B^t, X, -)}{\Pi_{i=1}^n \phi(Y_i | B, X, -)}\right).$$

Computational Efficiency

In order to determine the computational efficiency of PFRM, we ran simulations under various settings of genes and gene sets. We simulated sets of 30, 50, 100 and 200 genes and ran them using PFRM against pathways that ranged from 100 to 1500 by 50. As shown in 5, there does not appear to be a drastic change in computation time between these settings until the number of pathways is larger than 800 and the number of genes is larger than 100. Additionally, these simulations represent more extreme cases, meaning that the number of pathways and genes is much larger than the typical gene set analysis using enriched genes. As such, PFRM can be considered to be a more computationally efficient algorithm for gene set analyses.

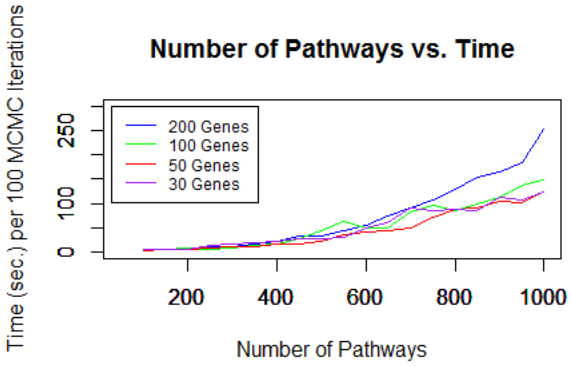
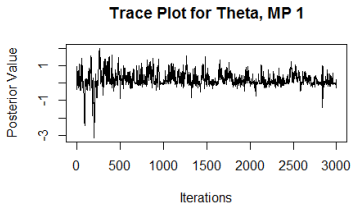
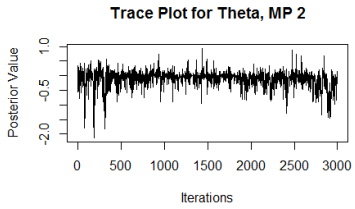


Figure 5: Time (seconds) per 100 MCMC iterations versus the number of pathways for four different numbers of genes. There is not a major difference in time between the number of genes and pathways analyzed until the number of pathways and number of genes increases above 500 and 100, respectfully.

Appendix B: Results from Melanoma Data



(a) Trace Plot for Theta in Metapathway 1



(b) Trace Plot for Theta in Metapathway 2

Figure 6: Plots of the posterior values of theta for each metapathway. It appears as if posterior convergence is reached, as the values do not degenerate to 0 and are centered around a single value. This indicates that Theta is identifiable.

Table 5: Posterior Values for each gene in each meta-pathway for the Melanoma example.

Gene	MP1	MP2
ACTN1	-1.031	-1.565
ARPC5L	-0.651	-0.986
ATG4A	-0.369	-0.562
ATP2B1	-0.37	-0.56
CCND1	5.809	-0.011
CCNG2	-0.37	-0.561
CD33	-0.369	-0.561
DUSP4	-0.37	-0.56
DUSP6	-0.369	-0.56
HCLS1	-0.526	-0.8
HLA.DMA	-3.215	5.759
HLA.DMB	-3.215	5.758
HYAL4	-0.369	-0.561
ID2	-0.369	-0.561
IRAK1	-0.758	-1.15
MYC	6.953	0.831
NPC2	-0.37	-0.561
NR3C1	-0.368	-0.561
PDE1C	-0.651	-0.988
POLR3G	-0.758	-1.15
PPAT	-0.527	-0.8
PRPF4	-0.369	-0.561
SDC3	-0.527	-0.8
SH2B3	-0.368	-0.562
SPRY2	-0.369	-0.562
ST3GAL5	-0.369	-0.561
THOC1	-0.37	-0.561
UGCG	-0.369	-0.561
VEGFB	-1.028	-1.566

