BINNING SOMATIC MUTATIONS BASED ON BIOLOGICAL KNOWLEDGE FOR PREDICTING SURVIVAL: AN APPLICATION IN RENAL CELL CARCINOMA

DOKYOON KIM, RUOWANG LI, SCOTT M. DUDEK, JOHN R. WALLACE, MARYLYN D. RITCHIE

Center for Systems Genomics, Department of Biochemistry and Molecular Biology, Pennsylvania State University, University Park, Pennsylvania, USA

Email: marylyn.ritchie@psu.edu

Enormous efforts of whole exome and genome sequencing from hundreds to thousands of patients have provided the landscape of somatic genomic alterations in many cancer types to distinguish between driver mutations and passenger mutations. Driver mutations show strong associations with cancer clinical outcomes such as survival. However, due to the heterogeneity of tumors, somatic mutation profiles are exceptionally sparse whereas other types of genomic data such as miRNA or gene expression contain much more complete data for all genomic features with quantitative values measured in each patient. To overcome the extreme sparseness of somatic mutation profiles and allow for the discovery of combinations of somatic mutations that may predict cancer clinical outcomes, here we propose a new approach for binning somatic mutations based on existing biological knowledge. Through the analysis using renal cell carcinoma dataset from The Cancer Genome Atlas (TCGA), we identified combinations of somatic mutation burden based on pathways, protein families, evolutionary conversed regions, and regulatory regions associated with survival. Due to the nature of heterogeneity in cancer, using a binning strategy for somatic mutation profiles based on biological knowledge will be valuable for improved prognostic biomarkers and potentially for tailoring therapeutic strategies by identifying combinations of driver mutations.

Keywords: Somatic mutation, pathway, somatic mutation burden, survival analysis, renal cell carcinoma

1. Introduction

Cancer is a complex and heterogeneous disease, many of which are caused by somatic mutations or structural alterations. Recent meta-dimensional omics data from The Cancer Genome Atlas (TCGA) or the International Cancer Genome Consortium (ICGC) have provided exceptional opportunities to investigate the complex genetic basis of disease for improving the ability to diagnose, treat, and prevent cancer [1,2]. Multiple alterations affecting cancer can be observed directly as somatic mutations or copy number changes, and indirectly as changes in epigenomic, transcriptomic, and proteomic dimensions. In particular, one of the main issues in cancer research is to distinguish between driver mutations and passenger mutations based on somatic mutation profiles. Massive efforts of whole exome/genome sequencing from hundreds to thousands of patients have provided the landscape of somatic genomic alterations in each specific cancer or across several cancer types [3,4].

Driver mutations show strong associations with survival in several different types of cancer [5]. Thus, evaluation of survival models to predict the disease trajectory of cancer patients based on somatic mutation profiles is one of the most imperative foci in the development of prediction models for cancer prognosis. However, due to the heterogeneity of tumors, somatic mutation profiles are exceptionally sparse whereas other types of genomic data such as miRNA or gene

expression contain nearly complete data for all genomic features with quantitative values measured in each patient. Somatic mutations, in contrast, occur with low to intermediate frequency among cancer patients (2-20%). Thus, it is common that patients do not share any somatic mutations even though they have same clinical features such as prognosis [6].

Previously, we proposed a framework for data integration to predict clinical outcomes in ovarian cancer [7]. However, somatic mutation profiles were not appropriate for predicting outcomes due to the sparseness. To overcome this challenge, we developed a new strategy to use somatic mutation profiles by performing a biologically based collapsing/binning of the mutations to look for an accumulation of somatic mutations in specific types of features based on biological knowledge such as pathways. Then, these somatic mutation burden features in specific pathways can be tested for association with cancer outcome such as survival. It may be desirable to focus on identifying driver pathways instead of driver mutations associated with survival because the patterns of altered pathways may be similar even though patients within a cancer subtype might have diverse mutations [8]. The hypothesis is that rather than looking for the shared mutation to be the driver, it is important to look for the shared pathway (or other biological feature) to be the driver. We applied grammatical evolution neural networks to identify not only specific pathways associated with cancer survival, but also interactions/combinations of pathways. In addition, we tested not only pathways as biological features, but also protein families, regulatory regions, and evolutionary conversed regions to test the association between different types of knowledge-based somatic mutation burden and survival. To test the utility of the proposed strategy, we applied our approach on somatic mutation profiles from renal cell carcinoma from TCGA, which is the most common type of kidney cancer.

