HHS Public Access

Author manuscript

Genomics. Author manuscript; available in PMC 2017 June 01.

Published in final edited form as:

Genomics. 2016 June ; 107(6): 223–230. doi:10.1016/j.ygeno.2016.04.005.

Integrated Analysis of Multidimensional Omics Data on Cutaneous Melanoma Prognosis

Yu Jiang[a,b,#], **Xingjie Shi**[c,#], **Qing Zhao**[d], **Michael Krauthammer**[e], **Bonnie E. Gould Rothberg**[f], and **Shuangge Ma**[b,g,*]

[a]Division of Epidemiology, Biostatistics, and Environmental Health, School of Public Health, University of Memphis, Memphis, TN, 38152, USA

[b]VA Cooperative Studies Program Coordinating Center, West Haven CT, 06516, USA

[c]Department of Statistics, Nanjing University of Finance and Economics, Nanjing, China

[d]Merck Research Laboratories, 126 East Lincoln Avenue, RY34, Rahway, NJ, 07065, USA

[e]Department of Pathology, Yale University, New Haven CT, 06520, USA

[f]Cancer Center, Department of Internal Medicine, Pathology, Chronic Disease Epidemiology, Yale University, New Haven CT, 06520, USA

[g]Department of Biostatistics, Yale University, New Haven CT, 06520, USA

Abstract

Multiple types of genetic, epigenetic, and genomic changes have been implicated in cutaneous melanoma prognosis. Many of the existing studies are limited in analyzing a single type of omics measurement and cannot comprehensively describe the biological processes underlying prognosis. As a result, the obtained prognostic models may be less satisfactory, and the identified prognostic markers may be less informative. The recently collected TCGA (The Cancer Genome Atlas) data have a high quality and comprehensive omics measurements, making it possible to more comprehensively and more accurately model prognosis. In this study, we first describe the statistical approaches that can integrate multiple types of omics measurements with the assistance of variable selection and dimension reduction techniques. Data analysis suggests that, for cutaneous melanoma, integrating multiple types of measurements leads to prognostic models with an improved prediction performance. Informative individual markers and pathways are identified, which can provide valuable insights into melanoma prognosis.

Corresponding author: Dr. Shuangge Ma, 60 college ST, LEPH 206, New Haven, CT, 06520, USA, Fax: +1 203-785-6912, ; Email: shuangge.ma@yale.edu

[#]The two authors contributed equally to this work

Keywords

multidimensional omics data; melanoma prognosis; integration; The Cancer Genome Atlas (TCGA)

1. INTRODUCTION

Cutaneous melanoma poses a major public health concern. In 2015, an estimated 73,870 new cases of invasive melanoma are expected in the U.S., with an estimated 9, 940 deaths (Siegel, et al., 2015). Cutaneous melanoma is the largest subtype, and Caucasians have a much higher risk and poorer prognosis. Despite extensive research, the understanding of melanoma prognosis is still very limited. Clinicopathologic features that have been suggested as prognostic include age at diagnosis, gender, Breslow tumor thickness, ulceration status, mitotic index, and presence of lymph node micrometatases (Balch, et al., 2009; Dickson and Gershenwald, 2011). Significant effort has been devoted to searching for omics markers that may contribute to melanoma prognosis independent of the aforementioned factors. Several multi-marker prognostic models have been published. Omics markers identified in the literature belong to the immunomodulation, DNA repair, signal transduction, melanoma endophenotypes, and other pathways.

Identifying prognostic omics markers has important implications. For basic scientists, it leads to a better understanding of the biological mechanisms underlying prognosis. For translational researchers and physicians, it assists patient stratification, treatment selection, and prediction of prognosis paths.

In the literature, multiple types of omics changes have been suggested as potentially associated with melanoma prognosis. For mRNA expression, Winnepenninckx and others (2006) identified 254 genes associated with distant metastasis-free survival. Gene expression studies also include Timar et al. (2010), Gerami et al. (2015), and others. Studies of tumor cells in melanoma patients have characterized prognostic alterations with a panel of five genes in copy number alteration (CNA; Chiu, et al. 2014). MicroRNA has also been implicated in melanoma prognosis. For example, the study by Streicher and others (2012) identified a fourteen-microRNA cluster on the X chromosome, the miRNA-506–514 cluster, and found that this cluster is critical in cancer cell growth and melanocyte transformation. DNA methylation profile has been investigated. Notable studies include Conway et al. (2011) and a review study by Schinke and other (2010). Sigalotti and others (2012) analyzed methylation data and constructed a seventeen-gene signature. For genetic mutations, the associations of several somatic variants – such as BRAF V600E and NRAS Q61R/L/H – with prognosis have been reported (Bucheit and Davies, 2014; Wu et al. 2014). A whole-genome sequencing study found the RAC1 mutation as the third most frequent in sun-exposed melanomas and suggested its potential role in prognosis (Krauthammer et al. 2012).

A common limitation shared by many of the existing studies, especially the early ones, is that they are “one-dimensional” in the sense that they profiled and analyzed only a single type of omics measurement. Multiple types of omics measurements are interconnected and have possibly overlapping but also independent information. For example, CNAs,

microRNAs, methylation, and other changes affect gene expressions, which affect cancer outcomes/phenotypes through proteins. On the other hand, they can also directly affect protein expressions and functionalities through channels other than gene expressions. That is, they contain independent information on cancer outcomes not reflected in gene expressions. Analyzing a single type of omics measurement cannot comprehensively and accurately describe the biological processes underlying prognosis and may lead to suboptimal prognostic models and uninformative marker identification (Zhao et al. 2015).

More recently, much effort has been devoted to multidimensional studies which profile multiple types of omics changes on the same subjects. A representative example is TCGA (The Cancer Genome Atlas) which is organized by NIH. For multiple cancer types such as breast cancer, ovarian cancer, and glioblastoma, the integrated analysis of TCGA data has been conducted. More accurate prognostic models have been constructed, and important markers missed by the existing studies have been identified (Cancer Genome Atlas, 2012; Cancer Genome Atlas Research, 2014; Cancer Genome Atlas Research, 2014). For cutaneous melanoma, the TCGA data were very recently published, making it possible to conduct integrated analysis and more accurately describe its prognosis.

For several cancer types, multiple approaches have been applied to conduct the integrated analysis of multidimensional data. Some of the existing studies focus on the regulations among multiple types of omics measurements. Of special interest is the regulation of mRNA gene expression by miRNA, CNA, methylation, and other mechanisms (Feng, et al., 2013; Wang, et al., 2013), as gene expression is the downstream product and can be more directly related to clinical outcomes and phenotypes. Different from these studies, the present one is more concerned with linking omics measurements with prognosis, which is of more practical interest. Some other studies have analyzed each type of omics measurement separately and then compare results across multiple types of measurements. This is basically a meta-analysis strategy and suitable for identifying “hot zones” that host multiple omics changes. However as prognosis is affected by the joint effects of multiple types of omics changes, such an approach may not be effective in building prognostic models.

