2. Specific Aims (one page maximum; separate PDF attachment)

The translation of information obtained from genomic data has generated many candidate biomarkers for disease with potential clinical value. These disease-specific gene products may be employed in diagnostics for early detection, disease staging, therapy selection as well as potential therapeutic targets to enhance patient care. Identifying, characterizing and translating knowledge of these disease-specific markers for these clinical applications is a major goal in medical research.

Deep sequencing is a relatively new, powerful technique for the identification of differentially expressed transcripts. Sequencing-based gene expression (RNA-Seq) studies have been touted as overcoming longstanding limitations of microarray-based studies and RNA-Seq has been shown to be effective for transcript discovery and analysis in large and complex mammalian genomes (Maher et al. 2008; Marioni et al. 2008; Mortazavi et al. 2008). These new biotechnologies raise statistical and computational challenges, as discussed at a recent conference of statisticians, geneticists, and computer scientists (*Statistical and Computational Challenges in Next-Generation Sequencing*, October 10th, 2008, Mathematical Sciences Research Institute, Berkeley, CA).

To help address these challenges we propose to test the hypothesis that ***the combination of Bayesian methods for expression modeling and variable screening with additive models of outcome will give an accurate predictive procedure for next generation sequencing data*.** The testing of this hypothesis will involve the development of (1) a statistical approach to expression measurement which can incorporate multiple isoforms and correlated read locations; (2) a variable selection technique for RNA-seq data which is effective when the true biological model is not sparse; and (3) statistical models for outcome that can be extended to outcome modeling from multiple data types (e.g., clinical, genetic, genomic). These challenges will be addressed through the following specific aims:

**Specific Aim 1: Implement a Bayesian method for RNA-seq expression inference which includes a dependence structure among isoforms.**

We propose a new approach based on the Beta-binomial model of Vencio et al. (2004) but extended to account for multiple isoforms and for dependence among isoforms of the same gene. The dependence among isoforms is expressed in terms of a multinomial model with a conjugate Dirichlet prior. In brief, our method is a fully conjugate Bayes solution which incorporates regional dependence and does not require any normal approximation of the data (unlike Jiang and Wong (2009)).

**Specific Aim 2: Implement a Bayesian approach to RNA-seq screening for variable selection which is robust to prior selection, using RNA-seq expression estimates from the method designed in Aim 1.**

We propose a search method which is an extension of sure independence screening (Fan and Lv 2008) to scenarios where the true model may not be sparse. Our method, named Bayesian screening, willconduct an intelligent statistical search for reads which may only be moderately informative by searching through groups of locations whose read counts may be related.

**Specific Aim 3: Develop additive statistical models for outcome prediction based on the results of Aims 1 and 2 which is accurate in the context of multitype data and heterogeneous samples.**

We produce a set of modelswhere each model is based on information from a single variable type (e.g., sequence, expression, clinical). The flexibility of each model is chosen to reflect the complexity of the information reflected in each variable type. In essence, we intelligently represent the information from thousands of variables in a handful of models. It is then possible to construct a single model of the outcome by constructing an overall additive model via backfitting (Hastie and Tibshirani 1990) which has as its components models representing each variable type.

The statistical methods developed within each aim will be evaluated in the context of multiple data sets obtained from either the NCBI Short Read Archive (SRA) or from research conducted at the University of Miami School of Medicine. All data is either currently publically available from the SRA or will be made publically available prior to the completion of the proposed research.

In order to facilitate future research involving next generation sequence data we will develop computational tools for the R statistical language (ref) which implement our methods.

5. Research Design and Methods (12 pages maximum; separate PDF attachment)

**Research Area:**

06: Enabling Technologies

New computational and statistical methods for the analysis of large data sets from next-generation sequencing technologies: 06-HG-101

**The Challenge and Potential Impact:** What is the research opportunity, scientific knowledge gap or technology that will be addressed? How broad is the potential impact in science and/or health? Which community (ies) will be affected? What is (are) the size(s) of the community(ies)? Will the potential impact be major?

