Published in final edited form as:

IEEE Int Workshop Genomic Signal Process Stat. 2013 November ; 2013: 5–8. doi:10.1109/GENSIPS.2013.6735914.

Integrative Analysis of Multi-modal Correlated Imaging-Genomics Data in Glioblastoma

Rolando J Olivares[*], **Arvind Rao**[‡], **Jeffrey S. Morris**[†], **Veerabhadran Baladandayuthapani**[†]

Rolando J Olivares: Olivares@stat.tamu.edu; Arvind Rao: aruppore@mdanderson.org; Jeffrey S. Morris: jefmorris@mdanderson.org; Veerabhadran Baladandayuthapani: veera@mdanderson.org

[*]Department of Statistics, Texas A&M University

[†]Department of Biostatistics, UT M.D. Anderson Cancer Center

[‡]Department of Bioinformatics and Computational Biology, UT M.D. Anderson Cancer Center

Abstract

We propose a method to integrate high-dimensional genomics datasets across multiple platforms with multiple imaging outcomes. This new statistical framework uses a hierarchical model to integrate biological relationships across platforms to identify genes that associate with multiple correlated imaging outcomes. Our two-stage hierarchical model uses the information shared across the platforms and thus increasing the predictive power to identify the relevant genes. We assess the performance of our proposed method through simulation and apply to data obtained from the Cancer Genome Atlas Glioblastoma Multiforme dataset. Our proposed method discovers multiple copy number and microRNA regulated genes that are related to patients' imaging outcomes in glioblastoma.

Index Terms

Bayesian Analysis; genomics; integrative analysis; Lasso Penalization; multiple outcomes; sensitivity; specificity

I. Introduction

Glioblastoma Multiforme (GBM) is the most lethal brain tumor with a patient survival rate of 15 months [1]. GBM develops for the most part in the cerebral hemisphere. If GBM is operable, then surgery is performed, followed by radiation therapy or a combination of radiation therapy and chemotherapy. Otherwise, radiation or radation/chemotherapy are administered. Radiological characterization/phenotyping has been carried out to identify survival-associated phenotypes. Typically, such studies have focused on single phenotypes such as T_1 or *flair* separately. In this study, we report analyses of combinatorial phenotypes that stratify survival in a subset of patients from the TCGA-GBM collection. Furthermore, we investigated molecular differences using messenger RNA (mRNA), microRNA (miRNA) expression and copy number alteration.

Cancer is a disease of the genome. Many profiling technologies have been developed and used to profile the cancer genome. Characterize the biology of cancer more comprehensibly the Cancer Genome Atlas (TCGA) created a project in 2006 to understand the molecular basis of cancer through genomic profiling. TCGA profiled genomic data using multiple platforms on more than 20 cancer types, and GBM was the first brain tumor to be selected because of its poor prognosis [2]. Thus, by combining all these various types of data, one can provide a unified view of the different molecular processes across the whole genome, which enables us to understand the gene functions across platforms. In addition, integrating all these various types of data in one statistical model will increase the statistical power to identify clinical relevant genes and/or better predict relevant patient-level clinical outcomes. Jennings [3] *et al.* proposed a general integrative Bayesian analysis of genomics data (iBAG) framework that models the biological relationships between several platforms to a single clinical outcome. This approach involves a global gene search, and uses variable selection via the Bayesian lasso-based shrinkage prior to deal with the high dimensionality of the data.

In this paper, we introduce a multivariate integrative analysis of high-dimensional genomic data with correlated imaging outcomes. This model uses a hierarchal approach to model biological relationships among and between different platforms and subsequently uses this information to the find associations with imaging outcomes. In this paper, we focus on aberrations of the RTK/PI3K, p53, and an RB cell signaling pathway which is known to be important pathways in cancer. We integrate mRNA expression, miRNA expression and copy number alteration to predict a combinatorial phenotype that stratifies survival for GBM patients. The two imaging phenotypes of interest in this paper are the volume of tumor T_1 and T_1 : *flair ratio*, where T_1 is the post contrast MRI image and *flair* is the hypertense volume. We also conduct simulation studies that show high sensitivity and specificity of our approach in gene selection.