Overall, this study may complement the existing literature and be warranted in the following aspects. First, it provides a timely integrated analysis of the TCGA cutaneous melanoma data, and the results may provide insights into this clinically important disease. Second, it describes in detail how to conduct effective integrated analysis of multiple types of omics data using advanced statistical techniques and proper statistical packages, which are potentially applicable to many other datasets and diseases.

2. METHODS

2.1 TCGA cutaneous melanoma data

TCGA is one of the largest and most comprehensive multidimensional cancer studies. For cutaneous melanoma, the goal was to collect data on about 500 samples. The protocols of TCGA sample and data collection have been described in detail elsewhere (Cancer Genome Atlas Research, 2015). Data analyzed in this study were downloaded either directly from the TCGA website or from cbioportal using the CGDS-R package. Brief data information is

provided in Table 1, and the flowchart of data processing is provided in the top part of Figure 1.

For clinical and pathological variables, the preprocessed level 3 data were downloaded. The number of samples with available data is 422. In the analysis, only white metastatic samples are included. Data on the normal samples are excluded, and multiple data records on the same samples are merged. Only variables with missing rates below 40% are considered. Among them, those that have been suggested as potentially associated with melanoma prognosis include: gender, age at diagnosis, tumor status, Breslow thickness at diagnosis, Clark level at diagnosis, primary melanoma tumor ulceration, AJCC tumor pathologic stage, AJCC nodes pathologic stage, new tumor event, percent of lymphocyte infiltration, percent of monocyte infiltration, percent of necrosis, percent of stromal cells, percent of tumor cells, and percent tumor nuclei. The following variable recoding is conducted to facilitate analysis (by reducing cells with very small counts). The AJCC tumor pathologic stage is coded as 0 for T0 and Ts, 1 for T1-T3, and 2 for T4. The AJCC nodes pathologic stage is coded as 0 for N0 and Nx, 1 for N1, 2 for N2, and 3 for N3. After processing, data are available for 16 variables and 317 samples. To accommodate the remaining missing measurements, multiple imputation is conducted using the package *Amelia* (Honaker et al. 2011).

Omics data were downloaded from cbioportal using the CGDS-R package. Mutation data are available on 278 samples. Following a recent study (Jayawardana et al. 2015), mutation data on NRAS and BRAF are included in analysis. For a sample, the mutation status is coded as 1 if there is at least one mutation in the specific gene, and as 0 otherwise. In addition, attempt has been made to incorporate all mutation data in analysis. It is found that, with the extremely high dimensionality and noisy nature of mutation data, including all mutations leads to inferior prediction performance (details omitted). Thus, only the two most important mutations are analyzed. CNA measurements were obtained using the Affymetrix Genome-wide Human SNP array 6.0 platform. The loss and gain levels of copy number changes of tumors compared to normal tissues were identified using segmentation analysis and expressed in the log2 transformed form. A total of 21,699 measurements are available on 366 samples. DNA methylation at CpG sites was measured using the Illumina Human Methylation 450 platform. The available data contain the beta values, which represent the percentages of methylation, for 15,589 genes and 373 samples. The range of the beta values is from 0 (fully unmethylated) to 1 (fully methylated). mRNA gene expressions were measured using the Illumina Hiseq RNAseq V2 platform. The downloaded data are the robust Z-scores which have been lowess-normalized, log-transformed, and median-centered and represent the gene expression status (up or down regulated) in tumor samples relative to normal tissues. A total of 19,626 measurements are available on 371 samples.

Besides the aforementioned omics measurements, TCGA also has miRNA data, which, however, are not available from cbioportal. The miRNA data are not analyzed in this study with the concern on data source consistency. In addition, both the TCGA website and cbioportal have protein data. However measurements are only available on 129 protein expressions and 204 samples. A closer examination suggests that including the protein data in analysis leads to a significant reduction in sample size and hence is not pursued in this study.

As shown in Figure 1, different types of data are merged using sample ID. A total of 253 samples have data on clinical variables, mutation status, gene expression, methylation, and CNA. Gene expression, methylation, and CNA measurements are high dimensional. The integrated analysis methods described below are capable of analyzing all available data. However, with concern on the stability of estimation and the fact that the number of prognosis-associated markers is expected to be small, we follow the literature (Zhao et al. 2015) and conduct a supervised prescreening. More specifically, we fit a Cox regression model for each gene expression, methylation, and CNA measurement. For each type of measurement, the p-values of regression coefficients are sorted, and the 2,500 measurements with the smallest p-values are selected for down-stream analysis. Note that to avoid bias, the prescreening needs to be conducted in each prediction evaluation run described below.

The prognosis outcome of interest is overall survival, which has been analyzed in recent studies (Mrazek and Chao, 2014). Among the 253 samples, 121 died during followup. For them, the median survival time is 103.2 months, with 95% confidence interval (72.0, 151.2) months. For those censored, the median followup time is 54.8 months, with standard deviation 59.5 months. Brief demographic information on the samples is provided in Appendix.

2.2 Methods for integrated analysis

Accommodating the high data dimensionality—Denote T as the survival time and C as the random censoring time. Under right censoring, one observes ($\tilde{T} = \min(T, C)$, $\delta = I(T \leq C)$). Assume n iid observations. Even with processing, the number of omics measurements remained for analysis is still dramatically larger than the sample size. We adopt the following variable selection and dimension reduction techniques to accommodate the high dimensionality. First consider a single type of omics measurement. Take gene expression as an example. Denote $X = (X_1, \ldots, X_d)'$ as the d gene expressions.

Enet: Elastic net is a *variable selection* technique. It applies penalization and shrinks the small coefficients in a regression model to zero, and only selected variables with nonzero regression coefficients are identified as important and used for model building. For detailed discussions on Enet and other penalized regularization techniques, refer to Zou and Hastie (2005) and others. Consider the Cox proportional hazard model, where the conditional hazard function is $\lambda(T|X) = \lambda_0(T)\exp(\beta' X)$ Here $\lambda_0(T)$ is the unknown baseline hazard function, and $\beta = (\beta_1, \ldots \beta_d)'$ is the d-vector of regression coefficients. Denote $l(\beta)$ as the log partial likelihood function. The Enet estimate is defined as

$$\hat{\beta}(\lambda) = \operatorname{argmin}\{-\ell(\beta) + \lambda\alpha\|\beta\|_1 + \lambda(1-\alpha)\|\beta\|^2\},$$

where $\|\beta\|_1 = \Sigma_j|\beta_j|$, $\|\beta\|^2 = \Sigma_j\beta_j^2$, and $0 \leq \alpha \leq 1$ and $\lambda > 0$ are tuning parameters. By changing the value of, Enet includes the popular Lasso and ridge penalties as special cases and are more flexible. Gene expressions corresponding to the nonzero components of $\hat{\beta}$ are identified as important and associated with prognosis. The number of identified genes is jointly controlled by λ and α, with a larger value of λ leading to fewer identified genes. In

our analysis, Enet is realized using the R package *glmnet*. α is set as 0.5 to balance the two penalties, as in the literature. The selection of λ is discussed below.