The translation of information obtained from genomic data has generated many candidate biomarkers for disease with potential clinical value. These disease-specific gene products may be employed in diagnostics for early detection, disease staging, therapy selection as well as potential therapeutic targets to enhance patient care. Identifying, characterizing and translating knowledge of these disease-specific markers for these clinical applications is a major goal in medical research.

Deep sequencing is a relatively new, powerful technique for the identification of differentially expressed transcripts. While sequencing-based gene expression (RNA-Seq) studies have been touted as overcoming longstanding limitations of microarray-based studies, and RNA-Seq has been shown to be effective for transcript discovery and analysis in large and complex mammalian genomes (refs), these new biotechnologies raise statistical and computational challenges (ref MSRI). We propose to address three of these challenges, namely, the development of (1) a statistical approach to expression measurement which can incorporate correlated read locations and is robust to assumptions; (2) a variable selection technique for RNA-seq data which is effective when the true biological model is not sparse; and (3) statistical models for outcomes that can accomodate multiple data types (e.g., clinical, genetic, genomic). These challenges must be addressed if information about the transcriptome is to be translated to clinical studies.

In reference to Specific Aim 1, several statistical approaches exist for estimating expression from RNA-seq data, including Poisson modeling (Marioni et al. 2008, Jiang and Wong 2009), categorical modeling (classifying tag counts as high, medium, low, or none), and independent Binomial modeling (‘t Hoen 2008). As discussed below each of these approaches has significant drawbacks. We propose a novel Bayesian approach which is similar to the Beta-Binomial approach of Vencio et al. (2004) but extends to genes with multiple isoforms and/or correlated tag reads.

In reference to Specific Aim 2, one of the most popular methods for variable selection/screening in ultra high-dimensional data spaces is Sure Independence Screening (SIS; Fan and Lv 2008). SIS can reduce the dimensionality of the dependent variable space from a large or huge scale (say, for some ξ>0) to a relatively large scale (*o(n)*) by a fast and efficient method. However, SIS is designed to consider the correlation of each dependent variable to outcome independent of all other variables, an excellent strategy if the true data generating model is sparse. We assume that the true data model is not sparse, e.g., there exist many transcripts whose expression is predictive of outcome but whose relationship to outcome may appear to not be significant if considered in isolation. To address this scenario we introduce Bayesian screening, an extension of SIS which considers interactions between variables and variables in groups when selecting variables for subsequent outcome modeling. We hypothesize that Bayesian screening is efficient in large dimensions and `sure’ in the sense that all of the important variables survive after variable screening with probability tending to 1.

In reference to Specific Aim 3, we emphasize the application of RNA-seq data to translational science by developing a statistical methodology for outcome modeling based transcript reads and their interactions. This method is designed specifically for ultra high-dimensional data spaces where the dimensionality of the relevant dependent variable space is larger than the sample size. It involves different model classes for main effect terms and interaction terms as well as model averaging to incorporate information from multiple transcripts. We intend to evaluate this approach in several publically available data sets and two data sets designed for detecting transcripts and gene fusion products relevant to breast cancer biology.

The potential impact of translating RNA-seq data to clinical practice is substantial. Gene expression arrays have generated discoveries relevant to diagnostics and treatment in many disease contexts (ref?). A gain in information is expected from RNA-seq as it has several advantages over traditional arrays. For example, unlike RNA-seq, expression microarrays suffer hybridization artifacts and design constraints that limit the detection of splicing events or previously unmapped genes. Although only recently available for large and complex transcriptomes, ultra-high-throughput sequencing has revealed novel gene fusions in prostate cancer cell lines and tissues (maher). Such fusions represent potential therapeutic targets, as similar fusion products have been developed as therapeutic targets in chronic myelogenous leukemia (CML).

We anticipate that this research will lead to substantial improvements in the analysis of next-generation transcriptome sequencing data and further the translation of RNA-seq data to clinical applications.