The outline of this paper is as follows. In Section 2, we describe our multivariate integrative analysis model construction. In Section 3, we describe the implementation of our model and a bootstrap lasso stability variable selection method to identify important predictors of imaging outcome(s). In Section 4 we present our simulation studies and Section 5 summarizes the application to TCGA-GBM data. Section 6 contains conclusions and discussions.

II. A Multivariate Integrative Analysis of Multiplatform Genomic Data

We construct a multivariate integrative analysis model that uses a two-level hierarchical approach to integrate multiplatform genomic data and evaluates the relationships to the imaging outcome(s) of interest. The first-level is the *genomic model* that assesses the biological relationship between and among the platforms, miRNA, mRNA and copy number in this case, and directly infers the effects of miRNA and copy number on mRNA expression. We the subsequently use this information in our second stage *imaging model* to asses and detect which genes affect the imaging outcome(s).

Each patient datum in our analysis consists of multiple imaging outcomes (e.g. T_1, *flair* and *ratio*), mRNA expression, miRNA binding sites, genes copy number expression and “other” prognostic clinical factors (e.g. age, Karnofsky score). In the subsections below, we describe the constructions of the genomic and imaging models.

A. Genomic Model

Suppose *n* is the number of patients, *j* indicates the type of the genome platform and p_j is the number of genes measured on platform *j*. The *genomic model* for each gene can be expressed as:

$$\mathrm{X}_{mRNA_g} = \mathrm{X}_{mRNA_g}\beta_{1_g} + \mathrm{X}_{CN_g}\beta_{2_g} + \mathrm{O}_g,$$

where the terms in the *genomic model* can be described as follows:

- X_{mRNA_g} is the mRNA gene expression for *g*th gene denoted by gene$_g$ and is of dimension n×1, $g = 1, \ldots, G$
- X_{mRNA_g} are the miRNAs binding sites on the complete sequence (promoter, five prime untranslated region (5′UTR), coding region (CDS) and three prime untranslated region (3′ UTR)) to gene$_g$ and is of dimension $n \times p_{j_g}$. In the miRNA data, we have multiple miRNA targets bind to any given gene. We use miRWalk [4], a comprehensive human database, that enables us to map miRNA-validated targets to the associated genes.
- X_{CN_g} is the expression of gene$_g$ that is attributed to changes in copy number and of dimension $n \times 1$
- O$_g$ represents the “other” (residual) part of the gene$_g$ expression unexplained by miRNA and CN and could be attributed to “other” biological mechanisms and is of dimension $n \times 1$

In essence, the genomic component models the multiple miRNA validated targets and copy number jointly on the mRNA expression. Thus, for any given gene, our *genomic model* regresses (independently) each gene’s mRNA expression on the miRNA validated targets sites and copy number alterations – resulting in estimation of specific platform specific components in the above equation. We then subsequently use the estimated copy number, miRNA and residual components in our second level of the hierarchicy – the *imaging model*.

B. Imaging Model

Our *imaging model* relates the gene expression effects from the *genomic model* on the imaging outcome(s). We include the estimated effects from the genomic model as factors on the *imaging model* to find which genes’ expression, copy number and/or miRNA influence the imaging outcome(s) across multiple genes and can be written as:

$$\mathbf{Y} = \mathbf{C}\gamma^{C} + \mathbf{miR}\gamma^{miR} + \mathbf{CN}\gamma^{CN} + \mathbf{R}\gamma^{R} + \varepsilon$$

where the terms of the *imaging model* have the following representations and interpretations:

- **Y** is a $n \times k$ matrix, where k is the number of imaging outcome(s)
- **miR** =$\mathbf{X}_{miRNA}\beta$ is a $n \times G$ matrix where the columns are the number of genes with miRNA expression.