2. Methods

2.1. *Data*

Somatic mutations from renal cell carcinoma patients were retrieved from the TCGA (http://tcga-data.nci.nih.gov/) on 1 July 2014. Due to the extreme sparseness of somatic mutation profile, we used all classes of somatic mutations generated from the mutation calling conducted by Baylor College of Medicine (BCM) as a mutation annotation format (MAF) and extracted patient mutation profiles where the analyte was a DNA sample from the tumor. Somatic mutations from 417 patients with renal cell carcinoma were retained for subsequent analysis. Based on chromosomal positions, there were 27,194 unique somatic mutations across all patients. As a clinical outcome, survival information was downloaded for 417 patients as well as sex and age for adjusting potential confounding factors when modeling.

2.2. *BioBin*

BioBin is a flexible collapsing or binning method using biological knowledge to automate the binning of low frequency variants for association tests [9,10]. The main function of BioBin

provides access to comprehensive knowledge-guided multi-level binning. For example, bin boundaries can be formed using genomic locations from: genes, regulatory regions, evolutionary conserved regions, and/or pathways. BioBin uses a built-in database called the Library of Knowledge Integration (LOKI), which is a repository of data assembled from public databases. LOKI contains multiple data resources [11].

Similar to germline low frequency variants, somatic mutations tend to occur with low or intermediate frequency among cancer patients. Thus, we used BioBin for binning somatic mutations based on biological knowledge, such as pathway, in order to overcome the sparseness of somatic mutation profiles. First, we converted MAF to variant call format (VCF) as an input for BioBin (Fig. 1). Then, we applied BioBin to generate KEGG pathway, Pfam, evolutionary conversed region (ECR), and regulatory bin profiles by accumulating somatic mutations in a specific bin. BioBin is open source and available at http://ritchielsb.psu.edu.

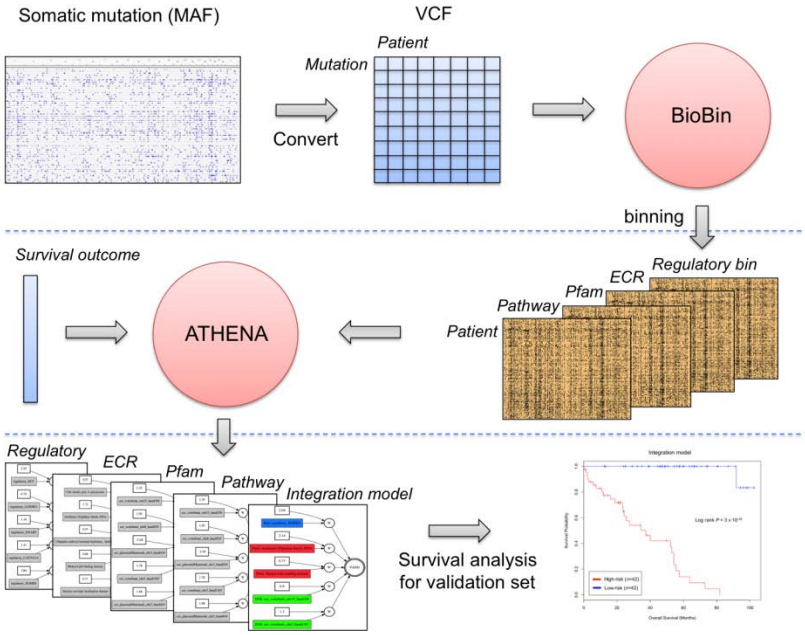


Fig. 1. Illustration of somatic mutation analysis using BioBin and ATHENA for predicting survival.

2.3. *ATHENA*

The Analysis Tool for Heritable and Environmental Network Associations (ATHENA) is a multi-functional software package designed to perform three essential functions to determine the meta-dimensional models of complex disease: (1) performing feature selections from categorical or continuous independent features; (2) building additive and interaction models that explain or predict categorical or continuous clinical outcomes; (3) interpreting the candidate models for use in further translational bioinformatics [7,12]. For this analysis, we used Grammatical Evolution Neural Networks (GENN) as the modeling component. ATHENA is open source and available at http://ritchielab.psu.edu.