SPCA: Principal component analysis (PCA) is one of the most commonly adopted *dimension reduction* techniques. It constructs k linear combinations of the original variables, denoted as $Z = (Z_1,\ldots,Z_k)'$, and uses them in downstream regression analysis. Here usually $k \ll n$, and thus standard model fitting techniques can be applied. Z_j's, which are referred to as the principal components (PCs), are orthogonal to each other and hence solve the collinearity problem in regression. Z_1 explains the most variation, followed by Z_2 The "classic" PCA can be realized using singular value decomposition (SVD). One problem of the PCA is that each PC is a linear combination of *all* original variables. Thus when PCs are used in regression, all original variables enter the model, and the results are difficult to interpret. In addition, with a large number of variables and a small number of samples, estimating the loadings of the PCs may not be reliable. To tackle these problems, the sparse PCA (SPCA) technique has been developed (Witten et al. 2009). It applies penalization to achieve sparsity in PCs, so that each PC is composed of a smaller number of variables. In data analysis, SPCA is realized using the R package *PMA*. With a slight abuse of notation, still use $Z = (Z_1,\ldots,Z_k)'$ to denote the sparse PCs. Consider the Cox model with conditional hazard function $\lambda(T|X) = \lambda_0(T)\exp(\gamma' Z)$, where γ is the length-k vector of regression coefficients. With $k \ll n$, this model can be fit in a standard manner using the R package *survival*.

SPLS: Partial least squares (PLS) is another dimension reduction technique. Different from PCA which constructs the new variables in an unsupervised manner, PLS takes a supervised approach. With a continuous response, PLS first searches for a linear combination of the original variables that has the highest correlation with the response. The remaining linear combinations are constructed in a similar manner, with the constraint of being orthogonal to the previous linear combination(s). PLS has been extended to censored survival data. Notably, in Nguyen and Rocke (2004), a two-step approach is developed. In the first step, linear regression is used to determine the PLS components; And in the second step, the Cox regression is applied. In Bastien (2004), the linear regression step is replaced by the Cox regression. Similar to PCA, a limitation of the standard PLS is that the constructed variables consist of all of the original variables. The sparse PLS (SPLS) has been developed (Chun and Keles, 2010), sharing a similar spirit with the SPCA. With high dimensional data, to reduce computational cost, the approach proposed in Chun and Keles (2010) is adopted which replaces the survival times by the deviance residuals in extracting the PLS components. This approach has been shown to have a good approximation performance. In data analysis, it is realized using the R package *plsRCox*.

Remarks: Penalized variable selection has been extensively applied to cancer genetic data. Compared to some other penalties, Enet has a ridge term and thus is capable of accommodating highly correlated markers, which are common in genetic analysis. It is also computationally simpler than some alternatives. It is noted that a few other penalization techniques may also be applicable for the present problem. PCA and PLS are the two most popular dimension reduction techniques. Tailored to the high dimensionality of omics data,

their sparse versions are adopted. As our goal is not to compare different penalization and dimension reduction methods, only the aforementioned three are adopted.

Integrating multiple types of omics measurements—To integrate clinical variables with multiple types of omics measurements, the following approach is adopted. First consider the Enet-based analysis. (a) For clinical variables with or without mutation status, fit a Cox model with Enet for variable selection. With the relatively low dimensionality of data, select the tuning parameter λ in a way such that five variables are identified. (b) For a single type of omics measurement, fit a Cox model with Enet. Select the tuning parameter so that ten variables are identified. (c) Consider a Cox model with the additive effects of variables selected in Steps (a) and (b). As shown in Table 2, multiple combinations are considered. As the number of variables in the model is not small compared to the sample size, ridge regression is adopted to generate stable estimation. This is also realized using the R package *glmnet* with the tuning parameter selected using cross validation. With SPCA and SPLS to accommodate the high dimensionality, the analysis approach is similar. The difference is that in Steps (a) and (b), the sparse PCs and sparse PLS components take the place of individual variables.

The above approach first extracts important features from each type of omics measurement. The numbers of features in Steps (a) and (b) may be somewhat subjective. They are chosen with the consideration that the numbers of prognosis-associated features are expected to be small. In data analysis, we have also experimented with a larger number of selected features but found less stable estimation and inferior prediction. An additive model is adopted in Step (c), which accommodates the contribution from all types of measurements. It is noted that some measurements may contain information heavily overlapping with other measurements. The ridge penalized estimation can appropriately accommodate such a scenario.

2.3 Assessing prediction performance of the integrated models

To assess the predictive power of the integrated models listed in Table 2, the ideal scenario is to apply them to independent testing data. As the TCGA data are very unique, comparable independent data are difficult to identify. Thus the following cross-validation based approach is adopted: (a) randomly split data into a training and a testing set with sizes 4:1; (b) apply the approach described in the above subsection, fit the Cox model, and obtain parameter estimates; (c) use the training set model and testing set samples to compute the predicted risk scores; (d) compute the C-statistic (Uno et al. 2011) to quantify prediction performance; (e) to avoid an extreme split, repeat Steps (a)-(d) 200 times and compute the mean and standard deviation of the C-statistics. The flowchart is shown in the bottom part of Figure 1.

The nonparametric C-statistic is a special case of the time-integrated AUC (area under curve) under the time-dependent ROC framework. It is essentially a rank correlation measure and takes values between 0.5 and 1. A larger value indicates better prediction, and a value of 0.5 corresponds to a model with no predictive power. It has been adopted in multiple studies (Riester et al. 2012; Schroder et al. 2011). In data analysis, it is realized using the R package *survAUC*.

Remarks—The procedure described above covers the whole spectrum of analysis, from processing to estimation to evaluation. It is more comprehensive than quite a few of the existing studies. In addition, the statistical techniques, all developed in the recent literature, are more effective than the "classic" techniques adopted in some existing studies. An advantage of the proposed procedure is that it can be realized using the existing software packages, and the computational cost is much affordable with practical data. To facilitate future applications, we have compiled code used in analysis and submitted along with this article.