- **CN**= $\mathbf{X}_{CN}\beta$ is a $n \times G$ matrix where the columns are the number of genes with copy number profiling.
- **R**= X_{mRNA}−(**miR**+**CN**) is a $n \times G$ matrix where the columns are the number of genes for the "other" genomic profiles.
- **C** is a $n \times L$ matrix, were the columns are the number non-genomic clinical variables (e.g age, demographics)
- γ^{miR} are the expression effects through miRNA
- γ^{CN} are the expression effects through copy number
- γ^{R} are the expression effects through "other" biological mechanism
- γ^{C} are the clinical effects

The number of estimated miRNA-validated gene targets and the copy number from our *genomic model* could be very large (on the order of thousands) compared to the number of patients (on the scale of hundreds). The vast majority of the genes have no substantial influence on the imaging outcome(s); we are require, to perform some variable selection and/or sparse regularization to effectively identified which variables are significant/ important.

III. Implementation

The implementation of our hierarchical model and the variable selection methods are described in this section. To conduct the estimation and the subsequent inference we first apply a Yeo-Johnson transformation [5] to satisfy the normality assumptions. Subsequently we use a covariance decomposition based projection approach to decorrelate the imaging outcomes and use a L1 (Lasso [6]) penalty for variable selection and shrinkage for a fast and scalable implementation of our approach.

Briefly, assuming the imaging features follow a multivariate normal distribution $\mathbf{Y} \sim$ Normal($\mathbf{X}\gamma$, Σ) where $\mathbf{X}$ is now stacked $\mathbf{X}$ = {**miR**, **CN**, **R**} and $\gamma = \{\gamma^{\mathrm{miR}}, \gamma^{\mathrm{CN}}, \gamma^{\mathrm{O}}, \gamma^{\mathrm{C}}\}$. To decorrelate the imaging outcomes we carry out the following steps. First we estimate the covariance matrix $\mathbf{\Sigma}$ using the empirical covariance matrix $\widehat{\sum} = \mathbf{Y}^{T}\mathbf{Y}/n$. Next, we decompose the estimated covariance matrix using singular value decomposition (SVD) as $\widehat{\sum} = \mathbf{P\Lambda P}^{T}$ and subsequently define a "projection" matrix $\mathbf{\Phi} = \mathbf{P\Lambda}^{-\frac{1}{2}}$. Now define $\mathbf{Y}^{*} = Y\Phi$ as the projections of $\mathbf{Y}$ on $\mathbf{\Phi}$, which now represent the decorrelated ("whitened") imaging outcomes, and now can be treated as independent random variables which enables fast fitting out our algorithms. Thus, $\mathbf{Y}^{*}$ follows a Normal($\mathbf{X}_{\gamma}{}^{*}$, $\mathbf{I}_{n}$) were $\gamma = \gamma\mathbf{\Phi}$. After the whitening procedure is perform, we use a L1 (Lasso) penalization for variable selection and shrinkage for each of the imaging outcomes $\mathbf{Y}^{*}$. Thus, the lasso penalization for each of the imaging outcomes i is:

$$\widehat{\gamma}^{*}_{i_{\lambda}} = argmin\left\{ \|y_i^{*} - X\gamma_i^{*}\|_2^2 + \lambda \|\gamma_i^{*}\|_1 \right\}$$

where $\lambda \in \mathfrak{R}^{+}$ is the regularization parameter, $\gamma_i^{*} = \left\{\gamma_i^{C*}, \gamma_i^{miR*} \gamma_i^{CN*}, \gamma^{R*}\right\}$ and λ is selected to be min(λ) that achieves 80% of the deviance/MSE. For interpretation we transform back ("un-whiten") the coefficients to the original scale by multiplying γ by $\mathbf{\Phi}^{-1}$

We expect the solution to be very sparse, since most of the genes will have no substantial contribution on the imaging outcome(s). To find the relative importance of each gene on the imaging outcome(s), we perform the bootstrap-based stability selection [7] procedure as described below:

Bootstrap Lasso Stability Selection

1. We pair ($y_{j1}, \ldots, y_{jk}, \mathbf{x}_j^T$) where $j = 1, \ldots n$ and $x_j^T = \left\{x_{j1}, \ldots, x_{jl}\right\}$.
2. for $b = 1, \ldots, B$, resample the $\left\{y_{1j}, \ldots, y_{kj}, x_j^T\right\}$
   - a. Decorrelate the k imaging outcomes
   - b. for $i = 1, \ldots, k$
   - c. $\widehat{\Gamma}_{i_\lambda}^{*(b)} = argmin\left\{\|y_i^{*(b)} - X^{(b)}\Gamma * \|_2^2 + \hat{\lambda}\|\Gamma * \|_1\right\}$, $\lambda = \min(\lambda \geq \text{Deviance}_{0.80})$ and $i = 1, \ldots, k$
   - d. "Un-whiten" $\widehat{\Gamma}_{\lambda}^{*(b)} = \left\{\widehat{\Gamma}_{1_\lambda}^{*(b)}, \ldots, \widehat{\Gamma}_{k_\lambda}^{*(b)}\right\} \Rightarrow \mathbf{\Gamma}^{(b)}$
3. repeat

Given the coefficients effect estimates for the B bootstrap samples, we determine the variable gene importance. We do this by first defining a practically minimum effect size δ ($\delta = 0.01$, our analysis). Note that for any bootstrap sample b, $\gamma_{i,l}^{(b)}$, is considered not important if it does not exceed the δ cut-off. Then, we compute a gene marker l importance "probability", $P(y_i, x_l) = \sum_{b=1}^{B} \mathbf{I}(|\gamma_{i,l}^{(b)}| \geq \delta)$ for each i imaging outcome. Finally, we select the gene markers for each of imaging outcomes as important if the $P(y_i, x_l) = \sum_{b=1}^{B} \mathbf{I}(|\gamma_{i,l}^{(b)}| \geq \delta) \geq 0.5, \forall i,l.$

Once the relevant genes are identified, we estimate the gene effects, by regressing the imaging outcome(s) on the relevant genes to learn which genes have a positive effect on the imaging imaging(s) and vice-versa.

IV. Simulations

In this section, we assess the performance of our models in terms of accuracy (sensitivity and specificity) of detecting associated genes. To better illustrate the basics features of our *imaging model*, we generate data with two distinct outcome variables based on our real data. We set $n = 100$, $X \sim \text{Normal}(0, 1)$ and $\mathbf{Y} \sim \text{Normal}(X\gamma, \mathbf{\Sigma})$. The correlation for the two imaging outcomes is set to $\rho = -0.5$. This correlation was chosen to mimic the correlation of the TCGA data imaging outcome(s). We simulated data for $p = (100, 250)$ predictors for γ_1 and γ_2, were 90% of γ_1 and γ_2 are set to 0, and the other 10% of Γ_1 and γ_2 are sample from Normal(4, 1) with probability 0.5 or from a Normal(−4, 1) with a probability 0.5. We allowed 5% both γ_1 and γ_2 to have an effect on the outcomes. For example, the regression coefficients for the first and second response could be:

$$\gamma_1 = (\overbrace{1, 1, \ldots, \overbrace{1, 1, \ldots, 1}^{5\%}}^{10\%}, \overbrace{0, 0, 0, 0, 0, 0, 0, 0, 0, \ldots, 0}^{90\%})$$
$$\gamma_2 = (\overbrace{0, 0, \ldots, 0}^{5\%}, \overbrace{1, 1, 1, \ldots, 1, 1, 1, 1}^{10\%}, \overbrace{0, 0, 0, \ldots, 0}^{85\%})$$

Now we apply our method to estimate the coefficients for our *imaging model* and evaluate the number of genes identified, the number of true positive rate and true negative rate which are shown in Table 1. Our simulation suggests that our method is very effective in selecting true positive and true negative genes. Our method mean γ_1 and γ_2 sensitivity is about 98% and mean specificity is about 92% for 200 parameters (100 each). The mean specificity and sensitivity for 500 parameters (250 each) is about 86% and is 82% respectively – thus demonstrating good performance in relevant gene selection.