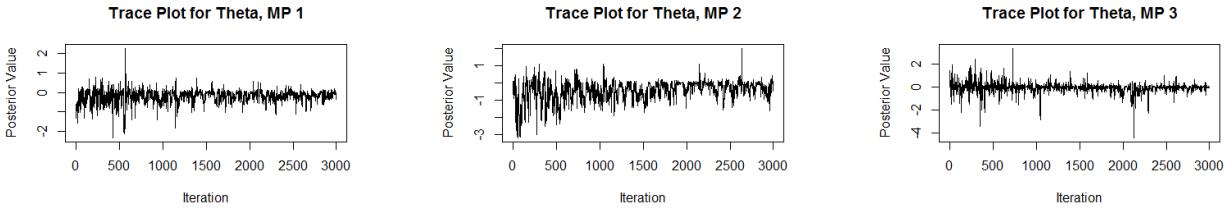
Table 6: Posterior values for the pathways in each meta-pathway in the Melanoma example.

Pathway	MP1	MP2
KEGG_ACUTE_MYELOID_LEUKEMIA	0.553	0.144
KEGG_ADHERENS_JUNCTION	-0.051	-0.077
KEGG_ALANINE_ASPARTATE_ AND_GLUTAMATE_METABOLISM	-0.024	-0.036
KEGG_ALLOGRAFT_REJECTION	-0.214	0.493
KEGG_ANTIGEN_PROCESSING_AND_PRESENTATION	-0.214	0.493
KEGG_APOPTOSIS	-0.032	-0.051
KEGG_ARRHYTHMOGENIC_RIGHT_ VENTRICULAR_CARDIOMYOPATHY_ARVC	-0.05	-0.075
KEGG_ASTHMA	-0.214	0.493
KEGG_AUTOIMMUNE_THYROID_DISEASE	-0.214	0.493
KEGG_BLADDER_CANCER	0.222	-0.009
KEGG_CALCIUM_SIGNALING_PATHWAY	-0.023	-0.037
KEGG_CELL_ADHESION_MOLECULES_CAMS	-0.092	0.117
KEGG_CELL_CYCLE	0.553	0.144
KEGG_CHRONIC_MYELOID_LEUKEMIA	0.553	0.145
KEGG_COLORECTAL_CANCER	0.553	0.144
KEGG_CYTOKINE_CYTOKINE_ RECEPTOR_INTERACTION	-0.048	-0.08
KEGG_CYTOSOLIC_DNA_SENSING_PATHWAY	-0.038	-0.05
KEGG_ECM_RECEPTOR_INTERACTION	-0.021	-0.038
KEGG_ENDOMETRIAL_CANCER	0.553	0.144
KEGG_ERBB_SIGNALING_PATHWAY	0.246	0.034
KEGG_FC_GAMMA_R_MEDIATED_PHAGOCYTOSIS	-0.033	-0.046
KEGG_FOCAL_ADHESION	0.019	-0.078
KEGG_GLIOMA	0.183	-0.007
KEGG_GLYCOSAMINOGLYCAN_DEGRADATION	-0.019	-0.025
KEGG_GLYCOSPHINGOLIPID_ BIOSYNTHESIS_GANGLIO_SERIES	-0.015	-0.027
KEGG_GRAFT_VERSUS_HOST_DISEASE	-0.214	0.493
KEGG_HEMATOPOIETIC_CELL_LINEAGE	-0.016	-0.021
KEGG_INTESTINAL_IMMUNE_ NETWORK_FOR_IGA_PRODUCTION	-0.214	0.493
KEGG_JAK_STAT_SIGNALING_PATHWAY	0.075	-0.004
KEGG_LEISHMANIA_INFECTION	-0.125	0.161

Table 6 continued.