2.4. *Grammatical Evolution Neural Networks (GENN)*

Various computational methods have been developed to identify non-linear interactions between genomic variables that have small or large main effects such as a multi-factor dimensionality reduction (MDR) [13]. However, MDR performs an exhaustive search of all possible combination of interacting loci to generate multi-locus predictor models. The search space increases exponentially with the number of variables and became infeasible when integrating meta-dimensional genomics data. Thus, stochastic methods using evolutionary algorithm have been developed and shown to utilize the full dimensionality of the data without exhaustively evaluating all possible combinations of variables [14,15].

Artificial Neural Network (ANN) is a flexible and robust machine learning technique designed to imitate neurons in the brain to solve complex problems. ANN is a good candidate for identifying complex and non-linear interactions that influence variance in an outcome of interest. Generally, the method for applying ANN to a classification problem is to use gradient descent algorithm such as backpropagation to fit the weights of the network given input variables and network architecture. However, the variables and network architecture are not known a priori. In order to simultaneously optimize the input variables, weights, and network structures, evolutionary algorithm approaches have been developed and applied [15,16]. Genetic programming, a specialization of genetic algorithms, is an evolutionary algorithm-based method that uses “survival of the fittest” to evolve optimal solutions from a population of random solutions [17]. Grammatical evolution is a more flexible version of genetic programming since it can also evolve functional solutions, or computer program, via grammar rules [15]. The details of the grammar rules were described in a previous study [15]. The GENN algorithm is briefly described as follows:

(1) The data is divided into five parts for five cross validation with 4/5 for training and 1/5 for testing.

(2) A random population of binary strings is generated to be ANNs using a Backus-Naur grammar. The total population is divided into demes as sub-populations across a user-defined number of CPUs for parallelization.

(3) All ANNs are evaluated with training data, and the solutions with the lowest prediction errors are selected for crossover and reproduction. The new population is composed of mutated original solutions and new random solutions.

(4) Step 3 is repeated for a set number of generations. Migration of best solutions also occurs between demes during evolution for a pre-defined number of times.

(5) The best solution at the final generation is tested using the remaining 1/5 test dataset and fitness is recorded.

(6) Steps 2-5 are repeated four more times, each time using a different 4/5 of the training data and 1/5 of testing data.

2.5. *Survival fitness function*

The goal of this study is to predict censored survival outcome based on somatic mutation burden generated from BioBin. In general, it is difficult to directly predict raw survival data via measuring of goodness-of-fit, such as R^2, due to the censored observations. Thus, an appropriate measure of goodness-of-fit should be required for predicting censored survival data. Martingale residuals are defined as the difference between the cumulative hazards assigned to an individual *i* with failure time t_i and its observed status, $\delta_i = 0$ censored, $\delta_i = 1$ event [18]. Thus, martingale residuals could be intuitively interpreted as the surplus deaths. Martingale residuals are calculated from the fitted Cox model as

$$M_i = \delta_i - \Lambda(t_i) \tag{1}$$

where Λ is a cumulative hazard function [18].

According to the model, the result of cumulative hazard function reflects the number of expected death events per individuals failing at t_i. The range of martingale residuals is between negative infinity and 1 because the cumulative hazard function does not have upper limit. However, the sum of all martingale residuals is zero. Each patient with a negative martingale residual is interpreted as a good prognosis, whereas one with a positive martingale residual is interpreted as a poor prognosis. The martingale residual of each patient is obtained from the reduced model with no genomic effects from somatic mutations. Thus, martingale residuals can be used as a new continuous outcome since they reflect the unexplained portion beyond what is explained by the adjusted clinical covariates excluding the genomic features [18]. In addition, another advantage of martingale residual is that the model can be adjusted by potential confounders such as sex and age. As a proof of concept, we adjusted for age and sex when calculating martingale residuals using *survival* R package.

Since the distribution of martingale residuals is more exponentially shaped, the assumption of R^2, which has normally distributed residuals, is not satisfied. Thus, a new fitness function was proposed for measuring the mean absolute differences (MAD) between observed martingale residuals (M_i) and predicted martingale residuals ($M_{i|x}$) from GENN with genomic covariate vector *x* [19]. The new fitness function is formulated as follows:

$$MAD = \frac{\sum_i \left| M_i - M_{i|x} \right|}{\sum_i \left| M_i \right|} \tag{2}$$

$$Fitness\ function = 1 - MAD \tag{3}$$

The output of the MAD fitness function will be from 0 to 1. The model with 1 fitness score represents the best predictive model whereas the one with 0 fitness score means the worst predictive model. We used MAD for the subsequent experiments.