V. Integrative analysis of GBM data

Radiological characterization/phenotyping was been done to identify survival-associated phenotypes on 83 TCGA-GBM patients. The two radiological characterization are T_1 and *ratio*. The T_1 feature represent the volume contrast enhancing region of the tumor as seen as a T_1-weighted post contrast MRI image. The ratio refers to the ratio of the T_1 post contrast image to the volume T_2-flair image. In our integrative analysis we use 83 matched tumor samples that have been assayed by expression, microRNA and copy number platforms.

A. Description of data

We focus on 49 genes of the RTK/PI3K, p53, and RB cell signaling pathways which is known to be an important pathway in cancer [8]. The level 3 processed data was downloaded from TCGA data portal by MD Anderson. The Karnofsky score (KPS) was downloaded from cBioPortal for Cancer Genomics [9] which resulted in the following data matrices:

- Radiological features (83×2) : T_1 is the volume enhancing region, *ratio* of the T_1 and *flair* hypertense volume
- Pathway-mRNA (83×49) contains the mRNA expression level for genes and patients.
- Pathway-CopyNumber (83×49) contains the genes log copy numnber and the patients
- Pathway-microRNA (83×534) contains the miRNA target sites and the patients

We removed six patients from our analysis. One TCGA patient is missing from the mRNA platform, another one is missing from the copy number platform, three more are missing from microRNA platform, and one more for being an outlier (TCGA-08-0352 flair=2.73E+05 T1=286 ratio=953.1). Thirteen KPS values were imputed. Therefore, at the end, we kept 77 patients for each platform. After performing a multivariate Yeo-Johnson transformation ($T_1^{0.29}$ and $ratio^{0.1}$), we apply our method to estimate the effects of our integrative model.

B. Results

The results indicate that 28 genes are significant from the 49 genes aberrations of the RTK/ PI3K, p53, and RB cell signaling pathways. In figure 1 and 2, we show the gene relative importance “probability” for the 49 genes group by copy number, miRNA and “other”. We consider a gene to be significant if the relative importance “probability” is greater than the cut-off 0.5, otherwise, it is classified as not significant. The genes with higher relative importance “probability” are blue in the heatmap and those that are not significant are gray. Few genes barely meet the cut-off, for example, the PDGFRB for the “other” biological mechanism importance “probability” is 0.51 for T_1 and .502 for *ratio*. We found 28 genes to be significant, 12 genes attributed through copy number (FGFR1, CDKN2B, CDKN2C, PIK3R1, CCND1, HRAS, PDGFRA, AKT3, CDK6, PIK3CD, MDM2 and AKT1), 2 genes attributed through miRNA (CCND2 and PDPK1) and 15 genes attributed through the “other” (MDM4, AKT2, ERBB2, PIK3CG, PDGFRA, ARAF, PIK3C2G, PIK3R1, PIK3CB, CBL, CDKN2C, TP53, MDM2, IRS1 and PDGFRB) biological mechanism in both T_1 and *ratio*. Only four genes CDKN2C, PIK3R1, PDGFRA and MDM2 were classified to be significant for both copy number and the “other” (e.g. methylation, transcriptional) mechanism.

After selecting the importance genes, we regress the two radiological phenotypes (T_1,*ratio*) on the estimated pieces for those relevant genes from the genomic model and the two clinical predictors (KPS, Age). In figure 3, we show that eight genes to be significant for the T_1 image outcome and eleven genes to be significant for the *ratio* image outcome. Six out of the eight significant genes from the T_1 image outcome have a negative effect (smaller tumor) and two have a positive effect (larger tumor) on T_1. Four out of the eleven significant genes from the *ratio* image outcome have a negative effect (greater infiltration) and seven have a positive effect (smaller infiltration) on *ratio*. We found that copy number attribute gene FGFR1 has antagonistic gene effect on the imaging T_1 and *ratio* outcomes.