Pathway	MP1	MP2
KEGG_LEUKOCYTE_TRANSENDOTHELIAL_MIGRATION	-0.049	-0.078
KEGG_LYSOSOME	-0.017	-0.026
KEGG_MAPK_SIGNALING_PATHWAY	0.017	-0.019
KEGG_MELANOMA	0.183	-0.006
KEGG_MTOR_SIGNALING_PATHWAY	-0.052	-0.078
KEGG_NEUROACTIVE_LIGAND_RECEPTOR_INTERACTION	-0.014	-0.023
KEGG_NEUROTROPHIN_SIGNALING_PATHWAY	-0.026	-0.037
KEGG_NON_SMALL_CELL_LUNG_CANCER	0.181	-0.009
KEGG_OLFACTORY_TRANSDUCTION	-0.033	-0.047
KEGG_P53_SIGNALING_PATHWAY	0.021	-0.019
KEGG_PANCREATIC_CANCER	0.049	-0.053
KEGG_PATHOGENIC_ESCHERICHIA_COLI_INFECTION	-0.03	-0.052
KEGG_PATHWAYS_IN_CANCER	0.224	-0.009
KEGG_PROSTATE_CANCER	0.176	-0.01
KEGG_PURINE_METABOLISM	-0.039	-0.065
KEGG_PYRIMIDINE_METABOLISM	-0.035	-0.053
KEGG_REGULATION_OF_ACTIN_CYTOSKELETON	-0.044	-0.072
KEGG_REGULATION_OF_AUTOPHAGY	-0.017	-0.025
KEGG_RENAL_CELL_CARCINOMA	-0.048	-0.074
KEGG_RNA_POLYMERASE	-0.037	-0.056
KEGG_SMALL_CELL_LUNG_CANCER	0.553	0.144
KEGG_SPHINGOLIPID_METABOLISM	-0.015	-0.024
KEGG_SPLICEOSOME	-0.018	-0.028
KEGG_SYSTEMIC_LUPUS_ERYTHEMATOSUS	-0.163	0.21
KEGG_TGF_BETA_SIGNALING_PATHWAY	0.03	-0.01
KEGG_THYROID_CANCER	0.553	0.144
KEGG_TIGHT_JUNCTION	-0.037	-0.056
KEGG_TOLL_LIKE_RECEPTOR_SIGNALING_PATHWAY	-0.033	-0.053
KEGG_TYPE_I_DIABETES_MELLITUS	-0.214	0.493
KEGG_VIRAL_MYOCARDITIS	-0.017	0.448
KEGG_WNT_SIGNALING_PATHWAY	0.553	0.144

Appendix C: Results from Colon Data



(a) Trace Plot for Theta in Metapathway 1 (b) Trace Plot for Theta in Metapathway 2 (c) Trace Plot for Theta in Metapathway 3

Figure 7: Plots of the posterior values of theta for each metapathway. It appears as if posterior convergence is reached, as the values do not degenerate to 0 and are centered around a single value. This indicates that Theta is identifiable.

Table 7: Posterior Values for each gene in each meta-pathway for the Colon Cancer example.

Gene	MP1	MP2	MP3
WNT2	-1.556	-2.118	5.367
GNG7	7.090	1.293	-0.905
CDH3	-1.319	-1.354	-0.768
INHBA	-1.319	-1.353	-0.767
TIMP1	-1.320	-1.353	-0.766
LIPC	-1.319	-1.353	-0.766
ABCG2	-1.173	-1.203	-0.682
COL11A1	-1.173	-1.202	-0.681
DAO	-1.174	-1.203	-0.681
SLCO4A1	-1.174	-1.203	-0.681
MMP7	-1.010	-1.037	-0.587
CLDN23	-1.012	-1.035	-0.586
ARNTL2	-0.820	-0.842	-0.478
MMP11	-0.821	-0.840	-0.477
LIFR	-0.821	-0.841	-0.476
SFRP1	-0.577	-0.592	-0.336
NEFM	-0.578	-0.592	-0.336
TNFRSF12A	-0.577	-0.591	-0.336
GCNT2	-0.578	-0.592	-0.336
PRDX6	-0.577	-0.591	-0.334
ADORA2A	-3.879	5.711	0.067

Table 8: Posterior Values for all of the pathways in each meta-pathway for the Colon Cancer example.