3. RESULTS

3.1 Prediction performance of the integrated models

For models with a single or multiple types of measurements, the summary C-statistics are shown in Table 2. Multiple insightful observations can be made. The first is that prediction performance depends on the approach taken to accommodate high data dimensionality. For this specific dataset, SPLS in general has better prediction than Enet and SPCA. For example with methylation, SPLS has a mean C-statistic of 0.702, compared to 0.605 of Enet and 0.668 of SPCA. The inferior prediction of Enet may be caused by both the shrinkage and sparsity properties. Unlike PLS, PCA is constructed in an unsupervised manner. Thus, the constructed PCs may not contain sufficient information on prognosis. The second observation is that the relative prediction performance of one type of measurement also depends on the analysis approach. With Enet, it is observed that the clinical variables have the best prediction performance with a mean C-statistic of 0.708, and CNA has almost no predictive power. With SPCA, the C-statistic of methylation is the largest (0.668). The third and the most important observation is that integrated analysis can lead to models with improved prediction. With Enet, the model with "clinical variables + mutation + methylation" has the best prediction performance, with a mean C-statistic of 0.724. With SPCA, the model with "clinical variables + mutation + methylation + CNA" has the best prediction, with a mean C-statistic of 0.718. Among the multiple constructed models, the one with the highest mean C-statistic (0.746) is obtained using SPLS and "clinical variables + mutation + gene expression + methylation". T-tests are conducted to compare the C-statistics of the best models against those of the models with only clinical variables. The corresponding p-values are 0.015 (Enet), <0.001 (SPCA), and <0.001 (SPLS), respectively, indicating significant differences. For all three methods, the DNA methylation profile is included in the models with the best prediction. Using the best models, we compute the samples' risk scores, dichotomize at the medians, and create two risk groups. The corresponding survival curves are shown in Figure 2, and the p-values are computed using the logrank tests. The separation between the two risk groups is clear.

In the existing analysis of prognosis data on melanoma and other cancer types, usually just a single analysis method is applied. Our analysis of the TCGA data suggests that prediction performance can depend on the specific analysis method adopted. Thus for melanoma and other cancers, it is prudent to experiment with multiple methods. For some other cancer types, it has been observed that integrated analysis can lead to improvement in prediction. It has also been suggested that the specific combinations of omics measurements with the best

prediction is cancer type-dependent (Zhao et al. 2015). With the TCGA melanoma data, our analysis confirms the improvement in prediction. The best prediction model contains clinical variables. The set of clinical variables collected by TCGA has been manually selected based on previous epidemiological studies and is expected to be prognostic. This model also contains gene expression. A recent Australian study suggested the superior predictive power of gene expression (Jayawardana et al. 2015). In the analyzed data, gene expression is the downstream product of other omics changes and "closest" to clinical outcomes. The model also includes BRAF and NRAS mutation status and methylation, which may affect prognosis through gene expression as well as other independent channels.

3.2 Identified individual markers and pathways

Enet—Using Enet, a small number of variables are identified, and their estimates and corresponding hazard ratios are shown in Table 3. Among the clinical variables and mutation status, the five identified are tumor status, Breslow thickness at diagnosis, Clark level at diagnosis, age at diagnosis, and gender. All have positive regression coefficients, suggesting that their higher values correspond to a higher risk and shortened survival. The findings are consistent with those in the literature (Jayawardana et al. 2015).

The identified top ten methylation loci are also shown in Table 3. The most interesting finding is HLA.C. A closer examination of data shows that 94 out of the 253 samples have a beta value larger than 0.7 (which is a commonly used cutoff value for hypermethlation). It is not surprising to observe that the gene expressions of samples with hypermethylated HLA.C status are significantly lower than those of the rest (p-value=1.687e-8). Samples with the hypermethylation of HLA.C tend to have a higher Clark level (p-value=0.007) and more advanced tumor status (p=0.012). Eight genes are found to be highly correlated with HLA.C, which are PSMB9, HLA.B, HLA.E, HLA.G, PSMB8, HLA.F, IRF1, and B2M. Their corresponding gene expressions are down-regulated due to hypermethylation. Samples that have hypermethylated HLA.C and correlated genes tend to have a poorer survival. The results are consistent with the finding of decreased HLA class I molecule in melanoma cells with the degree of de-differentiation of the tumor and increased malignancy (Carretero et al. 2008). The down-regulation of HLA class antigen can be linked to an important cancer immune escape mechanism. It is noted that most of the identified methylation are located on chromosome 6. The implication of instability of chromosome 6 has been studied (Santos et al. 2007). It has also been observed that genetic changes on chromosome 6 are highly associated with the expression of gene BCL2, which plays an important role in programmed cell death.

The identified gene expressions are also shown in Table 3. Beyond those genes, analyzing the original data also suggests a few genes expressions highly correlated with those identified. The network structure of the identified genes and their highly correlated ones are shown in the figure next to the table. Specifically, three clusters of genes are identified as associated with prognosis – the up-regulation of PI3K–related genes, up-regulation of TRIM32-related genes, and down-regulation of PARP-related genes are associated with poorer prognosis. The PI3K pathway has been suggested as a core pathway for melanoma development (Davies 2012). The PI3K gene has been widely studied as a potential

therapeutic target for melanoma (Russo et al. 2014). A phase 1/2 clinical trial of PI3K inhibitor BKM120 combined with Vemurafenib (PLX4032) is currently in progress (clinicaltrials.gov/ct2/show/NCT01512251). MAP2K2 (MEK2) is also a known therapeutic target (Flaherty et al. 2012). TRIM32 is an E3-ubiquitin ligase and has been found to be involved in both cancer and human development by negatively regulating tumor suppressor p53 (Liu et al. 2014).

SPCA—Even though the PCs have been sparsified, they still include a considerable number of variables. As opposed to looking into individual genes, we map the genes to pathways and use the estimated regression coefficients to compute the pathway norms for the best model. A better understanding of the pathway functions and networks may help identify the molecular targets for therapy (Smalley 2010). As shown in Table 2, the model that has the highest C-statistic is “clinical + mutation + methylation + CNA”. Thus, the pathway norms are computed based on the methylation and CNA measurements. The top pathways with the largest norms are shown in Table 4. The top three are the Mitogen Activated Protein Kinase (MAPK), integrin, and CSK pathway. The MAPK pathway plays an important role in melanoma development. Dysregulation of the MAPK pathway is partially due to the mutation of BRAF and RAS and other genetic modifications. Its activation leads to increased cell proliferation, metastasis, migration, and angiogenesis. The significant role of MAPK in melanoma prognosis has been studied intensively. This pathway has also been identified as a therapeutic target for melanoma treatment (Inamdar et al., 2010). The second pathway is integrin. Proteins in the integrin family are the major cell surface receptors that respond to extracellular matrix. Upon extracellular stimuli, such proteins activate cellular responses, such as cell proliferation, cytoskeletal reorganization, and cell survival. In addition, the integrin pathway has been found to crosstalk with the PI3K and MAPK pathways during tumor progression (Guo and Giancotti, 2004). The specific role of the integrin pathway in melanoma has been discussed in Bosserhoff (2011) and others. The CSK pathway is related to T cell activation. The pathway norms have been computed using both methylation and CNA measurements. In Table 4, we also decompose the two types of measurements. It is observed that for the top ten pathways, the norms of methylation measurements are considerably larger than those of CNAs. Thus, the contributions of the identified pathways to prognosis may have been more heavily driven by methylation. Such an observation has not been made in the literature and may demand additional attention.