2.6. *Experiment setup*

Figure 1 shows the overview of the analysis pipeline, which consists of a binning step using

BioBin and a modeling step using ATHENA. After converting MAF to VCF, BioBin was used to generate pathway, Pfam, ECR, regulatory bin profiles. Then, we used ATHENA to build additive/interaction models associated with survival. Martingale residuals and each bin profile can be used as an input for ATHENA. For building GENN models, we randomly split the input dataset into two groups, 4/5 dataset (n=333) for learning models and 1/5 dataset (n=84) for the validation. This is independent of the cross-validation (CV) procedure. The CV procedure was performed on the learning dataset which is 4/5 of the total dataset. Based on GENN results from 5-fold CV, the features from each model across 5 CVs were selected, and then, we reran GENN using selected features to generate the final model from entire training dataset. Lastly, the final GENN model can be used to predict survival from the validation dataset. To avoid over-fitting, the validation dataset was not used for the entire learning step. Table 1 shows the GENN parameters for the analysis. Based on the output of the final GENN model as predicted martingale residuals, the validation dataset was divided into two sub groups, low-risk group and high-risk group, by the median threshold of predicted martingale residuals. Then, survival analysis was performed using *survival* R package.

Table 1. GENN parameter settings

Parameter	Value
Number of demes (CPUs)	20
Population size/ Deme	5,000
Number of generations	1,000
Number of migrations	20
Probability of crossover	0.9
Probability of mutation	0.01
Fitness function	1 – MAD

3. Results and Discussion

3.1. *Binning somatic mutations using BioBin*

To predict survival based on somatic mutation burden, BioBin was used to generate KEGG pathway, Pfam, ECR, and regulatory bin profiles. Somatic mutation burden analysis can be biased when using bins consisting of extremely small number of mutations, thus bins from KEGG pathway, Pfam, ECR, and regulatory regions with more than 10 mutations were selected for the further study. The total number of KEGG pathway, Pfam, ECR, and regulatory bins were 272, 922, 250, and 41, respectively. Since somatic mutation profiles were conducted by whole-exome sequencing, regulatory bin profiles had a relatively small number of bins compared to other bins. Figure 2 shows the difference of sparseness between raw somatic mutation profiles and pathway bin profiles.

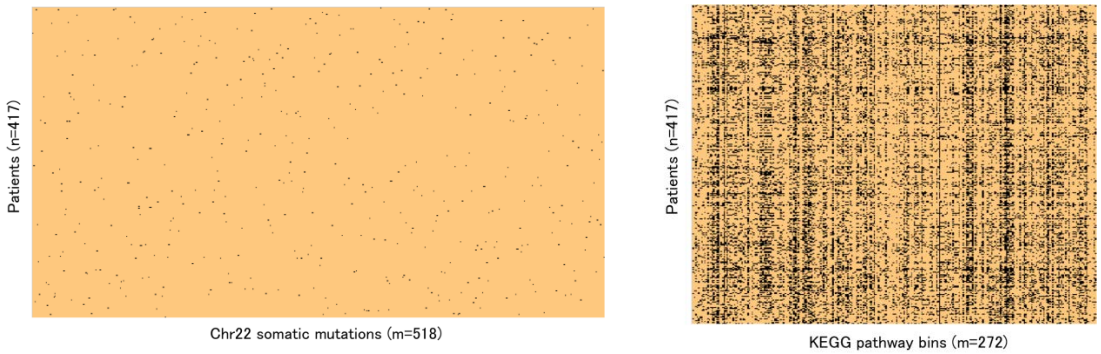


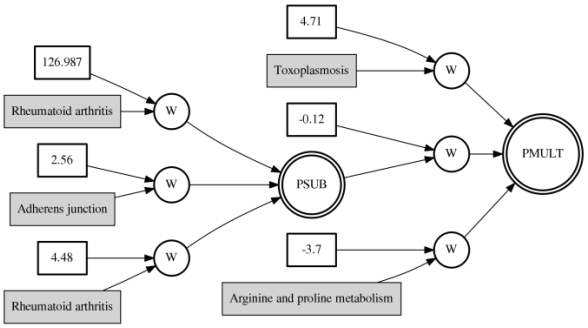
Fig. 2. Difference of sparseness between raw somatic mutation profiles and KEGG pathway bin profiles. For somatic mutation profiles, mutations from chromosome 22 were extracted for generating the heatmap figure. Each black dot represents the presence of either somatic mutation in the left heatmap or mutation burden in a pathway in the right heatmap.