Pathway	MP1	MP2	MP3
KEGG_ABC_TRANSPORTERS	-0.037	-0.033	-0.015
KEGG_AMYOTROPHIC_LATERAL_SCLEROSIS_ALS	-0.016	-0.018	-0.006
KEGG_ARGININE_AND_PROLINE_METABOLISM	-0.035	-0.034	-0.012
KEGG_BASAL_CELL_CARCINOMA	-0.038	-0.052	0.813
KEGG_CALCIUM_SIGNALING_PATHWAY	-0.335	0.561	0.014
KEGG_CELL_ADHESION_MOLECULES_CAMS	-0.046	-0.052	-0.022
KEGG_CHEMOKINE_SIGNALING_PATHWAY	0.549	0.131	-0.069
KEGG_CYTOKINE_CYTOKINE_RECEPTOR_INTERACTION	-0.041	-0.046	-0.02
KEGG_ECM_RECEPTOR_INTERACTION	-0.036	-0.029	-0.012
KEGG_FOCAL_ADHESION	-0.04	-0.032	-0.015
KEGG_GLYCEROLIPID_METABOLISM	-0.038	-0.041	-0.02
KEGG_GLYCINE_SERINE_AND_THREONINE_METABOLISM	-0.037	-0.032	-0.01
KEGG_GLYCOSPHINGOLIPID_BIOSYNTHESIS_ LACTO_AND_NEOLACTO_SERIES	-0.017	-0.012	-0.005
KEGG_HEDGEHOG_SIGNALING_PATHWAY	-0.039	-0.051	0.813
KEGG_JAK_STAT_SIGNALING_PATHWAY	-0.025	-0.019	-0.005
KEGG_LEUKOCYTE_TRANSENDOTHELIAL_MIGRATION	-0.032	-0.032	-0.014
KEGG_MELANOGENESIS	-0.038	-0.052	0.813
KEGG_NEUROACTIVE_LIGAND_RECEPTOR_INTERACTION	-0.335	0.561	0.014
KEGG_PATHWAYS_IN_CANCER	-0.038	-0.052	0.813
KEGG_PEROXISOME	-0.038	-0.036	-0.01
KEGG_PHENYLALANINE_METABOLISM	-0.015	-0.015	-0.006
KEGG_TGF_BETA_SIGNALING_PATHWAY	-0.039	-0.037	-0.014
KEGG_TIGHT_JUNCTION	-0.029	-0.027	-0.007
KEGG_VASCULAR_SMOOTH_MUSCLE_CONTRACTION	-0.334	0.56	0.014
KEGG_WNT_SIGNALING_PATHWAY	-0.032	-0.044	0.054
REACTOME_ABACAVIR_TRANSPORT_AND_METABOLISM	-0.034	-0.035	-0.014
REACTOME_ACTIVATION_OF_KAINATE_ RECEPTORS_UPON_GLUTAMATE_BINDING	0.549	0.131	-0.07
REACTOME_ADHERENS_JUNCTIONS_INTERACTIONS	-0.042	-0.038	-0.018
REACTOME_ADP_SIGNALLING_THROUGH_P2RY1	0.55	0.131	-0.07
REACTOME_ADP_SIGNALLING_THROUGH_P2RY12	0.55	0.131	-0.07
REACTOME_AQUAPORIN_MEDIATED_TRANSPORT	0.549	0.131	-0.069
REACTOME_BMAL1_CLOCK_NPAS2_ ACTIVATES_CIRCADIAN_EXPRESSION	-0.026	-0.024	-0.012
REACTOME_CELL_CELL_COMMUNICATION	-0.042	-0.043	-0.017
REACTOME_CELL_CELL_JUNCTION_ORGANIZATION	-0.039	-0.04	-0.016
REACTOME_CELL_JUNCTION_ORGANIZATION	-0.04	-0.04	-0.013
REACTOME_CHYLOMICRON_MEDIATED_LIPID_TRANSPORT	-0.046	-0.036	-0.01
REACTOME_CIRCADIAN_CLOCK	-0.023	-0.021	-0.004
REACTOME_CLASS_A1_RHODOPSIN_LIKE_RECEPTORS	-0.335	0.561	0.013
REACTOME_CLASS_B_2_SECRETIN_FAMILY_RECEPTORS	0.15	-0.016	0.646
REACTOME_COLLAGEN_FORMATION	-0.03	-0.039	-0.013
REACTOME_DEGRADATION_OF_THE_EXTRACELLULAR_MATRIX	-0.048	-0.062	-0.034
REACTOME_EXTRACELLULAR_MATRIX_ORGANIZATION	-0.063	-0.078	-0.032
REACTOME_G_ALPHA_I_SIGNALLING_EVENTS	0.549	0.131	-0.069
REACTOME_G_ALPHA_Q_SIGNALLING_EVENTS	0.55	0.131	-0.069
REACTOME_G_ALPHA_S_SIGNALLING_EVENTS	-0.046	0.509	-0.047
REACTOME_G_ALPHA_Z_SIGNALLING_EVENTS	0.549	0.131	-0.07
REACTOME_G_ALPHA1213_SIGNALLING_EVENTS	0.549	0.131	-0.07

Table 8 continued.