SPLS—The analysis is similar to that with SPCA. The model with the best prediction contains “clinical + mutation + methylation + gene expression”. Thus the pathway norms are computed using the methylation and gene expression measurements. The top ten pathways with the largest norms are shown in Table 4. The top three are the MAPK, IL6, and NFAT pathways. Same as SPCA, the top is the MAPK pathway. IL-6 is a pleiotropic cytokine, mainly secreted by T cells and macrophages. The binding of IL-6 to its receptors activates the downstream signaling of JAK/STA. Studies have shown that both IL6 and IL10 appear to be involved in the progression of melanoma. Higher levels of serum IL6 and IL10 are associated with poorer survival (Moretti et al., 2001). The NFAT pathway is important in T-cell development and function. In addition, the HIVNEF and NTHI pathways also play important roles in innate immune response. When decomposing the pathway norms, we find

that the contributions from methylation and gene expression are largely similar. Thus under this model, prognosis is found to be driven by both types of measurements.

4. DISCUSSION

In this study, we have described in detail an integrated analysis approach that can aggregate information from multiple types of omics measurements and build prognosis models. Cancer omics data have extremely high dimensionality, are noisy, and contain sparse signals. Performance of different methods is data-dependent. It is thus prudent to develop and apply multiple techniques. The proposed procedure is comprehensive in that it covers data processing, model construction, and evaluation. It is built on advanced variable selection, dimension reduction, and cross validation techniques. It is noted that although the proposed models and procedure are only applied to cutaneous melanoma in this study, they are relatively “data-independent”, and thus it is expected that they are also applicable to other datasets and cancer types, and possibly other types of diseases too.

The recently published TCGA data on the prognosis of cutaneous melanoma is analyzed. The most important finding is that data integration leads to a significant improvement in prediction. The inferiority of the models with only clinical variables suggests the necessity of omics profiling. It is noted that the observed increase in C-statistic is not “dramatic”, which is reasonable. The prognosis of cancer is a very complex process. Besides omics risk factors, environmental and socioeconomic factors, treatment, and others also play important roles. Thus omics measurements are only expected to be able to explain a certain percentage of variation in prognosis. On the other hand, the observed improvement can be of significant importance at the population level. In future studies, it can be of interest to collect and analyze data that can lead to more accurate prognosis models at the individual level.

In our data analysis, the SPLS integrated model has the best prediction performance and deserves further attention. We have examined individual findings. With Enet, meaningful individual markers have been identified. With SPLS and SPCA, the optimal models contain different sets of measurements. However, it is interesting to notice that the top pathways are similar, which partly supports the validity of findings. Different types of omics measurements are connected. Especially, gene expression is regulated by both methylation and copy number variation. Thus it is reasonable to expect that the model with “gene expression + methylation” and that with “methylation + CNA” contain overlapping information, which can explain the similarity in the top pathways in Table 4. It is interesting to note that with all three approaches, the DNA methylation profile is included in the models with the best prediction. Although methylation has been linked to the prognosis of melanoma and other cancer types, there is still a lack of study investigating its superior importance in prognosis.

Our data analysis can be potentially improved in multiple aspects. For example, there is a lack of cost-effectiveness analysis. In practice, the prognostic power of a model may need to be considered along with its cost. In addition, cutaneous melanoma, as other cancer types, is heterogeneous, and subset analysis may be needed. This is not conducted with concern on sample size. We have conducted comparison with the models built on only clinical variables.

A few such models have been developed in the literature. We did not compare with these models because of data unavailability.

The integrated analysis of melanoma prognosis has also been pursued by others. The most notable is Jayawardana et al (2014). The present study advances from Jayawardana et al (2014) in multiple ways. First, the dataset in Jayawardana et al (2014) is much smaller and thus has a lower power. In addition, Jayawardana et al (2014) analyzes a binary prognosis outcome (good or poor prognosis), which can be less informative than the actual survival time analyzed in this study. In addition, to be prudent, we have applied three approaches to accommodate the high data dimensionality. The data-dependent nature of performance may also be true for other datasets/cancer types. To the best of our knowledge, this study is the first to conduct the integrated analysis of TCGA melanoma data. This dataset has a higher quality and a larger sample size than the other datasets. In addition, the statistical techniques adopted in this study are more advanced and effective. The observed improvement in prediction of the integrated models and identified individual markers/pathways deserve further investigation. This study also demonstrates the necessity of developing and implementing multiple analysis approaches given the complexity of multidimensional data.

For TCGA data on other cancer types as well as other databases, the proposed procedure can also be rigorously applied. However, it is hard to “predict” what the optimal models and their performance are. We postpone examining other datasets to future studies.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

We thank the editor and reviewers for their careful review and insightful comments, which have led to a significant improvement of the article.

Funding: This work was supported by the National Institutes of Health (CA182984, CA142774, P50CA121974, P30CA016359), the National Social Science Foundation of China (13CTJ001, 13&ZD148), and the VA Cooperative Studies Program of the Department of Veteran Affairs, Office of Research and Development.

REFERENCES

Balch CM, et al. Final version of 2009 AJCC melanoma staging and classification. Journal of clinical oncology : official journal of the American Society of Clinical Oncology. 2009; 27(36):6199–6206. [PubMed: 19917835]

Bastien P. Deviance residuals based PLS regression for censored data in high dimensional setting. Chemometr Intell Lab. 2008; 91(1):78–86.

Bucheit AD, Davies MA. Emerging insights into resistance to BRAF inhibitors in melanoma. Biochemical pharmacology. 2014; 87(3):381–389. [PubMed: 24291778]

Bosserhoff, A. Melanoma development: molecular biology, genetics and clinical application. Springer Science & Business Media; 2011.

Cancer Genome Atlas, N. Comprehensive molecular portraits of human breast tumors. Nature. 2012; 490(7418):61–70. [PubMed: 23000897]

Cancer Genome Atlas Research, N. Comprehensive molecular characterization of urothelial bladder carcinoma. Nature. 2014; 507(7492):315–322. [PubMed: 24476821]

Cancer Genome Atlas Research, N. Comprehensive molecular profiling of lung adenocarcinoma. Nature. 2014; 511(7511):543–550. [PubMed: 25079552]

Cancer Genome Atlas Research, N. Genomic Classification of Cutaneous Melanoma. Cell. 2015; 161(7):1681–1696. [PubMed: 26091043]

Carretero R, et al. Analysis of HLA class I expression in progressing and regressing metastatic melanoma lesions after immunotherapy. Immunogenetics. 2008; 60(8):439–447. [PubMed: 18545995]

Chiu CG, et al. Genome-wide characterization of circulating tumor cells identifies novel prognostic genomic alterations in systemic melanoma metastasis. Clinical chemistry. 2014; 60(6):873–885. [PubMed: 24718909]

Chun H, Keles S. Sparse partial least squares regression for simultaneous dimension reduction and variable selection. J Roy Stat Soc B. 2010; 72:3–25.