3.2. *GENN modeling for somatic mutation burden*

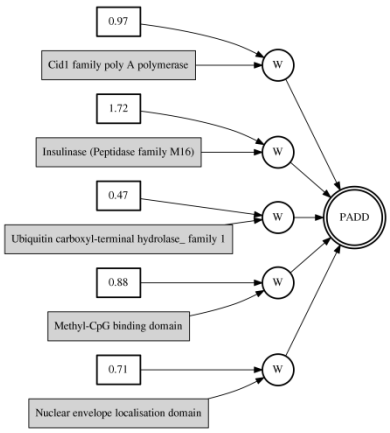
A simulation study was conducted to demonstrate the validity of the proposed survival fitness function and martingale residuals as a new outcome for predicting survival (data not shown) [Kim et al., submitted]. According to the results from the simulation data, martingale residuals performed well as a new outcome in terms of finding true survival genes and limited false positives using GENN. Next, somatic mutation profiles in renal cell carcinoma were analyzed to identify additive/interaction models based on knowledge-based somatic mutation burden. After generating pathway, Pfam, ECR, and regulatory bins using BioBin, GENN models were trained to predict survival from the validation dataset. The final model of GENN is the evolved neural network with optimized input variables, weights, and network structure to identify additive or interaction models that predict survival outcome. Figure 3 shows the best GENN models from each bin profile: KEGG pathway, Pfam, ECR, and regulatory bins, respectively. Finally, the final GENN model was used to predict survival from the validation dataset, which consisted of 84 patients. The fitness scores from the validation dataset for each of the best models with pathway, Pfam, ECR, and regulatory bin profiles were 0.641, 0.67, 0.665, and 0.654, respectively (Fig 3 and Table 2). Among four different bin profiles, Pfam bin profiles showed the best performance for predicting survival.

Table 2. Performance comparison between different types of bin profiles. Performance was measured from the validation dataset.

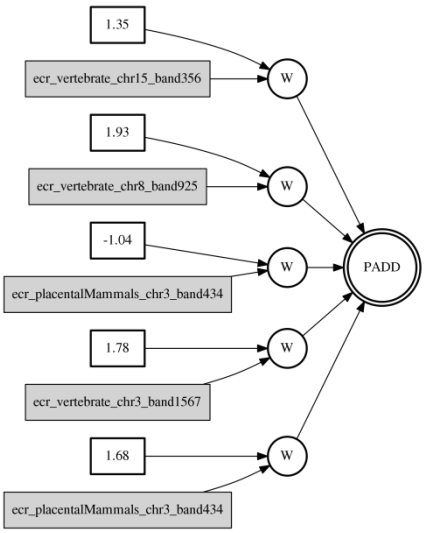
GENN model	1 - MAD	Permutation p-value
KEGG pathway bins	0.641	0.602
Pfam bins	0.67	0.108
ECR bins	0.665	0.2
Regulatory bins	0.654	0.423
Integration	0.685	0.026



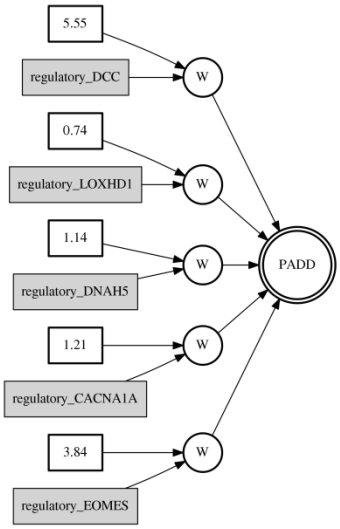
(a) Pathway bins (Fitness score: 0.641)



(b) Pfam bins (Fitness score: 0.67)



(c) ECR bins (Fitness score: 0.665)



(d) Regulatory bins (Fitness score: 0.654)

Fig. 3. Best GENN models from each knowledge-based somatic mutation profiles. PSUB, PMUL, and PADD are a subtraction, multiplication, and addition activation node, respectively. Knowledge features such as pathway, Pfam, ECR, regulatory regions are shown in the gray boxes. (a) pathway-based mutation profiles (b) Pfam-based mutation profiles (c) ECR-based mutation profiles (d) regulatory-based mutation profiles