**The Approach:** How will you attempt to explore or solve the stated research problem? How will your rationale and/or approach overcome existing challenges or barriers in the field? If you propose to improve existing technologies or to develop new technologies, which needs are being addressed and what is unconventional and exceptionally innovative about your approach? Provide enough information for reviewers to determine what you are proposing to do, but do not include a detailed experimental plan.

We propose to address statistical challenges, namely, the development of (1) a statistical approach to expression measurement which can incorporate correlated read locations and is robust to assumptions; (2) a variable selection technique for RNA-seq data which is effective when the true biological model is not sparse; and (3) statistical models for outcome that can be extended to outcome modeling from multiple data types (e.g., clinical, genetic, genomic). These challenges must be addressed if information about the transcriptome is to be translated to clinical studies.

**Specific Aim 1: Implement a Bayesian method for RNA-seq expression inference which includes a dependence structure among isoforms.**

If RNA-seq data is to yield reliable information about transcript levels then a statistical method is needed to provide robust estimates of expression and detect differential expression between samples of interest. Several methods have been proposed for expression estimation; of note, Poisson modeling (Marioni, Jiang) and Beta-binomial modeling (Vencio, tHoen). Marioni et al. (2008) modeled the number of reads mapped to a gene in a given sample as a Poisson random variable, assuming independence of read counts from different genes. This model was then cast as a generalized linear model to detect differentially expressed genes. The authors concluded that the data showed deviations from the assumptions of the Poisson approximation but, at a strict false discovery rate (FDR) level, these deviations did not lead to an appreciable number of false-positive identifications. We argue that a better model is needed there are deviations from the existing model and FDR is inappropriate for measuring the impact of these deviations. Their use of FDR assumes independence of transcript reads which is an assumption that, as with gene expression array data, does not hold. Also, most disease-related studies involving high-throughput expression have an exploratory component, or are looking for something to examine further. Hence a strict FDR level is inappropriate. In addition, as noted by Jiang and Wong (2009), many genes have very low levels of expression so the problem of estimating expression is a problem of estimating a parameter at the boundary of the parameter space (as expression is nonnegative). This is a problem for the Poisson model which the authors attempt to solve by using importance sampling for estimation and conditioning the covariance matrix to be more reliable. They do not examine, however, the impact of this conditioning on their expression estimates or verify that they have avoided problems with the boundary of the parameter space.

We propose a new approach based on the Beta-binomial model of Vencio et al. (2004) but extended to RNA-seq data in such a way as to account for multiple isoforms and for dependence among isoforms of the same gene. The dependence among isoforms is expressed in terms of a multinomial model with a conjugate Dirichlet prior. In brief, our method is a fully conjugate Bayes solution which incorporates regional dependence and does not require any normal approximation of the data (unlike Jiang and Wong (2009)). In addition, a Dirichlet will perform well up to the boundary of the parameter space while a Poisson breaks down as the parameter values approach the boundary. The Dirichlet can also be modified to handle the boundary by adding a term which places positive mass on the boundary. We perform a sensitivity analysis to examine the dependence of our estimation procedure on the choice of prior, an essential consideration for reliable estimation.

A basic outline of our proposed procedure is as follows. Let be the set of genes and for any let }, be the set of its isoforms. Let be the set of all possible isoforms of all genes in. For any isoform let be the length of the isoform and let be the number of copies of transcripts of isoform in the sample. Under simple random sampling, the probability that a read comes from isoform is . If we let represent the expression index of isoform , then . Let be the total number of mapped reads. By similar reasoning, given an isoform and a region of length, the number of reads from isoform which come from this region is where. Our goal is, given and, we want to estimate] for all .