Conway K, et al. DNA-methylation profiling distinguishes malignant melanomas from benign nevi. Pigment cell & melanoma research. 2011; 24(2):352–360. [PubMed: 21375697]

Davies MA. The Role of the PI3K–AKT Pathway in Melanoma. Cancer J. 2012; 18(2):142–147. [PubMed: 22453015]

Dickson PV, Gershenwald JE. Staging and prognosis of cutaneous melanoma. Surgical oncology clinics of North America. 2011; 20(1):1–17. [PubMed: 21111956]

Feng J, et al. Screening biomarkers of prostate cancer by integrating microRNA and mRNA microarrays. Genetic testing and molecular biomarkers. 2013; 17(11):807–813. [PubMed: 23984644]

Flaherty KT, et al. Improved Survival with MEK Inhibition in BRAF-Mutated Melanoma. New Engl J Med. 2012; 367(2):107–114. [PubMed: 22663011]

Gerami P, et al. Development of a prognostic genetic signature to predict the metastatic risk associated with cutaneous melanoma. Clinical cancer research : an official journal of the American Association for Cancer Research. 2015; 21(1):175–183. [PubMed: 25564571]

Guo W, Giancotti FG. Integrin signalling during tumour progression. Nat Rev Mol Cell Biol. 2004 Oct; 5(10):816–826. [PubMed: 15459662]

Honaker J, King G, Blackwell M. Amelia II: A Program for Missing Data. J Stat Softw. 2011; 45(7):1–47.

Inamdar GS, et al. Targeting the MAPK pathway in melanoma: why some approaches succeed and other fail. Biochem Pharmacol. 2010; 80(5):624–637. [PubMed: 20450891]

Jayawardana K, et al. Determination of prognosis in metastatic melanoma through integration of clinico-pathologic, mutation mRNA, microRNA, protein information International journal of cancer. Journal international du cancer. 2015; 136(4):863–874. [PubMed: 24975271]

Kim C, et al. Long-term survival in patients with metastatic melanoma treated with DTIC or temozolomide. The oncologist. 2010; 15(7):765–771. [PubMed: 20538743]

Krauthammer M, et al. Exome sequencing identifies recurrent somatic RAC1 mutations in melanoma. Nature genetics. 2012; 44(9):1006–1014. [PubMed: 22842228]

Liu J, et al. E3 ubiquitin ligase TRIM32 negatively regulates tumor suppressor p53 to promote tumorigenesis. Cell Death Differ. 2014; 21(11):1792–1804. [PubMed: 25146927]

Mrazek AA, Chao C. Surviving cutaneous melanoma: a clinical review of follow-up practices, surveillance, and management of recurrence. The Surgical clinics of North America. 2014; 94(5): 989–1002. vii–viii. [PubMed: 25245963]

Moretti S, et al. Serum imbalance of cytokines in melanoma patients. Melanoma Res. 2001; 11(4): 395–399. [PubMed: 11479428]

Nguyen DV, Rocke DM. On partial least squares dimension reduction for microarray-based classification: a simulation study. Comput Stat Data An. 2004; 46(3):407–425.

Riester M, et al. Combination of a novel gene expression signature with a clinical nomogram improves he prediction of survival in high-risk bladder cancer. Clinical Cancer Research. 2012; 18(5):1323–1333. [PubMed: 22228636]

Russo A, et al. Emerging targeted therapies for melanoma treatment (Review). Int J Oncol. 2014; 45(2):516–524. [PubMed: 24899250]

Santos GC, et al. Chromosome 6p amplification and cancer progression. J Clin Pathol. 2007; 60(1):1–7. [PubMed: 16790693]

Schinke C, et al. Aberrant DNA methylation in malignant melanoma. Melanoma research. 2010; 20(4): 253–265. [PubMed: 20418788]

Schroder MS, et al. survcomp: an R/Bioconductor package for performance assessment and comparison of survival models. Bioinformatics. 2011; 27(22):3206–3208. [PubMed: 21903630]

Siegel RL, Miller KD, Jemal A. Cancer statistics, 2015. CA: a cancer journal for clinicians. 2015; 65(1):5–29. [PubMed: 25559415]

Smalley KSM. Understanding Melanoma Signaling Networks as the Basis for Molecular Targeted Therapy. J Invest Dermatol. 2010; 130(1):28–37. [PubMed: 19571822]

Streicher KL, et al. A novel oncogenic role for the miRNA-506-514 cluster in initiating melanocyte transformation and promoting melanoma growth. Oncogene. 2012; 31(12):1558–1570. [PubMed: 21860416]

Tas F. Metastatic behavior in melanoma: timing, pattern, survival, and influencing factors. Journal of oncology. 2012; 2012:647684. [PubMed: 22792102]

Timar J, Gyorffy B, Raso E. Gene signature of the metastatic potential of cutaneous melanoma: too much for too little? Clinical & experimental metastasis. 2010; 27(6):371–387. [PubMed: 20177751]

Uno H, et al. On the C-statistics for evaluating overall adequacy of risk prediction procedures with censored survival data. Stat Med. 2011; 30(10):1105–1117. [PubMed: 21484848]

Uzdensky AB, Demyanenko SV, Bibov MY. Signal Transduction in Human Cutaneous Melanoma and Target Drugs. Curr Cancer Drug Tar. 2013; 13(8):843–866.

Wang W, et al. iBAG: integrative Bayesian analysis of high-dimensional multiplatform genomics data. Bioinformatics. 2013; 29(2):149–159. [PubMed: 23142963]

Winnepenninckx V, et al. Gene expression profiling of primary cutaneous melanoma and clinical outcome. Journal of the National Cancer Institute. 2006; 98(7):472–482. [PubMed: 16595783]

Witten DM, Tibshirani R, Hastie T. A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis. Biostatistics. 2009; 10(3):515–534. [PubMed: 19377034]

Wu S, et al. Association between BRAFV600E and NRASQ61R mutations and clinicopathologic characteristics, risk factors and clinical outcome of primary invasive cutaneous melanoma. Cancer causes & control: CCC. 2014; 25(10):1379–1386. [PubMed: 25048604]

Zhao Q, et al. Combining multidimensional genomic measurements for predicting cancer prognosis: observations from TCGA. Briefings in bioinformatics. 2015; 16(2):291–303. [PubMed: 24632304]

Zou H, Hastie T. Regularization and variable selection via the elastic net. J Roy Stat Soc B. 2005; 67:768–768.