VI. Conclusion

In this paper, we have introduced an innovative multivariate integrative model that integrates genomic data from various platforms with multiple imaging outcome(s). Our model uses a two level hierarchical approach to integrate biological relationships, and then uses this information to identify the significant genes that influence the imaging outcome(s). Our simulation shows that our model identifies true positive and true negative genes very effectively. We apply this model to integrate copy number, micorRNA and “other” expression from TCGA-GBM data. Our model helps us identify relevant genes effects from copy number, microRNA and “other” biological mechanisms influence the imaging outcome(s). In summary, the advantages from our model are to (i) identify genes effects that influence a disease, (ii) integrate a number of multiple platforms into one statistical model, (iii) model biological relationships among and between different platforms to the imaging outcome(s), (iv) handle high-dimensional data effectively and (v) relates effectively multiple correlated imaging outcome(s) to predictors from different genomic platforms. In the future, we plan to investigate non-linear extensions of our model which might be more biologically relevant in some instances. Another avenue of research might be to include a multiple pathways into our model to further explore the effects on imaging outcomes.

Acknowledgments

VB's research was partially supported by NIH grant R01 CA160736. Dr. Rao's research was supported by Institutional Start-up Funds. We would also like to thank Dr. Ganesh Rao, Mr. Juan Martinez and Dr. Payel Ghosh for help with volume segmentations and computations

References

1. Stupp R, Mason WP, Van Den Bent MJ, Weller M, Fisher B, Taphoorn MJ, Belanger K, Brandes AA, Marosi C, Bogdahn U, et al. Radiotherapy plus concomitant and adjuvant temozolomide for glioblastoma. New England Journal of Medicine. 2005.
2. The cancer genome atlas. [Online]. Available: http://cancergenome.nih.gov/abouttcga/overview
3. Jennings EM, Morris JS, Carroll RJ, Manyam G, Baladandayuthapani V. Hierarchical bayesian methods for integration of various types of genomics data. GENSiPS. 2012.
4. Dweep H, Sticht C, Pandey P, Gretz N. mirwalk–database: prediction of possible mirna binding sites by walking the genes of three genomes. Journal of Biomedical Informatics. 2011.
5. Yeo I-K, Johnson RA. A new family of power transformations to improve normality or symmetry. Biometrika. 2000.
6. Friedman J, Hastie T, Tibshirani R. Regularization paths for generalized linear models via coordinate descent. Journal of Statistical Software. 2010.
7. Meinshausen N, Bühlmann P. Stability selection. Journal of the Royal Statistical Society: Series B (Statistical Methodology). 2010.
8. McLendon R, Friedman A, Bigner D, Van Meir EG, Brat DJ, Mastrogianakis GM, Olson JJ, Mikkelsen T, Lehman N, Aldape K, et al. Comprehensive genomic characterization defines human glioblastoma genes and core pathways. Nature. 2008.
9. Gao J, Aksoy BA, Dogrusoz U, Dresdner G, Gross B, Sumer SO, Sun Y, Jacobsen A, Sinha R, Larsson E, et al. Integrative analysis of complex cancer genomics and clinical profiles using the cbioportal. Science Signaling. 2013.