REACTOME_G_BETA_GAMMA_ SIGNALLING_THROUGH_PI3KGAMMA	0.549	0.131	-0.069
REACTOME_G_BETA_GAMMA_ SIGNALLING_THROUGH_PLC_BETA	0.549	0.131	-0.07
REACTOME_G_PROTEIN_ACTIVATION	0.549	0.131	-0.07
REACTOME_G_PROTEIN_BETA_GAMMA_SIGNALLING	0.55	0.131	-0.07
REACTOME_GABA_B_RECEPTOR_ACTIVATION	0.549	0.131	-0.069
REACTOME_GABA_RECEPTOR_ACTIVATION	0.55	0.131	-0.07
REACTOME_GASTRIN_CREB_ SIGNALLING_PATHWAY_VIA_PKC_AND_MAPK	0.549	0.131	-0.069
REACTOME_GLUCAGON_SIGNALING_ IN_METABOLIC_REGULATION	0.549	0.131	-0.07
REACTOME_GLUCAGON_TYPE_LIGAND_RECEPTORS	0.55	0.131	-0.07
REACTOME_GLYCOPROTEIN_HORMONES	-0.043	-0.043	-0.017
REACTOME_GPCR_DOWNSTREAM_SIGNALING	-0.048	0.51	-0.048
REACTOME_GPCR_LIGAND_BINDING	-0.033	0.211	0.346
REACTOME_HEMOSTASIS	0.041	-0.025	-0.031
REACTOME_INHIBITION_OF_INSULIN_ SECRETION_BY_ADRENALINE_NORADRENALINE	0.549	0.131	-0.069
REACTOME_INHIBITION_OF_VOLTAGE_GATED_ CA2_CHANNELS_VIA_GBETA_GAMMA_SUBUNITS	0.55	0.131	-0.069
REACTOME_INTEGRATION_OF_ENERGY_METABOLISM	0.549	0.131	-0.07
REACTOME_INWARDLY_RECTIFYING_K_CHANNELS	0.549	0.131	-0.069
REACTOME_IRON_UPTAKE_AND_TRANSPORT	-0.035	-0.036	-0.016
REACTOME_LIPID_DIGESTION_ MOBILIZATION_AND_TRANSPORT	-0.038	-0.039	-0.017
REACTOME_LIPOPROTEIN_METABOLISM	-0.038	-0.042	-0.016
REACTOME_METABOLISM_OF_ AMINO_ACIDS_AND_DERIVATIVES	-0.053	-0.061	-0.033
REACTOME_METABOLISM_OF_ LIPIDS_AND_LIPOPROTEINS	-0.043	-0.036	-0.011
REACTOME_NEURONAL_SYSTEM	0.55	0.131	-0.07
REACTOME_NEUROTRANSMITTER_ RECEPTOR_BINDING_AND_DOWNSTREAM_ TRANSMISSION_IN_THE_POSTSYNAPTIC_CELL	0.55	0.131	-0.07
REACTOME_NGF_SIGNALLING_VIA_ TRKA_FROM_THE_PLASMA_MEMBRANE	-0.335	0.56	0.014
REACTOME_NUCLEOTIDE_LIKE_PURINERGIC_RECEPTORS	-0.335	0.561	0.014
REACTOME_OPIOID_SIGNALLING	0.549	0.131	-0.07
REACTOME_PEPTIDE_HORMONE_BIOSYNTHESIS	-0.038	-0.036	-0.017
REACTOME_PLATELET_ACTIVATION_ SIGNALING_AND_AGGREGATION	0.037	-0.025	-0.034
REACTOME_PLATELET_HOMEOSTASIS	0.55	0.131	-0.07
REACTOME_POTASSIUM_CHANNELS	0.55	0.131	-0.07

Table 8 continued.