Suppose gene has exons with lengths and isoforms with expressions []. Then the number of reads falling into exon , is where if isoform contains exon and 0 otherwise. Similarly, the number of reads falling into a junction between exons and is where is the length of the exon-exon region. Note in general that is a linear function of, i.e.. In the Bayesian context we can estimate from its joint posterior distribution. This requires the specification of a prior distribution, an issue not considered by Jiang and Wong (2009). We will implement several choices of prior, as discussed below, but initially we propose a conjugate prior, i.e., for

This prior makes sense as we recall that each represents a proportion where the numerator can be modeled as a Gamma random variable and the denominator can be modeled as a sum of Gamma random variables. Hence represents a Gamma random variable (as the sum of Gammas is Gamma) divided by a sum of Gamma random variables, i.e., a Beta random variable. The choice of a Beta prior provides the flexibility of choosing a prior close to a Uniform distribution for each , if we have no information regarding the expressions of the various isoforms, or a prior with most of its weight close to zero, if we suspect that the expression of an isoform is low. This will be useful for estimating the expression of low abundance transcripts, a known problem for next-gen mRNA data analysis (Marioni 08, Mortazavi 08).

As this prior is conjugate the posterior is a Beta distribution and the estimate of has a closed-form expression. For example, in the case of a gene with a single isoform with expression we have and . Thus the posterior estimate of is, i.e.,. Note that this is a Bayesian version of the RPKM measure of Mortazavi et al. (2008).

This approach avoids the problem of instability in the estimate of when the true value lies near the boundary of the parameter space, i.e., when an isoform is expressed at a very low level. However, it does not allow for the modeling of dependence among reads, as we would expect if their corresponding regions overlap. In this case we propose to replace our Beta prior and binomial likelihood on each with their multivariate generalizations, a Dirichlet prior and a multinomial likelihood, on the multiple parameters to be estimated. For example, we can jointly model the number of reads coming from the exons *j* and *k* and the exon-exon junction between exons *s* and *t* for a given gene, denoted , as a Multinomial random variable with parameters *w* and . With a prior on the posterior distribution of is

Although the above posteriors have closed forms, this may not always be the case, i.e., if the choice of prior is changed. If the prior is not conjugate samples from the posterior can be obtained by Markov Chain Monte Carlo (MCMC) simulation (Gelman et al. 1995; Liu 2001). Markov chain Monte Carlo is an established general computing technique that has been widely used in physics, chemistry, biology, statistics, and computer science (refs?). It simulates a Markov chain whose invariant states follow a given (target) probability in a very high dimensional state space. We propose initially to use the OpenBUGS or WinBUGS software (Lunn et al. 2000; Spiegelhalter et al. 2007) via the R programming language (ref) to generate posterior estimates of expression as well as uncertainty estimates. The convergence of the MCMC algorithm will be assessed by post-processing output through the CODA package in R (Plummer et al. 2006). CODA provides graphical analyses of MCMC results plus summary statistics and convergence diagnostics.

*Sensitivity Analysis*

A Bayesian model requires the specification of a prior distribution on the parameter(s) of interest. There will generally be some uncertainty in the choice of prior, especially when there is little information from which to construct such a distribution, or when there are several priors elicited, say, from different experts. It is of interest, then, to characterize the sensitivity of a posterior distribution (or posterior mean) to the prior (Hill and Spall 94). Both Vencio et al. (04) and Jiang and Wong (09) proposed Bayesian models of RNA-seq expression but did not examine prior sensitivity of their results.

One way to characterize this sensitivity is in terms of bounds on the difference between posterior distributions corresponding to different priors. For example, given a gene with isoforms we may have no prior information regarding their expression, so we choose a noninformative prior on. A generalization of this prior is , for . We will examine the sensitivity of our results to various Dirichlet priors by implementing several values for each. Other ways to characterize sensitivity include simply examining features of the posterior distribution. Traditionally, a sensitivity analysis would involve running a separate MCMC chain for each choice of prior. Given the size of RNA-seq data this is not a feasible approach. Instead we propose to use importance sampling (IS; Besag et al. 95) to examine sensitivity, as used by Nur et al. (2009) in the context of DNA sequencing data. By importance sampling we can efficiently calculate estimates of under different prior assumptions without running multiple MCMC chains. Sensitivity can be assessed by computing the sum of the squared distances between posterior distributions, i.e., where is the estimate of from the th alternate prior IS and is the estimate of from the base model MCMC.