To build an interaction model between different knowledge-guided bins associated with survival in renal cancer, we integrated pathway, Pfam, ECR, and regulatory bin profiles. The final integration model was generated using GENN with variables from the best models of each bin profile. The final integration model was also used to predict survival for the validation dataset. In terms of predictive power, the integration model showed the best performance with a fitness score of 0.685 (Table 2). The selected features in the final integration model are methyl-CpG binding domain and insulinase (Peptidase family M16) from Pfam bin profile, ecr_vertebrate_chr3_band1567 and

ecr_vertebrate_chr15_band356 from ECR bin profiles, and regulatory_EOMES from regulatory bin profiles (Fig. 4). To test the statistical significance of each GENN model, permutation testing was performed. The survival outcome for the validation dataset was randomly permuted 1000 times and permutation p-values of each GENN model were obtained from the 1000 random validation sets (Table 2). The integration model showed a significant result (P = 0.026) while other GENN models were not significant based on permutation testing. In addition, survival analysis was performed for two sub groups, low-risk and high-risk groups, which were divided by a median threshold of predicted martingale residuals for the validation dataset. Kaplan-Meier analysis showed that the two groups generated from the integration model were significantly different based on survival (Fig. 5).

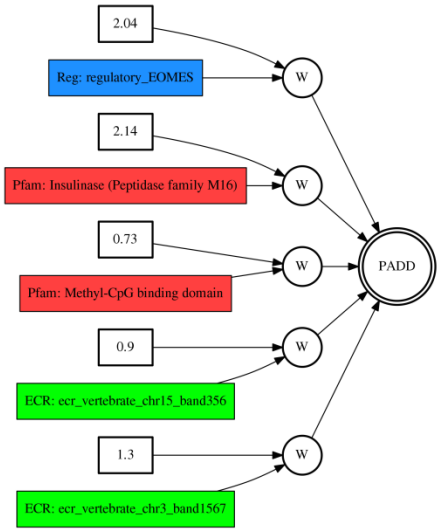


Fig. 4. Integration model containing variables from different knowledge-based mutation profiles. Red, green, and blue boxes represent Pfam, ECR, and regulatory region features, respectively. PADD represents an addition activation node. A fitness score of the integration model was 0.685.

3.3. *Biological interpretation*

Four pathways, arginine and proline metabolism, adherens junction, toxoplasmosis, and rheumatoid arthritis, were found in the GENN models. Adherens junctions are one of the most relevant junctional complexes in the kidney epithelium and adherens junction disruption is associated with cell proliferation, invasion, and angiogenesis in renal cell carcinoma [20]. In addition, arginine and proline metabolism has been shown to be important in renal cell carcinoma using proteomic and metabolic profiles [21]. Interestingly, rheumatoid arthritis was identified as one of the features in the final model. Associations between rheumatoid arthritis and kidney cancer have been reported in many studies. For example, infliximab, anti-tumor necrosis factor α (TNF-α) antibody, is licensed for use in rheumatoid arthritis and TNF-α might be a therapeutic target in renal cell carcinoma [22]. Notably, the top pathway model showed complex and non-linear interactions between pathways associated with survival. This suggests it is important to consider

interactions associated with survival, which might have crucial roles in molecular pathogenesis, progression, and prognosis of renal cell carcinoma, that are not easily detected by traditional pathway analysis approaches. For Pfam model, a combination of methyl-CpG binding domain, cid1 family poly A polymerase, ubiquitin carboxyl-terminal hydrolase family 1, insulinase (peptidase family M16), and nuclear envelope localization domain was found to be associated with survival. In particular, the epigenetic silencing of cancer-related genes such as *ABCG2* has been shown to be associated with renal cell carcinoma via being mediated through recruitment of a group of proteins, called methyl-CpG binding domain (*MBD*) [23]. Epigenetic control of the ubiquitin carboxyl terminal hydrolase family 1, which plays an important role in cell growth and differentiation, can be disturbed in renal cell carcinoma [24]. Several evolutionary conversed regions were also selected from ECR model. Many genes located in selected evolutionary conversed regions such as *BAP1*, *EIF4G1*, *EBAG9*, or *FBN1* were found as import genes involved in several cancers. In particular, *BAP1* loss defines a new class of renal cell carcinoma [25]. In addition, somatic mutation burden in the regulatory regions of *CACNA1A*, *LOXHD1*, *DNAH5*, *DCC*, or *EOMES* might play a functionally significant role in renal cell carcinoma survival. In the integration model, where we used multiple knowledge sources, methyl-CpG binding domain and insulinase (Peptidase family M16) from Pfam model, chr3_band1567 and chr15_band356 from ECR model, and *EOMES* from regulatory model were selected. Combination of somatic mutation burden based on multiple biological knowledge sources might reflect the complex molecular pathogenesis and progression of renal cell carcinoma.