REACTOME_PROSTACYCLIN_SIGNALLING_ THROUGH_PROSTACYCLIN_RECEPTOR	0.549	0.131	-0.07
REACTOME_REGULATION_OF_INSULIN_SECRETION	0.549	0.131	-0.07
REACTOME_REGULATION_OF_INSULIN_ SECRETION_BY_GLUCAGON_LIKE_PEPTIDE1	0.549	0.131	-0.07
REACTOME_REGULATION_OF_WATER_ BALANCE_BY_RENAL_AQUAPORINS	0.549	0.131	-0.07
REACTOME_RESPONSE_TO_ELEVATED_ PLATELET_CYTOSOLIC_CA2_	-0.042	-0.036	-0.015
REACTOME_SIGNAL_AMPLIFICATION	0.55	0.131	-0.07
REACTOME_SIGNALING_BY_GPCR	-0.032	0.208	0.349
REACTOME_SIGNALLING_BY_NGF	-0.335	0.56	0.014
REACTOME_SLC_MEDIATED_TRANSMEMBRANE_TRANSPORT	-0.041	-0.034	-0.013
REACTOME_THROMBIN_SIGNALLING_ THROUGH_PROTEINASE_ACTIVATED_RECEPTORS_PARS	0.549	0.131	-0.07
REACTOME_THROMBOXANE_ SIGNALLING_THROUGH_TP_RECEPTOR	0.55	0.131	-0.07
REACTOME_TRANSMEMBRANE_ TRANSPORT_OF_SMALL_MOLECULES	0.005	-0.043	-0.038
REACTOME_TRANSMISSION_ACROSS_CHEMICAL_SYNAPSES	0.55	0.131	-0.07
REACTOME_TRANSPORT_OF_ORGANIC_ANIONS	-0.033	-0.033	-0.01
REACTOME_TRANSPORT_OF_VITAMINS_ NUCLEOSIDES_AND_RELATED_MOLECULES	-0.039	-0.033	-0.012

References

[1] H. Shen and M. West, Bayesian modeling for biological annotation of gene expression pathway signatures. 2010.

[2] Nilsson, R., Björkegren, J., and Tegnér, J. (2009). On reliable discovery of molecular signatures. *BMC Bioinformatics, 10*(1), 38.

[3] A. Skarman, M. Shariati, L. Jans, L. Jiang, and P. Sorensen. A Bayesian variable selection procedure to rank overlapping gene sets. *BMC Bioinformatics*, **13**:73, 2012.

[4] Fridley, B. L., and Patch, C. (2011). Gene set analysis of SNP data: Benefits, challenges, and future directions. *Eur J Hum Genet European Journal of Human Genetics, 19*(8), 837-843.

[5] B. Shahbaba, R. Tibshirani, C. M. Shachaf, S. K. Plevritis. Bayesian gene set analysis for identifying significant biological pathways. *Journal of the Royal Statistical Society Series C, Applied statistics*, **60**(4):541-557, 2011.

[6] P. Carbonetto and M. Stephens. Integrated Enrichment Analysis of Variants and Pathways in Genome-Wide Association Studies Indicates Central Role for IL-2 Signaling Genes in Type 1 Diabetes, and Cytokine Signaling Genes in Crohn's Disease. *PLoS Genet*, 9(10), 2013.

[7] Hanahan, D., Weinberg, R. (2011). Hallmarks of Cancer: The Next Generation. *Cell, 144*(5), 646-674.

[8] Paavonen, K., Horelli-Kuitunen, N., Chilov, D., Kukk, E., Pennanen, S., Kallioniemi, O., . . . Alitalo, K. (1996). Novel Human Vascular Endothelial Growth Factor Genes VEGF-B and VEGF-C Localize to Chromosomes 11q13 and 4q34, Respectively. *Circulation, 93*(6), 1079-1082.

[9] M. West. Bayesian factor regression models in the "large p, small n" paradigm. *Bayesian Statistics*, **7**, 2003.