*Evaluation*

The methods that we develop for this Specific Aim will be assessed using publically available data from the NCBI's [Short Read Archive](http://www.ncbi.nlm.nih.gov/Traces/sra/sra.cgi?) (SRA) sequence database (http://www.ncbi.nlm.nih.gov/Traces/sra/sra.cgi). Both data sets have been used in recent publications on RNA-seq data. We briefly describe these data sets below.

1. **Marioni et al. (2008) liver/kidney data (SRA000299)**

This data was generated by extracting total RNA from liver and kidney samples of a single human male, purified the poly(A) mRNA, and sheared it prior to cDNA synthesis. The cDNA was then processed into a library of template molecules suitable for sequencing on the Illumina Genome Analyzer. Each sample was sequenced 5 times at 3 pM. Samples were also arrayed in triplicate to Affymetrix U133 Plus 2 arrays (GEO database with accession number GSE11045). To facilitate a direct comparison between the sequence and array data, the array probe sets were mapped to annotated

genes in the Ensembl database v.48. This identified a set of 17,708 probe sets, mapping uniquely to 17,708 genes. We propose to use this data to evaluate the results of our expression estimation procedure by comparing the expression estimates to those of Marioni et al. (2008).

1. **Mortazavi et al. (2009) brain/muscle/liver data (SRA001030)**

This data was generating by performing RNA-Seq, using Illumina/Solexa sequencing technology, on poly(A)-selected RNA from adult C57BL mouse brain, liver and skeletal muscle tissues. Two replicates were sequenced for each tissue sample. A comparison of mouse liver values to publicly available Affymetrix microarray intensities from GEO (GSE6850) was performed for genes called as present by Rosetta Resolver. This data was also analyzed by Jiang and Wong (2009) with a specific focus on genes with more than one isoform. We propose to use this data to evaluate the results of our expression estimation procedure by comparing the expression estimates to those of Mortazavi et al. (2008) and the expression estimates from genes with more than one isoform to those of Jiang and Wong (2009).

If the above Beta-binomial and multinomial models fail to provide robust and accurate estimates of RNA-seq expression, alternative models of expression will be implemented. Alternatives include the quasi-Poisson distribution or the negative binomial distribution, as suggested in Marioni et al. (2008). In the case of problems with the boundary of the parameter space, i.e., genes with many isoforms which have no expression, the Dirichlet prior can be adjusted by the addition of a probability mass on the boundaries of the parameter space.

**Specific Aim 2: Implement a Bayesian approach to RNA-seq screening for variable selection which is robust to prior selection, using RNA-seq expression estimates from the method designed in Aim 1.**

We assume that we are interested in building a model from RNAseq data to predict an outcome of interest, e.g., presence of disease or response to therapy. Most statistical approaches to this problem assume that the model underlying the observed data is sparse, i.e., the number of genomic locations whose expression (i.e., read counts) values are truly related to outcome is quite small and these locations are independent of one another. Such approaches often consider the predictive power of each location individually and use shrinkage methods to remove most locations from consideration as potential predictors. In this context shrinkage means that the values of the estimated model coefficients for a large number of variables are shrunken towards zero, with the intended effect of removing variables from consideration and reducing the mean square error of predicted values from the model when applied to new data (Copas 83). Shrinkage provides a method for variable selection in high dimensions, by selecting only those variables whose coefficients after shrinkage are greater than a chosen threshold. For example, sure independence screening (SIS; Fan and Lv 2008) searches through ultra high dimensional feature spaces by assuming independence of variables, examining each feature in isolation and using shrinkage to select features of interest. This approach is very powerful if the underlying model is sparse and the variables of interest are independent.