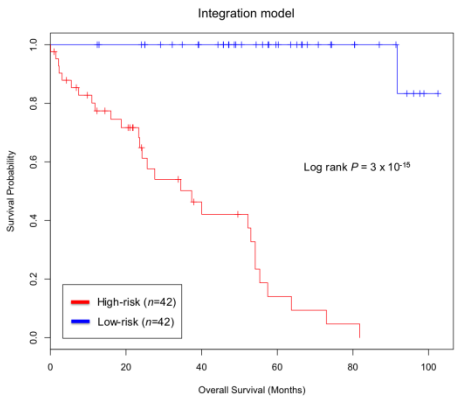


Fig. 5. Kaplan-Meier survival plots for the validation dataset. Validation dataset was divided into high-risk and low-risk groups based on a median value of predicted outputs from the integration model.

4. Conclusions

In this study, we proposed a new approach of binning somatic mutation based on biological knowledge for predicting survival in order to overcome the extreme sparseness of somatic mutation profiles. Through the analysis using renal cell carcinoma dataset, we identified interaction/combinations of somatic mutation burden based on pathway, protein families,

evolutionary conversed regions, and regulatory regions associated with survival. Knowledge guided binning/collapsing somatic mutations dramatically reduce the sparseness of profiles as well as search space, from 27,194 mutations to 272 pathways. In terms of predictive power, some of the GENN models were not significant based on the permutation test. However, it might be due to the fitness function based on mean absolute difference. Even though survival outcome of 84 patients were shuffled, a mean difference between observed martingale residuals and predicted martingale residuals could not be too large. In addition, a small number of patients in the validation dataset limit our power as sample size is often a limiting factor. Improving the methodology by incorporating directionality of the pathway when binning somatic mutations could also potentially increase the predictive power of pathway-based approach. Notably, the predictive power of the integration model outperformed other models from single types of biological knowledge sources. These results suggest that each biological knowledge source can be complementary to the prediction power of survival because each knowledge source has its specific biological context. Furthermore, the survival analysis for the validation dataset demonstrated that somatic mutation burden based on biological knowledge showed significant associations with cancer prognosis in renal cell carcinoma.

The present study underpins our on-going work. First, not only somatic mutations but also germline mutations can be regarded as important genomic features that are associated with cancer outcomes [5]. Thus, as one of promising future works, it would be valuable to combine both types of mutations to investigate the associations with cancer outcomes. It would be also interesting to investigate whether known somatic mutations influence the models. In addition, the proposed approach could be applied to explore associations with other cancer clinical outcomes such as stage, grade, recurrence, or metastasis. Furthermore, the current approach by biological-based bins such as pathways may unnecessarily inject noise into the model. It would be intriguing to apply an adquate filtering step to the mutations in future studies. Due to the nature of heterogeneity in cancer, using a binning strategy for somatic mutation profiles based on biological knowledge will be valuable for improved prognostic biomarkers and tailoring therapeutic strategies by identifying interactions/combinations of driver mutations.

Acknowledgments

This work was funded by NIH grant R01 LM010040, NHLBI grant U01 HL065962, and CTSI: UL1 RR033184-01. This work is also supported by a grant with the Pennsylvania Department of Health using Tobacco CURE Funds.

References

1. Cancer Genome Atlas Research N (2013) Comprehensive molecular characterization of clear cell renal cell carcinoma. Nature 499: 43-49.
2. International Cancer Genome Consortium (2010) International network of cancer genome projects. Nature 464: 993-998.
3. Cancer Genome Atlas Research N, Weinstein JN, Collisson EA, Mills GB, Shaw KR, et al. (2013) The Cancer Genome Atlas Pan-Cancer analysis project. Nat Genet 45: 1113-1120.