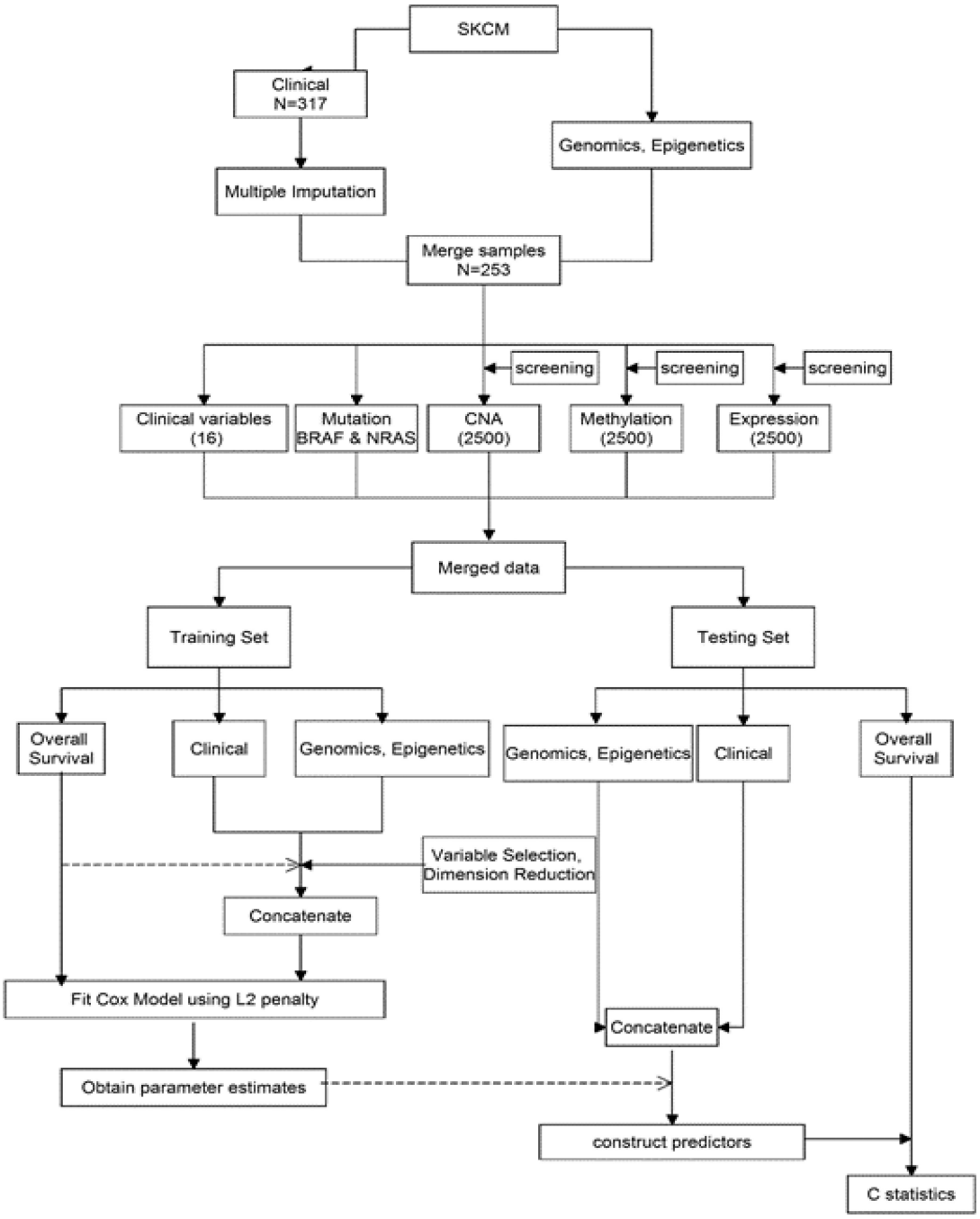

Figure 1.
Flowchart of data processing and analysis

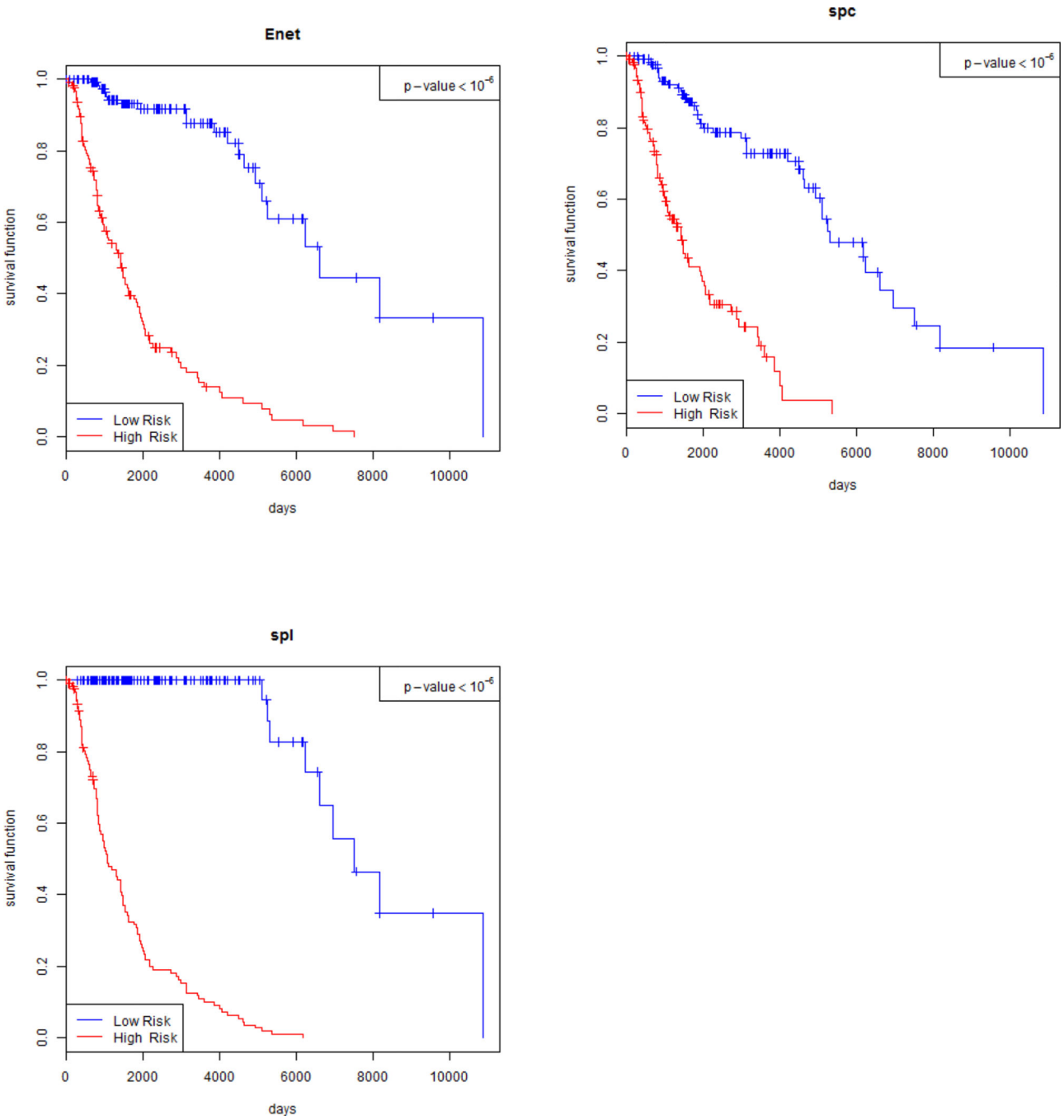

Figure 2.