If the true model contains greatly more variables than observations, and the variables are not independent, then an alternative approach is needed. In the case of RNAseq data it is probable that the number of reads from many locations are each moderately informative with respect to outcome, as opposed to a small set of locations which are highly informative. Susceptibility to common complex diseases probably involves the contribution of both common transcripts and rare transcripts (Kris … 08). In addition, there could exist a myriad of rare transcripts and in combination they might explain a considerable proportion of the variance in a trait of interest [56; kris et al.]. It is also plausible that reads may be dependent whether they come from adjacent locations (*cis*- effects) or not (*trans*- effects) (emilsson08nature). The above scenario is unlike that assumed by most statistical approaches to high-dimensional data analysis, including SIS, where the true model is assumed to be sparse () and the search method is focused on finding a small set of highly informative, independent variables.

Given the large amount of data available the task is to conduct an intelligent statistical search for reads which may only be moderately informative by searching through groups of locations whose read counts may be related. Once informative locations have been identified, they can be combined in a statistical model whose performance is measured by its ability to predict outcome.

*Bayesian Screening*

We propose a search method which is an extension of SIS to scenarios where the true model may not be sparse. Let ) be the number of reads coming from specified genomic locations, and let be the outcome of interest, e.g., presence/absence of disease, measured on *n* subjects. In Bayesian Screening we screen subsets of of size. For each subset j,, we build a model and form the posterior. Given a (noninformative) prior on**,** denoted the information about the outcome contained in can be summarized by the relative entropy between the posterior and the prior, i.e.). This measure can be interpreted as the distance of the posterior from the distribution we would *expect* if none of the variables were informative, i.e., the prior distribution. Using relative entropy as the measure of information places all subsets of variables on a common code length scale for easy comparison. The statistical significance of) can be estimated by resampling-based permutations, as has been done with other high-dimensional biological data (Yan et al. 2005; Vacic et al. 2007). A significant value for the relative entropy is evidence that the variables in are informative and should be considered as possible predictors for; subsets of variables with nonsignificant relative entropy can be excluded (a method of thresholding where the threshold is statistical significance).

This approach has several advantages relative to SIS screening. First, Bayesian screening takes into account the whole shape of the posterior distribution, not just the mode. Second, it considers variables in groups, making it more computationally efficient and allowing it to account for dependence among variables. SIS will have difficulty when screening dependent variables as shrinkage methods are known to break down quickly in the presence of dependence (ref?). Third, we submit that Bayesian Screening will perform better then SIS when the true model is not sparse, as it will detect variables which are only moderately informative when considered individually but significantly informative when considered in groups. This is due to the sensitivity of relative entropy and the thresholding of relative entropy values in order to exclude variables.

Bayesian screening may be computationally intensive for a very large data set such as RNAseq data. The computational time required for screening can be reduced considerably by using parallel computation, as the process is easily adaptable to parallel implementation. We plan to implement Bayesian screening on a cluster in CCS (details).

*Sensitivity Analysis*

Bayesian screening selects variables based on a relative entropy measure of information. This measure involves the posterior distribution of so we must specify a prior distribution for . We specify a non-informative prior, as we expect that given RNAseq data we will not know a priori which transcripts will be significant. In addition, we would like all transcripts to be given equal weight in the screening process if our screening process is to be unbiased.

As in Specific Aim 1, the results of Bayesian screening should be evaluated with respect to sensitivity to the choice of the prior on. This sensitivity analysis will parallel the sensitivity analysis conducted in Specific Aim 1. We will conduct Bayesian screening using a variety of noninformative priors, as well as other dependence priors, to examine the impact of the prior choice on the variable selection process. Other choices of prior include shrinkage priors, i.e., priors which correspond to shrinkage penalties. The use of such priors is one way to perform shrinkage when there is dependence among the predictor variables (a context where usual shrinkage methods break down). Zellner’s g-prior (1986) is a popular choice because it does not require any prior specification of the covariances between the elements of. We also plan to implement a data augmentation prior (DAP), which is related to Zellner’s g-prior (bedrick 96), and a new weakly informative prior proposed by Gelman et al. (2008) for logistic regression models.