4. Brennan CW, Verhaak RG, McKenna A, Campos B, Noushmehr H, et al. (2013) The somatic genomic landscape of glioblastoma. Cell 155: 462-477.
5. Yang D, Khan S, Sun Y, Hess K, Shmulevich I, et al. (2011) Association of BRCA1 and BRCA2 mutations with survival, chemotherapy sensitivity, and gene mutator phenotype in patients with ovarian cancer. JAMA 306: 1557-1565.
6. Lawrence MS, Stojanov P, Polak P, Kryukov GV, Cibulskis K, et al. (2013) Mutational heterogeneity in cancer and the search for new cancer-associated genes. Nature 499: 214-218.
7. Kim D, Li R, Dudek SM, Ritchie MD (2013) ATHENA: Identifying interactions between different levels of genomic data associated with cancer clinical outcomes using grammatical evolution neural network. BioData Min 6: 23.
8. Vogelstein B, Papadopoulos N, Velculescu VE, Zhou S, Diaz LA, Jr., et al. (2013) Cancer genome landscapes. Science 339: 1546-1558.
9. Moore CB, Wallace JR, Frase AT, Pendergrass SA, Ritchie MD (2013) BioBin: a bioinformatics tool for automating the binning of rare variants using publicly available biological knowledge. BMC Med Genomics 6 Suppl 2: S6.
10. Moore CB, Wallace JR, Wolfe DJ, Frase AT, Pendergrass SA, et al. (2013) Low frequency variants, collapsed based on biological knowledge, uncover complexity of population stratification in 1000 genomes project data. PLoS Genet 9: e1003959.
11. Pendergrass SA, Frase A, Wallace J, Wolfe D, Katiyar N, et al. (2013) Genomic analyses with biofilter 2.0: knowledge driven filtering, annotation, and model development. BioData Min 6: 25.
12. Holzinger ER, Dudek SM, Frase AT, Pendergrass SA, Ritchie MD (2013) ATHENA: the analysis tool for heritable and environmental network associations. Bioinformatics.
13. Ritchie MD, Hahn LW, Roodi N, Bailey LR, Dupont WD, et al. (2001) Multifactor-dimensionality reduction reveals high-order interactions among estrogen-metabolism genes in sporadic breast cancer. Am J Hum Genet 69: 138-147.
14. Ritchie MD, White BC, Parker JS, Hahn LW, Moore JH (2003) Optimization of neural network architecture using genetic programming improves detection and modeling of gene-gene interactions in studies of human diseases. BMC Bioinformatics 4: 28.
15. Motsinger-Reif AA, Dudek SM, Hahn LW, Ritchie MD (2008) Comparison of approaches for machine-learning optimization of neural networks for detecting gene-gene interactions in genetic epidemiology. Genet Epidemiol 32: 325-340.
16. Turner SD, Dudek SM, Ritchie MD (2010) ATHENA: A knowledge-based hybrid backpropagation-grammatical evolution neural network algorithm for discovering epistasis among quantitative trait Loci. BioData Min 3: 5.
17. Ritchie MD, Motsinger AA, Bush WS, Coffey CS, Moore JH (2007) Genetic Programming Neural Networks: A Powerful Bioinformatics Tool for Human Genetics. Appl Soft Comput 7: 471-479.
18. Therneau TM, Grambsch PM, Fleming TR (1990) Martingale-Based Residuals for Survival Models. Biometrika 77: 147-160.
19. Müller M (2004) Goodness-of-fit criteria for survival data. Sonderforschungsbereich Paper 382.
20. Peruzzi B, Athauda G, Bottaro DP (2006) The von Hippel-Lindau tumor suppressor gene product represses oncogenic beta-catenin signaling in renal carcinoma cells. Proc Natl Acad Sci U S A 103: 14531-14536.
21. Perroud B, Lee J, Valkova N, Dhirapong A, Lin PY, et al. (2006) Pathway analysis of kidney cancer using proteomics and metabolic profiling. Mol Cancer 5: 64.
22. Harrison ML, Obermueller E, Maisey NR, Hoare S, Edmonds K, et al. (2007) Tumor necrosis factor alpha as a new target for renal cell carcinoma: two sequential phase II trials of infliximab at standard and high dose. J Clin Oncol 25: 4542-4549.
23. To KK, Zhan Z, Bates SE (2006) Aberrant promoter methylation of the ABCG2 gene in renal carcinoma. Mol Cell Biol 26: 8572-8585.
24. Seliger B, Handke D, Schabel E, Bukur J, Lichtenfels R, et al. (2009) Epigenetic control of the ubiquitin carboxyl terminal hydrolase 1 in renal cell carcinoma. J Transl Med 7: 90.
25. Pena-Llopis S, Vega-Rubin-de-Celis S, Liao A, Leng N, Pavia-Jimenez A, et al. (2012) BAP1 loss defines a new class of renal cell carcinoma. Nat Genet 44: 751-759.