[10] Schneikert, J., Behrens, J. (2007). The canonical Wnt signalling pathway and its APC partner in colon cancer development. *Gut*, **56**(3), 417–425.

[11] Segditsas, S., Tomlinson, I. (2006). Colorectal cancer and genetic alterations in the Wnt pathway. *Oncogene*, 25. doi:10.1038/sj.onc.1210059.

[12] Crawford, L., Wood, K. C., and Mukherjee, S. Scalable Bayesian kernel models with variable selection. stat.ME. 2015.

[13] P. R. Hahn, C. M. Carvalho, and S. Mukherjee. Partial factor modeling: Predictor-dependent shrinkage for linear regression. *J. Am. Stat. Assoc.*, **808**:999-1008, 2013.

[14] F. Liang, S. Mukherjee, and M. West. The use of unlabeled data in predictive modeling. *Statistical Science,* **22**(2):189-205, 2007.

[15] H. F. Lopes and M. West. Bayesian model assessment in factor analysis. *Statistica Sinica*, **14**:41-67, 2004.

[16] A. Rahimi and B. Recht. Random features for large-scale kernel machines. *Neural Information Processing Systems (NIPS)*, 2007.

[17] E. G. Bazavan, F. Li, and C. Sminchisescu. Fourier kernel learning. *Computer Vision - ECCV*, 459-473, 2012.

[18] Nikolopoulos, S. N., Spengler, B. A., Kisselbach, K., Evans, A. E., Biedler, J. L., and Ross, R. A. (2000). The human non-muscle -actinin protein encoded by the ACTN4 gene suppresses tumorigenicity of human neuroblastoma cells. *Oncogene, 19*(3), 380-386.

[19] Dhillon, A. S., Hagan, S., Rath, O., and Kolch, W. (2007). MAP kinase signalling pathways in cancer. *Oncogene, 26*(22), 3279-3290.

[20] Genes and mapped phenotypes. (2016, March 6). Retrieved from http://www.ncbi.nlm.nih.gov/gene/135.

[21] Hotelling, H. (1957), "The Relationship of the Newer Multivariate Statistical Methods to Factor Analysis," *British Journal of Statistical Psychology*, 10, 69–79.

[22] Jolliffe, I. T. (1982), "A Note on the Use of Principal Components in Regression," *Journal of the Royal Statistical Society*, Series C, 31, 300– 303.

[23] Cox, D. (1968), "Notes on Some Aspects of Regression Analysis," *Journal of the Royal Statistical Society*, Series A, 131, 265–279.

[24] Carvalho, C. M., Polson, N. G., and Scott, J. G. (2010), "The Horseshoe Estimator for Sparse Signals," *Biometrika*, 97, 465–480.

[25] Zhihua Zhang, Guang Dai, and Michael I. Jordan. Bayesian generalized kernel mixed models. *Journal of Machine Learning Research*, 12:111139, January 2011.

[26] Hoti, F., Sillanpaa, MJ. (2006). Bayesian mapping of genotype expression interactions in quantitative and qualitative traits. *Heredity, 97*:4-18.

[27] Goel, H. L. and Mercurio, A. M. (2013). VEGF targets the tumour cell. *Nature Reviews Cancer Nat Rev Cancer, 13*(12), 871-882.

[28] Hamill, K. J., Hiroyasu, S., Colburn, Z. T., Ventrella, R. V., Hopkinson, S. B., Skalli, O., and Jones, J. C. (2015). Alpha Actinin-1 Regulates Cell-Matrix Adhesion Organization in Keratinocytes: Consequences for Skin Cell Motility. *Journal of Investigative Dermatology, 135*(4), 1043-1052.

[29] Alvaro-Benito, M., Wieczorek, M., Sticht, J., Kipar, C., and Freund, C. (2015). HLA-DMA Polymorphisms Differentially Affect MHC Class II Peptide Loading. *The Journal of Immunology, 194*(2), 803-816.

[30] Holling, T. M., Schooten, E., and van Den Elsen, P. J. (2004). Function and regulation of MHC class II molecules in T-lymphocytes: Of mice and men. *Human Immunology, 65*(4), 282-290.