*Evaluation*

The methods that we develop for this Specific Aim will be assessed using the same data sets as in Specific Aim 2, namely, the liver/kidney data of Marioni et al. (2008) (SRA000299) and the mouse brain/liver/muscle data of Mortazavi et al. (2008). We will compare the performance of Bayesian Screening to Sure Independence Screening (Fan and Lv 2008) with respect to the selection of variables that are significantly associated with outcome.

**Specific Aim 3: Develop additive statistical models for outcome prediction based on the results of Aims 1 and 2 which is accurate in the context of multitype data and heterogeneous samples.**

*Batch modeling*

Statistical models of outcome will be built from the variables selected by Bayesian screening. If the number of variables selected by Bayesian Screening,,is less than the sample size then a logistic regression model of the outcome can be built using these variables as predictors. If as we expect, since the true underlying model is not sparse, then the variables will be clustered. In other words, let ) represent the variables selected by Bayesian screening. We propose to cluster the elements of **’** into clusters by either *k*-means or multinomial clustering, as described in Hamerly et al (2005) and Bouguila (2008). Both clustering approaches have shown to yield reasonable results for relatively large sets of count data (). Once clustered, if the number of variables in a given cluster, , is less than then a model of the outcome can be fit using the variables in cluster . If the clustering algorithm yields clusters of size greater than, the centroid of each cluster will be calculated. For each cluster we will choose those variables that are closest to the centroid, as measured by Euclidean distance, as representative of the cluster; all other variables from the cluster are not retained. The result will be clusters of size less than so that an additive model of outcome can be built from the variables in each cluster.

*Composite modeling of Multitype data*

We assume that the number of variables selected by Bayesian Screening is less than the sample size or that the selected variables have been clustered into groups of size less than . Let represent the set of selected variables (i.e. the main effects) and consider all of the possible two-way interactions among elements of We propose to build models of outcome on main effects or interactions whose flexibility reflects the complexity of the effects being modeled. Main effects () capture less complex effects than interaction terms so the model class for main effects can be relatively simple, e.g., linear regression or simple additive models. As interaction terms capture more complex effects than main effect terms, they require a more flexible model class, e.g., recursive partitioning models. Our modeling process will yield a model of outcome from each variable type; if a variable type has multiple clusters then a model will be build from each cluster and these cluster models will be averaged into an overall model). A model representing both classes of variables can be constructed by constructing an additive model with two components, one for each model, and fitting the model via backfitting (ref). This algorithm uses partial residuals to fit each component model in an iterative fashion, avoiding any dependence on the order in which the models are fit to the data.

This process can be extended multitype data. Suppose ) represents the number of reads coming from specified genomic locations and represents single nucleotide polymorphism (SNP) measurements taken on the same *n* subjects. We are interested in screening and for variables with information regarding an outcome. In addition, we can consider screening the interactions between reads, between SNP measurements, as well as between reads and SNPs . After Bayesian screening (and clustering, if necessary) within each of these 5 variable types we will have a set of selected variables of each type. We can repeat the above modeling process, building linear or logistic regression models on the main effects, denoted and , and recursive partitioning models on the interactions within a data type (and ), denotedand .Finally, the interaction terms between data types, being the most complex effects, require the most flexible modeling class, e.g., projection pursuit regression. We build a projection pursuit regression model of the outcome using only these interaction terms, denoted

**Table 1: List of Variable Types and Model Classes**

|  |  |  |
| --- | --- | --- |
| **Variable Type** | **Complexity** | **Model Class** |
| **Main Effects** | Least | Logistic regression or additive models |
| **Interaction within Type** | Moderate | Recursive partitioning models |
| **Interaction between Types** | Highest | Projection pursuit regression |

We have produced a set of models of the outcome where each model is based on information from a single variable type and the flexibility of each model was chosen to reflect the complexity of the information reflected in each variable type. In essence, through Bayesian screening, clustering, and model averaging we have intelligently represented the information from thousands of variables in a handful of models. It is now possible to construct a single model of the outcome. This can be done by constructing an additive model for with the 5 models as components, i.e., where is an additive model constructed via the backfitting algorithm (ref).