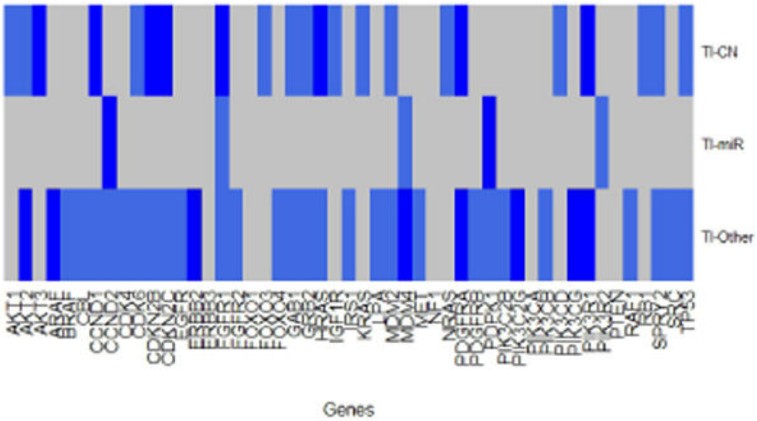

Fig. 1.

T_1 significant importance “probability” heatmap plot. A gene is consider significant if the “probability” of being chosen is greater than 0.5. The gray shaded markers are the genes that are not significant and the blue shaded markers are the significant genes.

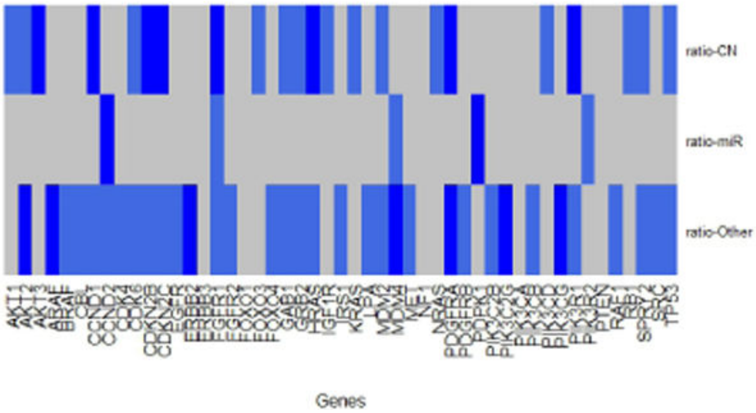

Fig. 2.

ratio significant importance “probability” heatmap plot. A gene is considered significant if the “probability” of being chosen is greater than 0.5. The gray shaded markers are the genes that are not significant and the blue shaded markers are the significant genes.

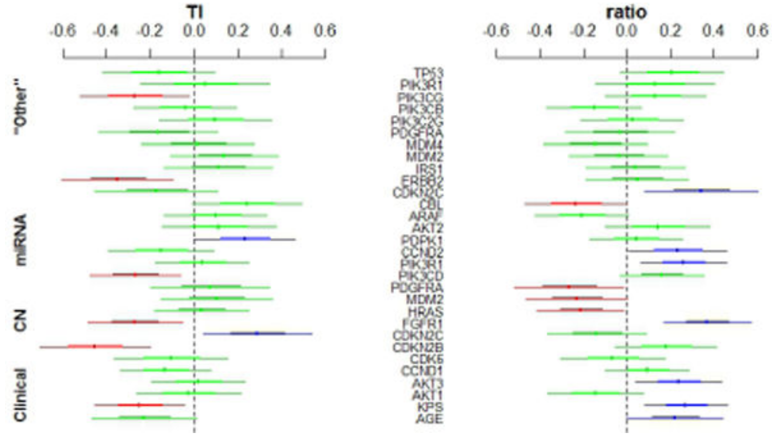

Fig. 3.

The direction of the important gene effects for the T_1 and *ratio* outcomes. Red: significant negative effect on T_1 (smaller tumor) and on *ratio* (greater infiltration); Blue: significant positive effect on T_1(lager tumor) and on *ratio* (less infiltration) ; Green: not significant

Table I

Simulations: sensitivity and specificity for various sample sizes and regression parameters.

		Sensitivity		Specificity	
# Parameters		γ_1	γ_2	γ_1	γ_2
200	mean	0.986	0.987	0.926	0.927
	stdev	0.0377	0.0368	0.0171	0.0171
500	mean	0.861	0.858	0.823	0.822
	stdev	0.07012	0.0719	0.0172	0.0284