Survival curves for the low-risk (blue lines) and high-risk (red lines) samples using the three methods. P-values are from the log-rank tests.

Table 1

Brief data information, before and after processing

	Platform /Method	Number of samples	Number of features before processing	Number of features after processing
Clinical-pathological	N.A.	317	83	16
Mutation	Mutation Calling	278	15861	2
CNA	Affymetrix Genome-wide Human SNP array 6.0	336	21699	2500
Methylation	Illumina Human Methylation 450	373	15589	2500
Gene Expression	Illumina Hiseq RNAseq V2	371	19626	2500

Table 2

Prediction performance of different combinations of clinical and omics measurements: mean (sd) of C-statistic.

Model	Enet	SPCA	SPLS
Cln	0.708 (0.077)	0.610 (0.076)	0.714 (0.023)
Cln+Mu	0.707 (0.078)	0.612 (0.072)	0.710 (0.021)
Gen	0.575 (0.071)	0.662 (0.063)	0.665 (0.022)
Met	0.605 (0.068)	0.668 (0.067)	0.702 (0.021)
CAN	0.501 (0.069)	0.570 (0.071)	0.586 (0.026)
Cln+Mu+Gen	0.705 (0.076)	0.661 (0.072)	0.713 (0.022)
Cln+Mu+Met	**0.724 (0.071)**	0.707 (0.060)	0.743 (0.020)
Cln+Mu+CNA	0.691 (0.083)	0.626 (0.069)	0.722 (0.023)
Cln+Mu+Gen+Met	0.717 (0.070)	0.714 (0.059)	**0.746 (0.021)**
Cln+Mu+Gen+CNA	0.686 (0.076)	0.660 (0.073)	0.714 (0.022)
Cln+Mu+Met+CNA	0.713 (0.073)	**0.718 (0.061)**	0.743 (0.020)
Cln+Mu+Gen+Meth+CNA	0.706 (0.072)	0.707 (0.059)	0.746 (0.021)

Cln: clinical variables; Mu: mutation; Gen: mRNA gene expression; Met: methylation; CNA: copy number alternation. A larger value of C-statistic indicates better prediction. A value of 0.5 corresponds to a model with no predictive power.

Table 3

Integrated analysis based on Enet: identified prognostic markers. (A) Top five for clinical variables + mutation; (B) Top ten for methylation. (C) Top ten for gene expression. In the figure, the identified genes (red dots) and those with highly correlated expressions (green dots).

(A)

Clinic+Mutation	Coefficient	Hazard Ratio
Tumor status	2.243	9.423
Breslow thickness at diagnosis	0.0564	1.058
Clark level at diagnosis	0.075	1.078
Age at diagnosis	0.020	1.020
gender	0.053	1.054

(B)

Methylation	Coefficient	Hazard Ratio
ZNF503.AS2	−0.192	0.825
HLA.C	0.272	1.313
IFITM1	0.051	1.052
LIMA1	−0.122	0.885
SLC4A2	−0.005	0.995
PPTC7	−0.199	0.819
MYADML	0.399	1.490
ASCL1	−0.064	0.938
SYT7	−0.035	0.966

(C)

Gene Expression	Coefficient	Hazard Ratio
PIK3R2	0.025	1.025
GATAD2A	0.057	1.058
TRIM32	0.018	1.018
NUP188	0.051	1.052
LAPTM4B	0.009	1.009
ZDHHC13	-0.096	0.908
PARP9	-0.008	0.992
NXT2	-0.041	0.960
MDP1	-0.006	0.994

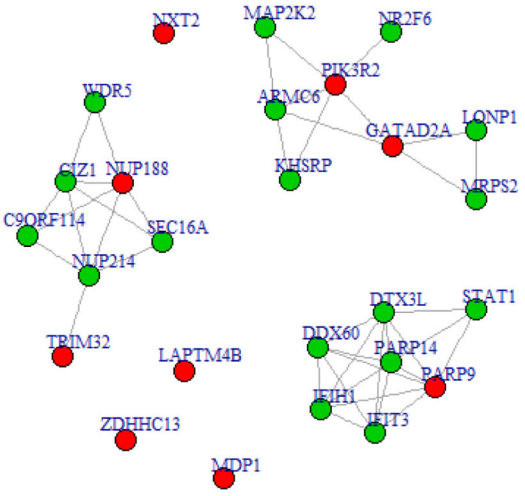

Table 4

Top ten pathways with the largest norms identified by SPCA and SPLS. (Methylation, CNA/GE): norms of the methylation and CNA/GE measurements, respectively.

SPCA		**(Methylation, CNA)**
1	MAPK_PATHWAY	(0.0648, 0.0060)
2	INTEGRIN_PATHWAY	(0.0624, 0.0035)
3	CSK_PATHWAY	(0.0601,0.0028)
4	CELL2CELL_PATHWAY	(0.0574,0.0011)
5	SRCRPTP_PATHWAY	(0.0562,0.0011)
6	CARM_ER_PATHWAY	(0.0519,0.0039)
7	KERATINOCYTE_PATHWAY	(0.0492,0.0060)
8	NFAT_PATHWAY	(0.0502,0.0032)
9	TCYTOTOXIC_PATHWAY	(0.0502,0.0021)
10	CTL_PATHWAY	(0.0481,0.0023)
SPLS		**(Methylation, GE)**
1	MAPK_PATHWAY	(0.0323, 0.0439)
2	IL6_PATHWAY	(0.0243, 0.0389)
3	NFAT_PATHWAY	(0.0262, 0.0365)
4	TEL_PATHWAY	(0.0175, 0.0432)
5	NO1_PATHWAY	(0.0210, 0.0396)
6	IL10_PATHWAY	(0.0259, 0.0319)
7	ACH_PATHWAY	(0.0131, 0.0432)
8	HIVNF_PATHWAY	(0.0192, 0.0365)
9	STATHMIN_PATHWAY	(0.0224, 0.0327)
10	NTHI_HWAY	(0.0189, 0.0344)