*Evaluation*

The methods that we develop for this Specific Aim will be assessed using the publically available data from the NCBI's [Short Read Archive](http://www.ncbi.nlm.nih.gov/Traces/sra/sra.cgi?) (SRA) sequence database (as previously described). In addition, we will analyze two data sets collected at the University of Miami for breast cancer research. We briefly describe these two additional data sets below.

1. **Breast Cancer Cell Lines and Tumor Samples**

Primary (n=10) and secondary (n=10) breast tumors have been obtained from the Tissue Banking Core Facility (TBCF) of the Sylvester Comprehensive Cancer Center at the University of Miami. Due to the diverse breast cancer patient population managed at UM/Sylvester, the TCBF is uniquely positioned to obtain and store very large primary and secondary breast tumors that are not observed at other treatment centers. These large tumor samples provide sufficient material to complete the analysis outlined in this proposal that could not be readily completed by research groups at other institutes. The TBCF furnishes investigators who have IRB approved protocols with de-identified tumor samples and clinical history including pathology. The TBCF has full approval from the IRB and an NIH certificate of confidentiality. Sample preparation (generation of cDNAs) and sequencing of breast cancer cell transcriptomes was completed at the Center for Genome Technology of the University of Miami Institute for Human Genomics (MIHG-CGT). Long-read sequencing (200-500 nt) of cDNA fragments was completed using 454 FLX Sequencing (Roche) and short-read sequencing (20-75 nt) was performed on the Illumina Genome Analyzer II. The Illumina next-generation sequencing platform was based on single molecule extension, also known as sequencing by synthesis.

Long-read sequences will be compared to a human reference genome (Human Genome Build 18: <http://genome.ucsc.edu>) and categorized as mapping, partially aligned or non-mapping reads. Alignment will be performed using the Enhanced Read Analysis of Gene Expression (ERANGE) program, as described in Mortazavi et al. 2008. This program will (1) assign reads that map uniquely in the genome to their site of origin and assign sites that match equally well to several sites to their most likely site(s), (2) detect splice-crossing reads, (3) cluster reads that do not map to a known exon and suggest candidate exons, and (4) calculate the prevalence of transcripts.

**Table-2: List of Breast Cancer Cell Lines and Tumor Samples**

|  |  |  |
| --- | --- | --- |
| **Sample Type** | **Number of Samples** | **Details** |
| **Breast Cancer Cell Lines** | 10 | ER+: MCF-7, T47D, BT474, ZR751, SUM44  ER-: MDA-MB-231, SKBR3, BT549, MDA-MB-436, MDA-MB-453 |
| **Primary Tumors** | 10 | ER+, ER- |
| **Secondary Tumors** | 10 | ER+, ER- |
| **Normal Breast Tissue** | 5 | Reduction Mammoplasty |
| **Control Cell Lines** | 2 | MCF-10a, HMECs |

1. **Breast Cancer Xenograft Titration**

A titration series was designed by combining normal mouse lung RNA with human breast cancer cell line (MDA231) RNA. Each titration series had 5 samples with different proportions of each RNA source, specifically, 100% MDA, 75% MDA/25% mouse, 50% MDA/50% mouse, 25% MDA/75% mouse, and 100% mouse. All samples were sequenced in duplicate as described for the above data. In addition the same samples were arrayed on both the Illumina Mouse WG-6 BeadChip and the Illumina Human Ref-8 BeadChip. The outcome of interest is the organism of origin.

**Timeline and Milestones:**

The timeline provided indicates the start time and length of time required (in months) to complete each step of the proposed research plan. Time allotted to ‘Alternate Models’ relates periods of time when intermediate objectives will be assessed and decisions will be made regarding the course and direction of the research effort.

**Inclusion of Women, Minorities, and Children in Challenge Grant Studies**

Letters of Support: Nick T.; M. Lippman; MIHG?

Hires/job support: self, Bertrand, Stuart, Dale, Pearl, post-doc

Resource Sharing Plan: website and code