[31] Lázár, V., Ecsedi, S., Szöllősi, A. G., Tóth, R., Vízkeleti, L., Rákosy, Z., . . . Balázs, M. (2009). Characterization of candidate gene copy number alterations in the 11q13 region along with BRAF and NRAS mutations in human melanoma. *Mod Pathol Modern Pathology, 22*(10), 1367-1378.

[32] Mellman, I., Coukos, G., and Dranoff, G. (2011). Cancer immunotherapy comes of age. *Nature, 480*(7378), 480-489.

[33] Nilsson, J. A., and Cleveland, J. L. (2003). Myc pathways provoking cell suicide and cancer. *Oncogene, 22*(56), 9007-9021.

[34] Parsons, S. J., and Parsons, J. T. (2004). Src family kinases, key regulators of signal transduction. *Oncogene, 23*(48), 7906-7909.

[35] Renaud, M., Praz, V., Vieu, E., Florens, L., Washburn, M. P., L'hote, P., and Hernandez, N. (2013). Gene duplication and neofunctionalization: POLR3G and POLR3GL. *Genome Research, 24*(1), 37-51.

[36] Schadendorf, D., Fisher, D. E., Garbe, C., Gershenwald, J. E., Grob, J., Halpern, A., . . . Hauschild, A. (2015). Melanoma. *Nature Reviews Disease Primers*. doi: 10.1038/nrdp.2015.3.

[37] Srivastava, R., Geng, D., Liu, Y., Zheng, L., Li, Z., Joseph, M. A., . . . Davila, E. (2012). Augmentation of Therapeutic Responses in Melanoma by Inhibition of IRAK-1,-4. *Cancer Research, 72*(23), 6209-6216.

[38] Vízkeleti, L., Ecsedi, S., Rákosy, Z., Orosz, A., Lázár, V., Emri, G., . . . Balázs, M. (2012). The role of CCND1 alterations during the progression of cutaneous malignant melanoma. *Tumor Biol. Tumor Biology, 33*(6), 2189-2199.

[39] Schlagbauer-Wadl, H., Griffioen, M., Elsas, A. V., Schrier, P. I., Pustelnik, T., Eichler, H., . . . Jansen, B. (1999). Influence of Increased c-Myc Expression on the Growth Characteristics of Human Melanoma. *J Invest Dermatol Journal of Investigative Dermatology, 112*(3), 332-336.

[40] Zhang, J., Han, S., Zhang, B., and Zhang, Y. (2014). Cancer Immunology and Cancer Immunodiagnosis. *Journal of Immunology Research, 2014*, 1-2.

[41] Hartmann, S., Szaumkessel, M., Salaverria, I., Simon, R., Sauter, G., Kiwerska, K., . . . Giefing, M. (2011). Loss of protein expression and recurrent DNA hypermethylation of the GNG7 gene in squamous cell carcinoma of the head and neck. *Journal of Applied Genetics J Appl Genetics, 53*(2), 167-174.

[42] Ohta, M., Mimori, K., Fukuyoshi, Y., Kita, Y., Motoyama, K., Yamashita, K., . . . Mori, M. (2008). Clinical significance of the reduced expression of G protein gamma 7 (GNG7) in oesophageal cancer. *Br J Cancer British Journal of Cancer, 98*(2), 410-417.

[43] Pardoll, D. M. (2012). The blockade of immune checkpoints in cancer immunotherapy. *Nature Reviews Cancer Nat Rev Cancer, 12*(4), 252-264.

[44] Park, J., Song, J., He, T., Nam, S., Lee, J., and Park, W. (2009). Overexpression of Wnt-2 in colorectal cancers. *Neo Neoplasma, 56*(2), 119-123.

[45] Jung, Y., Jun, S., Lee, S., Sharma, A., and Park, J. (2015). Wnt2 complements Wnt/-catenin signaling in colorectal cancer. *Oncotarget, 6*(35).

[46] Katoh, M. (2003). WNT2 and human gastrointestinal cancer (Review). *International Journal of Molecular Medicine, 12*(5), 811-816. http://dx.doi.org/10.3892/ijmm.12.5.811